cgNA+: A sequence-dependent coarse-grain model of double-stranded nucleic acids

Submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

by

RAHUL SHARMA

under the supervision of

Prof. John H. Maddocks

Laboratory for Computation and Visualization in Mathematics and Mechanics
Institute of Mathematics
École Polytechnique Fédérale de Lausanne
Lausanne 1015, Switzerland

July, 2022

ABSTRACT

DNA mechanics plays a crucial role in many biological processes, including nucleosome positioning and protein-DNA interactions. It is believed that nature employs epigenetic modifications in DNA to further regulate gene expression. Moreover, double-stranded RNA and DNA:RNA hybrid (DRH) are also important in biology, and their mechanics play a significant role. It is now well established that the mechanics of double-stranded nucleic acid (dsNA) is a function of its sequence. In particular, the sequence-dependent mechanics of DNA is often considered as the "secondary genetic code" owing to its quintessential role in DNA readout. However, a comprehensive understanding of sequence-dependent mechanics of dsNAs is still lacking, primarily due to enormous sequence space, which is unexplorable using either experiment or atomistic molecular dynamics (MD) simulation, and, thus, requires an accurate and efficient alternative.

This thesis extends the cgDNA+ model, a sequence-dependent coarse-grained model of dsDNA, to cgNA+ by estimating parameters for various dsNAs, including dsRNA, DRH, and dsDNA with epigenetic base modifications. The model is trained on atomistic MD simulations generated with state-of-the-art MD protocols. For an arbitrary sequence, the model efficiently predicts sequence-dependent equilibrium distributions, treating bases and phosphates as rigid bodies. The model is thoroughly assessed for mechanically diverse test sequences and various modeling choices are explained and justified by quantifying the associated error.

Moreover, as exhibited in the protein-DNA X-ray structure data, flanking contexts are essential for dimer mechanics. We compared X-ray observation with model predictions for dimers in all tetramer contexts and found a reasonable agreement for average shape, stiffness, direction of variation of groundstate in sequence space, and direction of dsDNA deformation in configuration space. Remarkably, we also found an excellent alignment between the direction of variation of groundstate in sequence space and the direction of dsDNA deformation in configuration space, implying that, for various sequences/flanking contexts, dimer adopts groundstate by compromising more in the soft modes of configuration space.

The cgNA+ model efficiency enables the study of interesting properties of dsNAs, such as average shape, persistence length, backbone conformations, and groove widths for millions of sequences, thereby, drawing statistical conclusions over sequence space. It allows addressing questions including (a) which single nucleotide polymorphisms influence dsDNA mechanics the least/most and its sensitivity to flanking sequence, (b) the role of sequence in narrowing/widening of grooves, and (c) the role of flanking sequence in epigenetic modifications. Other

applications include scanning genomes for mechanically exceptional sequences, understanding sequence-dependent nucleosome (un)wrapping, predicting protein binding affinity, and studying dsNA response to external load.

Lastly, we develop a deep learning tool to predict the location of sugar atoms in any cgNA+ coarse-grained configuration. It allows generating an ensemble of atomistic configurations for any sequence comparable to MD simulations but with little computational effort and studying backbone and sugar conformations. Furthermore, a fine-grain sequence-dependent equilibrium structure can be used to start MD simulations, particularly useful for dsDNA mini-circles.

Keywords: *Coarse-graining*, *MD simulations*, *DNA mechanics*, *RNA*, *DNA:RNA hybrid*, *Epigenetics*, *Neural network*, *Sugar puckering*, *Groove widths*, *Persistence length*.

Résumé

La mécanique de l'ADN joue un rôle crucial dans de nombreux processus biologiques, notamment le positionnement des nucléosomes et les interactions protéine-ADN. On pense que la nature utilise des modifications épigénétiques de l'ADN pour réguler davantage l'expression des gènes. De plus, l'ARN double brin et l'hybride ADN:ARN (DRH) sont également importants en biologie, et leur mécanique joue un rôle important. Il est maintenant bien établi que la mécanique d'un acide nucléique double brin (ADNd) est fonction de sa séquence. En particulier, la mécanique de l'ADN séquence-dépendant est souvent considérée comme le "code génétique secondaire" en raison de son rôle primordial dans la lecture de l'ADN. Cependant, une compréhension complète de la mécanique séquence-dépendant des acides nucléiques (AN) fait toujours défaut, principalement en raison de l'énorme espace des séquences, qui est inexplorable en utilisant des expériences ou des simulations atomistiques de dynamique moléculaire (MD), et nécessite donc une alternative précise et efficace.

Cette thèse étend le modèle cgDNA+, un modèle à gros grains séquence-dépendant de l'ADNdb, à cgNA+ en estimant les paramètres pour divers ADNdb, y compris l'ARN, le DRH et l'ADN avec des modifications épigénétiques. Le modèle est entraîné sur des simulations MD atomistiques en utilisant les protocoles MD les plus récents. Pour une séquence arbitraire, le modèle prédit efficacement les distributions d'équilibre séquence-dépendant, en traitant les bases et les phosphates comme des corps rigides. Le modèle est évalué de manière approfondie pour des séquences de test mécaniquement diverses et divers choix de modélisation sont expliqués et justifiés en quantifiant l'erreur associée.

De plus, ayant d'abord montré dans les données de structure aux rayons X protéine-ADN que les contextes flanquants sont essentiels pour la mécanique des dimères, nous avons comparé les observations aux rayons X aux prédictions du modèle pour les dimères dans tous les contextes de tétramères et avons trouvé un accord raisonnable pour la forme moyenne, la rigidité, la direction de variation de l'état fondamental dans l'espace de séquence, et la direction de la déformation de l'ADN dans l'espace de configuration. De manière remarquable, nous avons également trouvé un excellent alignement entre la direction de variation de l'état fondamental dans l'espace des séquences et la direction de la déformation de l'ADN dans l'espace de configuration, ce qui implique que, pour diverses séquences/contextes de flanquement, le dimère adopte l'état fondamental en faisant plus de compromis dans les modes souples de l'espace de configuration.

L'efficacité du modèle permet d'étudier des propriétés intéressantes des AN, telles que la

forme moyenne, la longueur de persistance, les conformations du chaîne principale et la largeur des sillons pour des millions de séquences, ce qui permet de tirer des conclusions statistiques sur l'espace des séquences. Elle permet d'aborder des questions telles que (a) quel polymorphisme nucléotidique unique influence le moins/le plus la mécanique de l'ADN et sa sensibilité à la séquence flanquante, (b) le rôle de la séquence dans le rétrécissement/l'élargissement des sillons, et (c) le rôle de la séquence flanquante dans les modifications épigénétiques. D'autres applications comprennent le scanning des génomes à la recherche de séquences mécaniquement exceptionnelles, la compréhension de l'enroulement/déroulement des nucléosomes séquence-dépendant, la prédiction de l'affinité de liaison des protéines et l'étude de la réponse de AN aux charges externes.

Enfin, nous développons un outil d'apprentissage profond pour prédire l'emplacement des atomes de sucre dans toute configuration à gros grains d'ADNc+. Il permet de générer un ensemble de configurations atomistiques pour toute séquence comparable aux simulations MD mais en un temps très court et d'étudier les conformations du chaîne principale et des sucres. De plus, une structure d'équilibre à grain fin séquence-dépendant peut être utilisée pour démarrer les simulations MD, ce qui est particulièrement utile pour les mini-cercles d'ADN.

Dedicated to my parents

ACKNOWLEDGEMENT

With great pleasure and gratitude, I express my sincere thanks to Prof. John H. Maddocks for his supervision, continuous encouragement, and indispensable feedback. Along with scientific learning, I hope to remember his entertaining and, most of the times, insightful stories.

I am thankful to all the jury members, Prof. Pablo Dans, Dr. Oliver Henrich, and Prof. Jiri Vanicek, for their discerning and insightful remarks, which have substantially improved this thesis and introduced me to several other relevant scientific questions.

This work is a contribution to the cgDNA family of models, which have been developed by several LCVMM members (and collaborators). I am thankful to all of them, particularly Alessandro, who developed the cgDNA+ model and initially helped me understand various codes. I cannot imagine a kinder and more patient teacher. Moreover, I am thankful to Prof. Wilma Olson, Dr. Luke Czapla, and Dr. Helen Lindsay for their contributions and insights in the 5^{th} chapter of this thesis. For running MD simulations, I would like to acknowledge SCITAS, HPC facilities at EPFL and, in particular, Gilles for quickly troubleshooting cluster-related problems and Philippe for other computer-related assistance. I had the pleasure to directly work and collaborate on scientific as well as non-scientific activities (somewhat more cherishable to me) with other LCVMM members, including Chakri, Daiva, Giulio, Harmeet, Jannes, Rasa, Raushan, and Thomas. Lastly, Carine's contribution to protecting us from non-trivial administrative work deserves much credit. I am thankful to her for being the kindest person around.

Even though Covid was the only one to be always out there for me (and everyone), I was blessed with many friends in Lausanne and around the world (connected via Zoom). Firstly, I am most thankful to Raushan and Lucky for being companions in countless activities, out of which playing cards, swimming, and cooking are closest to my heart. I am also thankful to Abhishek, Challenger, and Vikranth for many and long online discussions on scientific and non-scientific aspects. Lastly, I would be remiss in not mentioning Ankit, Anurag, Neeraj, Omkar, Ritesh, and Saloni, whose presence and experiences enriched my life.

I would also like to acknowledge all the Alpine mountains for their abundant beauty and serenity. Weekend hikes in these mountains rejuvenated and filled me with fresh inspiration to work on my research endeavors.

I thank Saumya for being the most amazing person and partner in everything. Her presence, support, understanding, and countless discussions on science, life, and somehow other remaining things made my life delightful and motivating. I am glad she was there to help polish the thesis and the final presentation.

Last but not least, I am indebted to my parents for constant encouragement and love throughout my life, which is impossible to express in words. Finally, a big thanks to my dear sister, Parag, for being the most cheerful and lovely person and for endless mischiefs and arguments.

This work was funded by the Swiss National Science Foundation, project 200020_182184.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

The protagonist in this thesis, deoxyribonucleic acid (**DNA**) was first isolated [122] in 1869 by F. Miescher, and following several indispensable breakthroughs, its double-stranded helical structure was finally deciphered by J. Watson and F. Crick in 1953 [207]. The Watson-Crick model describes dsDNA as a helical structure with a sugar-phosphate bi-chain, also known as Crick and Watson strands, on the outside, held together by hydrogen-bonding between nitrogenous base-pairs inside. This base-pairing is highly specific (Adenine[A]:Thymine[T], Guanine[G]:Cytosine[C]), so that either strand has the necessary information to replicate itself and is thus capable of transferring information from one generation to the next. Nevertheless, the story did not end there. Contemporary to the discovery of DNA structure, Conrad H. Waddington found that environmental factors also influence phenotypic features in fruit flies [205] and termed it "**epigenetics**" which means "in addition to genetics". It is formally defined as "An epigenetic trait is a stably heritable phenotype resulting from changes in a chromosome without alterations in the DNA sequence" [19]. The most common epigenetic modifications include base modifications, non-coding RNAs, and histone modifications. For instance, around 70-80% of CpG steps in mammalian cells [77] are methylated and in promoter regions, it is anti-correlated with gene expression [32, 151]. Epigenetics has implications in gene silencing [32, 86, 151], X-chromosome inactivation [66], and in humans, epigenetic aberrations are associated with diseases such as cancer, autoimmune disease, and neurodevelopmental disorders [93, 161] and, therefore, is of great interest for research. Moreover, some epigenetic aberrations are reversible and, thus, targeted in therapeutic approaches [87].

With DNA as the protagonist, ribonucleic acid (**RNA**) is the deuteragonist. It acts as a genetic carrier in viruses and has other diverse roles in biology such as reaction catalysis, genetic information processing, and gene regulation [55, 114, 192]. Chemically, RNA only differs from DNA in deoxyribose sugar and Thymine, instead has ribose sugar and Uracil (U) base. In biology, RNA is often present in a single-stranded form; however, double-stranded RNA is also vital in gene regulation via RNA interference (sequence-specific gene suppression) [55], components of several tertiary structures such as riboswitches [114], hairpins, and transfer RNA and as the genome of some viruses. Moreover, in many biological processes [2, 24, 121, 167, 185, 204], unique heterogeneous nucleic acids (NAs) are formed with one DNA strand and one RNA strand known as DNA:RNA hybrid (**DRH**). For example, during reverse transcription, RNA viruses create transient DRH whose stability is consequential in their replication cycle [24, 116, 204]. Furthermore, DRHs are considered as potential medicinal agents for HIV or other retroviruses diseases [185, 200] and are crucial in the CRISPR-Cas9 technology [194], genome

stability, and DNA repair [135].

In parallel, it became evident that not only the chemistry but also the mechanics of double-stranded nucleic acids, specifically their shape and flexibility, plays a crucial role in their function. For example, DNA mechanics is pivotal in nucleosome positioning [179, 180], indirect readout [36, 126], DNA looping [3, 11, 176] and protein-DNA interactions [36, 85, 126, 146, 171]. In particular, owing to its quintessential role in DNA readout, sequence-dependent mechanics of DNA is often regarded as a "secondary genetic code". Moreover, epigenetic modifications in DNA further regulate gene expression, supposedly by changing the mechanics of DNA. For instance, methylation of CpG steps in promoter regions leads to gene silencing [32, 151] by reducing flexibility and, thus, reducing the ability of DNA to interact with transcription factors, modulating DNA accessibility and making them less prone to wrap around nucleosomes [38, 72, 156, 162, 172]. Sequence-dependent mechanics is also determining in the functioning of RNA and DRH as well [132, 133, 185, 193, 196, 209]. Such direct evidence piqued significant interest in understanding the sequence-dependent mechanics of nucleic acids.

The primary experimental tools for studying the mechanics of dsNAs are cyclization experiments [188], optical tweezers [189], small-angle X-ray scattering [120], and cryogenic electron microscopy [15, 48]. More details on various techniques for *in vivo* and *in vitro* characterization of dsNAs mechanics can be found in ref. [157]. Even though experiment is an excellent approach to explore the role of mechanical properties of dsNAs in their functioning, designing and performing experiments is highly time-consuming and, thus, can only be performed for a few sequences. Recently, Basu et al. [11] developed high-throughput methods to measure the tendency for DNA looping and computed intrinsic cyclizabilities of roughly 300,000 50 base-pairs (bps) DNA fragments (flanked both sides by 25 bps fixed duplex and single-stranded complementary overhangs) and found an intricate role of sequence in the overall mechanics of DNA that can not be sufficiently described by basic sequence descriptors such as GC content, A-tracts, and dimer steps. Such experimental methods can provide an overall picture, but lack a finer description.

A promising alternative is provided by computational modeling and simulation. In particular, atomistic molecular dynamics (MD) simulations have been widely used to study various structural, mechanical, and functional aspects of nucleic acids [9, 12, 28, 50, 56, 74, 97, 116, 132, 133, 142, 147, 152, 153, 155, 156, 196] and have become an indispensable tool in general for studying bio-molecules [74]. However, due to the immense sequence space of DNA, it is not feasible to investigate all sequences (even for DNA dodecamers); for instance, the most extensive analysis using atomistic MD simulations published so far is only for the 136 independent tetramers [50, 147] by the ABC consortium. Also, MD simulations are extremely slow for simulating longer dsNA fragments as most of the simulation time are consumed to model relatively uninteresting water molecules. Therefore, a systematic investigation requires computationally efficient alternatives.

There have been several attempts to model DNA, starting with worm-like chain models [92, 186]. One of the first and widely applied models for the coarse-grained sequence-dependent DNA model was a rigid base-pair model [139] with dimer-dependent parameters obtained from X-ray crystal data from protein-DNA complexes (which have been a great source of information to study the structure and flexibility of DNA [128]). Similar rigid base-pair models were

also trained on atomistic MD simulations [61, 97]. One major drawback of rigid base-pair models is that the non-local sequence dependence in the model requires non-local sequence-dependent parameters, which are almost impractical to obtain, particularly for a model trained on experimental data. Therefore, most rigid base-pair models only have local dimer sequence dependence. However, it has been observed multiple times that sequence dependence limited to the local dimer step is not always sufficient to explain all properties of specific DNA sequences, and non-local sequence dependence (i.e., depend on flanking sequence) plays a pivotal role in DNA mechanics [6, 9, 56, 102, 145, 155, 208]. In particular, Balaceanu et al. [9] using MD simulations demonstrated that the structure and flexibility of the central TA step are significantly modulated by hexamer or even beyond flanking context.

More recently, a few other models have been proposed to study sequence-dependence properties in DNA [33, 71, 143, 203]; however, the parameters in these models have often been fit analytically to experimental data and have limited and local sequence-dependence. The first and only model, to our knowledge, that predicts non-local sequence dependence in the shape of DNA is cgDNA [62, 159]. cgDNA is a coarse-grained model of B-DNA which predicts the probability distribution function of an arbitrary DNA sequence (in standard A, T, C, G alphabets) under pre-specified physical solvent conditions. Originally, in this model, DNA was coarse-grained as a bi-chain of explicit bases, which is now extended to "cgDNA+" [149] with explicit representation of phosphates in addition to the bases. The non-local sequence-dependence in the cgDNA(+) model using dimer-dependent parameters originates from the fact that individual base-pair steps cannot achieve their local minima simultaneously, and frustration of energy surfaces arises in the nearest neighbors; thus, the cgDNA(+) model naturally captures the non-local sequence-dependence in the DNA mechanics, but only using dimer dependent parameters. Both models are trained on molecular dynamics time-series of a comprehensive set of DNA sequences, and the model predictions are shown to be almost indistinguishable from the corresponding MD statistics of first and second moments.

The cgDNA(+) model has been successfully implemented to explore sequence-dependent persistence lengths [123] of DNA, sequence-dependent unwrapping pathways of DNA from the nucleosome core particle [119], crystal structure packing forces [149], and the role of histone tails in nucleosome stability [16]. Moreover, the cgDNA+ model has been used to scan entire genomes searching mechanically exceptional sequences [213] and obtain sequence-dependent shapes of DNA minicircles [13, 60]. Other exciting applications of the cgDNA+ model actively pursued (in LCVMM and with collaborators) include the response of DNA to external loading and twisting, the calculation of the nucleosome wrapping energy for DNA, and the prediction of protein-DNA binding affinity. Moreover, this model can potentially contribute to fine-tuning rapidly evolving DNA applications, for example, DNA nanotechnology. Thus, the model has shown great potential for diverse applications involving DNA mechanics with the overarching goal of deciphering how DNA mechanics facilitate its functioning in biology.

One particular limitation is that the current model only allows standard DNA bases. However, in biology, bases in the DNA sequences are often modified, particularly, Cytosine (C), which is either methylated or hydroxymethylated at the 5-position of Cytosine in CpG steps. Furthermore, RNA and DRH are also ubiquitous in biology and essential for several other appli-

cations. Therefore, in this thesis, we have extended the cgDNA+ model by estimating parameter sets for RNA, DRH, and DNA with base modifications and called it the cgNA+ model. Moreover, we have compared cgNA+ model predictions with experimental findings, particularly with the existing protein-DNA X-ray structure database for dimers in all tetramer flanking contexts, and further emphasized the model's potential to help understand the functioning of DNA in biology. Another limitation is that the cgNA+ model treats sugar rings implicitly and does not provide any information on sugar pucker modes and backbone conformations. Notably, unlike phosphate and base groups, the sugar groups can not be treated as rigid bodies due to their high intrinsic flexibility. Therefore, we have introduced a machine learning approach that predicts the position of all sugar atoms from the knowledge of position of neighboring phosphate and base atoms which can be obtained from the cgNA+ model.

This thesis has been divided into eight chapters. The first two chapters are dedicated to background material with basic details of nucleic acids in chapter 1 and the description of the prior cgDNA+ model in chapter 2 including the discussion of model training, various methods to assess the model, and the Monte Carlo code to sample from the predicted Gaussian pdf. In chapter 3, we have reported the MD simulation protocol and the training library used to parameterize the model. Furthermore, we have analyzed the various aspects of MD simulations performed for various nucleic acids. In particular, we have examined the convergence of MD simulations and the distributions of cgNA+ internal coordinates for various nucleic acids. In chapter 4, we have introduced the cgNA+ model with a systematic assessment of the model's predictive capabilities. Furthermore, we have applied the cgNA+ model to explore various applications of the cgNA+ model, such as computing persistence length and groove widths, and comparing the statistics for DNA, RNA, and DRH. Moreover, we have also compared cgNA+ predictions with the available protein-DNA X-ray structure database in chapter 5. In chapter 6, we have discussed the extension of the cgNA+ model for epigenetically modified bases, along with the illustration of various applications. In chapter 7, we have introduced a neural network module to predict the location of sugar atoms in any cgNA+ coarse-grained configuration from the knowledge of neighboring base and phosphate atomic positions. It allows computing sequence-dependent backbone conformations and sugar pucker modes for an arbitrary DNA sequence where the positions of base and phosphate atoms can be accurately obtained from the cgNA+ model. Finally, chapter 8 discusses the conclusions of this thesis, outlines limitations, and proposes directions for future work.

1.1 Nucleic acids

In this section, we describe the basic structural and chemical aspects of various NAs modeled in this thesis. A NA is a polymer of repeating nucleotides composed of three basic units: phosphate, sugar, and bases. The sugar in NA is 5-carbon sugar (pentose) which is ribose and deoxyribose (as shown in figure 1.1) in the case of RNA and DNA, respectively. The sugar and phosphate are alternately connected through a phosphodiester bond and make the NA backbone (see figure 1.3). The phosphate group is attached to the $5'-$ and $3'-$ carbon of the pentose sugar that provides directionality to the NA, and the corresponding ends of the NA are called $5'-$ and $3'-$ ends. Bases, the third component of NAs, are nitrogenous base compounds that act as ele-

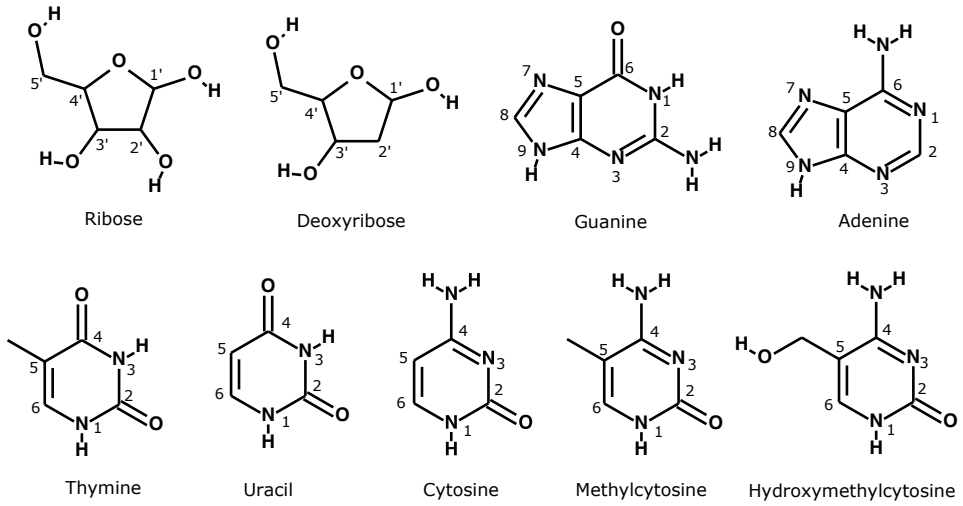


Fig. 1.1 Chemical structure and labeling of various sugar and bases in nucleic acids.

mental units of the genetic code. There are primarily two nucleobases categories: purine (R) and pyrimidine (Y) bases. Purine bases are larger with two fused ring structures, while pyrimidine has a single ring. The canonical purine bases are Adenine (A) and Guanine (G), while canonical pyrimidine bases are Thymine (T), Cytosine (C), and Uracil (U) (as shown in figure 1.1). The base is connected to the sugar, forming an N-glycosidic bond between base-nitrogen (N1 for Y and N9 for R) and $1'-$ carbon of the sugar ring.

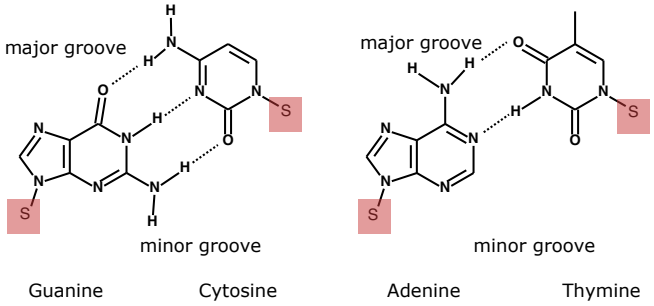


Fig. 1.2 Base-pairing and grooves in DNA

Double-stranded NA (dsNA), which is the primary focus of this thesis, is composed of two such anti-parallel chains which interact through the bases in which purine bases in one chain form hydrogen bonds with pyrimidine bases in other or vice-versa as shown in figure 1.2. The direction of one chain is $5'-$ to $3'-$ end while $3'-$ to $5'-$ end for the other. These interacting bases (through H-bonds) in the anti-parallel strands are complementary. The following discussion is relevant to dsNA.

Now, the structure of dsNA is defined at four levels, namely, primary, secondary, tertiary, and quaternary. The primary structure of a dsNA is defined as the list of nucleotides (denoted using the base name) read in the $5'-$ to $3'-$ end direction. To write down this sequence of nucleotides (S), one of the strands is selected as the reading (or Watson) strand, while the other strand is called the complementary (or Crick) strand. A dsNA oligomer comprising of N base-pairs as $S = X_1X_2X_3...X_{N-1}X_N$ where X_i are in the alphabets representing bases and the

sequence is written in the $5'$ to $3'$ direction. The complementary base of X_i on the Crick strand is $\bar{X}_i$. Following the notation, $\bar{S}$, is written as $\bar{X}_N\bar{X}_{N-1}\bar{X}_{N-2}...\bar{X}_2\bar{X}_1$ in the $5'$ to $3'$ direction on the Crick strand. The secondary structure defines the interactions between the bases and, thus, defines the basic shape of dsNA. For example, in canonical dsNA, the complementary base-pairing and wrapping of the two strands lead to a double-helical structure. There may be other types of secondary structures, such as bulges and loops but they are outside the scope of this thesis.

The main interest in this work is the tertiary structure of dsNA which refers to its intrinsic shape and flexibility. The tertiary structure includes key structural features, including handedness of the helix (left or right), length of helix per turn, the number of base-pairs per turn, and size of major and minor grooves. These features depend mainly on the physical conditions and the type and sequence of dsNA. Finally, the quaternary structure of dsNAs describes their interactions with other molecules, e.g., protein-DNA complexes or interactions of different RNAs in the ribosome. The following subsections provide basic information on the various dsNAs studied in this work.

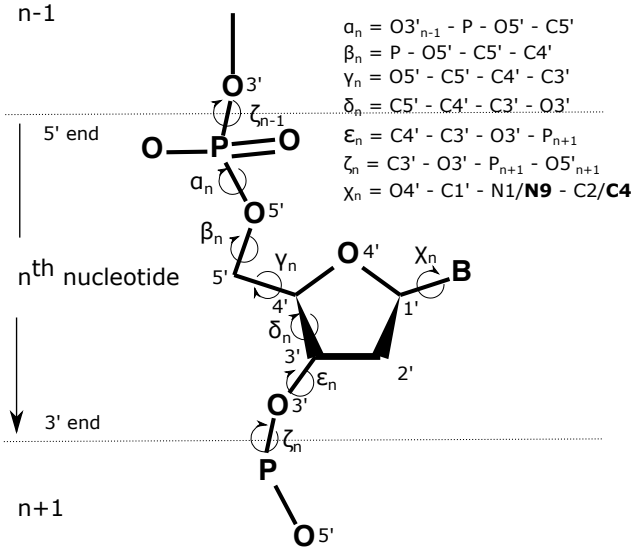


Fig. 1.3 DNA backbone and the torsional angles as defined in ref. [177]. For χ_n, the third and fourth atoms of the torsional angle depend on the kind of base. For pyrimidine bases, the atoms are N1 and C2, while for purine bases, the atoms are N9 and C4 as shown in bold. In the figure, the base is denoted as B.

1.1.1 Deoxyribonucleic Acid

The three components of a DNA nucleotide are deoxyribose sugar, phosphate, and bases (A, T, C, and G), as shown in figure 1.1. In the case of double-stranded DNA (dsDNA), complementary bases on opposite strands form H bonds, as shown in figure 1.2. In this base-pairing, C forms three H-bonds with G while A forms two H-bonds with T. Furthermore, the distance between the backbones of two strands is not symmetrical and forms two grooves of different sizes called major and minor grooves, as shown in figure 1.2. The major and minor grooves arise because the glycosidic bonds (base-sugar bonds) in the base-pair are not diametrically op-

posite. Notably, for a given base-pair, the minor groove contains O2 of Y and N3 of R, and the major groove is on the opposite side. Moreover, the methyl group of Thymine also lies in the major groove. Each groove consists of potential H-bond donor and acceptor atoms that enable specific interactions with proteins, and the sequence modulates their precise chemistry. Thus, the sequence-dependent groove widths play a crucial role in the protein-DNA interaction [134, 171, 172, 195].

The sugar-phosphate backbone is, in general, quite flexible and therefore requires more parameters to characterize its conformations. In figure 1.3, we have shown a typical DNA backbone and standard torsional angles used to characterize its conformation. It requires six torsional (or dihedral) angles (defined as the angle between planes passing through two sets of three atoms that have two atoms in common) to describe the backbone of DNA. In this case, the two sets of three atoms are consecutive first three and last three of the four covalently bonded atoms as described in figure 1.3. Furthermore, there is an additional torsional angle involving base and sugar atoms called χ. The involved base atoms depend on the type of base (R or Y) and are indicated in figure 1.3. χ torsional angle determines the nucleoside conformation, namely anti and syn. For anti conformation, $\chi \in [+90^{\circ} \cdots + 270^{\circ}]$ while for the rest of the χ values, nucleoside is in the syn conformation. Primarily, nucleoside stays in anti-conformation but sometimes base flips from anti-to-syn conformation, known as base-flipping. We would like to emphasize that the six torsional parameters in the DNA backbone are not entirely free to rotate and are quite restricted by steric constraints. In particular, there exist two common backbone conformations BI and BII, which are inter-convertible. These BI-BII conformations are found to play an important role in protein-DNA recognition [65]. BI or BII conformation is identified by using the difference between the torsional angle $\epsilon - \zeta$ which is negative for the BI conformation and positive for the BII conformation.

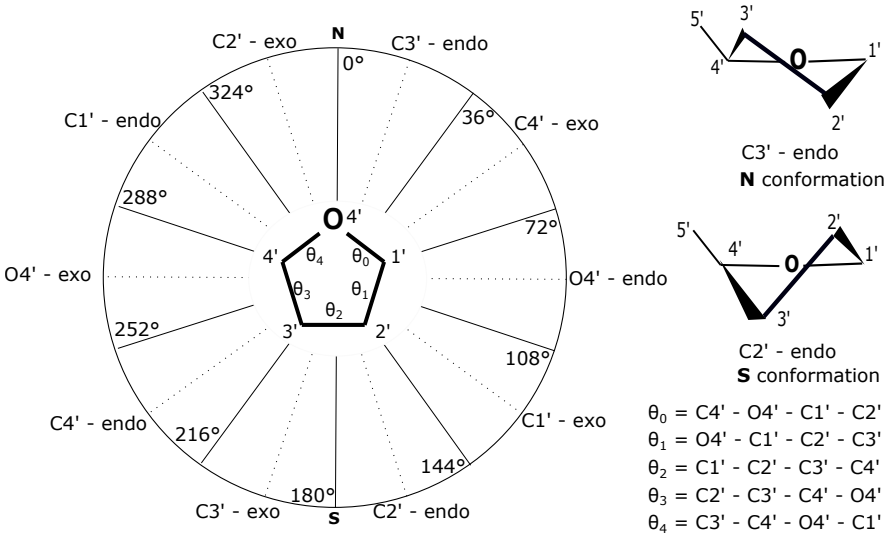


Fig. 1.4 Pseudorotation wheel (on the left) adopted from ref. [4] with sugar pucker notations defined based on the pseudorotation phase angle (**P**). **P** is computed using various dihedral angles $\theta_i \ \forall \ i \in [0, 1, 2, 3, 4]$ as given in equation (1.1) and the label of atoms from which dihedral angles are computed is shown in the figure. The two most common conformations adopted by the sugar in DNA are shown on the left.

The deoxyribose sugar ring in DNA is inherently non-planar (due to ring strains) and adopts puckered conformations as shown in figure 1.4. To fully describe the conformations of the sugar ring, five torsional angles are required. The different ring torsional angles give rise to different puckered conformations, and the thermodynamic stability of these conformations is governed by the substituents on the ring carbon atoms and how far those substituents are from each other in a given puckered conformation. These puckered conformations can inter-convert into each other and can be concisely explained using two parameters [4], namely, pseudorotation phase angle (**P**) and maximum degree of pucker, θ_{max}

$$tan(\mathbf{P}) = \frac{(\theta_4 + \theta_1) - (\theta_3 + \theta_0)}{2\theta_2\,(sin(36^\circ) + sin(72^\circ))} \text{ and } \theta_{max} = \frac{\theta_2}{cos(\mathbf{P})} \tag{1.1}$$

where **P** can be anything between $0 - 360^\circ$ and if $\theta_2 < 0$ then $\mathbf{P} = \mathbf{P} + 180^\circ$. Now, puckered conformations corresponding to **P** can be best represented on a pseudo-rotation wheel, as shown in figure 1.4 with the description of $\theta_i \ \forall \ i \in [0, 1, 2, 3, 4]$.

One of the most common pucker types is a conformation in which two of the ring atoms are out of the plane (on either side) formed by the other three atoms. The name of such a conformation is based on the direction of major deviation of a non-planar atom, which is if on the opposite side as $C4'$–$C5'$ bond and base, then the atom is called exo; otherwise, if it occurs in the same direction, then endo. The most frequently observed conformations are $C2'$–endo (with **P** values $\in 140$ to 185°) or $C3'$–endo (with **P** values $\in -10$ to $+40^\circ$) [127] and are shown in figure 1.1. Even though various pucker modes of deoxyribose are inter-convertible, $C3'$–endo pucker conformation is primarily dominated in the case of A-form DNA while $C2'$–endo in case of B-form DNA. In these two conformations, $C2'$–endo and $C3'$–endo, the relative distance between the $3'$ and $5'$ phosphates as well as the orientation of the phosphate group with respect to base/sugar is significantly different, giving rise to very different A- and B-form of DNA.

The DNA conformations are a function of the sequence, direction and amount of super-coiling, and the physical conditions of the solution. For example, the transition of B-DNA to A-DNA can be promoted under reduced humidity or by adding organic solvents. DNA also exists in a few other forms, such as Z-DNA, which is a left-handed helical structure form in alternating RY tracts under high salt, the presence of certain divalent cations, or DNA super-coiling. Z-DNA is structurally quite different from B-DNA in terms of sugar puckering, glycosidic bond configuration, and relative bp orientation. A concise summary of various DNA helical structures can be found in [202].

Compared to the DNA backbone, the bases are quite rigid and planar because of conjugation in the rings. We have heavily exploited this property of the bases by coarse-graining them. More details on this are discussed later in this thesis.

1.1.2 Ribonucleic Acid

The key chemical differences in RNA from DNA are a) the sugar molecule is ribose instead of deoxyribose and b) the pyrimidine base is U instead of T, as shown in figure 1.1. The ribose sugar preferably adopts $C3'$–endo conformations, and the preferred geometry of the RNA is the A-form. Discussion concerning the presence of two grooves, backbone, and sugar pucker

modes analysis is analogous to DNA.

1.1.3 Epigenetic modifications in DNA

Most common epigenetic modifications in DNA are methylation or hydroxymethylation of Cytosine at the $5-$ position, as shown schematically in figure 1.1. Other base modifications, such as 5-formyl-C, 5-carboxyl-C, and N6-methyl-A, are comparatively rare in biology and are not studied in this work. Most often, Cytosine substitution occurs at CpG dinucleotide steps[77], which can be di-substituted if both strands are symmetrically modified or hemi-substituted if only one of the strands is asymmetrically modified. In this thesis, we have used the letter M for 5-methylated-Cytosine, and N for Guanine when the complementary Cytosine is methylated. Similarly, the letters H and K are used for 5- hydroxymethylated-Cytosine and Guanine complementary to 5-hydroxymethylated-Cytosine, respectively.

1.1.4 DNA-RNA hybrid

A double-stranded DNA-RNA hybrid (abbreviated as DRH in this thesis) has one DNA strand and another RNA strand. We always take the DNA strand as the Watson or reading strand for simplicity in writing and coding. DRHs are important intermediates in many biological processes [2, 24, 121, 167, 185, 204]. Compared to DNA and RNA, the structure and mechanics of DRH are much less explored. Initial crystallographic studies indicated that DRH adopts a A-form geometry [206] but soon challenged by several other experimental techniques such as fiber diffraction [212], circular dichroism [168], NMR [54, 58, 96, 198] finding a mix A- and B-form geometries in DRH. Moreover, NMR results suggested that the DNA strand adopts a geometry closer to the B-form while the RNA strand is close to the A-form geometry. Discussion concerning the presence of two grooves, backbone, and sugar pucker modes analysis is analogous to DNA.

1.2 Methods

1.2.1 Sequence logos

Sequence logos [178] are a popular graphical representation to plot sequence characteristics in DNA, RNA, or protein sequences. It is widely used to visualize sequence motifs in multiple sequence alignments. In sequence logos, each position has a stack of characters (A/T/C/G) on top of each other, with the character height proportional to the relative frequency of the nucleic base at that position, and the total height of the stacks tells the information content at that position. In this work, we have exploited this visualization technique to understand the role of sequence in dsNA mechanics.

In standard sequence logos, the height of the stack is the information content (in bits), which

for nucleic acids (say DNA) is given as

$$\begin{aligned} H_i &= \sum_b f_{b,i} \times log_2(f_{b,i}) \\ R_i &= 2 - H_i - e \\ h_{b,i} &= f_{b,i} \times R_i \end{aligned} \tag{1.2}$$

where $b \in \{\text{A, T, C, G}\}$, i is the i^{th} position in the sequence, H_i is the uncertainty or Shannon entropy at position i, $f_{b,i}$ is the frequency of b^{th} base at position i, R_i is the total information at position i defined as the loss in uncertainty, and lastly, $h_{b,i}$ tells the height of base b at position i. e is a small sample correction taken as zero in this work. Note that the maximum value of R_i can be 2, which implies no uncertainty at that position, and the minimum value is 0, which implies the highest uncertainty at that position, i.e., all possible b are equally probable.

To better explain sequence logos, we have generated a set of sequences of length five bps. In figure 1.5(top), we have plotted a variant of sequence logos with the y-axis as the probability of alphabet b at i^{th} position. In the bottom plot, we have plotted the corresponding standard sequence logos with information content (in bits) on the y-axis. From the probability plot, it can be observed that the uncertainty in base type increases from left to right. At position-1: all sequences have A, at position-2: 75% G and 25% C, and at position-5: all alphabets are equally probable. In the corresponding information content plot, left to right, the information decreases with position-1 containing the maximum and position-5 containing zero information.

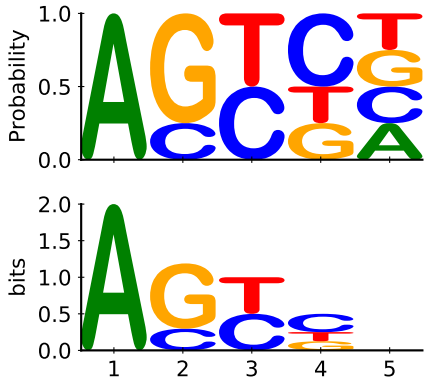


Fig. 1.5 Sequence logos plot for an artificial dataset with probability (top) and the information content (bottom) on the y-axis and base position in the sequence on the x-axis.

1.3 Codes and data availability

Details of all the codes and data used in this thesis are provided appendix F.

CHAPTER 2

cgDNA+ model

The central theme of this chapter is to provide an overview of the cgDNA+ model and its training [149]. First, we have discussed the coarse-graining of DNA atomistic configuration by fitting frames in phosphate and base units and then defining internal coordinates used in the cgDNA+ model. The next part introduces the cgDNA+ model, the underlying assumptions, and mathematical details of how the model reconstructs groundstate and stiffness matrix given a parameter set and a sequence. The subsequent section describes the cgDNA+ model training from atomistic molecular dynamics (MD) times-series of a set of sequences. After model description and training, this chapter discusses methods to assess the model performance and quantify various approximation errors in the model. The last part of the chapter is a brief discussion on the cgDNA+ Monte Carlo code [123], which allows efficient sampling of cgDNA+ predicted Gaussian probability density function (pdf) and thus, computing the expectation of any interesting physical observables for an ensemble of configurations.

2.1 Coarse-graining of atomistic structure of dsDNA

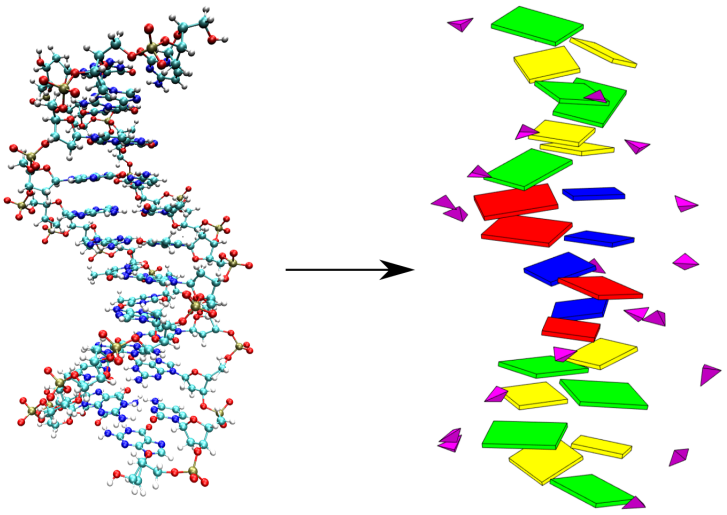


Fig. 2.1 Coarse-graining of atomistic structure of DNA oligomer to cgDNA+ cartoon representation by embedding frames in bases and phosphates.

This section describes the methodology to coarse-grain the atomistic representation of DNA observed in MD simulations by fitting frames in bases and phosphates. A typical atomistic

1	Compute weighted centroids	$\bar{\mathbf{x}} = \frac{\sum_{i=1}^{n} w_i \mathbf{x}_i}{\sum_{i=1}^{n} w_i}$ and $\bar{\mathbf{y}} = \frac{\sum_{i=1}^{n} w_i \mathbf{y}_i}{\sum_{i=1}^{n} w_i}$
2	Compute centred vectors	$\mathbf{p}_i := \mathbf{x}_i - \bar{\mathbf{x}}$ and $\mathbf{q}_i := \mathbf{y}_i - \bar{\mathbf{y}}$
3	Compute covariance matrix	$\mathbb{R}^{3\times3} \ni S = PWQ$ where $P, Q \in \mathbb{R}^{3\times n}$ with p_i, q_i as their columns, respectively and $W = diag(w_1, w_2, ..., w_n)$
4	Compute singular value decomposition	$S = U\Sigma V^T$
5	Compute $R, \mathbf{r}$	$R = V \begin{bmatrix} 1 & & \\ & 1 & \\ & & \lvert VU^T \rvert \end{bmatrix} U^T$, $\mathbf{r} = \bar{\mathbf{y}} - R\bar{\mathbf{x}}$

Table 2.1 Algorithm to find a rigid body transformation (translation r and rotation R) that best aligns two rigid bodies $X(x_1, x_2, ..., x_n)$ and $Y(y_1, y_2, ..., y_n)$ in terms of least-squares error, where $W(w_1, w_2, ..., w_n)$ is the weight matrix. This algorithm is used to fit frames in bases and phosphates.

structure of dsDNA (shown in figure 2.1) is an anti-parallel double-stranded helix whose primary structure is defined as the list of nucleotides denoted by the base name along a chosen strand from $5'-$ to $3'-$ direction. One strand is chosen as a reading strand or Watson strand, and the other is a complementary strand or Crick strand. Using this notion, we can write the sequence of a DNA oligomer comprising of N base-pair as $S = X_1X_2X_3...X_{N-1}X_N$ where $X_i \in$ {A, T, C, G}. The complementary base of X_i on the Crick strand is written as $\bar{X}_i$. Following the notation, $\bar{S}$, is written as $\bar{X}_N\bar{X}_{N-1}\bar{X}_{N-2}...\bar{X}_2\bar{X}_1$ in $5'$ to $3'$ direction on the Crick strand.

The cgDNA+ model explicitly treats phosphates and bases as rigid units and fits $SE(3)$ frame (see appendix section C.3) to describe their position ($\mathbf{r} \in \mathbb{R}^3$) and orientation ($R \in SO(3)$). In order to fit the $SE(3)$ frame, we have used cgFrame, a C++ code, analogous to CURVE+ [101], and based on the algorithm described in ref. [191]. cgFrame fits frames to bases and phosphates in the atomistic structure of DNA with respect to the ideal atoms definition formalized in the Tsukuba convention [140]. The details of the ideal coordinates are provided in table A.1. cgFrame input is a .PDB or .nc (binary) format of the MD trajectory and the outputs are two text files, namely .fra and .pfra files, containing frames, $(R, \mathbf{r}) \in SE(3)$, for bases and phosphates, respectively.

The general fitting problem can be stated as, let $X(x_1, x_2, ..., x_n)$ and $Y(y_1, y_2, ..., y_n)$ be two sets of corresponding points in $\mathbb{R}^3$ and $R \in SO(3)$ and $\mathbf{r} \in \mathbb{R}^3$ be a rotation matrix and translation vector, respectively, aligning X and Y such that

$$\sum_{i=1}^{n} w_i \left\| \mathbf{r} + R\mathbf{x}_i - \mathbf{y}_i \right\|^2, \tag{2.1}$$

is minimum. $w_i > 0$ represents the weight for a particular pair of points (for cgDNA+, $w_i = 1 \; \forall \; i = 1 \cdot\cdot n$). The algorithm [191] to find $R, \mathbf{r}$ is summarized in table 2.1. $(R, \mathbf{r})$ represents the rigid body transformation and forms an element of the special Euclidean group $SE(3)$. More details on $SE(3)$ are provided in appendix section C.3.

Using the algorithm described in table 2.1, we have obtained best-fit frames for bases and phosphates of a given MD snapshot and transformed atomistic MD time-series into a coarse-grain time-series where each snapshot is in $SE(3)^{4N-2}$ (N is the number of base-pairs in the given oligomer). Note that the first $5'-$phosphates on both the strands are not present in the MD time-series, thus, $4N-2$. A typical coarse-grain cartoon for an atomistic MD snapshot is shown in figure 2.1 which can be perceived as a tetra-chain representation of DNA.

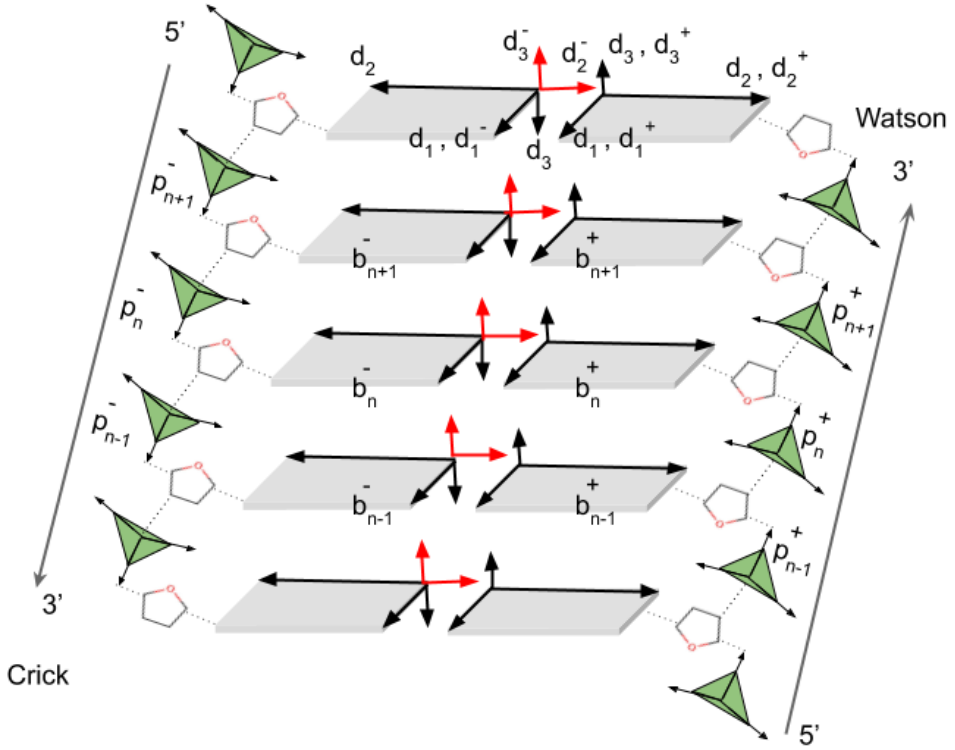


Fig. 2.2 A schematic view of coarse-grain dsDNA with rigid bases and rigid phosphates. The sugar molecule is shown in the image but is modeled only implicitly in the cgDNA+ model. $\{d_1, d_2, d_3\}$ is the orthonormal frame as per Tsukuba convention [140] while for modeling purposes we flip the Crick frame to align with Watson frame to give the final orientations of Watson and Crick frame as $\{d_1^+, d_2^+, d_3^+\}$ and $\{d_1^-, d_2^-, d_3^-\}$, respectively.

This frame fitting leads to a reference point $\mathbf{r}$ and a right-handed, orthonormal frame $\mathbb{R}^{3\times3} \ni D = \{d_1, d_2, d_3\}$ to each base as per Tsukuba convention [140]. The vector d_1 is in the direction of major-groove, d_2 in the direction of the reference strand or away from the complementary base and, $d_3 = d_1 \times d_2$ and is approximately in the direction pointing from base n to $n+1$ while reading from that strand. It implies that the frames associated with the two bases in a given base-pair are not aligned, as shown in figure 2.2 in black color and denoted as $\{d_1, d_2, d_3\}$. However, in the context of dsDNA modeling, it is convenient to model if both bases in the base-pair are aligned in the same direction. Therefore, we have introduced a matrix $P_{\text{flip}} \in O(3)$,

$$P_{\text{flip}} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix} \tag{2.2}$$

which flips d_2 and d_3 directions of Crick frame. It gives the final orientation of Crick frame as $D^- = DP_{\text{flip}} = \{d_1, -d_2, -d_3\} = \{d_1^-, d_2^-, d_3^-\}$ and Watson frame as $D^+ = D = \{d_1, d_2, d_3\} = \{d_1^+, d_2^+, d_3^+\}$. Now, $d_1^\pm, d_2^\pm$, and $d_3^\pm$ are in the direction of major-groove, Watson (chosen as reading) strand, and n to $n+1$ base while reading from the Watson strand, respectively. Note that after flipping base frames associated with the Crick strand, we have denoted base and phosphate frames on the Watson strand (chosen as the reading strand) with + superscript and on the Crick strand with − superscript.

Similarly, we have defined the position of phosphate atom as the reference point, $r^p \in \mathbb{R}^3$ and three vectors, $\mathbb{R}^{3\times3} \ni D^p = \{d_1^p, d_2^p, d_3^p\}$ in the direction of $d_3^p = \text{O5}' - \text{O3}'$, $d_2^p = \text{P} - (\text{O5}' + \text{O3}')/2$ and, $d_1^p = d_2^p \times d_3^p$. More details on the ideal coordinates for the phosphate are in table A.1. A schematic diagram of coarse-grain bases and phosphates is shown in figure 2.2.

2.2 Internal coordinates

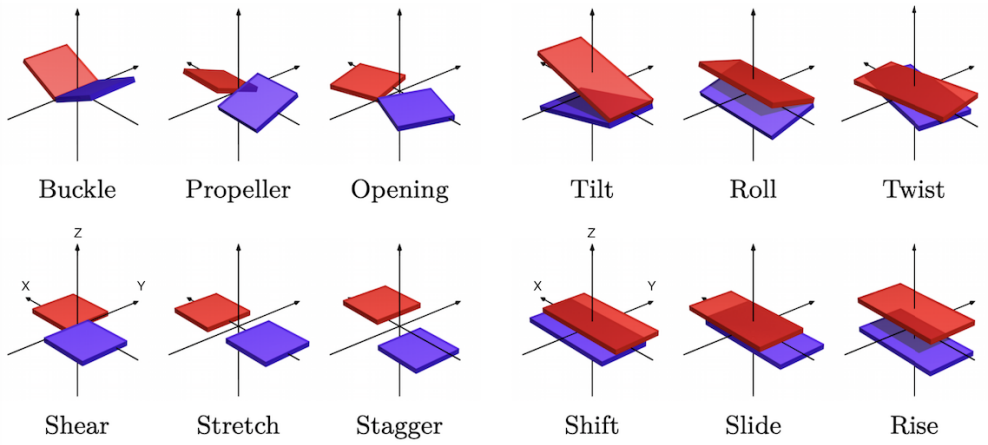


Fig. 2.3 CURVES+ coordinates for a coarse-grain DNA configuration. Intra base-pair (left) and Inter base-pair (right). X, Y, and Z are in the direction of the reading strand, major-groove, and from base n to n + 1 while reading the sequence from the reference strand, respectively.

The frames obtained in the last section are in absolute coordinates, i.e., in a fixed lab frame, which are challenging to work with due to global translations and rotations of the whole molecule during the MD simulation. A convenient alternative is to use a coordinate system fixed on the dsDNA molecule itself, i.e., internal coordinates. In the following discussion, we have first introduced the internal coordinates for the bases and then for the phosphates while reading the sequence ($5'$ to $3'$ direction) from a chosen strand. The cgDNA+ internal coordinates for the same configuration while reading the sequence from two different strands are different but related by a linear transformation. This linear transformation for change of reading strand is reviewed in section 2.2.2. Furthermore, the inverse transformation from cgDNA+ internal coordinates to the absolute frames is detailed in section 2.2.3. Lastly, we discussed H-bond filtering to discard MD snapshots with broken H-bonds.

2.2.1 Frames to cgDNA+ internal coordinates

We first defined the base-pair and the junction frames to introduce internal coordinates for bases. The base-pair frame is defined as the average orientation (B) of two complementary base frames of a base-pair with an average reference point, g (equation (2.3)). While the junction frame is defined as the average orientation (J) of the two neighboring base-pair frames (n^{th} and $(n+1)^{\text{th}}$) with average reference point t given in equation (2.4).

$$\{B_n, g_n\} = \{D_n^- \sqrt{\Lambda_n}\ ,\ \frac{1}{2}(r_n^+ + r_n^-)\} \quad \text{where} \quad \Lambda_n = (D_n^-)^T D_n^+ \tag{2.3}$$

$$\{J_n, t_n\} = \{B_n \sqrt{\Gamma_n}\ ,\ \frac{1}{2}(g_{n+1} + g_n)\} \quad \text{where} \quad \Gamma_n = (B_n)^T B_{n+1} \tag{2.4}$$

where Λ_n describes the relative orientation of the base frame D_n^+ with respect to D_n^- and Γ_n describes the relative orientation of base-pair frame B_{n+1} with respect to B_n.

Now, the internal coordinates for the bases are classified into intra and inter base-pair coordinates. The rotational components of intra base-pair coordinates are defined as Cayley parameters (refer appendix section C.2) of the relative rotation matrix Λ_n and the intra translational coordinates, ζ_n are defined in the reference frame of base-pair frame B_n (equation (2.3)) and are given as,

$$\mathbb{R}^6 \ni x_n = \{\lambda_n,\ \zeta_n\} = \{cay_\alpha^{-1}(\Lambda_n),\ B_n^T(r_n^+ - r_n^-)\} \tag{2.5}$$

where λ_n and $\zeta_n \in \mathbb{R}^3$ and $cay_\alpha^{-1}(\cdot)$ is defined in appendix equation (C.6) and $\alpha \in \mathbb{R}$ is a scaling parameter which is discussed later in this section. For the cgDNA+ model, we have chosen $\alpha = 5$. The inter base-pair coordinates are defined between two neighboring base-pairs. The rotational component $\gamma_n \in \mathbb{R}^3$ is the Cayley parameters of the relative rotation matrix Γ_n between n^{th} and $(n+1)^{\text{th}}$ base-pairs. While the translational coordinates $\eta_n \in \mathbb{R}^3$ are defined as the relative translation between two neighboring base-pair frames but in reference of the junction frame J_n. The inter base-pair coordinates are given as,

$$\mathbb{R}^6 \ni y_n = \{\gamma_n,\ \eta_n\} = \{cay_\alpha^{-1}(\Gamma_n),\ J_n^T(g_{n+1} - g_n)\}. \tag{2.6}$$

Notably, intra rotational and translational coordinates are commonly known as (Buckle, Propeller, Opening) and (Shear, Stretch, Stagger), respectively, and inter rotational and translational coordinates are commonly known as (Tilt, Roll, Twist) and (Shift, Slide, Rise), respectively. A schematic diagram for intra and inter coordinates is depicted in figure 2.3.

The internal coordinates for a given phosphate are defined relative to the base to which the phosphate is attached. It can be defined in two ways: base to $3'-$ phosphate and base to $5'-$ phosphate. cgDNA+ model adopts base to $5'-$ phosphate as this parameterization is through the covalent bond between base and phosphate, which involves BI-BII backbone conformations. Both parameterizations were investigated in ref. [149] with the findings that base to $5'-$ phosphate parameterization can capture bi-modality in the backbone conformation. This parameterization is essentially defining coordinates for the phosphate with respect to the corresponding base in the same nucleotide. More details about the choice of parameterization can be found in A. Patelli's thesis [149]. Finally, the internal coordinates for phosphate are defined in reference

to the corresponding base and are given as

$$\mathbb{R}^6 \ni z_n^{\pm} = \{\tau_n^{\pm},\ \xi_n^{\pm}\} = \{cay_{\alpha}^{-1}(D_n^{\pm}D_n^{p\pm}),\ D_n^{\pm T}(r_n^{p\pm} - r_n^{\pm})\}. \tag{2.7}$$

Thus, the coarse-grain configuration of dsDNA oligomer of length N bps in terms of internal coordinates (independent of global translation and rotation of the molecule) can be written as a vector w in $24N - 18$ dimensions made up of $6N$ intra base-pair (x_n), $6N - 6$ inter base-pair (y_n) and $12N - 12$ base to $5'$−phosphate coordinates ($z_n^{\pm}$) given in equation (2.8),

$$\mathbb{R}^{24N-18} \ni w = (x_1, z_1^-, y_1, z_2^+, x_2, z_2^-, ..., y_{n-1}, z_n^+, x_n), \tag{2.8}$$

where ± denote Watson and Crick strands, respectively. The transformations in equations (2.3) to (2.7) that transform bases and phosphates frames into internal coordinates for a given dsDNA configuration is denoted by $T_{F\rightarrow I} : SE(3)^{4N-2} \rightarrow \mathbb{R}^{24N-18}$.

Lastly, we have chosen a scale of 1 Å for translational coordinates and rad/5 for rotational coordinates and transformed the coordinates into dimensionless form and scaled them so that the magnitudes of the rotational and translational coordinates can be compared more directly. More details on the scaling parameters can be found in refs. [62, 158, 159].

2.2.2 Change of reading strand

In the previous section, we introduced cgDNA+ internal coordinates to describe the configuration of dsDNA while reading the sequence from a particular strand (chosen first). However, the same dsDNA configuration read from the Watson strand (S) or the Crick strand ($\bar{S}$) leads to two different cgDNA+ internal coordinates, which are related by a linear map [149, 158] as,

$$\begin{aligned} w(\bar{S}) &= E_N w(S) \\ \mathcal{K}(\bar{S}) &= E_N \mathcal{K}(S) E_N \end{aligned} \tag{2.9}$$

where

$$\mathbb{R}^{24N-18\times 24N-18} \ni E_N = \begin{bmatrix} & & & & E^{5'} \\ & & & E^{int} & \\ & & \iddots & & \\ & E^{int} & & & \\ E^{3'} & & & & \end{bmatrix} \tag{2.10}$$

and $\mathcal{K}$ is the stiffness matrix discussed in section 2.3. The entries in E_N are given as

$$\mathbb{R}^{36\times 36} \ni E^{5'} = \begin{bmatrix} & & & & & E \\ & & & & I_6 & \\ & & & E & & \\ & & I_6 & & & \\ & E & & & & \\ I_6 & & & & & \end{bmatrix}, \tag{2.11}$$

$$\mathbb{R}^{36\times 36} \ni E^{3'} = [E^{5'}]^{-1} = [E^{5'}]^T, \text{ and} \tag{2.12}$$

$$\mathbb{R}^{42\times42} \ni E^{int} = \begin{bmatrix} & & & & & & I_6 \\ & & & & & E & \\ & & & & I_6 & & \\ & & & E & & & \\ & & I_6 & & & & \\ & E & & & & & \\ I_6 & & & & & & \end{bmatrix} = \left[E^{int}\right]^T = \left[E^{int}\right]^{-1} \tag{2.13}$$

where $E = \text{diag}(-1, 1, 1, -1, 1, 1) \in \mathbb{R}^{6\times6}$ and I_6 is a 6×6 identity matrix.

2.2.3 cgDNA+ internal coordinates to frames

The inverse transformation to obtain the position and orientation of base and phosphate frames from cgDNA+ internal coordinates is denoted as $T_{I\to F} : \mathbb{R}^{24N-18} \to SE(3)^{4N-2}$. For a given configuration in cgDNA+ internal coordinates, $w = (x_1, z_1^-, y_1, z_2^+, x_2, z_2^-, \cdots, y_{n-1}, z_n^+, x_n) \in \mathbb{R}^{24N-18}$ where each component can further be split into rotational and translational components as $x_n = (\lambda_n, \zeta_n), y_n = (\gamma_n, \eta_n),$ and $z_n^\pm = (\tau_n^\pm, \xi_n^\pm)$, bases and phosphates frames can be obtained in three steps. First the n^{th} base-pair frame (B_n, g_n) can be obtained from the corresponding inter coordinates (y_{n-1}) using recursive relation as

$$\begin{aligned} \begin{bmatrix} B_n & g_n \\ \mathbf{0}^T & 1 \end{bmatrix} &= \begin{bmatrix} B_{n-1} & g_{n-1} \\ \mathbf{0}^T & 1 \end{bmatrix} \begin{bmatrix} cay_\alpha(\gamma_{n-1}) & cay_\alpha(\gamma_{n-1})\eta_{n-1} \\ \mathbf{0}^T & 1 \end{bmatrix} \\ &= \prod_{i=1}^{n-1} \begin{bmatrix} cay_\alpha(\gamma_i) & cay_\alpha(\gamma_i)\eta_i \\ \mathbf{0}^T & 1 \end{bmatrix} \end{aligned} \tag{2.14}$$

where $cay_\alpha(\cdot)$ is the Cayley transformation defined in equation (C.5) with α as the scaling factor discussed in section 2.2.2, and (B_1, g_1) is taken as $(\mathbf{I}, \mathbf{0})$ with $\mathbf{I} \in \mathbb{R}^{3\times3}$ an identity matrix and $\mathbf{0} \in \mathbb{R}^{3\times1}$ a zero vector. Note that absolute coordinates for frames require a reference point, which is chosen to be the first base-pair frame. Subsequently, base and phosphate frames can be obtained as

$$\begin{bmatrix} D_n^\pm & r_n^\pm \\ \mathbf{0}^T & 1 \end{bmatrix} = \begin{bmatrix} B_n(cay_\alpha(\lambda_n))^{\pm\frac{1}{2}} & g_n \pm \frac{1}{2}B_n\zeta_n \\ \mathbf{0}^T & 1 \end{bmatrix} \tag{2.15}$$

$$\begin{bmatrix} D_n^{p\pm} & r_n^{p\pm} \\ \mathbf{0}^T & 1 \end{bmatrix} = \begin{bmatrix} D_n^\pm L^\pm cay_\alpha(\tau_n^\pm) & g_n^\pm + \frac{1}{2}D_n^\pm L^\pm \xi_n^\pm \\ \mathbf{0}^T & 1 \end{bmatrix} \tag{2.16}$$

where $L^+ = \mathbf{I}$, and $L^- = P_{\text{flip}}$ (see equation (2.2)).

Lastly, to obtain the atomistic PDB structure one can re-embed ideal atom coordinates (table A.1) in bases and phosphates frames using the transformation $T_{F\to C} : SE(3) \to \mathbb{R}^{3\times K}$. For a given frame with (R, r) as orientation and position, the positions of the associated atoms can be obtained as,

$$\mathcal{C}_{X^k} =: R\mathcal{A}_{X^k} + r, \ \forall\ k = 1 \cdots K \tag{2.17}$$

where K is the number of atoms in base or phosphate, X is kind of rigid body (base or phosphate), $\mathcal{A}_{X^k} \in \mathbb{R}^{3\times1}$ is the coordinate of k^{th} ideal atom in X type rigid body, and $\mathcal{C}_{X^k} \in \mathbb{R}^{3\times1}$ is

the coordinate of k^{th} atom embedded in the frame.

2.2.4 H-bond filtering

We have used the Cayley vector to parameterize rotational coordinates between two rigid bodies, whose norm tends to infinity if the relative rotation between the rigid bodies is close to π. This case can happen in MD time-series due to broken H-bonds, especially toward the ends of a dsDNA molecule. Such cases may lead to massive bias in the statistics of the rotational internal coordinates, which may create optimization issues in model training. Moreover, MD snapshots after broken H-bond filtering may better represent the dsDNA properties than the raw data. Thus, we have introduced a filtering step that removes snapshots with one or more broken H-bonds. We have declared (consistent with [62, 99, 109, 159]) an H-bond broken if a) distance between the heavy atoms involved in the H-bond is greater than four Å or b) the angle between the heavy atoms via H-atom is less than $120°$. The latter condition, in general, can not be applied to dsDNA data obtained from X-ray techniques (as the H atom is missing); therefore, only the first criterion is used [109].

2.3 cgDNA+ model

cgDNA+ model is a predictive model that given a sequence S of length N bps along the reading strand and a parameter set $\mathcal{P}$ delivers a Gaussian pdf in configuration space by reconstructing a ground-state $\hat{w}(S,\mathcal{P}) \in \mathbb{R}^{24N-18}$ and a positive-definite stiffness matrix $\mathcal{K}(S,\mathcal{P}) \in \mathbb{R}^{24N-18\times 24N-18}$:

$$\rho(w;S,\mathcal{P}) = \frac{1}{Z} exp\{-\frac{1}{2}(w-\hat{w})\cdot\mathcal{K}(w-\hat{w})\}. \tag{2.18}$$

In cgDNA+ model, a parameter set $\mathcal{P}$ for dsDNA in standard alphabets is made up of a) ten independent interior dinucleotide-dependent $\mathcal{K}^{\text{XY}}$ blocks $\in \mathbb{R}^{42\times 42}$ plus sixteen independent $\mathcal{K}^{5'\text{XY}}$ end blocks $\in \mathbb{R}^{36\times 36}$, and b) ten independent interior dinucleotide-dependent stress vectors $\sigma^{\text{XY}} \in \mathbb{R}^{42}$ plus sixteen independent $\sigma^{5'\text{XY}}$ end stress vectors $\in \mathbb{R}^{36}$,

$$\mathcal{P} = \{\sigma^{5'\text{XY}}, \sigma^{\text{XY}}, \mathcal{K}^{5'\text{XY}}, \mathcal{K}^{\text{XY}}\} \in \mathbb{P} = [\mathbb{R}^{36}]^{16} \times [\mathbb{R}^{42}]^{10} \times [\mathbb{R}^{36\times 36}]^{16} \times [\mathbb{R}^{42\times 42}]^{10}, \tag{2.19}$$

where $5'\text{XY} \in \{16 \text{ dimer steps}\}$ and $\text{XY} \in \{10 \text{ independent dimer steps}\}$. The parameter for $3'$ ends and dependent dimer steps can be obtained using Crick-Watson (CW) symmetry as;

$$\begin{aligned} \sigma^{\bar{\text{Y}}\bar{\text{X}}} &= E^{int}\sigma^{\text{XY}} \\ \mathcal{K}^{\bar{\text{Y}}\bar{\text{X}}} &= E^{int}\mathcal{K}^{\text{XY}}E^{int} \\ \sigma^{\bar{\text{Y}}\bar{\text{X}}3'} &= E^{5'}\sigma^{5'\text{XY}} \\ \mathcal{K}^{\bar{\text{Y}}\bar{\text{X}}3'} &= E^{5'}\mathcal{K}^{5'\text{XY}}E^{5'} \end{aligned} \tag{2.20}$$

Note that these symmetry conditions put additional constraints on the parameter blocks for palindromes $\text{X}\bar{\text{X}}$ (AT, TA, CG, and GC for dsDNA), which are exploited in the parameter set extraction. The details on the parameter set estimation are provided in section 2.4.

2.3.1 cgDNA+ model assumptions

The assumptions or approximations in the cgDNA+ model are the following:

i) MD time-series are stationary

ii) base and phosphates are rigid

iii) pdfs of internal coordinates follow Gaussian distribution, i.e., internal energy for any oligomer assumes a shifted quadratic (see equation (2.18))

iv) total energy of the dsDNA oligomer is approximated as the sum of local junction energies, i.e., nearest-neighbor interactions only,

$$U(w,S) = \frac{1}{2}(w-\hat{w}).\mathcal{K}(w-\hat{w}) \approx \frac{1}{2}\sum_{n=1}^{N-1}(w_n-\hat{w}_n).\mathcal{K}_n(w_n-\hat{w}_n) \tag{2.21}$$

where $w_n, \hat{w}_n,$ and $\mathcal{K}_n$ represents local junction energy contribution.

v) coefficients in the local junction energies depend on the local dimer step.

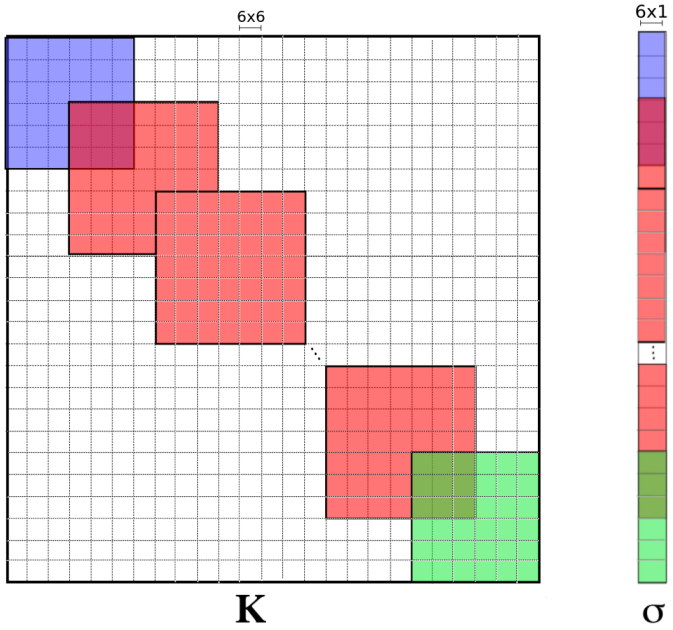


Fig. 2.4 Construction of banded oligomer stiffness matrix $\mathcal{K}$ and stress vector σ by overlapping dimer-step dependent parameter set blocks shown for poly(A). The parameters for $3'$ end, $5'$ end, and interior blocks are different and are shown in different colors. Each cell of the matrix is of dimension 6×6. Each cell in the vector is of dimension 6×1.

Solving equation (2.21) with appropriate algebraic considerations (for more details refer [158]), the groundstate for the dsDNA oligomer is given as

$$\hat{w}(S) = \mathcal{K}^{-1}(S)\sigma(S) \tag{2.22}$$

where $\mathcal{K}(S)$ and $\sigma(S)$ can be computed by overlaying dimer-dependent parameter blocks (as shown in figure 2.4 and discussed in section 2.2.4) and $\sigma_n = \mathcal{K}_n\hat{w}_n$. It leads to one of the most important features in the cgDNA+ model; that is, even though $\mathcal{K}$ matrix and σ vector have local sequence dependence, the groundstate configuration which involves the inversion of

overlapping banded stiffness matrix ($\mathcal{K}$) has a non-local (often strongly non-local) sequence dependence. Moreover, a non-zero constant energy term $\hat{U}(S)$ naturally arises in the solution of equation (2.21) reflecting that all bases and phosphates cannot simultaneously achieve absolute zero energy minima in the ground–state and achieve an equilibrium configuration with some non-zero frustration energy. This physical phenomenon of frustration observed here is only possible in double chain rigid base model (cgDNA) or higher hierarchy models like cgDNA+ but not in single-chain models such as rigid base-pair models.

2.3.2 cgDNA+ reconstruction

Given a sequence S and parameter set $\mathcal{P}$, the model $\mathcal{K}$ matrix and σ vector are constructed by overlaying dinucleotide-step parameter set blocks as shown in figure 2.4 and given as

$$\mathcal{K}(\mathcal{P},S) = R_d^T \mathcal{K}_d R_d \;\text{ and }\; \sigma(\mathcal{P},S) = R_d^T \sigma_d \tag{2.23}$$

where $\mathcal{K}_d = \text{diag}(\mathcal{K}^{5'X_1X_2}, ..., \mathcal{K}^{X_iX_{i+1}}, ..., \mathcal{K}^{X_{N-1}X_N3'})$ and $\sigma_d = (\sigma^{5'X_1X_2}, ..., \sigma^{X_iX_{i+1}}, ..., \sigma^{X_{N-1}X_N3'})$ and $R_d \in 42N - 12 \times 24N - 18$ is a matrix defined in equation (2.24).

$$R_d = \begin{bmatrix} I_{18} & & & & & \cdots & \\ & I_{18} & & & & & \\ & I_{18} & & & & & \\ & & I_6 & & & & \\ & & & I_{18} & & & \\ & & & I_{18} & & & \\ & & & & I_6 & & \\ \vdots & & & & & \ddots & \vdots \\ & & & & & & I_{18} \end{bmatrix} \tag{2.24}$$

where I_k is k-dimensional identity matrix.

Thus, given a sequence S and parameter set $\mathcal{P}$, we can define a reconstruction rule $\mathcal{R}(\mathcal{P},S)$ which reconstructs a banded stiffness matrix $\mathcal{K}(S)$ and a $\sigma(S)$ vector as

$$\mathcal{R}(\mathcal{P},S) = (\sigma(\mathcal{P},S), \mathcal{K}(\mathcal{P},S)). \tag{2.25}$$

It is important to note that the reconstruction rule $\mathcal{R}(\mathcal{P},S)$ is not invertible due to overlapping blocks in the stiffness matrix $\mathcal{K}$ and stress vector σ.

Moreover, we have defined $\mathcal{R}_{\text{vec}}(\mathcal{P}_{\text{vec}},S)$ which is equivalent to $\mathcal{R}(\mathcal{P},S)$ but allows a linear relation between $\sigma(\mathcal{P},S)$, $\mathcal{K}(\mathcal{P},S)$ and the vector form of parameter set $\mathcal{P}_{\text{vec}} \in \mathbb{R}^L$ where L is the total number of independent entries in the parameter set $\mathcal{P}$. It can be defined as

$$\mathcal{R}_{\text{vec}}(\mathcal{P}_{\text{vec}},S) = R_{\mathcal{P}}(S)\mathcal{P}_{\text{vec}} = (\sigma(\mathcal{P},S), \mathcal{K}(\mathcal{P},S)) \tag{2.26}$$

where $R_{\mathcal{P}}(S) \in N + N^2 \times L$ is parameter reconstruction matrix which maps the entries in parameter set in vector form $\mathcal{P}_{\text{vec}}$ to the entries of $\sigma(S)$ and $\mathcal{K}(S)$. This vectorized form is convenient to code and explain the mathematics of extracting the cgDNA+ parameter set.

2.4 cgDNA+ parameter set estimation

In this section, we have discussed the protocol for extraction of cgDNA+ model parameter set from MD time-series. This protocol was initially developed for the cgDNA model [159] and was optimized and extended for the cgDNA+ model [149]. The key steps in this protocol are:

i) long-enough atomistic MD time-series for a set of sequences referred to as the training library, Lb (discussed in section 3.3)

ii) fit rigid bodies in atomistic MD snapshots to obtain snapshots in base and phosphate frames as discussed in section 2.1 and discard snapshots with broken H-bond.

iii) transform base and phosphate frames into internal coordinates using $T_{F\rightarrow I}$ (see section 2.2)

iv) estimation of first and second moments (i.e., fit Gaussian pdf) for each oligomer in the training library, assuming that MD time-series are converged (section 2.4.1).

v) train dimer-dependent σ vector and $\mathcal{K}$ stiffness matrix (i.e., parameter set $\mathcal{P}$ as defined in equation (2.19)) using Gaussian pdfs obtained in the previous step. This step is explained in sections 2.4.2 to 2.4.4

vi) parameter set stiffness blocks obtained in last step may not be positive-definite, so search an element in null space to make stiffness blocks positive-definite (see section 2.4.5).

All the above steps in the parameter set estimation are briefly discussed in the following subsections and a detailed explanation can be found elsewhere [149].

2.4.1 Estimation of oligomer-level statistics

Once the internal coordinates are obtained for each MD snapshot for all sequences in the training library (Lb = $\{S_i\}_{i=1}^{L}$), the next step is to compute oligomer-level statistics, $\{\bar{w}(S), C(S)\}$. For a given MD time-series ($[w^m(S)]_{m=1}^{M}$) where m represents the m^{th} snapshot in the time-series and $M \sim 10^6$ is the total number of snapshots in the time-series, the first moment (mean) and second central moment (covariance matrix) can be computed as,

$$\begin{aligned} \bar{w}(S) &= \frac{1}{M}\sum_{m=1}^{M} w^m(S) \\ C(S) &= \frac{1}{M}\sum_{m=1}^{M} \left(w^m(S) - \bar{w}(S)\right)\left(w^m(S) - \bar{w}(S)\right)^T \quad \forall\ S \in \text{Lb}. \end{aligned} \tag{2.27}$$

Furthermore, we have exploited the CW symmetry of dsDNA (only if the sequence is a palindrome, i.e., $S = \bar{S}$) to enhance the quality of first and second centred moments by defining palindromically symmetrized estimators as:

$$\begin{aligned} \bar{w}_p(S) &= \frac{1}{2}(\bar{w} + E_N\bar{w}) \\ H(S) &= C + \bar{w}\bar{w}^T \\ H_p(S) &= \frac{1}{2}(H + E_N H E_N) \\ C_p(S) &= H_p - \bar{w}_p\bar{w}_p^T \end{aligned} \tag{2.28}$$

where H is the second moment and E_N defines a linear map for reading strand transformation (refer section 2.2.2). Note that all sequences used to train cgDNA+ model parameters are palindromes (see table B.1). Therefore, we have dropped the subscript notation for convenience, i.e., $\bar{w}_p(S)$ is written as $\bar{w}(S)$. Lastly, the cgDNA+ model only considers the first and second moments, as the computation of higher moments in such high dimensions $\in \mathbb{R}^{24N-18}$ is non-trivial.

Now, the maximum entropy principle [78, 79] allows computing the least biased probability distribution for precise prior data. The absolute entropy for a density $\rho(w)$ is given as:

$$D(\rho) = -\int_{\mathbb{R}^{24N-18}} \rho(w) \log \rho(w) dx. \tag{2.29}$$

By assumption, in our case, the probability distribution is a Gaussian pdf with $\bar{w}(S)$ and $C(S)$ $\forall\ S \in$ Lb as mean and covariance, respectively. We have computed the least biased observed Gaussian probability distribution $\rho_o(w; \bar{w}(S), C(S))$ using equation (2.29) under the following constraint,

$$\mathcal{C}(S) = \left\{\rho(w) | \langle 1 \rangle_\rho = 1, \langle w \rangle_\rho = \bar{w}, \langle (w - \bar{w})^T (w - \bar{w}) \rangle_\rho = C\right\} \tag{2.30}$$

Note that using maximum likelihood estimation to find the observed Gaussian pdf would have resulted in the same probability distribution. More details can be found in ref. [63, 149].

2.4.2 Definition of best-fit parameter set

In the previous step, we obtained oligomer-level Gaussian pdfs ($\rho_o(w; \bar{w}(S), C(S))$ abbreviated as $\rho_o(w; S)$) observed in MD time-series for all $S \in$ Lb. The best-fit cgDNA+ parameter set $\mathcal{P}$ to these observed Gaussian pdfs is defined as

$$\mathcal{P} = \underset{P \in \mathbb{P}}{\operatorname{argmin}} \sum_{i=1}^{L} D_{\mathrm{KL}}(\rho_P(w; S_i, P)), \rho_o(w; S_i)) \tag{2.31}$$

where L is the number of sequences Lb, D_{KL} is the Kullback-Leibler (KL) divergence defined in equation (C.11), $\rho_o = \rho_o(w; S)$ is the observed Gaussian obtained in the previous step from the raw MD data, and $\rho_P = \rho_P(w; P, S)$ is Gaussian pdf predicted by the parameter set (P) for a given sequence. $\mathbb{P}$ contains all admissible parameter sets for the cgDNA+ model in which σ^{XY} and $\mathcal{K}^{\mathrm{XY}}$ satisfy CW symmetry constraints for palindromic dimer steps as described in equation (2.20) and $\mathcal{K}^{\mathrm{XY}}$ reconstructs a positive definite stiffness matrix for any $S \in$ Lb. Lastly, note that two notations are used for parameter set, $\mathcal{P}$ is the best-fit parameter set while P is any parameter set in $\mathbb{P}$.

Note that KL divergence [94] is not symmetric (details in section 2.4), and therefore, there exist two orderings of arguments in equation (2.31). In the model training, Gaussian predicted by the model (see equation (2.18)) is in the first argument, while the observed banded Gaussian (in MD simulations) is in the second. The parameter set estimation in this ordering is the maximum entropy estimation of the parameter set. Swapping the arguments (Maximum Likelihood estimation of the parameter set) has noticeable changes in the parameter set; however, the general predictions from the parameter set obtained from either setting are close. Comparison of these choices is not the main focus of this thesis, and a detailed discussion on this will be published elsewhere.

2.4.3 Computation of initial solution for the parameter set

The next step is to numerically solve equation (2.31) which requires an initial guess solution for the parameter set. This section briefly describes how to obtain the initial guess solution using the Fisher information matrix, and more details can be found elsewhere [149].

The Fisher information [80] can be defined as the second centred moment of $\log \rho(x;\theta)$ conditional to parameter $\mathbb{R}^N \ni \theta = \{\mu, \mathcal{K}\},\ N > 1$,

$$[\mathcal{F}(\theta)]_{ij} = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta_i \partial\theta_j}(\log \rho(x;\theta))\middle|\theta\right] \tag{2.32}$$

where $\rho(x;\theta)$ is a pdf conditional on parameter $\theta \in \mathbb{R}^N$, and μ and $\mathcal{K}$ are the mean and inverse covariance matrix. Furthermore, for two parametric pdfs that are close in parametric space, say $\rho(x;\theta')$ and $\rho(x;\theta)$ where $\theta = \theta' + \delta\theta,\ \theta,\ \theta',\ \delta\theta \in \mathbb{R}^N$ and $\delta\theta << 1$, there exists a relation between KL divergence (defined in appendix section C.4) and Fisher information,

$$\mathcal{F}(\theta) = -\int_\Omega \rho(x,\theta')\frac{\partial^2}{\partial\theta'^2}\log(\rho(x,\theta')) = \frac{\partial^2}{\partial\theta^2} D_{\mathrm{KL}}(\rho(x,\theta'),\rho(x,\theta))\big|_{\theta=\theta'}. \tag{2.33}$$

Now, using Taylor expansion at $\theta = \theta'$ gives,

$$D_{\mathrm{KL}}(\rho(x;\theta'),\rho(x;\theta)) = \frac{1}{2}\delta\theta.\mathcal{F}(\theta)\delta\theta + \mathcal{O}(|\delta\theta^3|) = D_{\mathrm{KL}}(\rho(x;\theta),\rho(x;\theta')). \tag{2.34}$$

The training of cgDNA+ model (refer equation (2.31)) requires computation of the KL divergence between the observed Gaussian for each oligomer in the training library, $\rho(w;\theta_o(S))$ and banded Gaussian predicted by cgDNA+ reconstruction, $\rho(w;\theta_{P_{\mathrm{vec}}}(S,P))$. Now, equation (2.34) can be approximated as,

$$D_{\mathrm{KL}}(\rho(w;\theta_{P_{\mathrm{vec}}}),\rho(w;\theta_o)) \approx \frac{1}{2}\theta_{P_{\mathrm{vec}}}.\mathcal{F}(\theta_o)\theta_{P_{\mathrm{vec}}} - \theta_{P_{\mathrm{vec}}}.\mathcal{F}(\theta_o)\theta_o + \frac{1}{2}\theta_o.\mathcal{F}(\theta_o)\theta_o \tag{2.35}$$

Using the relation $\theta_{P_{\mathrm{vec}}}(P,S) = R_P(S)P_{\mathrm{vec}}$ (equation (2.26)) in equation (2.35) gives,

$$\begin{aligned} D_{\mathrm{KL}}(\rho(w;\theta_{P_{\mathrm{vec}}}),\rho(w;\theta_o)) \approx & \frac{1}{2}P_{\mathrm{vec}}.R_P^T(S)\mathcal{F}(\theta_o)R_P(S)P_{\mathrm{vec}} \\ & - \frac{1}{2}P_{\mathrm{vec}}.R_P^T(S)\mathcal{F}(\theta_o)\theta_o + \frac{1}{2}\theta_o.\mathcal{F}(\theta_o)\theta_o. \end{aligned} \tag{2.36}$$

Now, using linear change of variable $\mathcal{F}_{(P,S)}(\theta_o) = R_P^T(S)\mathcal{F}(\theta_o)R_P(S)$ in above equation gives

$$\begin{aligned} D_{\mathrm{KL}}(\rho(w;\theta_{P_{\mathrm{vec}}}),\rho(w;\theta_o)) \approx & \frac{1}{2}P_{\mathrm{vec}}.\mathcal{F}_{(P,S)}(\theta_o)P_{\mathrm{vec}} \\ & - \frac{1}{2}P_{\mathrm{vec}}.R_P^T(S)\mathcal{F}(\theta_o)\theta_o + \frac{1}{2}\theta_o.\mathcal{F}(\theta_o)\theta_o. \end{aligned} \tag{2.37}$$

Now, in order to find the best-fit parameter set, the following equation needs to be minimized,

$$\sum_{i=1}^{L} D_{\mathrm{KL}}(\rho(w;\theta_{P_{\mathrm{vec}}}(S_i)),\rho(w;\theta_o(S_i))) \approx \sum_{i=1}^{L} \frac{1}{2} P_{\mathrm{vec}}.\mathcal{F}_{(P,S_i)}(\theta_o) P_{\mathrm{vec}} - \sum_{i=1}^{L} \frac{1}{2} P_{\mathrm{vec}}.R_P^T(S_i)\mathcal{F}(\theta_o)\theta_o + \sum_{i=1}^{L} \frac{1}{2}\theta_o.\mathcal{F}(\theta_o)\theta_o. \tag{2.38}$$

Thus, differentiating equation (2.38) with respect to P_{vec} gives,

$$\sum_{i=1}^{L} \mathcal{F}_{(P,S_i)}(\theta_o) P_{\mathrm{vec}} - \sum_{i=1}^{L} R_P^T(S_i)\mathcal{F}(\theta_o)\theta_o = 0, \tag{2.39}$$

which can be solved using the least squares method. Equation (2.39) can be rewritten as

$$\begin{aligned} &\mathcal{F}_{(P,\mathrm{Lb})} P_{\mathrm{vec}} = B \\ &\text{where } \mathcal{F}_{(P,\mathrm{Lb})} = \sum_{i=1}^{L} \mathcal{F}_{(P,S_i)}(\theta_o), B = \sum_{i=1}^{L} R_P^T(S_i)\mathcal{F}(\theta_o)\theta_o. \end{aligned} \tag{2.40}$$

However, the matrix $\mathcal{F}_{(P,\mathrm{Lb})}$ is not invertible due to non-injectivity of the reconstruction rule (section 2.3.2). Therefore, to find the least squares solution for equation (2.40), Moore-Penrose pseudo-inverse has been used to obtain the initial guess as

$$\mathcal{P}_{\mathrm{vec}}^{\mathrm{lsq}} = \mathcal{F}_{(P,\mathrm{Lb})}^{\dagger} B, \tag{2.41}$$

where $\mathcal{F}_{(P,\mathrm{Lb})}^{\dagger}$ is pseudo-inverse of $\mathcal{F}_{(P,\mathrm{Lb})}$. However, there are no known methods to ensure whether $\mathcal{P}_{\mathrm{vec}}^{\mathrm{lsq}} \in \mathbb{P}$. Therefore, the following two tests should be done a) do reconstructions using $\mathcal{P}_{\mathrm{vec}}^{\mathrm{lsq}}\ \forall\ S \in \mathrm{Lb}$ have positive-definite stiffness matrices?, and b) how close are reconstructions using $\mathcal{P}_{\mathrm{vec}}^{\mathrm{lsq}}\ \forall\ S \in \mathrm{Lb}$ to the observed Gaussian pdfs in MD time-series in terms of KL divergence? If the answers to both questions are affirmative, then $\mathcal{P}_{\mathrm{vec}}^{\mathrm{lsq}}$ can be taken as $\mathcal{P}_{\mathrm{vec}}^{\mathrm{ini}}$. Fortunately, in training the cgDNA+ model, the answer has always been affirmative.

2.4.4 Fisher-informed gradient flow to find best-fit parameter set

Once the initial guess for parameter set $\mathcal{P}_{\mathrm{vec}}^{\mathrm{ini}}$ is obtained, one can use Fisher-informed gradient descent algorithm to solve equation (2.31) as:

$$\mathcal{P}_{\mathrm{vec}}^{k+1} = \mathcal{P}_{\mathrm{vec}}^{k} - \alpha \mathcal{F}_{(P,\mathrm{Lb})}^{\dagger} \nabla_P \left(\sum_{i=1}^{L} D_{\mathrm{KL}}(\rho_P(w;S_i,P)),\rho_b(w;S_i)) \right) \tag{2.42}$$

where $\alpha \in [0,1]$ is the step-size, $\mathcal{P}_{\mathrm{vec}}^{0} = \mathcal{P}_{\mathrm{vec}}^{ini}$ and $\mathcal{F}_{(P,\mathrm{Lb})}^{\dagger}$, pseudo-inverse of the Fisher matrix, is used as pre-conditioner. However, computing $\nabla_P \left(\sum_{i=1}^{L} D_{\mathrm{KL}}(\rho_P(w;S_i,P)),\rho_b(w;S_i)) \right)$ is non-trivial. Details can be found in ref. [149].

2.4.5 Proving positivity of the best-fit parameter set

Solving equation (2.42) leads to a best-fit parameter set $\mathcal{P}^{gf}$ for a given MD data. The final step is now to prove that stiffness matrix reconstruction using $\mathcal{P}^{gf}$ is positive-definite for any

arbitrary sequence, i.e., $\mathcal{K}(\mathcal{P}^{gf}, S) > 0 \ \forall \ S$ of length greater than 3 base-pairs. One sufficient condition is to prove that the stiffness blocks $\mathcal{K}^{\mathrm{XY}} \in \mathcal{P}^{gf}$ for any dimer step XY is positive definite and, thus, for any arbitrary sequence S, $\mathcal{K}(\mathcal{P}^{gf}, S)$ is the overlaying sum of positive-definite stiffness blocks $\mathcal{K}^{\mathrm{XY}}$ and will be positive-definite. However, stiffness blocks $\mathcal{K}^{\mathrm{XY}} \in \mathcal{P}^{gf}$ computed in the last sub-section are often indefinite. Notably, the parameter set $\mathcal{P}^{gf}$ obtained is non-unique due to non-injectivity in the reconstruction rule in equation (2.23) which allows searching for the blocks in the null space such that the stiffness blocks are positive-definite.

Thus, using the null-space, one can find $\Gamma^{\mathrm{X}}, \Gamma^{\mathrm{Y}} \in \mathbb{R}^{18\times 18}$ such that $\bar{\mathcal{K}}^{\mathrm{XY}}, \bar{\mathcal{K}}^{5'\mathrm{XY}} \in \mathcal{P}'$ are positive definite,

$$\bar{\mathcal{K}}^{\mathrm{XY}} = \mathcal{K}^{\mathrm{XY}} + \mathrm{diag}(\Gamma^{\mathrm{X}}, \mathbf{0}_6, \Gamma^{\mathrm{Y}}), \ \bar{\mathcal{K}}^{5'\mathrm{XY}} = \mathcal{K}^{5'\mathrm{XY}} + \mathrm{diag}(\mathbf{0}_{18}, \Gamma^{\mathrm{X}}) \tag{2.43}$$

and satisfy

$$\mathcal{R}(\mathcal{P}', S) = \mathcal{R}(\mathcal{P}^{gf}, S) \ \forall \ S \in \mathrm{Lb} \tag{2.44}$$

where $\mathcal{R}$ is the reconstruction rule (equation (2.23)). However, this search in null-space is non-trivial and therefore, this computation is only performed if $\mathcal{P}^{gf}$ reconstructs positive-definite $\mathcal{K}$ for all physical decamers. More details on the algorithm to search in the null-space are provided in ref. [149]. If blocks $\Gamma^{\mathrm{X}}, \Gamma^{\mathrm{Y}} \in \mathbb{R}^{18\times 18}$ are found such that $\bar{\mathcal{K}}^{\mathrm{XY}}, \bar{\mathcal{K}}^{5'\mathrm{XY}}$ are positive definite, $\mathcal{P}^{gf}$ is declared as the best-fit positive-definite parameter set, $\mathcal{P}$.

2.5 How to quantify errors in the model?

There are several assumptions in the model (see section 2.3.1) which lead to certain approximation errors in cgDNA+ reconstructions/predictions. This section presents methods to quantify these approximation errors and set a scale for the comprehension of these errors.

2.5.1 Error due to non-convergence of MD time-series

The cgDNA+ model assumes that the MD time-series are stationary, which leads to convergence error in the MD statistics. In ideal conditions (stationary MD time-series), for palindromic sequences ($S = \bar{S}$), equation (2.9) can be rewritten as

$$\begin{aligned} \bar{w}(S) &= E_N \bar{w}(S) \\ C(S) &= E_N C(S) E_N. \end{aligned} \tag{2.45}$$

It implies that the mean estimators for a converged MD time-series of a palindromic sequence are independent of the reading strand. Thus, this property of the mean estimators can be used to define the approximation error (referred to as palindromic error) due to the non-convergence of the MD time-series. Quantitatively, the palindromic error is defined as symmetric KL divergence between pdfs while reading the dsDNA oligomer from the Watson strand and the Crick strand:

$$\mathcal{E}^{\mathrm{palin}}_{\mathrm{KL}}(\rho_w, \rho_c) := D_{\mathrm{KLS}}(\rho_w, \rho_c) = \mathcal{M}_S(\rho_w, \rho_c) + \mathcal{S}_S(\rho_w, \rho_c) \tag{2.46}$$

where ρ_w and ρ_c are Gaussian pdfs while reading dsDNA oligomer from the Watson and Crick strand, respectively and D_{KLS} is symmetrised KL divergence, and $\mathcal{S}_S$, and $\mathcal{M}_S$ (also denoted as $\mathcal{E}^{\mathrm{palin}}_{\mathcal{M}}$) are corresponding stiffness and shape contribution defined in equation (C.12). In the ideal case of fully converged MD time-series both $\mathcal{E}^{\mathrm{palin}}_{\mathrm{KL}}$ and $\mathcal{E}^{\mathrm{palin}}_{\mathcal{M}}$ will be equal to zero.

2.5.2 Error due to Gaussianity imposition on the helical coordinate distributions

Another assumption in the model is that the pdfs of internal coordinates follow a Gaussian distribution, which is not exactly the case. This assumption leads to an inevitable modeling error which is quantified in terms of KL divergence and is referred to as $\mathcal{E}^{\mathrm{Gauss}}_{\mathrm{KL}}$. $\mathcal{E}^{\mathrm{Gauss}}_{\mathrm{KL}}$ is defined as symmetric KL divergence between the pdf observed in the MD time-series with the best-fit Gaussian approximation. Since one of the involved pdfs is not Gaussian, KL divergence can only be computed numerically, which is non-trivial for multi-dimensional pdfs. Therefore, $\mathcal{E}^{\mathrm{Gauss}}_{\mathrm{KL}}$ is only computed between two 1D-pdfs (for instance, between two 1D distributions of the Twist) obtained from MD time-series and corresponding best-fit Gaussian approximation and is separately reported for each of the cgDNA+ variables.

2.5.3 Error due to nearest-neighbor interactions assumption

Another important assumption in the cgDNA+ model is the nearest-neighbor interactions as described in equation (2.21) in which the total energy of the oligomer is approximated as the sum of local junction energy contributions. Moreover, another closely associated approximation is that these local junction energy parameters depend only on the corresponding junction dinucleotide step sequence. In terms of modeling, these two approximations are implemented in one step (equation (2.31)) that computes the dimer-dependent parameters from observed Gaussian pdfs in the MD time-series for Lb. Therefore, it is impossible to individually determine errors associated with these two assumptions. One possible way to approximate the errors associated with these two assumptions is by hypothesizing that the parameter set estimation proceeds in two steps; a) first, banded Gaussian pdfs are obtained from observed Gaussian pdfs, and then b) dimer step dependent parameters for junction energy contributions are computed.

To compute a banded Gaussian pdf from the observed Gaussian pdf $\rho_o(w; \hat{w}(S), C(S))$ that corresponds to nearest-neighbor interactions assumption, the maximum entropy principle [78, 79] can be used. In other words, the best-fit density via the maximum entropy principle can be found such that the observed stiffness matrix (i.e., inverse covariance matrix) becomes a banded matrix of 42×42 block diagonal with 18×18 specific sparsity pattern. Thus, given $\rho_o\left(\bar{w}(S), C(S)\right)\ \forall\ S \in \mathrm{Lb}$, the best-fit density is defined as:

$$\rho_b(w(S); \bar{w}_b(S), \mathcal{K}_b(S)) := \underset{\rho_o \in \mathbb{R}}{\operatorname{argmax}}\, D(\rho_o(S)) \tag{2.47}$$

under the constraint

$$\bar{w}_b = \bar{w},\ [\mathcal{K}_b]_{\mathcal{N}^c} = 0 \text{ and } \left[\mathcal{K}_b^{-1}\right]_{\mathcal{N}} = [C]_{\mathcal{N}} \tag{2.48}$$

where $\mathcal{N}$ is a set of all indices associated to the sparsity pattern of 42×42 block diagonal with 18×18 overlaps (as shown in figure 2.4), $\mathcal{N}^c$ is the complement of $\mathcal{N}$ and $D(\cdot)$ is defined in

equation (2.29). This banded stiffness matrix in equations (2.47) and (2.48) can be computed using an analytical algorithm presented elsewhere [60, 149].

Lastly, the error associated with nearest-neighbor interactions approximation can be quantified as the symmetric KL divergence between observed unbanded Gaussian pdf and corresponding banded Gaussian pdf obtained using the Maximum entropy principle:

$$\mathcal{E}_{\mathrm{KL}}^{\mathrm{Trunc}}(\rho_o, \rho_b) = D_{\mathrm{KLS}}(\rho_o, \rho_b) = \mathcal{M}_S(\rho_o, \rho_b) + \mathcal{S}_S(\rho_o, \rho_b) \tag{2.49}$$

where ρ_o and ρ_b are observed Gaussian pdf and banded Gaussian pdf, respectively and D_{KLS}, $\mathcal{S}_S$, and $\mathcal{M}_S$ (also denoted as $\mathcal{E}_{\mathcal{M}}^{\mathrm{Trunc}}$) are defined in equation (C.12). Note that in the bandedness imposition, the average shape remains the same; therefore, $\mathcal{E}_{\mathcal{M}}^{\mathrm{Trunc}}$ is zero.

2.5.4 Error due to local dimer sequence dependence in junction energy coefficients

The last assumption in the cgDNA+ model is the locality in sequence dependence in the local junction energy parameters. It assumes that the local junction energy parameters only depend on the corresponding dimer sequence of that junction. One can also contemplate a model where local junction energy parameters are sequence-average (a relatively simple model) or local tetramer sequence dependence (a more complex model). However, the observations in MD data and cgDNA/cgDNA+ model conclude that dimer level sequence dependence in the local junction energy parameters is sufficient [62, 149, 159] and the gain in accuracy by increasing the complexity of the model is negligible compared to the increase in the number of parameters.

As described in the previous section, this assumption is implemented in the cgDNA+ model in one step along with the nearest-neighbor interactions assumption, and therefore, the associated error can only be approximated. Following the hypothesis explained in the previous section, the error corresponding to the locality of the sequence dependence can be quantified by comparing banded Gaussian pdf (i.e., truncated observed Gaussian pdf in MD) and cgDNA+ predicted Gaussian pdf in terms of KL divergence ($\mathcal{E}_{\mathrm{KL}}^{\mathrm{local}}$) and Mahalanobis distance ($\mathcal{E}_{\mathcal{M}}^{\mathrm{local}}$) as,

$$\mathcal{E}_{\mathrm{KL}}^{\mathrm{local}}(\rho_b, \rho_{\mathcal{P}}) := D_{\mathrm{KLS}}(\rho_b, \rho_{\mathcal{P}}) = \mathcal{M}_S(\rho_b, \rho_{\mathcal{P}}) + \mathcal{S}_S(\rho_b, \rho_{\mathcal{P}}) \tag{2.50}$$

where ρ_b and $\rho_{\mathcal{P}}$ are observed Gaussian pdf and cgDNA+ predicted Gaussian pdf, respectively and D_{KLS}, $\mathcal{S}_S$, and $\mathcal{M}_S$ are defined in equation (C.12).

2.5.5 How accurate are cgDNA+ reconstructions?

Lastly, the total modeling error in the cgDNA+ model (referred as reconstruction or prediction error) can be quantified in terms of symmetrized KL divergence between the observed pdf in MD time-series and pdf predicted by cgDNA+,

$$\mathcal{E}_{\mathrm{KL}}^{\mathrm{res}}(\rho_o, \rho_{\mathcal{P}}) := D_{\mathrm{KLS}}(\rho_o, \rho_{\mathcal{P}}) = \mathcal{M}_S(\rho_o, \rho_{\mathcal{P}}) + \mathcal{S}_S(\rho_o, \rho_{\mathcal{P}}) \tag{2.51}$$

where ρ_o and $\rho_{\mathcal{P}}$ are observed Gaussian pdf and cgDNA+ predicted Gaussian pdf, respectively and D_{KLS}, $\mathcal{S}_S$, and $\mathcal{M}_S$ (also denoted as $\mathcal{E}_{\mathcal{M}}^{\mathrm{res}}$) are defined in equation (C.12).

Furthermore, to gain more confidence in the cgDNA+ model, it has been tested for various sequences not present in the training library. More details with specific examples are discussed later in chapters 3, 4 and 6.

2.5.6 How large is the error?

While discussing the performance of cgDNA+ using the above-described approximation errors, a natural question arises: how to set up a scale to comprehend the magnitude of the error. In an ideal case, both KL divergence and Mahalanobis distance for any error should be zero, which tells that the two pdfs are identical. However, this is never the case. So, in this scenario, it is necessary to set a scale in order to understand or visualize what KL divergence or Mahalanobis distance per dof means, in particular, for multi-dimensions pdfs.

The model primary purpose is to predict a given sequence's groundstate and stiffness, thus accurately capturing the features that depend on the sequence. Therefore, the errors must not be larger than the changes introduced in the mechanics of dsDNA by changing the sequence. To quantify this error, a scale is set by computing average pair-wise differences in all the sequences in the training library, which provides a robust scale for comparing various errors in the model. This scale is computed in terms of both symmetric KL divergence and Mahalanobis distance. Using this scale, the performance of the cgDNA+ model has been evaluated and discussed in detail in later chapters. Moreover, to visualize KL divergence between two 1D Gaussian pdfs, figure 2.5 plots an envelope of 1D Gaussian $\mathcal{N}(\epsilon_1, 1+\epsilon_2)$ around $\mathcal{N}(0,1)$ such that the symmetric KL divergence between two 1D Gaussian is a particular value for a family of ϵ_1 and ϵ_2. Impressively, in the cgDNA+ model, the various errors (in terms of KL divergence) are of the order of 10^{-2} or 10^{-3}, and visually they belong to a very tiny difference in the 1D Gaussian pdfs.

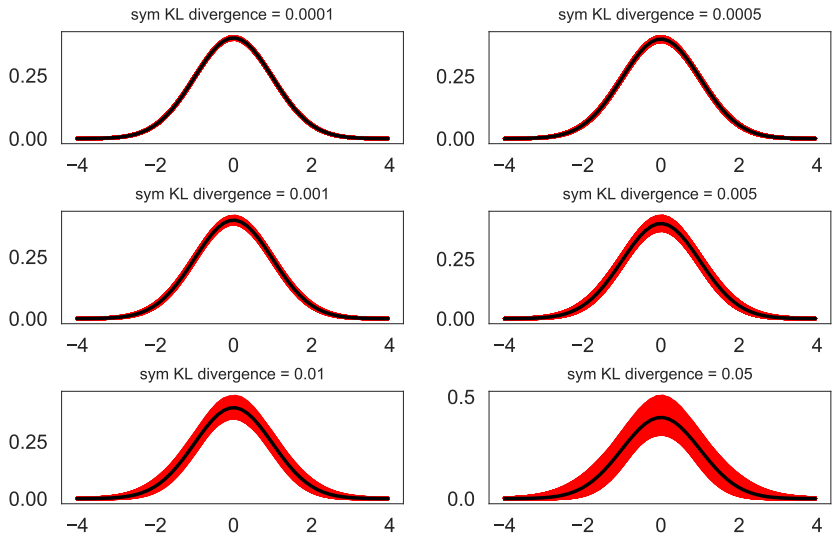


Fig. 2.5 An envelope of $\mathcal{N}(\epsilon_1, 1+\epsilon_2)$ 1D Gaussian around an $\mathcal{N}(0,1)$ Gaussian for a family of ϵ_1 and ϵ_2 corresponding to various symmetric KL divergences.

2.6 cgDNAmc+

The cgDNA+ model predicts a Gaussian pdf (groundstate and stiffness matrix) for a given S and $\mathcal{P}$. This section briefly describes how to efficiently sample cgDNA+ Gaussian pdf and then use that ensemble of configurations to compute expectations of any function, particularly of various interesting physical observables such as persistence length and end-to-end distribution. Mitchell et al. [123] developed a very efficient Monte-Carlo sampling method that allows generating a million samples in only a few minutes on a single processor for a given sequence of length 300 bps. The efficient sampling is due to the efficient Cholesky decomposition of the sparse banded stiffness matrix, $\mathcal{K} = LL^T$. The cgDNA+ Gaussian can be rewritten as;

$$\rho(w; S, \mathcal{P}) = \frac{1}{Z} e^{-\beta E(w)} = \frac{1}{Z} e^{-\beta y^T y/2} \tag{2.52}$$

where $y = L^T(w - \hat{w})$ which can be easily sampled directly as product of independent uni-variate normal Gaussian distributions and the configuration in cgDNA+ variables can be obtained by solving $L^T z = y$ and $w = z + \hat{w}$. More details can be found in ref. [123]. Subsequently, using these configurations in cgDNA+ internal variables, the ensemble expectation of any function, $f(w)$ can be approximated as $\frac{1}{n}\sum_{i=1}^{n} f(w_i)$ where n is the total number of configurations generated.

2.6.1 Persistence length

One of the interesting physical observables in the context of DNA is persistence length, which represents the inclination of a polymer to "persist" in a given direction. It has been a popular and traditional measure to quantify the rigidity of DNA and is defined as the length scale over which correlations in the direction of tangent along a polymer centerline are lost [68]. Mathematically, using the discrete version of Kratky-Porod Worm-Like Chain model [92], the persistence length (ℓ_p) for a linear chain of N rigid bodies with the position $r_i|_{i=1,\ldots,N}$ can be defined as;

$$\langle t_i \cdot t_1 \rangle_{\text{WLC}} = e^{-i/\ell_p} \tag{2.53}$$

where $\langle\cdot\rangle$ represents the ensemble average, t_1 and t_i are the unit vectors for the base-pair index 1 and i along the DNA. t_i can be defined as $t_i := (r_{i+1} - r_i)/b$ where rigid bodies are separated at length b and r_i is the position of the i$^{\text{th}}$ rigid body.

In the context of dsDNA (and other dsNAs), the definition of persistence length has been frequently used in the sequence-average sense, which has two crucial governing factors [201], stiffness and intrinsic shape, which can be deconvoluted as;

$$\frac{1}{\bar{\ell}_p} = \frac{1}{\bar{\ell}_s} + \frac{1}{\bar{\ell}_d} \tag{2.54}$$

where $\bar{\ell}_p$, $\bar{\ell}_s$, and $\bar{\ell}_d$ are sequence-average apparent, static, and dynamic persistence length, respectively and $\{\hat{t}_i \cdot \hat{t}_1\} = e^{-i/\bar{\ell}_s}$ where $\hat{t}_i$ is tangent at the i$^{\text{th}}$ position in the groundstate and $\{\cdot\}$ represents the average taken over an ensemble of sequences (thus, sequence-averaged persistence lengths). Mitchell et al. [123] generalized these definitions introducing sequence-

dependent dynamic persistence length as

$$\langle t_i \cdot t_1 \rangle = \{\hat{t}_i \cdot \hat{t}_1\} e^{-i/\ell_d} \tag{2.55}$$

where $\{\hat{t}_i \cdot \hat{t}_1\} = e^{-i/\ell_s}$ is only evaluated at the groundstate of a given sequence, thus resulting in sequence-dependent ℓ_s and ℓ_d.

cgDNAmc+ efficiently code this using just inter-coordinates. Monte Carlo code samples a configuration, $w \in \mathbb{R}^{24N-18}$ as described in equation (2.8) and from w extracts inter coordinates, y and subsequently, the base-pair frames with orientation and position for i$^{\text{th}}$ base-pair frame given as $R_i \in SO(3)$ and r_i (refer to section 2.2.3 for details). This work further simplifies the computations by approximating t_i as base-pair normal (third column of R_i). A detailed discussion of the various possible definitions of t_i and their implications can be found in the original cgDNAmc article [123]. Now, $t_i \cdot t_1$ can be written as $(R_i^T R_1)_{33}$ and is computationally obtained as the inner-product of third column of R_i and third column of R_1. The corresponding ensemble average can be computed as $\langle t_i \cdot t_1 \rangle = \frac{1}{n} \sum_i^n t_i \cdot t_1$ where n is typically chosen as 10^5 which provides sufficiently converged statistics [149]. Lastly, ℓ_p and ℓ_d can be computed as the -1/slope of the linear fit, $\{i, log(\langle t_i \cdot t_1 \rangle)\} | i = 1, ..., N$ and $\{i, log(\langle t_i \cdot t_1 \rangle) - log(\langle \hat{t}_i \cdot \hat{t}_1 \rangle)\}|_{i=1,...,N}$, respectively, where N is the number of the base-pairs. Note that, in this work, the reference terminal base-pairs (i.e., $i = 1$ and $i = N$) have been chosen away from the ends to avoid any end effects (i.e., six base-pairs from both sides have been dropped in the computation of persistence length).

CHAPTER 3

Molecular Dynamics simulations

This chapter briefly describes the basic details of molecular dynamics (MD) simulations and the simulation protocol used in this work. Then, we discuss the palindromic library introduced in ref. [149] for training the cgDNA+ model. Note we also use the same library (in different alphabets) for training corresponding parameter sets for double-stranded RNA (dsRNA) and DNA:RNA hybrid (DRH). Moreover, we introduce a new library to train parameter sets that allows epigenetically modified bases and all non-GC ends in dsDNA. Lastly, we brief the MD data processing, followed by a detailed discussion on the convergence of MD simulations and the distributions of helical coordinates in MD time-series.

3.1 Molecular Dynamics Simulations

MD simulations, based on Newtonian equations, give a dynamic evolution of the system. In a typical MD simulation, the trajectories of a system of interacting particles (atoms and molecules) are determined by integrating Newton's equations. MD simulations have become state-of-the-art for studying biomolecules and are often employed to complement several experimental techniques [74, 192]. With the advance in computational power as well as the development of better force fields, MD simulations can provide insights into how bio-molecules behave or interact. In particular, for nucleic acids (NA), the first MD simulations for DNA [105] were performed about four decades ago, and since then, MD simulations have contributed significantly to the understanding of the NAs [39, 142, 155, 192].

In the simplest terms, a typical MD simulation starts with a system containing N particles (atoms in this case) that can interact with each other based on a given forcefield. With chosen starting positions of the particles, i.e., initial configuration and forcefield, one can solve Newtonian equations to obtain a dynamic evolution of the system. This temporal evolution or trajectory of each particle in the system (containing N particles) is determined using Newton's second law (equation (3.1)) where $F_i \in \mathbb{R}^3$ is the force on each particle in the system with mass $m_i \in \mathbb{R}$ and position $r_i \in \mathbb{R}^3$ in Cartesian coordinates. The force, F_i on each particle is defined as the derivative of the potential energy $U(r_1, r_2, ..., r_n) \in \mathbb{R}$ as given in (equation (3.2)).

$$F_i = m_i \frac{d^2}{dt^2} r_i \tag{3.1}$$

$$F_i = -\frac{\partial U(r_1, r_2, ..., r_n)}{\partial r_i} \tag{3.2}$$

The potential in MD, also known as forcefield, consists of bonded (first three terms in equation (3.3)) and non-bonded potential terms (last two terms in equation (3.3)). The first two terms represent the stretching energy of a covalent bond and the bending energy of a valence angle, which is modeled using the harmonic potential. k_b and k_θ are the stiffness constant for bond stretching energy and angle bending energy with $\hat{x}_b$ and $\hat{\theta}_a$ as equilibrium values and x_b and θ_a as observed values for the bond and valence angles. The third term in equation (3.3) models the torsional energy, which is defined as the strain when the angle (ϕ_d) between planes through two sets of three bonded atoms (with two atoms in common) deviates from the minimum torsional energy determined by the phase factor δ_d. n is a multiplicity representing the total number of energy minima when the torsional angle rotates from 0 to 2π and V_n is the barrier height. The last two non-bonded terms represent the Van der Waals and Coulombic interactions between the particles, which are not directly bonded. The Van der Waals interaction in MD is popularly approximated by Lennard-Jones or 12-6 potential, a sum of short-range repulsive and attractive force. ε_{ij} is the depth of the potential well, $\hat{r}_{ij}$ is the distance at which inter-particle potential is zero, and r_{ij} is the observed distance. The last term in equation (3.3) is the electrostatic energy between two particles defined by Coulombic interactions where ϵ is Coulomb constant and r_{ij} is the distance between two particles with charge q_i and q_j. The different parameters (such as stiffness constant and equilibrium bond distances and angles) in this potential, U depend on the particles involved and are usually obtained from Quantum Mechanical calculations or experiments.

$$\begin{aligned} U = \sum_{bonds} k_b(x_b - \hat{x}_b)^2 + \sum_{angles} k_\theta(\theta_a - \hat{\theta}_a)^2 + \sum_{dihedrals} \frac{V_n}{2}[1 + cos(n\phi - \delta_d)] \\ + \sum_{i<j}\left[\frac{\varepsilon_{ij}\hat{r}_{ij}^{12}}{r_{ij}^{12}} - \frac{2\varepsilon_{ij}\hat{r}_{ij}^{6}}{r_{ij}^{6}}\right] + \sum_{i<j}\left[\frac{q_i q_j}{\epsilon r_{ij}}\right] \end{aligned} \tag{3.3}$$

Along with the challenges in obtaining a good potential, there are several other challenges in performing MD simulations, such as boundary effects, computational cost, and discretization errors. The simulation box size must be large enough to avoid boundary effects, usually achieved by imposing periodic boundary conditions. This method attempts to emulate the bulk conditions by looping back one side of the simulation box to another side. The most intensive part of MD simulations of large systems is the computation of potential energy U, particularly non-bonded interactions. Ideally, Coulombic and Van der Waals interactions should be calculated for every pair of particles in the system, but infeasible due to highly intensive computations. Thus, various approximation techniques are used to reduce computational efforts. One of the most popular methods used to reduce the computation of the Coulombic part is the Particle Mesh Ewald (PME) method [46]. Alternatives to PME method include fast multipole method [67]. The basic idea behind PME technique is to replace the direct computation of interaction energies of the point particles with two components, a) the short-range potential in real space and b) the long-range potential in Fourier space. Both components converge quickly with a minor loss of accuracy. Similarly, Van der Waals interactions are approximated via a continuum model beyond a cut-off. Lastly, the integration time step is crucial for the total computational cost. To

avoid discretization errors, the integration time step must be chosen to be smaller than the fastest vibrational frequency. The fastest internal vibrations are due to the lightest element, Hydrogen, which is about one femtosecond. To speed up the MD simulation, typically, algorithms like SHAKE [173], and RATTLE [5] are used, which fix the vibrations of Hydrogen atoms. Precise details about cut-offs and algorithms used in our simulations are provided in the next section.

Once a forcefield, simulation box-size and time step are chosen, the final step in performing MD simulations is to numerically solve N second-order differential equations in equation (3.1) which can be simplified into $2N$ first-order differential equations as in equation (3.4).

$$\begin{aligned} \frac{dv_i}{dt} &= \frac{F_i}{m_i} \\ v_i &= \frac{dr_i}{dt} \end{aligned} \tag{3.4}$$

where v_i is the velocity of i^{th} particle. Several integration algorithms such as Verlet, velocity Verlet, and Leapfrog have been developed to solve the above equations. The AMBER module (used in this work) uses the Leapfrog algorithm as elaborated in equation (3.5).

$$\begin{aligned} r_i^{k+1} &= r_i^k + v_i^k \Delta t + \frac{1}{2} a_i^k \Delta t^2 \\ v_i^{k+1} &= v_i^k + \frac{1}{2}(a_i^k + a_i^{k+1})\Delta t \end{aligned} \tag{3.5}$$

where Δt is the integration time-step.

Thus, a temporal evolution based on the provided potential can be obtained with a given initial position and velocity of the particles in the system. The key steps in performing MD simulations are the following:

- The choice of initial positions of the particle is crucial as it must not be very far from the potential energy minima. There are several experimental databases or theoretical software available to obtain an acceptable initial configuration for MD simulation. We have used nucleic acid builder (NAB) by AMBERTOOLS [29] to generate the initial configurations of dsNAs.
- In our case, a solvated dsNA molecule with ions is a complete description of the initial setup. Therefore, the next step is to solvate the dsNA molecule obtained in the last step and then add cations to neutralize the system, followed by adding ion pairs to reach the desired salt concentration comparable to physiological conditions.
- The next step is the energy minimization of the solvent.
- The temperature in MD simulation is equivalent to the system's kinetic energy, thus the velocity of the particles. The final desired temperature is achieved in several small steps of random velocities addition and minimization steps.
- Once the desired temperature is reached, an equilibration step follows and, finally, the production step.

Usually, in experiments, several macroscopic parameters, such as pressure, temperature, volume, and energy, are kept fixed. In MD simulations, this can be achieved by re-scaling the velocities of the particles. Algorithms, such as Nosé-Hoover, Andersen, and Berendsen [18]

thermostats are available to obtain different conditions.

Lastly, special GPU-based modules are available to speed up the calculations to perform large-scale MD simulations. For production simulations, we have used the pmemd.cuda code by AMBER [29] on the high-performance computing facilities at EPFL. There exist several other platforms to run MD simulations, such as CHARMM, GROMACS, and GROMOS, while AMBER [29, 30, 150] and CHARMM [25, 26] are the most popular for DNA simulations.

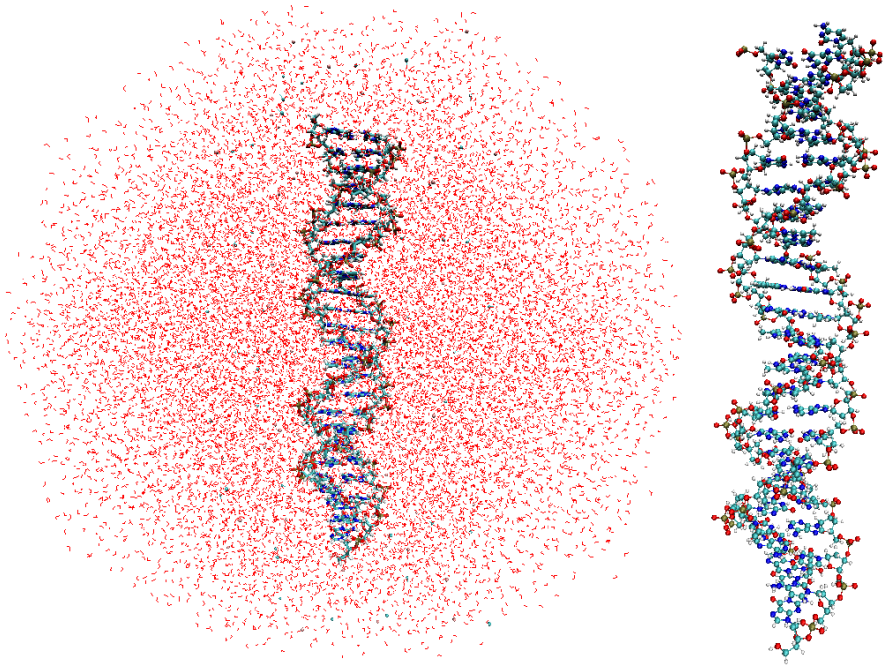


Fig. 3.1 A typical snapshot of the molecular dynamics simulation setup of a 24mer dsDNA. On the left, the dsDNA molecule and ions are solvated in water and on the right, a snapshot of dsDNA.

3.2 Simulation details

The initial structure of a given sequence has been generated using NAB in AMBERTOOLS 18 [29]. We have used Arnott right-handed B-DNA fiber parameters [7] and Arnott right-handed A-RNA fiber parameters [8] for DNA and RNA strands, respectively. Furthermore, to obtain the initial geometry of DRH, we have first generated an A-form RNA (i.e., start with pure A-form) and then modified all Uracil to Thymine and changed the sugar molecule from ribose to deoxyribose (i.e., replace the OH group at $2'$ with an H) in one of the strands. Then, in the first few nanoseconds of production run, we observed that the molecule changes its geometry to a mixed A-B form. Our analysis has ignored the first 100 nanoseconds of the production simulation. Lastly, to methylate or hydroxymethylate the desired cytosine in the structure, we have modified the initial structures obtained from NAB using the leap program in AMBERTOOLS 18.

To describe the DNA and RNA strands in the MD simulations, we have used the PARMBSC1 [76] and OL3 forcefield corrections [210], respectively along with PARMBSC0 [154] and parm99 force field [35]. Furthermore, we have used additional parameters for methylated and hydroxymethylated cytosine [12, 156]. For each sequence, once the initial dsNA structure was generated, the molecule was solvated in a truncated octahedral water box of explicit TIP3P water molecules [83] with a water layer of a minimum 10 Å surrounding the dsNA molecule. Subsequently, the solvated dsNA was neutralized with K^+ ions, and then K^+ and Cl^- ion-pairs were added to make the salt concentration approximately 150 mM. The ions were described using the Joung and Cheatham model [84] and added with the constraint that the ions are at least 5 Å away from the dsNA molecule and at least 3.5 Å away from each other. For training sequences in libraries (tables B.1 and B.2), this complete initial setup contains ≈ 45,000 atoms, of which ≈ 40,000 are from water molecules. A typical snapshot of such an initial system is shown in figure 3.1. The system size is smaller for the sequences in the 12mer library (table B.3).

The system preparation is followed by an energy minimization step to minimize the solvent energy. Then the system temperature is slowly raised to 300 K, followed by an equilibration step of 50 picoseconds. All simulations were performed using AMBER 18 modules [29]. Simulations were carried out in the NPT ensemble using the Berendsen algorithm [18] to maintain the temperature (at 300K) and the pressure (1 atm) with a coupling constant of 5 picoseconds. Furthermore, we have used SHAKE [173] algorithm to freeze the motion of bonds involving Hydrogen allowing a larger simulation time-step of 2 femtoseconds. Simulations were carried out under periodic boundary conditions, and long-range electrostatic interactions were treated using the particle mesh Ewald method [46] with 9 Å real space cut-off. The short-range Lennard-Jones interactions were also truncated at 9 Å. Note that most simulation parameters have been motivated by the choices made by the Ascona B-DNA Consortium (ABC) [102]. Finally, for each system, 10 μs of production run simulations were carried out at EPFL HPC facilities using GPUs with each node containing 2 Xeon-Gold processors and 2 NVIDIA V100 PCIe 32 GB GPUs. Production run simulation trajectories were stored at two picosecond intervals. Each production run simulation of length 10 μs took approximately two months for a 24mer.

3.3 Training library

In this section, we have discussed the various training and test sequences that are used to train the cgNA+ model parameter sets.

3.3.1 Training library for interior blocks and GC ends of dsNA parameter sets

We have simulated a comprehensive set of palindromic sequences to train the parameter sets for dsDNA and dsRNA called the palindromic training library, originally proposed in ref. [149]. It contains 16 training sequences (with GC ends) of length 24 base-pairs such that the library has almost similar instances for all monomers, dimers, and trimer, and all 256 tetramers appear at least once on both strands. The sequences in both libraries, Lb_{DNA} and Lb_{RNA} are provided in table B.1. The palindromic nature of these training libraries allows quantifying the convergence of MD simulations, which is otherwise non-trivial as discussed in section 3.5.

Sequence index	Acceptance rate				
	Lb_{DNA}	Lb_{RNA}	Lb_{DRH}	Lb_{Met}	Lb_{Hmet}
1	85.29	77.51	78.78	83.25	86.39
2	83.10	76.27	74.31	85.90	83.52
3	86.61	80.78	81.57	84.28	86.71
4	85.00	74.41	78.69	86.18	84.07
5	82.90	80.32	80.19	86.11	87.74
6	88.83	84.37	78.24	87.93	87.08
7	84.40	74.86	75.46	88.44	85.16
8	85.24	78.40	80.72	82.89	84.10
9	79.94	76.77	71.90	74.86	87.44
10	88.97	75.89	75.24	88.30	84.69
11	84.29	80.03	74.41	83.20	82.67
12	82.15	82.94	76.74	91.40	89.72
13	88.47	78.89	79.60		
14	88.86	76.92	80.99		
15	79.58	81.24	75.65		
16	82.46	83.00	81.37		

Table 3.1 % MD snapshots left after discarding snapshots with broken H-bonds. The total number of configurations before filtering is $10 \cdot 5 \cdot 10^5$ ($10\ \mu s$) for each training sequence listed in tables B.1 and B.2.

Sequence index	Acceptance rate	Sequence index	Acceptance rate	Sequence index	Acceptance rate	Sequence index	Acceptance rate
1	24.49	2	27.86	3	29.72	4	30.13
5	28.33	6	32.89	7	30.02	8	32.96
9	17.53	10	28.93	11	28.20	12	29.07
13	35.47	14	25.39	15	28.96	16	35.76
17	19.90	18	24.30	19	16.33	20	24.91
21	24.17	22	30.86	23	13.62	24	29.54
25	33.59	26	31.29	27	21.46	28	38.57
29	27.02	30	29.85	31	23.71	32	25.38
33	61.08	34	59.17	35	66.19	36	62.76
37	57.17	38	65.35	39	49.92	40	62.90
41	54.78	42	53.06	43	71.07	44	66.19
45	39.36	46	43.43	47	54.22	48	46.07
49	64.36	50	65.49	51	67.62	52	60.02
53	64.96	54	64.26	55	49.42	56	68.65
57	58.70	58	61.25	59	52.68	60	59.07

Table 3.2 % MD snapshots left after discarding snapshots with broken H-bonds. The total number of configurations before filtering is $3 \cdot 5 \cdot 10^5$ ($3\ \mu s$) for each sequence listed in table B.3.

Using these training sequences with GC ends, we have estimated model parameters for interior blocks and GC ends. Along with the training sequences, Lb_{DNA} and Lb_{RNA} also contain test sequences comprising a random palindrome, some mechanically exceptional sequences such as poly(A), poly(AT), sequences with point mutations and A-tracts. We have used the same training library (as dsDNA or dsRNA) to train coarse-grain parameters for DRH, but the sequences are not palindrome anymore. More details on the parameter sets are provided in section 4.2.1.

3.3.2 Training library for dsDNA non-GC ends parameters

Furthermore, we have used an additional 60 sequences of length 12 base-pairs (refer table B.3) to train parameters for non-GC ends. For each non-GC end, we have four sequences starting with the non-GC end (while another end is GC) followed by one of the RR, RY, YR, and YY steps (where R represents the purine base and Y represents the pyrimidine base) for a diverse training set, and the rest of the sequence is chosen randomly. Thus, to train model parameters for a given non-GC end, we have MD time-series data from four sequences with a non-GC end followed by different contexts. For each sequence in the end library (Lb_{End}), we have generated an MD time-series of 3 μs. More details on how these libraries are used to calculate coarse-grained model parameters are provided in section 2.4.

3.3.3 Training library for epigenetically modified

In this work, the objective is to obtain a parameter set that allows epigenetically modified CpG steps, in particular, methylated or hydroxymethylated, which can be symmetric or asymmetric. For training such parameters for modified CpG steps, we have again designed a palindromic training library that contains symmetrically and asymmetrically modified CpG steps in diverse sequence contexts, as well as modified steps next to each other. The libraries are referred to as Lb_{Met} and Lb_{Hmet} for training sequences containing methylated and hydroxymethylated CpG steps, respectively, and details are provided in table B.2. Moreover, table B.2 also contains a few test sequences, in particular, typical CpG islands with CpG step modifications.

3.4 MD data processing

As described earlier in section 3.2 for each sequence (except the sequences in Lb_{End}), we have run 10 μs of the production run. The data are stored in binary format (.nc) provided by AMBER, and it takes $\approx$ 3.3 TB and $\approx$ 96 GB of storage to save the data for one sequence with water and without water, respectively. The first step after the production run is to strip the water using CPPTRAJ [169, 170] which makes further analysis easier due to the smaller size of the trajectories. Subsequently, we fit the frames in the MD trajectories, compute the cgNA+ model internal coordinates from frames and discard snapshots with broken H-bonds. The last step is to compute the first and second moments, i.e., oligomer level Gaussian statistics (refer equation (2.27)) for the internal coordinates distributions in (filtered) MD time-series for each sequence.

H-bond filtering is one of the most crucial steps in MD analysis, and in tables 3.1 and 3.2, we have provided the % of accepted snapshots for each training sequence in various training libraries along with the total number of snapshots. For each sequence with GC ends used for the

training of interior block parameters, the acceptance is $\approx 70 - 90\%$. In contrast, the acceptance after H-bond filtering in the training sequences for end-block parameters is comparatively lower, i.e., $\approx 18 - 71\%$ and highly depends on its non-GC end. A similar acceptance of MD snapshots is observed for the test sequences; therefore, details of those sequences are omitted for brevity.

In figures 3.2 to 3.5, we have plotted marginal normalized histograms for various internal coordinates observed in MD time series before and after H-bond filtering for sequence index 1 in Lb_{DNA}. The histograms are plotted for the half-sequence while reading the sequence from both the Crick and Watson strands. The primary objective of the plots is to visualize the effect of H-bond filtering in the MD data (by comparing histograms in dotted and solid lines for MD data before and after filtering, respectively). Moreover, since the sequence is palindromic, the two dotted histograms corresponding to reads from the Crick and the Watson strand comment on the convergence of MD simulations (details are provided in sections 2.5.1 and 3.5).

Notably, for all internal coordinates, the two histograms in dotted and solid lines for broken H-bond filtered and unfiltered MD data, respectively, are indistinguishable except for intra-translational coordinates (Shear, Stretch, and Stagger) for terminal G and phosphate rotational parameters for the terminal base-pair step GC on both Crick and Watson strands. It highlights that the rejected MD snapshots with broken H-bond are primarily due to fraying of terminal base-pairs and do not affect the distribution of coordinates for internal base-pairs and base-pair steps. We observed similar patterns for other sequences in Lb_{DNA} and other dsNA libraries.

3.5 Convergence of MD simulations

How to determine whether a simulation is long enough is a challenging task? Several studies have investigated the convergence of MD simulations. Traditionally, the decay of the average root mean square deviation values from some reference structure over time has been used to assess the convergence of MD simulations. Alternatively, one can run multiple production runs starting from different initial configurations and compare statistics for those independent trajectories. In particular, for dsDNA (ignoring terminal base-pairs), it has been suggested that 1-5 μs of MD simulations are sufficient for converging its structure and dynamics [57]. In this work, to quantify the convergence of MD time series, we have exploited the palindromic nature of dsNA sequences and defined the palindromic error ($\mathcal{E}_{\text{KL}}^{\text{palin}}$) as the symmetric Kullback-Leibler divergence between Gaussian pdfs while reading the dsNA sequence from the Crick and Watson strands. The shape contribution in the palindromic error is the Mahalanobis distance denoted as $\mathcal{E}_{\mathcal{M}}^{\text{palin}}$. More details on these computations are provided in section 2.5.1.

Firstly, in figures 3.2 to 3.5, each panel has two marginal normalized histograms (in solid line) for various internal coordinates observed in the MD time series while reading the sequence from the Crick and Watson strands. The deviations in the two histograms (plotted in the same solid color) highlight the convergence error, and it is evident from the plots that the two solid plots are on top of each other for intra and inter variables. In the case of phosphate coordinates, one can find examples of observing two solid lines, for example, WRot and WTra coordinates for AA ([3,2] panel). In general, it can be observed that the phosphate coordinates are slightly less converged than the base coordinates, which is consistent with the earlier observations [149].

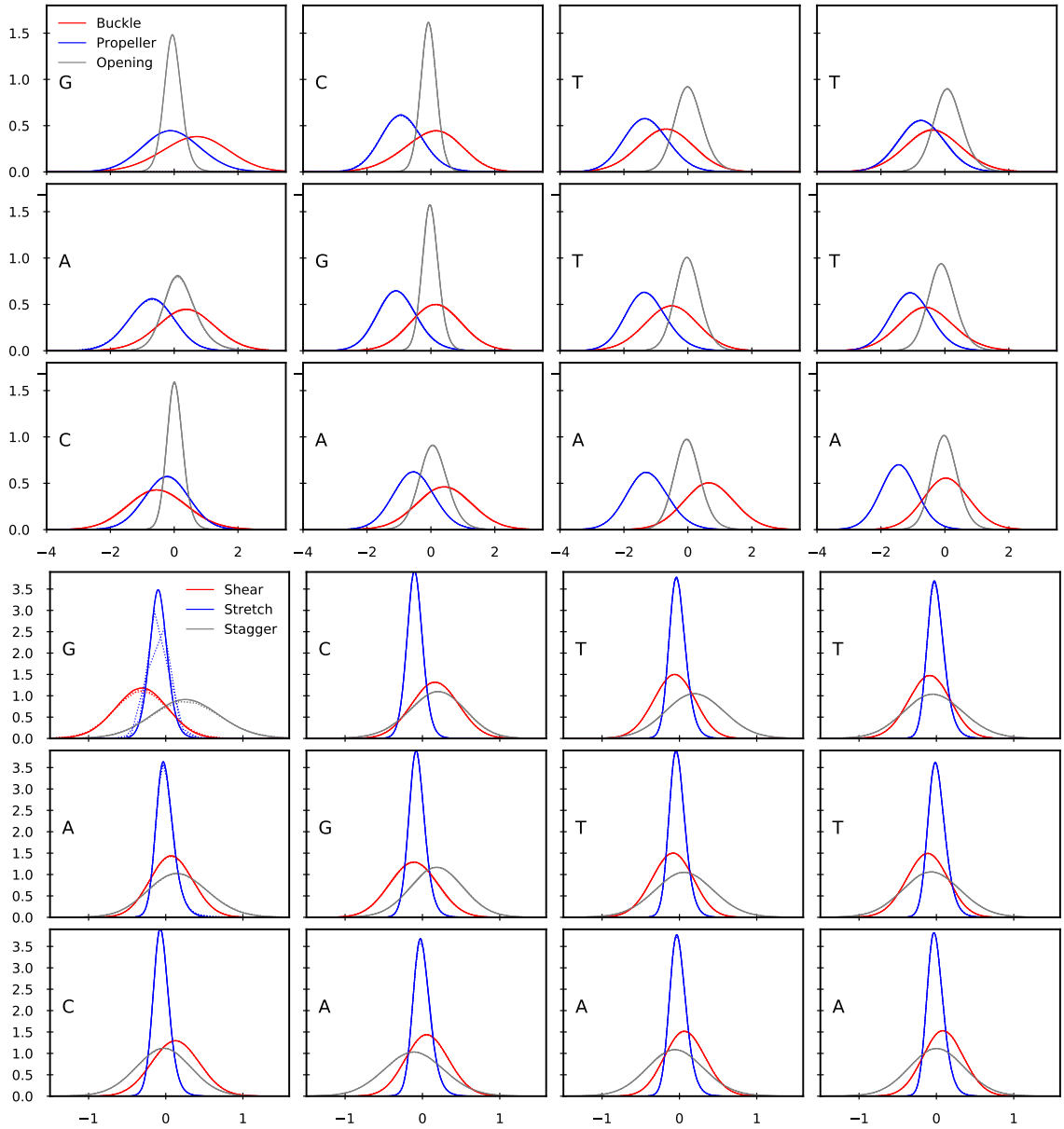


Fig. 3.2 Marginal normalized histograms for intra base-pair rotational (top figure) and translational (bottom figure) coordinates for sequence index 1 in $\mathrm{Lb_{DNA}}$. The coordinates are plotted from left to right and from top to bottom for base-pairs 1 to 12 while reading the strands from both Crick and Watson strands. The histograms in solid and dotted lines are for filtered (snapshots without broken H-bonds) and unfiltered MD data, respectively.

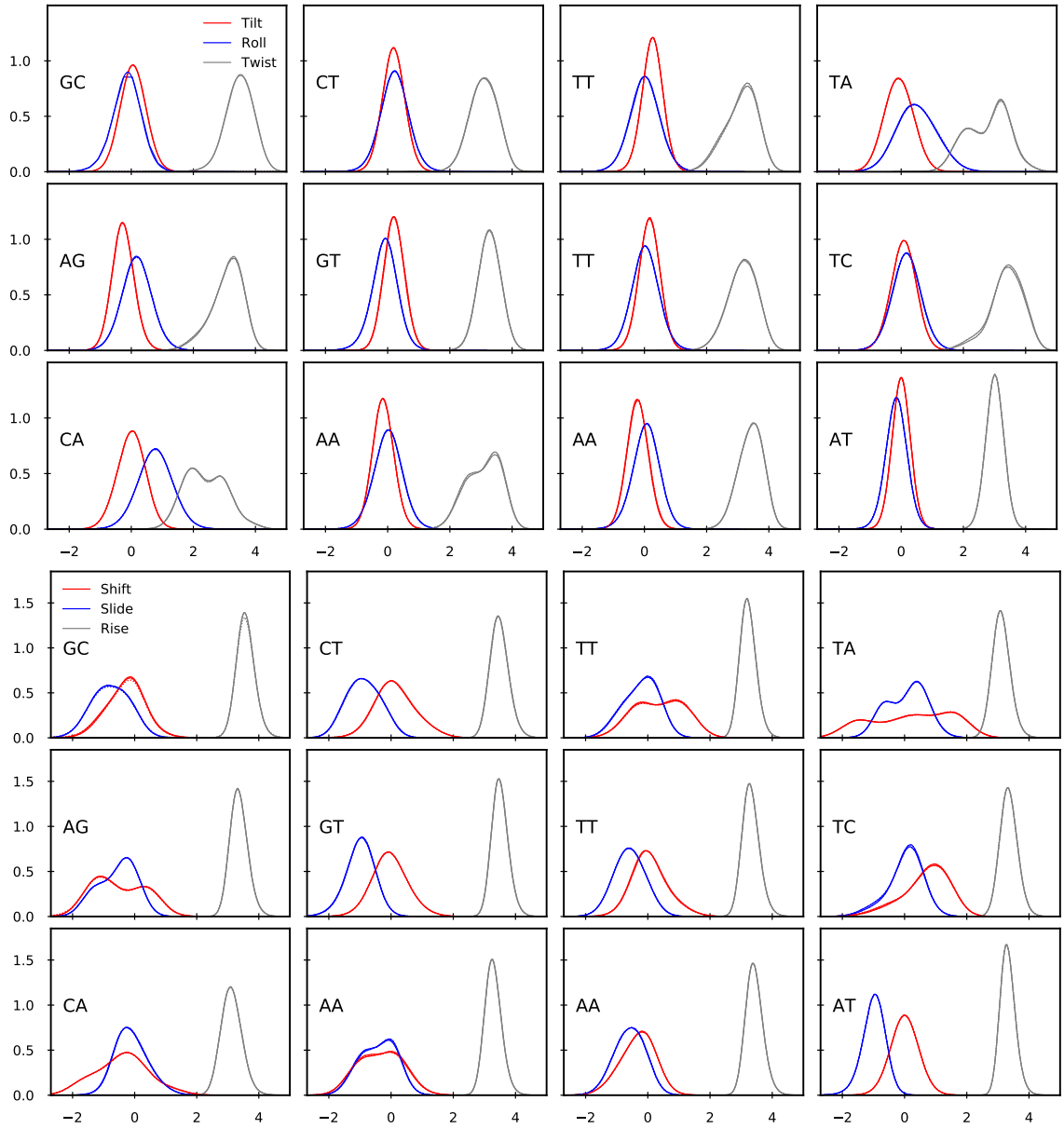


Fig. 3.3 Marginal normalized histograms for inter base-pair rotational (top figure) and translational (bottom figure) coordinates for sequence index 1 in Lb_{DNA}. The coordinates are plotted from left to right and from top to bottom for base-pair steps 1 to 12 while reading the strands from both Crick and Watson strands. The histograms in solid and dotted lines are for filtered (snapshots without broken H-bonds) and unfiltered MD data, respectively.

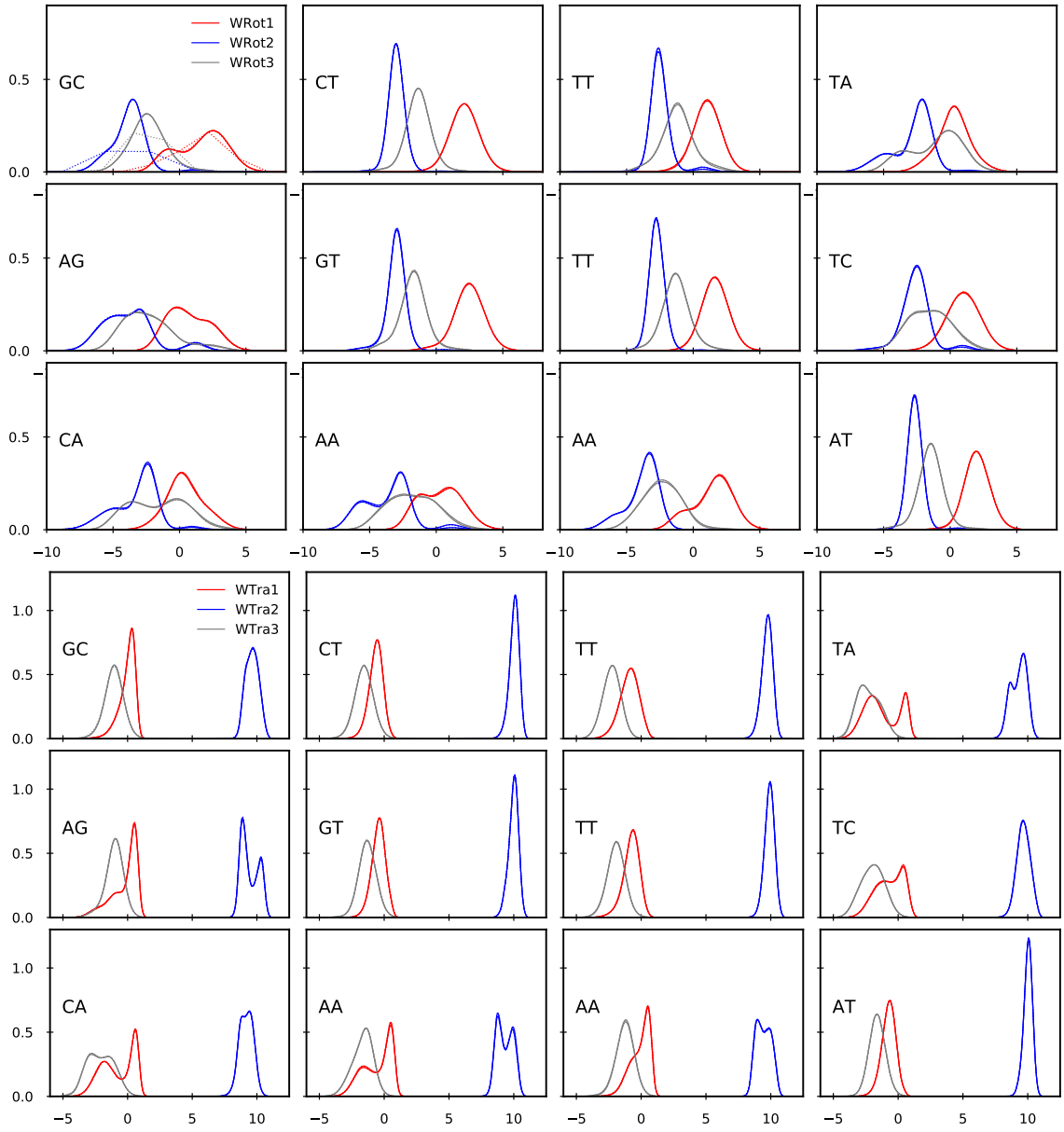


Fig. 3.4 Marginal normalized histograms for Watson phosphate rotational (top figure) and translational (bottom figure) coordinates for sequence index 1 in $\mathrm{Lb}_{\mathrm{DNA}}$. The coordinates are plotted from left to right and from top to bottom for base-pair steps 1 to 12 while reading the strands from both Crick and Watson strands. The histograms in solid and dotted lines are for filtered (snapshots without broken H-bonds) and unfiltered MD data, respectively.

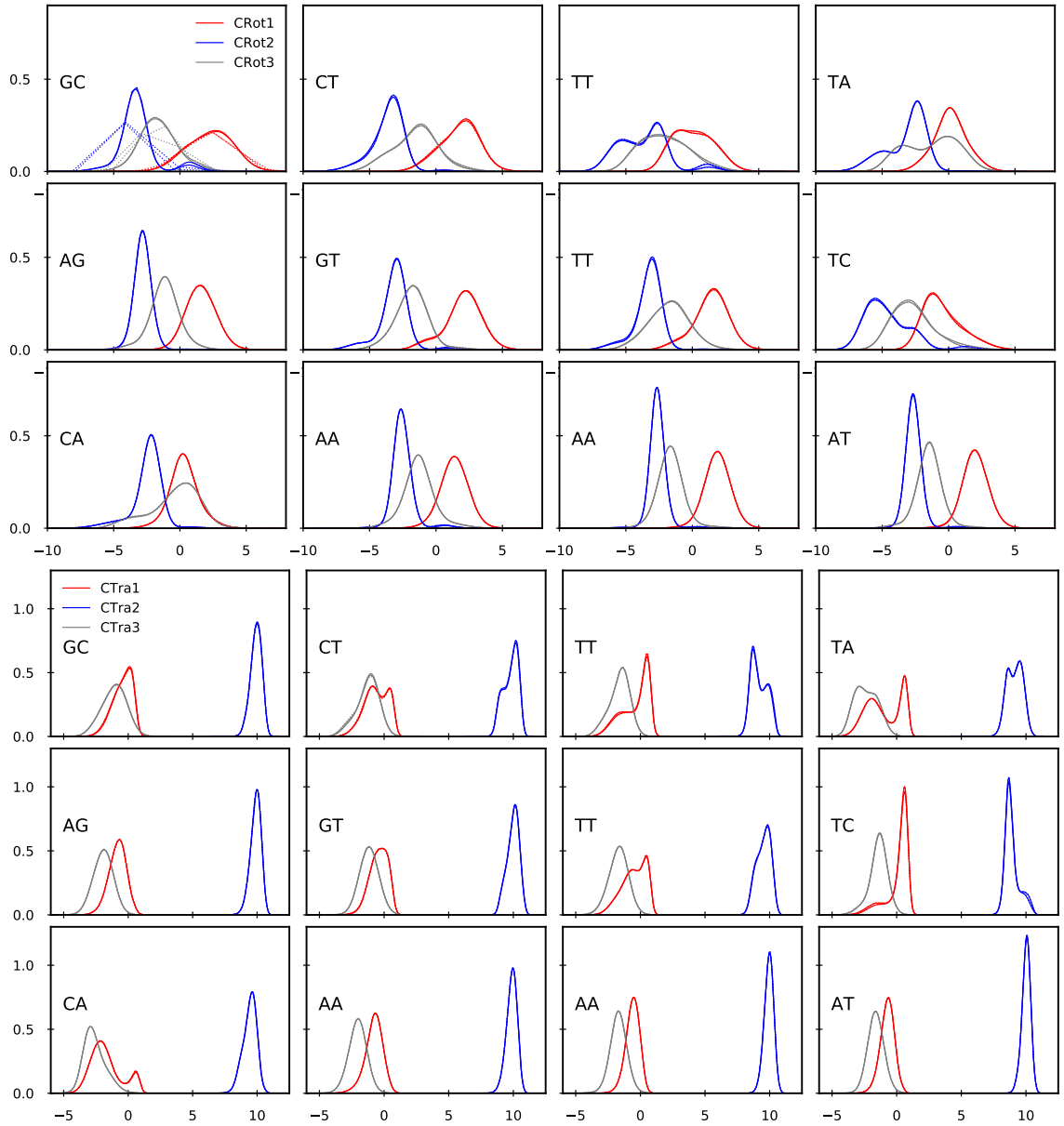


Fig. 3.5 Marginal normalized histograms for Crick phosphate rotational (top figure) and translational (bottom figure) coordinates for sequence index 1 in Lb_{DNA}. The coordinates are plotted from left to right and from top to bottom for base-pair steps 1 to 12 while reading the strands from both Crick and Watson strands. The histograms in solid and dotted lines are for filtered (snapshots without broken H-bonds) and unfiltered MD data, respectively.

In table 3.3, we have provided $\mathcal{E}_{\text{KL}}^{\text{palin}}$ and $\mathcal{E}_{\mathcal{M}}^{\text{palin}}$ per degree of freedom (dof) for the training sequences in Lb_{DNA} over the simulation time. The dof is the number of internal coordinates required to describe the configuration of a given dsDNA sequence, i.e., for a sequence of length N base-pairs, the dofs are $24N - 18$. Note that we have also included one test sequence (index 17) with 20 μs simulation data. In table 3.3, it can be observed that a) different sequences converge differently, for example, the convergence error (both $\mathcal{E}_{\text{KL}}^{\text{palin}}$ and $\mathcal{E}_{\mathcal{M}}^{\text{palin}}$) in sequence index 10 is almost double that of 11, and b) with longer simulation times, the convergence error decreases for all sequences. Here, we have provided the convergence error for 1-10 μs simulation time. Notably, after ≈ 5 μs simulation time, the decrease in convergence error is relatively tiny. Moreover, for sequence index 17, we have generated 20 μs of MD time-series. Although the convergence error decreases for 20 μs simulation data from 10μs, it is still comparable to the convergence error in other training sequences. Lastly, the average of $\mathcal{E}_{\text{KL}}^{\text{palin}}$ for 10 μs data for all training sequences is 0.0050 which is very small, as can be seen in figure 2.5. The corresponding $\mathcal{E}_{\mathcal{M}}^{\text{palin, avg}}$ is 0.0009 which can be considered equal to 0.03 Å or rad/5 per dof which is tiny.

Furthermore, to set a *scale* for this convergence error, we have computed the average pairwise symmetric KL divergence and symmetric Mahalanobis distance between the training sequences in Lb_{DNA} (as described in section 2.5.6), which is 0.4395 and 0.0245 per dof, respectively. It sets a *scale* quantifying the average difference in various sequences in the training library, i.e., quantifies variation over sequence. This *scale* is approximately 27 and 88 times larger than $\mathcal{E}_{\mathcal{M}}^{\text{palin}}$ and $\mathcal{E}_{\text{KL}}^{\text{palin}}$ observed for 10 μs data. Thus, we conclude that 10 μs of the MD simulations are well converged and sufficient to train the cgNA+ model.

Moreover, we again observed that convergence trends are very similar for all palindromic sequences in Lb_{RNA}, Lb_{Met}, and Lb_{Hmet} and therefore, for brevity, we have provided the average of convergence statistics for training sequences in the various libraries. In table 3.4, we have provided the average of $\mathcal{E}_{\text{KL}}^{\text{palin}}$ and $\mathcal{E}_{\mathcal{M}}^{\text{palin}}$ taken over for all the training sequences in Lb_{RNA}, Lb_{Met}, and Lb_{Hmet} and corresponding *scales*. Firstly, the convergence trends in Lb_{Met} and Lb_{Met} are very similar to the trends in Lb_{DNA}, which can be expected as the two libraries differ from Lb_{DNA} slightly in terms of epigenetic modifications in some of the bases. Similarly, the *scales* in the two libraries are comparable to the *scale* in Lb_{DNA}. In contrast, dsRNA sequences appear to converge with a palindromic error similar to that observed in dsDNA, but in a relatively shorter simulation time, and after that, the palindromic error decreases very slowly. It might be attributed to the smaller conformational space of dsRNA compared to dsDNA [131, 132].

Lastly, DRH sequences are not palindromes, and therefore, we can not use palindromic error to quantify the convergence in the MD simulations of DRH. As mentioned earlier, we have generated 10 μs of simulation data for each sequence, which are essentially two independent trajectories of 5 μs started with two different random initial configurations. Therefore, for DRH sequences, we have compared the statistics for these two independent trajectories, which are extremely close and, thus, concluded that the time-series are sufficiently converged.

Index	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
							$\mathcal{E}_{\mathcal{M}}^{\text{palin}}$										
1 μs	0.0041	0.0049	0.0040	0.0030	0.0033	0.0076	0.0016	0.0045	0.0024	0.0034	0.0031	0.0084	0.0031	0.0045	0.0072	0.0016	0.0037
2 μs	0.0021	0.0028	0.0028	0.0022	0.0020	0.0027	0.0017	0.0022	0.0025	0.0014	0.0020	0.0033	0.0028	0.0030	0.0037	0.0014	0.0025
3 μs	0.0016	0.0018	0.0018	0.0014	0.0015	0.0018	0.0010	0.0014	0.0015	0.0013	0.0016	0.0029	0.0016	0.0017	0.0030	0.0013	0.0022
4 μs	0.0014	0.0014	0.0016	0.0013	0.0012	0.0014	0.0011	0.0012	0.0012	0.0013	0.0012	0.0021	0.0015	0.0017	0.0023	0.0011	0.0026
5 μs	0.0013	0.0012	0.0015	0.0009	0.0012	0.0012	0.0012	0.0012	0.0014	0.0011	0.0011	0.0020	0.0014	0.0015	0.0020	0.0011	0.0026
6 μs	0.0012	0.0010	0.0013	0.0009	0.0010	0.0011	0.0009	0.0011	0.0010	0.0012	0.0009	0.0015	0.0011	0.0012	0.0018	0.0009	0.0021
7 μs	0.0011	0.0010	0.0012	0.0009	0.0013	0.0011	0.0009	0.0010	0.0009	0.0013	0.0009	0.0012	0.0010	0.0011	0.0018	0.0007	0.0019
8 μs	0.0011	0.0009	0.0010	0.0008	0.0010	0.0009	0.0008	0.0011	0.0008	0.0013	0.0008	0.0010	0.0009	0.0010	0.0019	0.0007	0.0017
9 μs	0.0011	0.0009	0.0010	0.0007	0.0009	0.0008	0.0009	0.0011	0.0008	0.0014	0.0007	0.0009	0.0008	0.0009	0.0017	0.0007	0.0016
10 μs	0.0009	0.0009	0.0010	0.0007	0.0009	0.0008	0.0008	0.0011	0.0008	0.0012	0.0006	0.0009	0.0008	0.0009	0.0014	0.0007	0.0015
20 μs																	0.0008
							$\mathcal{E}_{\text{KL}}^{\text{palin}}$										
1 μs	0.0570	0.0634	0.0560	0.0445	0.0402	0.0849	0.0211	0.0532	0.0342	0.0470	0.0363	0.1102	0.0426	0.0684	0.1076	0.0239	0.0420
2 μs	0.0213	0.0318	0.0354	0.0259	0.0254	0.0299	0.0192	0.0261	0.0235	0.0139	0.0149	0.0286	0.0311	0.0383	0.0419	0.0144	0.0221
3 μs	0.0118	0.0153	0.0199	0.0129	0.0147	0.0169	0.0092	0.0142	0.0148	0.0112	0.0138	0.0227	0.0165	0.0166	0.0271	0.0116	0.0178
4 μs	0.0084	0.0110	0.0140	0.0112	0.0111	0.0102	0.0094	0.0113	0.0092	0.0080	0.0099	0.0146	0.0142	0.0137	0.0162	0.0083	0.0238
5 μs	0.0085	0.0087	0.0114	0.0070	0.0096	0.0100	0.0091	0.0096	0.0091	0.0070	0.0080	0.0126	0.0144	0.0125	0.0126	0.0065	0.0221
6 μs	0.0078	0.0080	0.0120	0.0061	0.0082	0.0100	0.0066	0.0085	0.0061	0.0070	0.0063	0.0099	0.0090	0.0079	0.0105	0.0048	0.0176
7 μs	0.0072	0.0068	0.0102	0.0056	0.0083	0.0077	0.0058	0.0071	0.0059	0.0074	0.0065	0.0072	0.0084	0.0068	0.0102	0.0040	0.0148
8 μs	0.0076	0.0057	0.0082	0.0047	0.0062	0.0063	0.0050	0.0074	0.0049	0.0072	0.0050	0.0058	0.0069	0.0058	0.0103	0.0037	0.0118
9 μs	0.0070	0.0056	0.0070	0.0042	0.0051	0.0045	0.0051	0.0068	0.0050	0.0079	0.0041	0.0049	0.0056	0.0055	0.0095	0.0035	0.0107
10 μs	0.0055	0.0052	0.0065	0.0037	0.0048	0.0041	0.0042	0.0063	0.0050	0.0062	0.0033	0.0046	0.0052	0.0051	0.0072	0.0032	0.0092
20 μs																	0.0035

Table 3.3 Palindromic error, $\mathcal{E}_{\text{KL}}^{\text{palin}}$ and Mahalanobis error, $\mathcal{E}_{\mathcal{M}}^{\text{palin}}$ per dof for the sequences in the training sequences for cgDNA+ model and a test palindrome sequence. The details of the sequences are in table B.1.

	Lb_{RNA}		Lb_{Met}		Lb_{Hmet}	
Simulation time (μs)	$\mathcal{E}^{palin}_{\mathcal{M},\,avg}$	$\mathcal{E}^{palin}_{KL,\,avg}$	$\mathcal{E}^{palin}_{\mathcal{M},\,avg}$	$\mathcal{E}^{palin}_{KL,\,avg}$	$\mathcal{E}^{palin}_{\mathcal{M},\,avg}$	$\mathcal{E}^{palin}_{KL,\,avg}$
0.25	0.0010	0.0118				
0.50	0.0008	0.0076				
0.75	0.0008	0.0071				
1	0.0007	0.0061	0.0036	0.0492	0.0035	0.0457
2	0.0008	0.0068	0.0022	0.0245	0.0022	0.0231
3	0.0008	0.0063	0.0017	0.0156	0.0017	0.0244
4	0.0009	0.0070	0.0014	0.0114	0.0014	0.0176
5	0.0008	0.0070	0.0012	0.0094	0.0012	0.0142
6	0.0008	0.0060	0.0011	0.0077	0.0011	0.0120
7	0.0008	0.0057	0.0010	0.0067	0.0010	0.0101
8	0.0007	0.0057	0.0010	0.0058	0.0009	0.0086
9	0.0007	0.0050	0.0009	0.0052	0.0009	0.0076
10	0.0006	0.0047	0.0009	0.0048	0.0008	0.0070
scale	0.0177	0.2185	0.0211	0.3378	0.0214	0.3449

Table 3.4 Average palindromic error, $\mathcal{E}^{palin}_{KL,\,avg}$ and average Mahalanobis error, $\mathcal{E}^{palin}_{\mathcal{M},\,avg}$ per dof for training sequences (error is averaged over all training sequences) in Lb_{RNA}, Lb_{Met}, and Lb_{Hmet}. The details of the sequences are given in tables B.1 and B.2. The *scale* (which quantifies variation over sequence) is obtained by computing the average pair-wise difference between all the training sequences.

3.6 Distribution of internal coordinates in MD simulations

This section discusses the distributions of internal coordinates in MD time series. For dsDNA, the distribution of helical coordinates (base internal coordinates) has been extensively studied before, and it is well known that the helical coordinates often show a non-Gaussian distribution. For instance, in studies by the ABC consortium [22, 50, 102, 147] as well as in other studies [9, 44, 153], it was observed that inter base-pair parameters (in particular, Shift, Slide, and Twist) often show the multi-peak distribution, and for a given dimer step, the behavior also depends on the flanking tetramer context. Similar observations were also found true for experimental X-ray data [88, 112]. In ref. [158], the distributions for both inter and intra base-pair coordinates have been investigated with the findings that non-Gaussian behavior is only dominant in Shift, Slide, and Twist. Also, it was argued that the bimodality in internal coordinates is not the result of the choice of configuration parameterization or MD forcefield but is an inherent physical property of dsDNA. Lastly, it was concluded that the Gaussian approximation on the internal coordinates for bases is a reasonable choice.

Again, referring to figures 3.2 to 3.5, one can observe that the distributions for a) intra coordinates are close to Gaussian, b) inter coordinates, several are non-Gaussian distributions, and c) lastly, most phosphate coordinates are non-Gaussian. Note that one of the model assumptions (see section 2.3.1) is that the distributions of internal coordinates are Gaussian, and for phosphates, it is certainly not the case. We have quantified this approximation error in section 3.7.

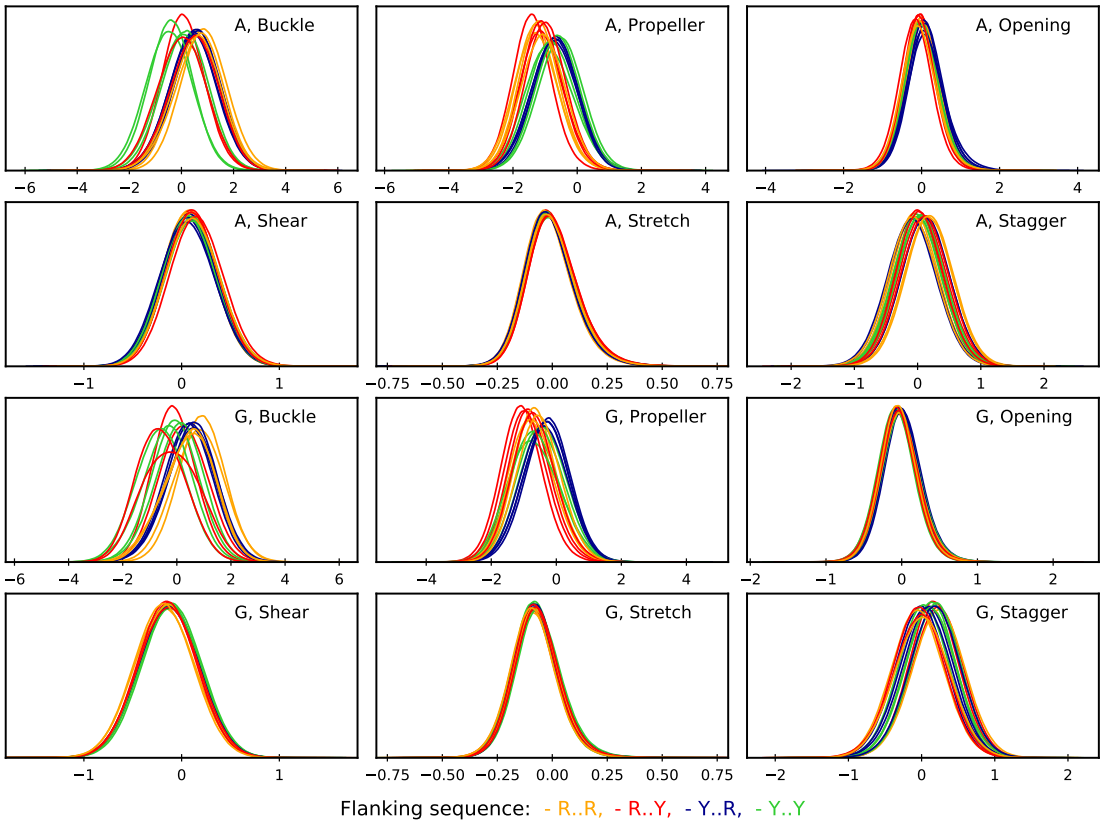
Here in this section, we have provided plots for internal coordinates distribution for monomer (intra coordinates) and dimers (inter and phosphate coordinates) in all independent trimer and tetramer flanking contexts, respectively. We have plotted the intra coordinates for A and G (T

and C are dependent) in all possible trimer contexts for dsDNA and dsRNA in figure 3.6. The various trimer contexts are plotted in different colors based on Y/R classification. It can be easily observed from figure 3.6(a) that the distribution of intra coordinates is almost Gaussian for all intra coordinates for A and G in both dsDNA and dsRNA. Moreover, for a given base-pair, the distribution of intra coordinates is influenced by its flanking context, and different coordinates are influenced differently. For example, distributions for Buckle, Propeller, and Stagger depend more on trimer context than other intra coordinates. Note that in figures 3.6 to 3.10, we have plotted MD time-series data for the training sequences of Lb_{DNA}, Lb_{RNA}, and Lb_{DRH}.

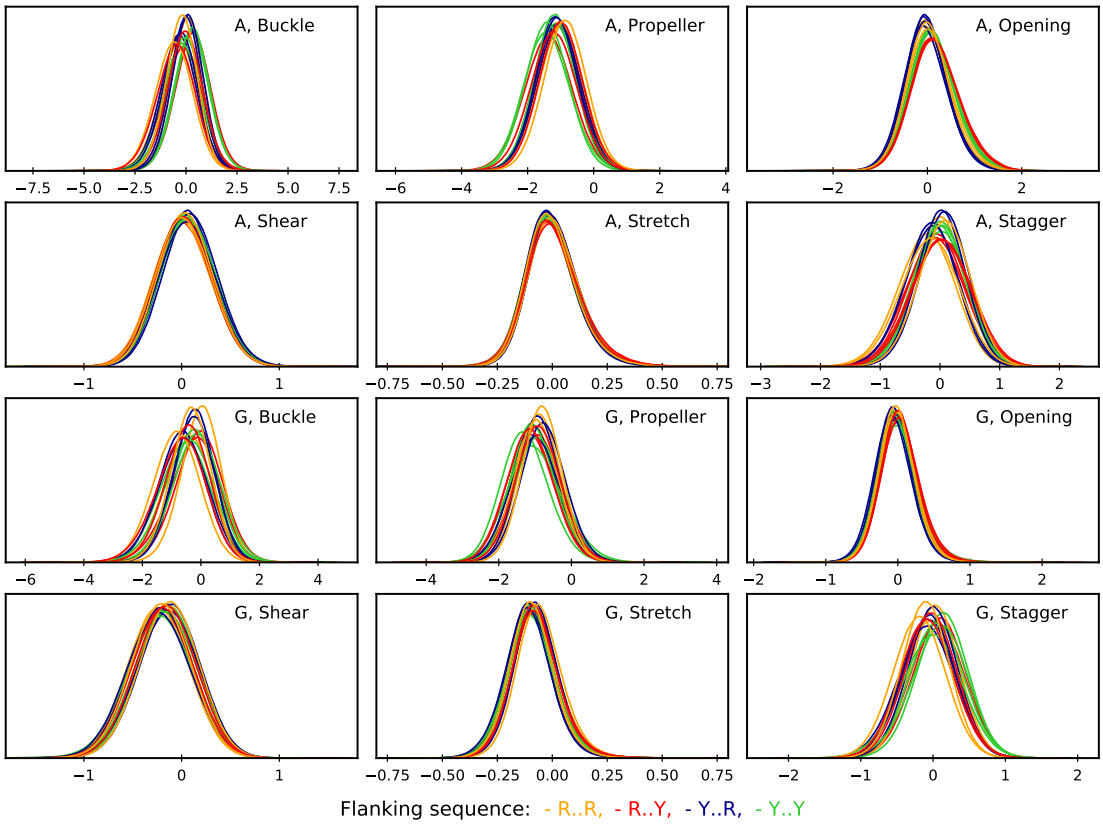
In figure 3.7, we have plotted inter base-pair step and phosW coordinates for CG and AT. For CG, the distributions for inter and phosW coordinates are often non-Gaussian, in contrast, the corresponding distributions for AT are close to Gaussian. Here we have only shown two typical contrasting examples, but as previously observed, particularly for YR steps, Twist, Shift, and Slide have non-Gaussian distribution. Moreover, it can be observed that non-Gaussian behavior in inter coordinates appears to be correlated with phosphate coordinates. A systematic investigation of the relation between inter coordinates and backbone conformations [44, 147] revealed that multi-modality in inter coordinates (dominant in YR steps and certain tetramer contexts) is strongly coupled with BI-BII backbone conformational states.

Moreover, in figures 3.6 and 3.8, we have plotted intra, inter and phosW coordinates for dsRNA for some example cases. Firstly, the distributions for intra coordinates, similar to dsDNA, are close to Gaussian, and Buckle, Propeller, and Stagger are most sensitive to the flanking sequence contexts. In contrast to the observations for dsDNA, the distributions for inter and phosphate coordinates for dsRNA are close to Gaussian. Notably, all the internal coordinates distributions depend on flanking sequence context, but the sensitivity is relatively less as compared to dsDNA. As mentioned earlier, in the cgNA+ model, we have assumed that the underlying distributions of the internal coordinates in MD simulations are Gaussian, and for dsRNA, the corresponding approximation error in the model should be relatively smaller than for dsDNA. A detailed quantification of this approximation error is in section 3.7.

Lastly, in figures 3.9 and 3.10, we have plotted the corresponding plots for DRH. In DRH, one of the strands is DNA, and the other is RNA (more details are provided in section 1.1.4), and the DNA strand is chosen as the reading strand. Once again, in figure 3.9, one can observe that the distributions for intra coordinates are almost Gaussian. The distributions of inter coordinates are non-Gaussian, particularly Twist and Shift, for the CG step, whereas they are close to Gaussian for the AT step. Notably, the deviation from the Gaussian behavior in the distributions of inter coordinates for DRH is less than the corresponding distributions for dsDNA (figure 3.7) and more than dsRNA (figure 3.8). Furthermore, for the phosphate coordinates, the distributions are very interesting. The DNA strand behaves like pure dsDNA, while the RNA strand behaves like pure dsRNA. It is consistent with prior literature but quantified using different metrics [132, 133, 196]. Figure 3.10(a) plots the distribution of phosphate coordinates for Watson (reading strand), which is the DNA strand by choice, while (b) plots the corresponding phosphate coordinates for the Crick strand, which is the RNA strand. For AT, distributions for phosC and phosW coordinates show Gaussian behavior; in contrast, for CG, the distributions are non-Gaussian for the DNA (Watson) strand while close to Gaussian for RNA (Crick) strand.

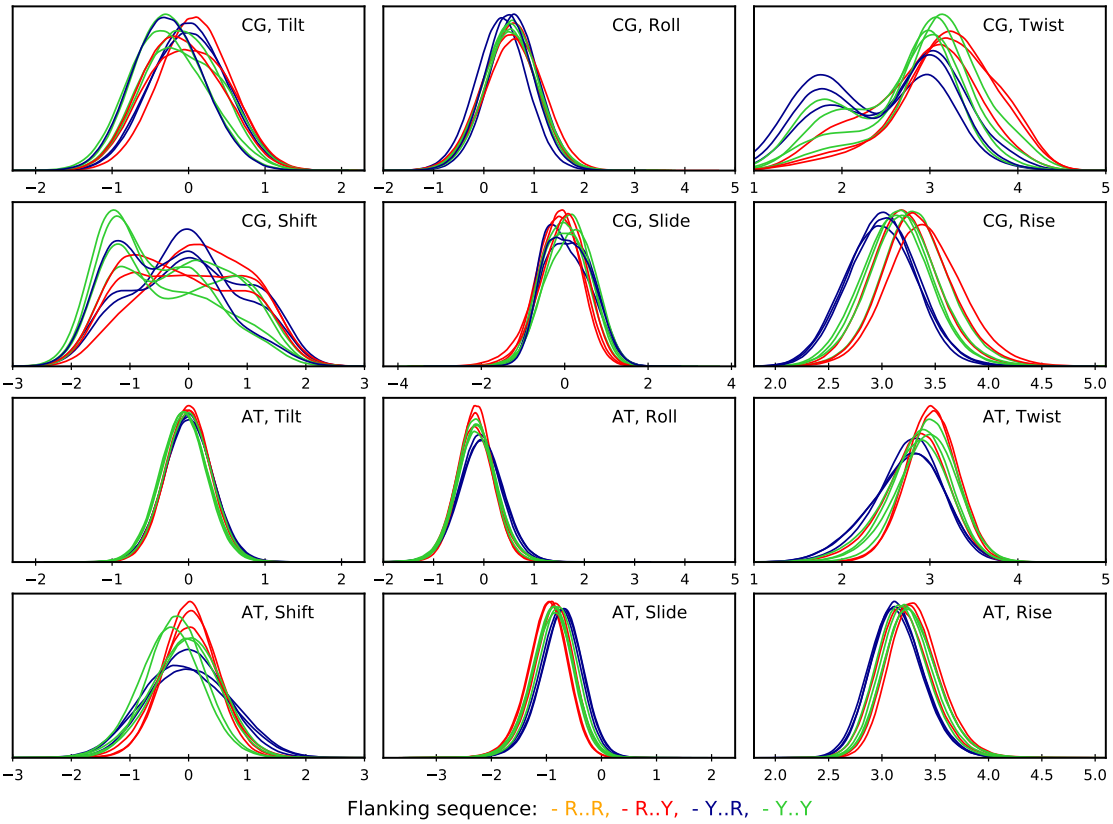


(a) Intra base-pair coordinates in dsDNA

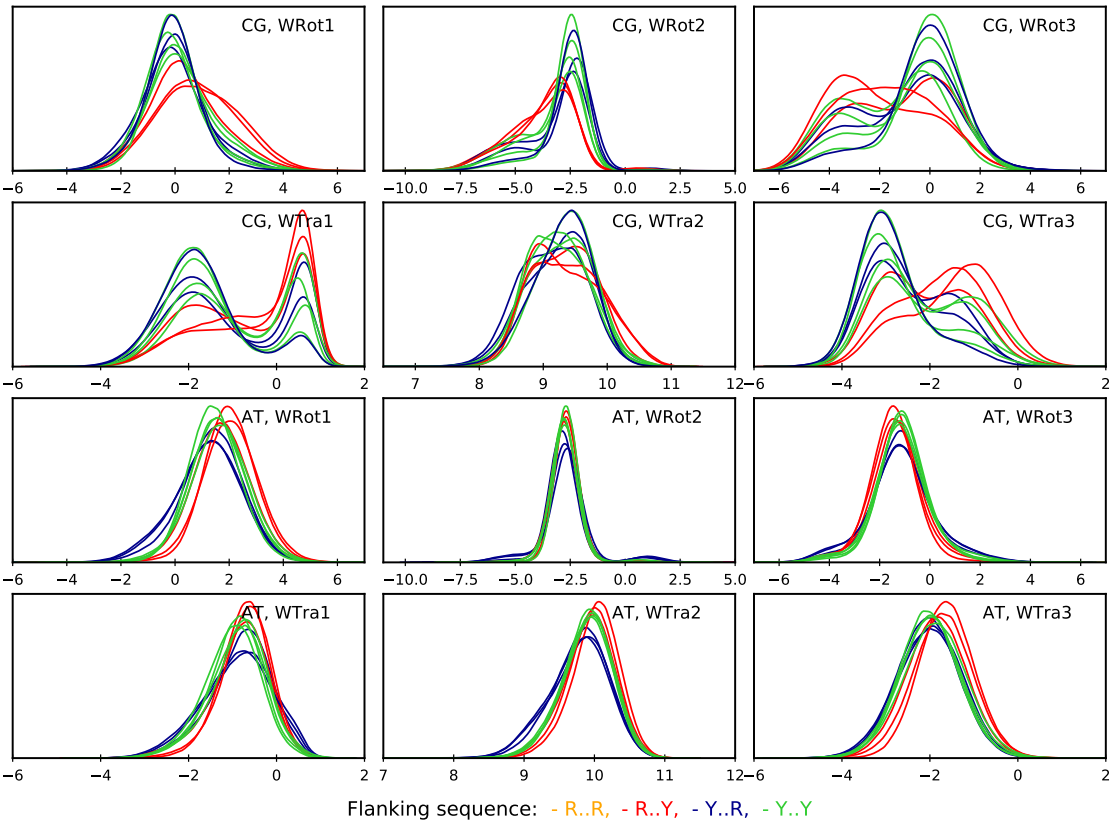


(b) Intra base-pair coordinates in dsRNA

Fig. 3.6 The normalized histograms for intra base-pair coordinates for A and G in all 16 trimer contexts in (a) dsDNA and (b) dsRNA as observed in MD time series of training sequences. The various contexts are plotted in different colors based on Y and R classification.

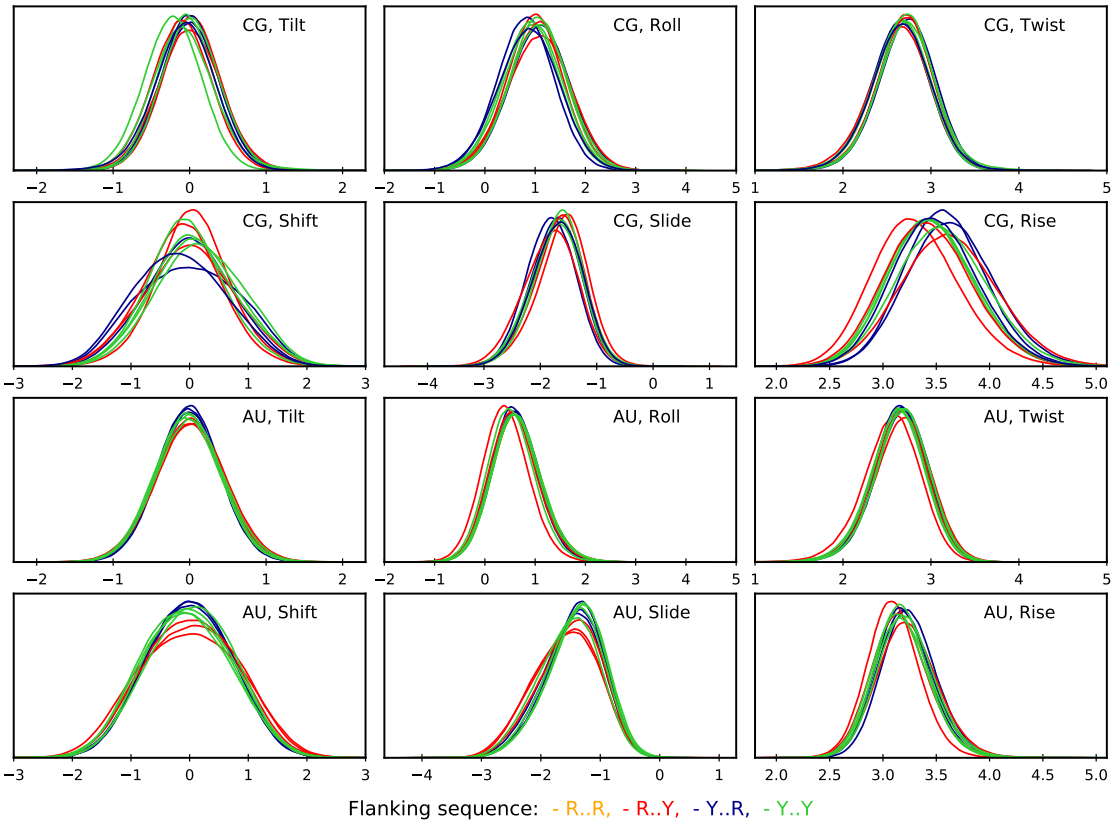


(a) Inter base-pair step coordinates in dsDNA

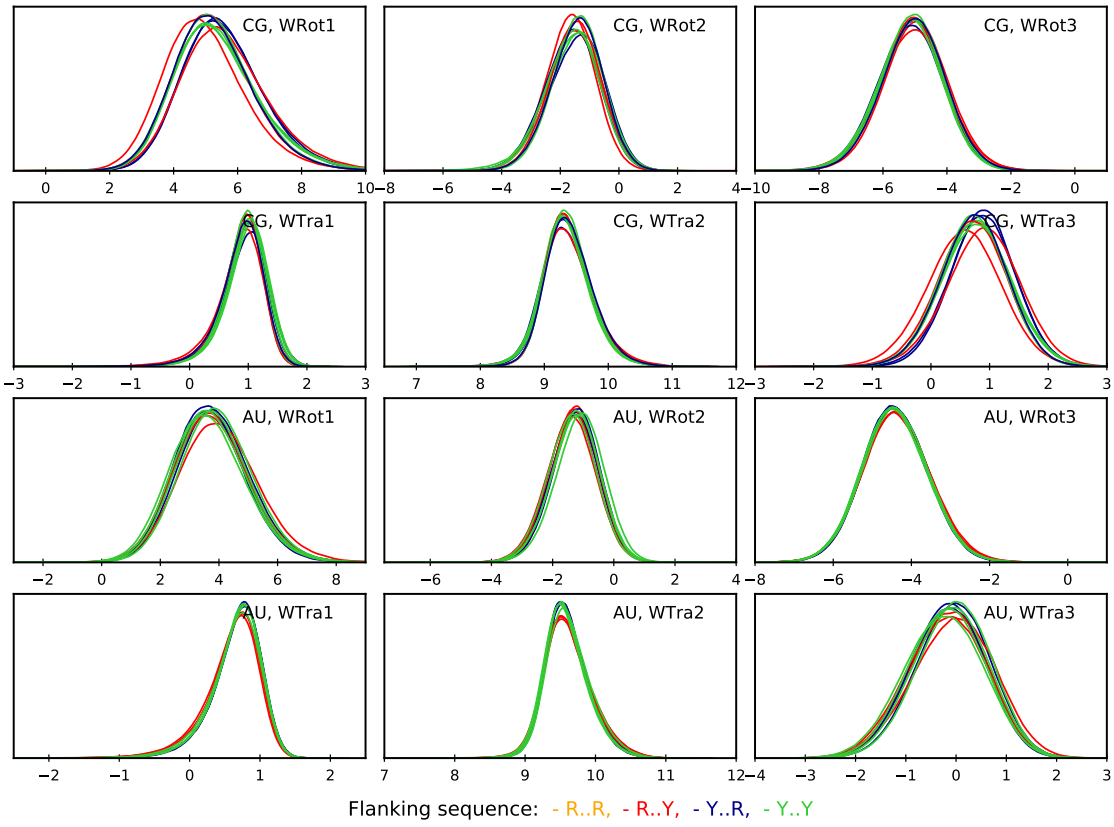


(b) PhosW coordinates in dsDNA

Fig. 3.7 The normalized histograms for (a) inter base-pair step and (b) phosW coordinates for CG and AT in all 10 independent tetramer contexts for dsDNA observed in MD time series of the training sequences in Lb_{DNA}. The various contexts are plotted in different colors based on Y and R classification.

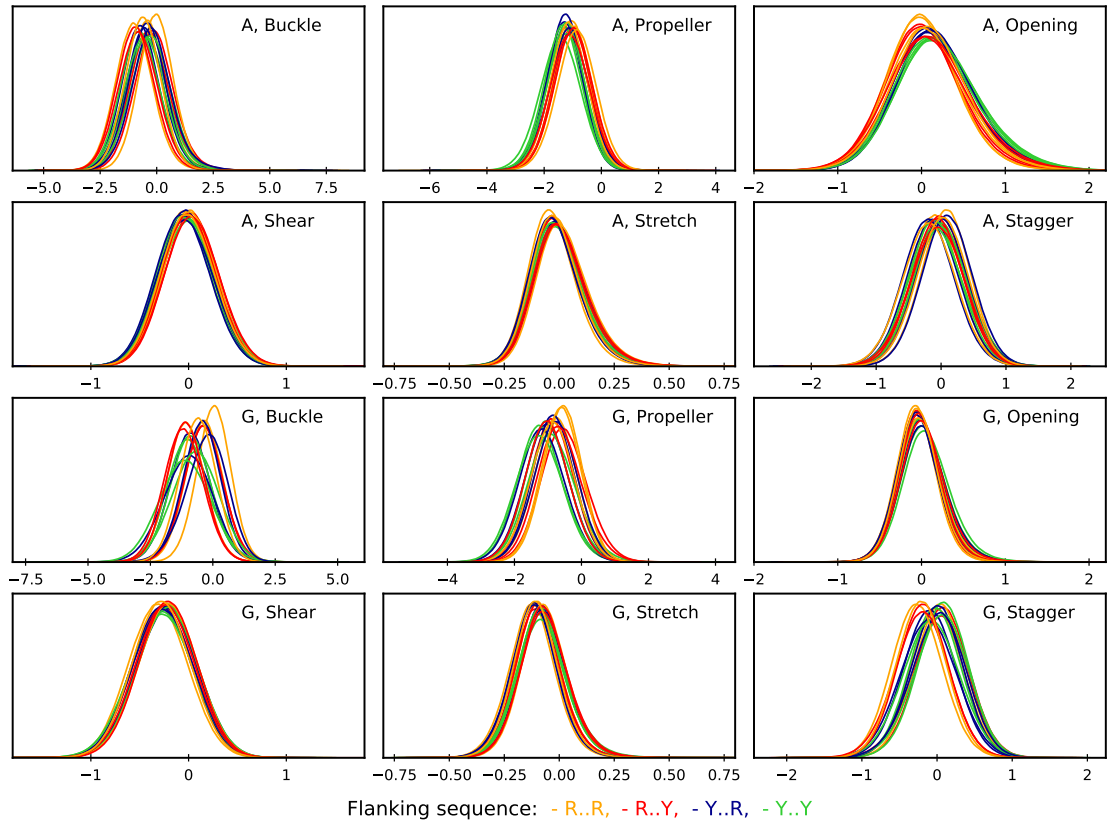


(a) Inter base-pair step coordinates in dsRNA

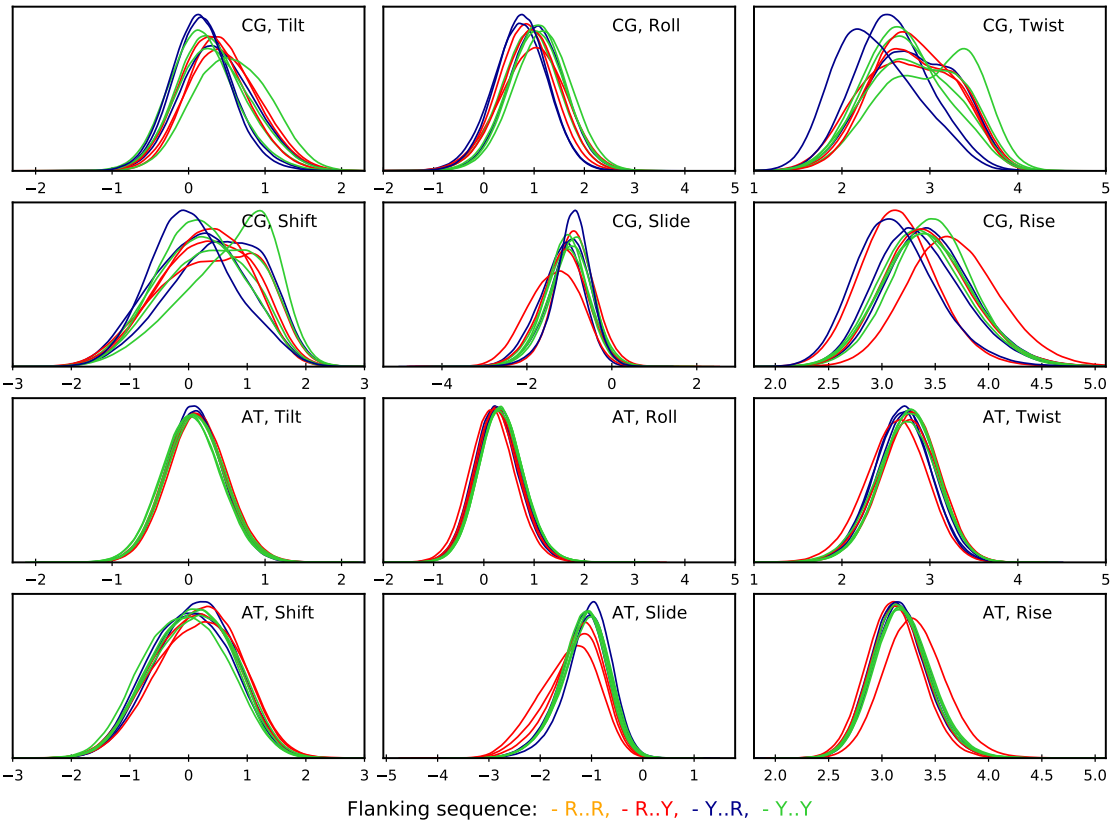


(b) PhosW coordinates in dsRNA

Fig. 3.8 The normalized histograms for (a) inter base-pair step and (b) phosW coordinates for CG and AU in all 10 independent tetramer contexts for dsRNA observed in MD time series of all the training sequences in Lb_{RNA}. The various contexts are plotted in different colors based on Y and R classification.



(a) Intra base-pair coordinates in DRH



(b) Inter base-pair step coordinates in DRH

Fig. 3.9 The normalized histograms for (a) intra base-pair coordinates for A and G and (b) inter base-pair step coordinates CG and AT in all immediate flanking contexts for DRH observed in MD time series of all the training sequences in Lb_{DRH}. The various contexts are plotted in different colors based on Y and R classification.

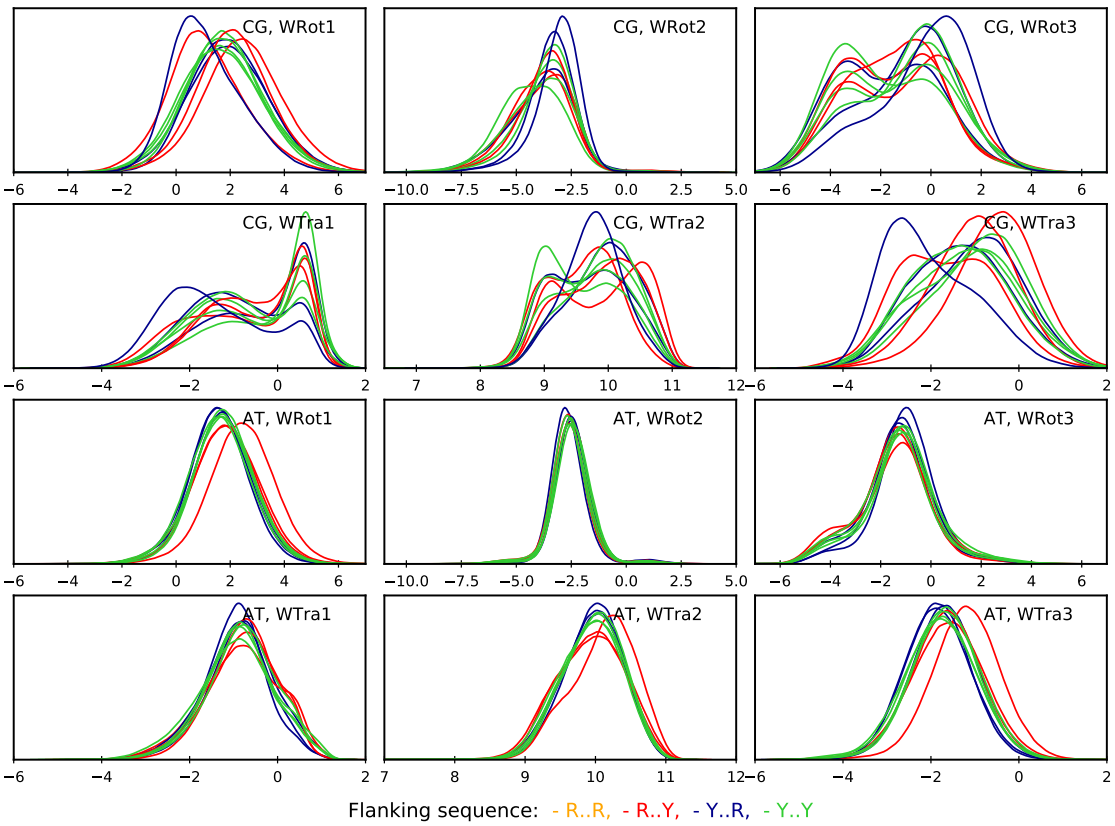
3.7 Gaussian approximation error

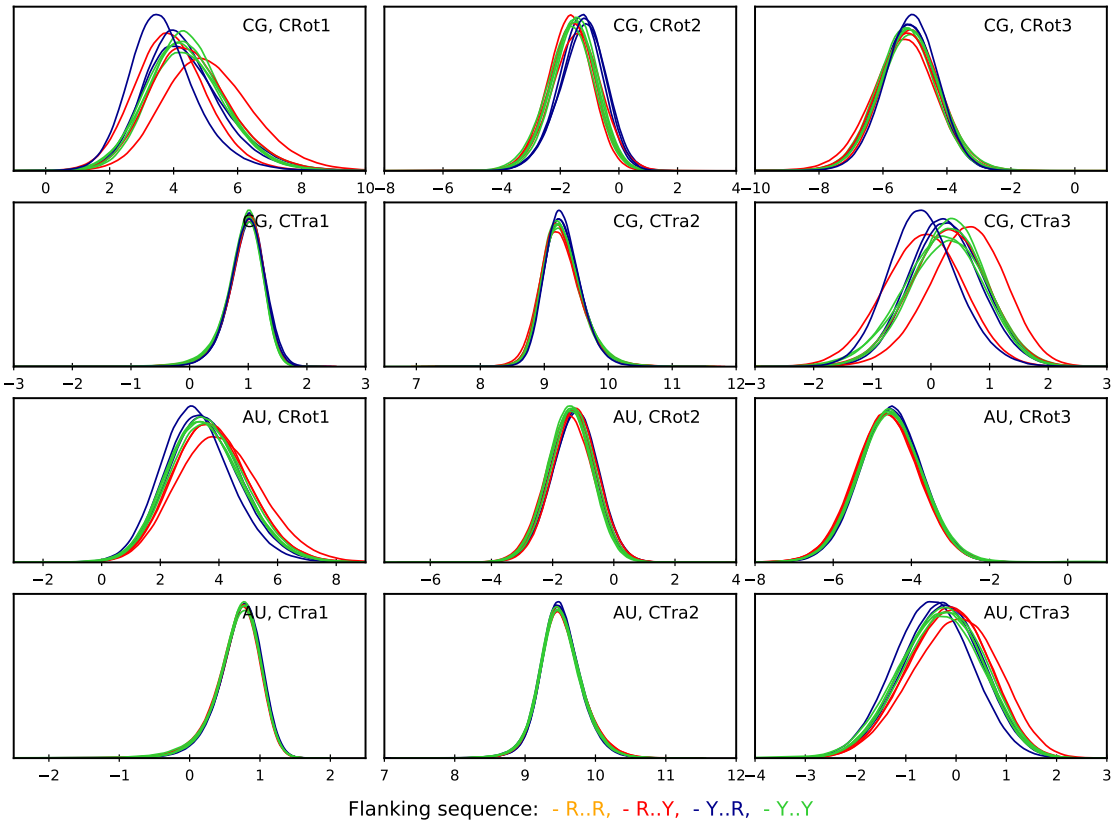
In this final section, we have quantified the error associated with the Gaussian approximation of the model in the underlying distributions of the internal coordinates. In figures 3.2 to 3.10, we have shown that the observed distributions of internal coordinates in MD simulations often deviate from Gaussian behavior, in particular, phosphate coordinates and Shift, Slide, and Twist in the inter-coordinates. Here, we have quantified the error, $\mathcal{E}_{\mathrm{KL}}^{\mathrm{Gauss}}$ corresponding to the assumption in the model that the internal coordinates follow Gaussian behavior by numerically computing symmetric KL divergence between the observed internal coordinate distribution in MD simulations and the corresponding best-fit Gaussian as defined in section 2.5.2.

In figure 3.11, we have plotted $\mathcal{E}_{\mathrm{KL}}^{\mathrm{Gauss}}$ for each internal coordinate of sequence index 1 in (a) $\mathrm{Lb}_{\mathrm{DNA}}$, (b) $\mathrm{Lb}_{\mathrm{RNA}}$, and (c) $\mathrm{Lb}_{\mathrm{DRH}}$ as heat map with the sequence shown on the labels. The plots for the corresponding distributions for sequence index 1 in $\mathrm{Lb}_{\mathrm{DNA}}$ are shown in figures 3.2 to 3.5 where it can be visually concluded that intra coordinates are close to Gaussian, some of the inter coordinates deviate from Gaussian behavior and almost all phosphate coordinates show non-Gaussian distributions. The same observation can be confirmed quantitatively in terms of $\mathcal{E}_{\mathrm{KL}}^{\mathrm{Gauss}}$ from figure 3.11(a) in which KL divergence between the observed distribution and corresponding best-fit Gaussian for intra coordinates is approximately 0.004 (average), for inter coordinates is 0.011 (average) and for Crick/Watson phosphate coordinates is 0.057 (average). Notably, $\mathcal{E}_{\mathrm{KL}}^{\mathrm{Gauss}}$ for any particular internal coordinate, in general, depends on the dimer step (or monomer for intra coordinates) as well as on the flanking context. For instance, $\mathcal{E}_{\mathrm{KL}}^{\mathrm{Gauss}}$ for Wtra1 for AG steps is much larger than any TT steps as well as $\mathcal{E}_{\mathrm{KL}}^{\mathrm{Gauss}}$ for AG steps at 5^{th} or 22^{nd} in different flanking contexts are considerably different. In the same plot, one can also observe a Crick-Watson symmetry in inter/intra coordinates (the two half of the plots look similar) and in Crick and Watson phosphates (the two are similar when looked at from opposite directions).

In figure 3.8, we have plotted the distributions for internal coordinates as observed in the MD simulations of training sequences in $\mathrm{Lb}_{\mathrm{RNA}}$, highlighting that the distributions are very close to Gaussian behavior. In figure 3.11(b), we have shown the corresponding $\mathcal{E}_{\mathrm{KL}}^{\mathrm{Gauss}}$ in a heat map with similar conclusions. The average $\mathcal{E}_{\mathrm{KL}}^{\mathrm{Gauss}}$ is approximately 0.004, 0.003, and 0.019 in intra, inter, and phosphate coordinates, respectively. The average $\mathcal{E}_{\mathrm{KL}}^{\mathrm{Gauss}}$ in intra coordinates are comparable for dsDNA and dsRNA. In contrast to dsDNA, the average $\mathcal{E}_{\mathrm{KL}}^{\mathrm{Gauss}}$ is considerably lower for dsRNA in inter and phosphate coordinates. Moreover, in figure 3.11(c), we have shown the corresponding $\mathcal{E}_{\mathrm{KL}}^{\mathrm{Gauss}}$ for sequence index 1 in $\mathrm{Lb}_{\mathrm{DRH}}$. Firstly, it can be noted that there is no Crick-Watson symmetry in phosphates, inter, or intra coordinates. The average $\mathcal{E}_{\mathrm{KL}}^{\mathrm{Gauss}}$ in intra coordinates is approximately 0.005, which is comparable to the corresponding observations in dsDNA or dsRNA. The corresponding error in inter coordinates is approximately 0.006 which is almost double than $\mathcal{E}_{\mathrm{KL}}^{\mathrm{Gauss}}$ for dsRNA (0.003) and half than $\mathcal{E}_{\mathrm{KL}}^{\mathrm{Gauss}}$ for dsRNA (0.011). Lastly, the two phosphate coordinates (on Crick and Watson strands) in DRH, as shown in figure 3.10 behave differently depending on the strand type (DNA or RNA). The average $\mathcal{E}_{\mathrm{KL}}^{\mathrm{Gauss}}$ for Crick phosphates (RNA strand) and Watson phosphates are approximately 0.036 and 0.029, which are comparable and closer to the observed values for dsRNA (0.019) than dsDNA (0.056). However, note that Wtra1 is often multi-modal, in particular, for TG and GA steps.

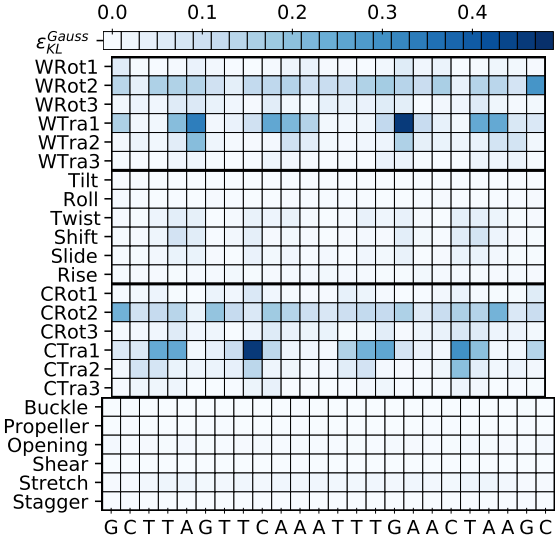
Thus, in this section, we have quantified the approximation error due to the Gaussianity imposition on the observed internal coordinates distributions in MD simulations. One can visualize the magnitude of this error expressed in terms of the KL divergence using figure 2.5. In general, we can conclude that for intras and most inters, the Gaussianity imposition is a natural choice. In contrast, for phosphate coordinates, $\mathcal{E}_{\mathrm{KL}}^{\mathrm{Gauss}}$ is considerable (average for dsDNA ≈ 0.057) which goes as high as 0.459 for GA step as shown in figure 3.11(a). For such cases, the Gaussianity imposition is questionable. Non-Gaussian models are certainly a better approach for treating such cases, but introduce several modeling challenges, e.g., a significant increase in model parameters and finding a parameter set in such high dimensions. In this work, we have continued with the Gaussianity approximation, and non-Gaussian models are left for future research. In principle, such a non-Gaussian model can be realized using quartic free-energy; however, the complexity of quartic models could easily explode for such large dimensions. A feasible compromise can be introducing sequence-independent non-Gaussian perturbation in the sequence-dependent Gaussian cgDNA+ model.



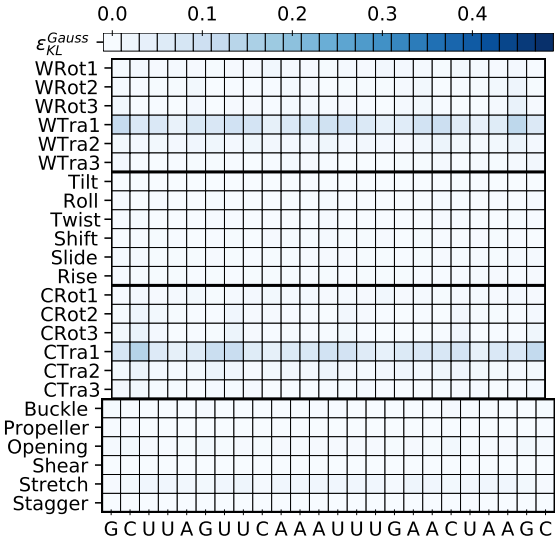


(b) PhosC coordinates in DRH

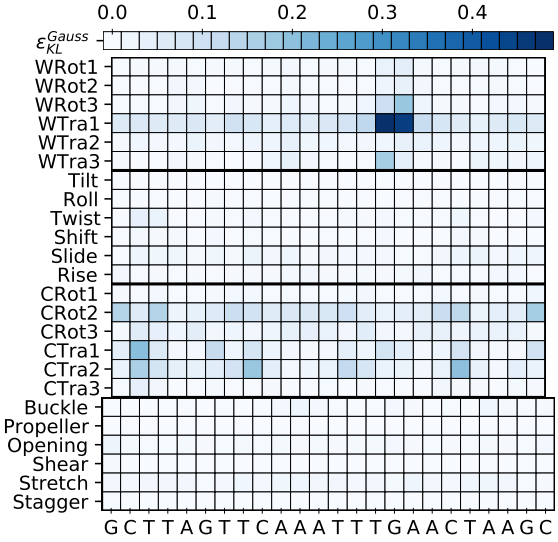
Fig. 3.10 The normalized histograms for (a) phosW and (b) phosC coordinates for CG and AT/AU in all flanking tetramer contexts for DRH were observed in the MD time series of all the training sequences in $\mathrm{Lb}_{\mathrm{DRH}}$. The various contexts are plotted in different colors based on Y and R classification.



(a) $\mathcal{E}_{\mathrm{KL}}^{\mathrm{Gauss}}$ for sequence index 1 in $\mathrm{Lb}_{\mathrm{DNA}}$



(b) $\mathcal{E}_{\mathrm{KL}}^{\mathrm{Gauss}}$ for sequence index 1 in $\mathrm{Lb}_{\mathrm{RNA}}$



(c) $\mathcal{E}_{\mathrm{KL}}^{\mathrm{Gauss}}$ for sequence index 1 in $\mathrm{Lb}_{\mathrm{DRH}}$

Fig. 3.11 Gaussian approximation error, $\mathcal{E}_{\mathrm{KL}}^{\mathrm{Gauss}}$ in the internal coordinate distribution in MD simulations for sequence index 1 in (a) $\mathrm{Lb}_{\mathrm{DNA}}$, (b) $\mathrm{Lb}_{\mathrm{RNA}}$, and (c) $\mathrm{Lb}_{\mathrm{DRH}}$ which is numerically computed as the symmetric KL divergence between the observed internal coordinate distribution in MD simulations and the corresponding best-fit Gaussian.

CHAPTER 4

cgNA+ parameter sets for double-stranded nucleic acids

This chapter extends the cgDNA+ model introduced in A. Patelli's thesis [149] to the cgNA+ model by estimating parameter sets for various other double-stranded nucleic acids (dsNAs). cgNA+ is a coarse-grained model of dsNA (including dsDNA, dsRNA, and DNA:RNA hybrid) to predict the probability distribution function (pdf) of an arbitrary dsNA sequence (in standard A, T, C, G, U alphabets) at pre-specified physical solvent conditions. cgNA+ model explicitly considers phosphates and bases as rigid bodies in $\in SE(3)$ and uses modified CURVES+ [101] helicoidal coordinates for their configuration (see section 2.2). The model is trained on extensive molecular dynamics (MD) time-series (refer chapters 2 and 3) of a comprehensive set of rationally designed sequences. Given a sequence S along the reading strand and a parameter set $\mathcal{P}_{\text{NA}}$ (i.e., different parameter sets are trained for different kind of dsNAs), the cgNA+ model predicts a Gaussian pdf in the configuration space by reconstructing a ground-state $\hat{w}(S, \mathcal{P}_{\text{NA}}) \in \mathbb{R}^{24N-18}$, and a positive-definite stiffness matrix $\mathcal{K}(S, \mathcal{P}_{\text{NA}}) \in \mathbb{R}^{24N-18 \times 24N-18}$:

$$\rho(w; S, \mathcal{P}_{\text{NA}}) = \frac{1}{Z} exp\{-\frac{1}{2}(w - \hat{w}) \cdot \mathcal{K}(w - \hat{w})\}, \tag{4.1}$$

where $\mathcal{P}_{\text{NA}}$ is the parameter set and is discussed in detail in section 4.2. A summary of the cgDNA+ model and its training procedure is in chapter 2 and more details can be found in ref. [149].

In the cgNA+ model, we have made a few modifications in the parameter set estimation techniques from the original cgDNA+ model to simplify the training procedure as discussed in section 4.1 and updated the training library used to train end-block dsDNA parameters, which allows prediction of sequences with any ends (previously not possible in the cgDNA+ model). Moreover, we have significantly enhanced the quality of the parameter set in the cgNA+ model using more extensive MD training data. In the following section 4.2, we have introduced the cgNA+ model, and in section 4.3, we have assessed the accuracy of the model and discussed various sources of errors in the model and their quantification. The last part presents the model's applications. Finally, we recall that all of the discussion in this work is pertinent to dsNAs.

Details of all the codes and data used in this chapter are provided appendix F.

4.1 Updates in the cgNA+ model

The modeling aspects, including coarse-graining, various modeling assumptions, and parameter estimation in the cgNA+ model, remain the same as in the prior cgDNA+ model. However, extending the model parameter sets for various dsNAs is a non-trivial task. In the section, we have described the various updates or changes made in the parameter estimation procedure, MD protocol, and the training library. These updates in the MD protocol and the training data have significantly enhanced the quality of the cgNA+ parameter sets.

4.1.1 Modifications in parameter set estimation techniques

The parameter set estimation procedure remains similar to that used for training the cgDNA+ model, except for a few changes. In particular, in the cgDNA+ model, first, from the Gaussian pdf observed in the MD simulations (after filtering snapshots with broken H-bond), a banded Gaussian pdf was computed. Then, the best-fit parameter set is estimated by minimizing the sum of KL divergences between the model reconstructions and banded Gaussian pdfs in the MD simulations for all training sequences. In the cgNA+ model, the best-fit parameter set is directly computed from the observed Gaussian pdf in the MD simulations without computing the banded Gaussian pdfs. This step simplifies the training procedure.

Moreover, estimating the parameters for various dsNAs requires several adaptations in the parameter estimation step, for example, expansion of the parameter set for DNA:RNA hybrid (DRH) as there is no Crick-Watson (CW) symmetry or the parameter set for epigenetically modified DNA requires additional (to standard) parameter blocks for modified steps. More details are provided in sections 2.4.2, 4.2 and 6.1.4.

4.1.2 Updates in the MD protocol

In the cgNA+ model, we have different parameter sets for various dsNAs trained on the extensive MD simulations of the corresponding dsNAs. The MD protocol to simulate training data for the cgNA+ model is described in section 3.2. As required, we have used different MD force-fields to simulate different dsNAs. However, we have made some changes to the other MD simulation parameters. In particular, we have replaced the water and ion model from SPC/E [17] and Dang ions parameters [40] to TIP3P [83] and Joung and Cheatham ions models [84], respectively. The previous choice of MD protocol in the cgDNA+ model was inspired by the MD simulations protocol used by the Ascona B-DNA Consortium [147]. However, using Dang ions parameters is no longer recommended in the Amber user manual [31]. Therefore, we decided to change the ions parameters and move to a widely used water model, TIP3P. These updated choices are used for all dsNAs.

More importantly, we have also extended the duration of MD simulations from 3 μs (used to train the cgDNA+ model) to 10 μs for each training sequence, which reduces the MD convergence error in terms of symmetric KL divergence and Mahalanobis distance by a factor of approximately 3.2 and 1.9 (details in table 3.3).

4.1.3 Expansion of the training library for end-blocks parameters

One particular limitation of the cgDNA+ model is that for some non-GC ends, the reconstructed/predicted stiffness matrix was non-positive definite. As discussed earlier in section 3.3, non-GC ends parameters in the cgDNA+ model are trained on a library of 15 sequences with one non-GC end followed by a random sequence, while the other end of that sequence is GC. It implies that in the training library for end parameters, each non-GC end is followed by a particular kind of dimer step, unlike for GC ends and interior dimer steps, where all possible flanking contexts are present in the training sequences of $\mathrm{Lb}_{\mathrm{DNA}}$. To better understand the kind of sequences that lead to positive-definite or non-positive definite reconstruction, we rigorously analyzed all sequences of lengths 3 to 12 and observed that GC ends always lead to a positive-definite reconstruction of the stiffness matrix, in contrast, the non-positive definite reconstructions appear only when non-GC ends are followed by a dimer that is absent in the training library. Moreover, we observed that only one dimer case is sufficient for a positive-definite reconstruction of all the steps in that Y/R alphabets. For example, if a training sequence is present for a non-GC end followed by AG, then the reconstructions for all sequences containing that non-GC end followed by any of the {AG, AA, GA, GG} are positive definite. It suggests that the lack of diversity in the training sequences might be a possible reason for non-positive definite reconstructions. Even though we do not have a pure mathematical rationale, we decided to expand the training library for non-GC ends parameters empirically. The extended end library, $\mathrm{Lb}_{\mathrm{End}}$ is provided in table B.3 in which for each non-GC end, we have four training sequences such that the four sequences have one non-GC end followed by one random example dimer from each YR, RR, YY, and RY step. In this way, we enriched the training library for non-GC ends. We found that using this comprehensive library, we could obtain the parameter set that guarantees positive definite reconstructions for all sequences of any length (≥ 3).

4.2 From cgDNA+ to cgNA+ parameter sets

4.2.1 cgNA+ parameter sets

cgNA+ is a coarse-grained model for dsDNA, dsRNA, and DRH that allows computing sequence-dependent Gaussian pdfs for an arbitrary sequence at pre-specified solvent conditions. cgNA+ model is developed over the cgDNA+ model by estimating analogous parameters for dsRNA and DRH and improving the original cgDNA+ parameter set for dsDNA (updating MD protocol, more extensive and diverse training data, and simpler parameter estimation procedure) as described in the previous section. To train the interior and GC end blocks of the cgNA+ parameter sets for dsDNA, dsRNA, and DRH, we have used identical training sequences (referred to as palindromic library [149]) listed in table B.1, and the same MD protocol except for force-fields to describe dsNAs. More details on training sequences, MD protocol, and rigorous analysis of MD data are provided in chapter 3.

Thus, using analogous MD time-series data for three kinds of dsNAs and parameter estimation protocol described in chapter 2, we have obtained three different parameter sets (one for

each dsNAs), which can be written as:

$$\mathcal{P}_{\text{DNA}} = \{\sigma^{5'\text{XY}}, \sigma^{\text{XY}}, \mathcal{K}^{5'\text{XY}}, \mathcal{K}^{\text{XY}}\} = [\mathbb{R}^{36}]^{16} \times [\mathbb{R}^{42}]^{10} \times [\mathbb{R}^{36\times 36}]^{16} \times [\mathbb{R}^{42\times 42}]^{10} \tag{4.2}$$

where $5'\text{XY} \in \{16 \text{ end dimer steps}\}$ and $\text{XY} \in \{10 \text{ independent dimer steps}\}$,

$$\mathcal{P}_{\text{RNA}} = \{\sigma^{5'\text{XY}}, \sigma^{\text{XY}}, \mathcal{K}^{5'\text{XY}}, \mathcal{K}^{\text{XY}}\} = [\mathbb{R}^{36}]^{1} \times [\mathbb{R}^{42}]^{10} \times [\mathbb{R}^{36\times 36}]^{1} \times [\mathbb{R}^{42\times 42}]^{10} \tag{4.3}$$

where $5'\text{XY} \in \{\text{GC step}\}$ and $\text{XY} \in \{10 \text{ independent dimer steps}\}$,

$$\begin{aligned}\mathcal{P}_{\text{DRH}} &= \{\sigma^{5'\text{XY}}, \sigma^{\text{XY}}, \sigma^{3'\text{XY}}, \mathcal{K}^{5'\text{XY}}, \mathcal{K}^{\text{XY}}, \mathcal{K}^{3'\text{XY}}\} \\ &= [\mathbb{R}^{36}]^{1} \times [\mathbb{R}^{42}]^{16} \times [\mathbb{R}^{36}]^{1} \times [\mathbb{R}^{36\times 36}]^{1} \times [\mathbb{R}^{42\times 42}]^{16} \times [\mathbb{R}^{36\times 36}]^{1}\end{aligned} \tag{4.4}$$

where $5'\text{XY}$ and $3'\text{XY} \in \{\text{GC step}\}$ and $\text{XY} \in \{16 \text{ independent dimer steps}\}$. The parameter for $3'$ ends, and dependent dimer steps in $\mathcal{P}_{\text{DNA/RNA}}$ can be obtained using CW symmetry. Furthermore, it must be noted that there is no CW symmetry in DRH, as reading the sequence from the DNA strand (in DRH) is chemically different from reading the sequence from the RNA strand; therefore, different parameter blocks are required for all dimers, $3'-$GC end and $5'-$GC end. To avoid confusion in the writing and code implementation, we have always chosen the DNA strand (in DRH) as the reading strand, and the sequence is written in A/T/C/G alphabets.

4.3 cgNA+ reconstructions and associated modeling errors

As mentioned earlier, the cgNA+ model predicts the non-local sequence-dependent groundstate of any dsDNA, dsRNA, and DRH sequence. In this section, we have assessed the accuracy of the cgNA+ model by plotting the predicted groundstate for a given sequence along with the average shape obtained from the MD simulations and then quantifying the modeling error in terms of KL divergence and Mahalanobis distance. Moreover, we have also quantified the contributions of various modeling assumptions in the total modeling error (described in section 2.3.1).

4.3.1 Test library

To demonstrate the generalizability of the cgNA+ model for any sequence, we have tested the model for a diverse set of sequences (listed in table B.1) not present in the training library. The test sequences contain random palindromes, A-tracts, sequences with single-nucleotide polymorphism (SNP), poly(A), poly(AT), typical CpG islands, and long random sequences of length double that of those in the training library. Note that some of the sequences in the test library, such as A-tracts (intrinsically bent fragments), poly-A (stiffest in terms of persistence length), and CpG islands are mechanically exceptional sequences, and the model is not directly trained on such sequences. For instance, sequence indices 22 and 23 in Lb_{DNA} are two A-tracts of class $(\text{XA}_4\text{T}_4\text{Y})_n$ and $(\text{XT}_4\text{A}_4\text{Y})_n$ where $\text{X,Y} \in \{\text{G, C}\}$ which are similar in chemical composition but show contrasting differences in their super-helical structure. To compare the cgNA+ model predictions with the MD estimates for the test sequences, we have generated the same length of MD time-series using the same protocol for each sequence. Lastly, we only have an exten-

sive test library for dsDNA and dsRNA sequences, while test sequences are limited for DRH, a choice to optimize resources.

4.3.2 Reconstruction or prediction error in cgNA+

In this subsection, we first plotted the groundstate for a few selected sequences along with the observed MD estimates to visualize the accuracy of the cgNA+ model. In figure 4.1(a), we have plotted the groundstate (w) for the sequence indices 20 and 21 of Lb_{DNA} (see table B.1). Sequence index 20 is carefully chosen to contain all independent dimer steps, while sequence index 21 is the point mutation (SNP) of the same sequence. Moreover, along with cgNA+ predicted groundstate (in dashed line), we plotted the corresponding average shape from the MD time-series (in solid line). The following observations can be made from figure 4.1(a) i) the cgNA+ model predicts the groundstate almost indistinguishable from the corresponding MD statistics. Remarkably, the examples provided here are not in the training sequences used to obtain cgNA+ model parameters; ii) The two sequences differ by only a point mutation at the middle position; however, the change in groundstate due to that point mutation is highly non-local, i.e., up to three to four base-pairs on both sides of the mutation. More importantly, the cgNA+ model accurately captures this non-local sequence dependence in the groundstate while only using dimer-dependent parameters. This feature is only possible in a rigid-base model (cgDNA) or finer models (cgNA+) for which individual base-pair steps cannot achieve their local minima simultaneously, and frustration arises between the nearest-neighbors; thus, naturally capturing the non-local sequence-dependence in the mechanics of dsDNA but only using dimer dependent parameters [62, 149, 159].

Furthermore, in figure 4.1(b), we have plotted the predicted groundstate of two A-tracts (sequence indices 22 and 23 in Lb_{DNA}) along with the corresponding MD average shape. Note that the A-tracts are intrinsically bent fragments and the two A-tracts shown here have distinct superhelical structures. It can be observed in the figure that the cgNA+ model accurately captures the groundstate for such mechanically exceptional sequences. However, it is worth noting that the predictions for both sequences are not equally accurate. For example, the predicted Propeller for sequence index 22 is equal to the value observed in MD; in contrast, for sequence index 23, the prediction for Propeller deviates from the MD observations at the TA step. It is challenging to understand its reason precisely, and we have left a more detailed investigation for future studies. Moreover, in figure 4.2(a), we have compared poly(A) and poly(AU) embedded in the GC ends, which are sequence indices 18 and 19 in Lb_{RNA} and shown that for dsRNA sequences, cgNA+ model predictions are highly accurate. Lastly, in figure 4.2(b), we have highlighted the influence of beyond tetramer context by plotting the groundstate of two dimer steps in two different beyond tetramer flanking contexts along with the corresponding MD estimates and shown that the cgNA+ model accurately captures such strongly non-local changes in the groundstate.

In figures 4.1 and 4.2, we have demonstrated that the cgNA+ model predicts the groundstate for any dsDNA/dsRNA/DRH sequence with negligible error and is visually almost indistinguishable from the corresponding MD estimates. Now, to quantify this error, we have defined the reconstruction or prediction error, $\mathcal{E}^{\text{res}}$ as the deviation of the predicted Gaussian pdf from the corresponding observed Gaussian pdf in MD simulations. We have computed this reconstruction

error in terms of symmetric KL divergence and symmetric Mahalanobis distance as defined in section 2.5.5. Note that $\mathcal{E}_{\mathrm{KL}}^{\mathrm{res}}$ (reconstruction error in terms of KL divergence) describes the total reconstruction error in the predicted groundstate and stiffness matrix while $\mathcal{E}_{\mathcal{M}}^{\mathrm{res}}$ (reconstruction error in terms of Mahalanobis distance) highlights the difference in the predicted groundstate and MD average shape scaled by the stiffness. In table 4.1, we have tabulated the reconstruction errors per degree of freedom, dof (which is $24N - 18$, i.e., the number of internal coordinates required to describe a given sequence of length N bp) in the training and test sequences for $\mathrm{Lb}_{\mathrm{DNA}}$, $\mathrm{Lb}_{\mathrm{RNA}}$, and $\mathrm{Lb}_{\mathrm{DRH}}$. Firstly, the average model reconstruction errors in $\mathrm{Lb}_{\mathrm{DNA}}$ training sequences are 0.0020 and 0.0313 in terms of $\mathcal{E}_{\mathcal{M}}^{\mathrm{res}}$ and $\mathcal{E}_{\mathrm{KL}}^{\mathrm{res}}$, respectively, which are approximately one order smaller than the corresponding *scale* (which quantifies variation over sequence) obtained by computing the average pair-wise difference in the training sequences. It highlights the precision of the cgNA+ model in capturing the non-local sequence-dependent mechanics of dsDNA. Similar observations can be made for dsRNA and DRH. The average reconstruction error in test sequences ($\mathcal{E}_{\mathcal{M}}^{\mathrm{res}} \approx 0.0027$ and $\mathcal{E}_{\mathrm{KL}}^{\mathrm{res}} \approx 0.0316$) is slightly higher than in training sequences, as most test sequences possess exceptional mechanical behavior, and such sequences are not directly present in the training set. Therefore, an accuracy comparable to that in the training set for such exceptional sequences is highly impressive.

Note that the *scale* obtained for three types of dsNAs are in the order dsDNA > DRH > dsRNA, even though computed identically on similar training sequences. It can be attributed to the larger conformational space of dsDNA compared to dsRNA [131, 132] (refer to section 3.6). The Gaussian pdfs for two dsDNA sequences are farther from each other in conformational space than the identical two dsRNA sequences. Unsurprisingly, DRH lies between dsDNA and dsRNA. Similarly, the reconstruction errors are also in the same order, since it is easier to train a model (with a fixed number of parameters) on pdfs in a smaller conformational space.

This total reconstruction error in the cgNA+ model results from several modeling assumptions as listed in section 2.3 and the error associated with each assumption can be quantified as described in section 2.5. We have discussed the contributions of various modeling assumptions to the reconstruction error in the following subsections.

4.3.3 Approximation error in the training data

The first modeling assumption is that the MD time-series is stationary, which is not the case. The associated convergence error (referred to as palindromic error) is discussed in section 2.5.1, and details on the quantification of this error are provided in section 3.5. For the training sequences in $\mathrm{Lb}_{\mathrm{DNA}}$, $\mathrm{Lb}_{\mathrm{RNA}}$, and $\mathrm{Lb}_{\mathrm{DRH}}$, the average palindromic errors in terms of KL divergence $\mathcal{E}_{\mathrm{KL}}^{\mathrm{palin}}$ and Mahalanobis distance $\mathcal{E}_{\mathcal{M}}^{\mathrm{palin}}$ are of the order 10^{-4} and 10^{-3}, respectively, which are approximately two orders smaller than the corresponding *scales*.

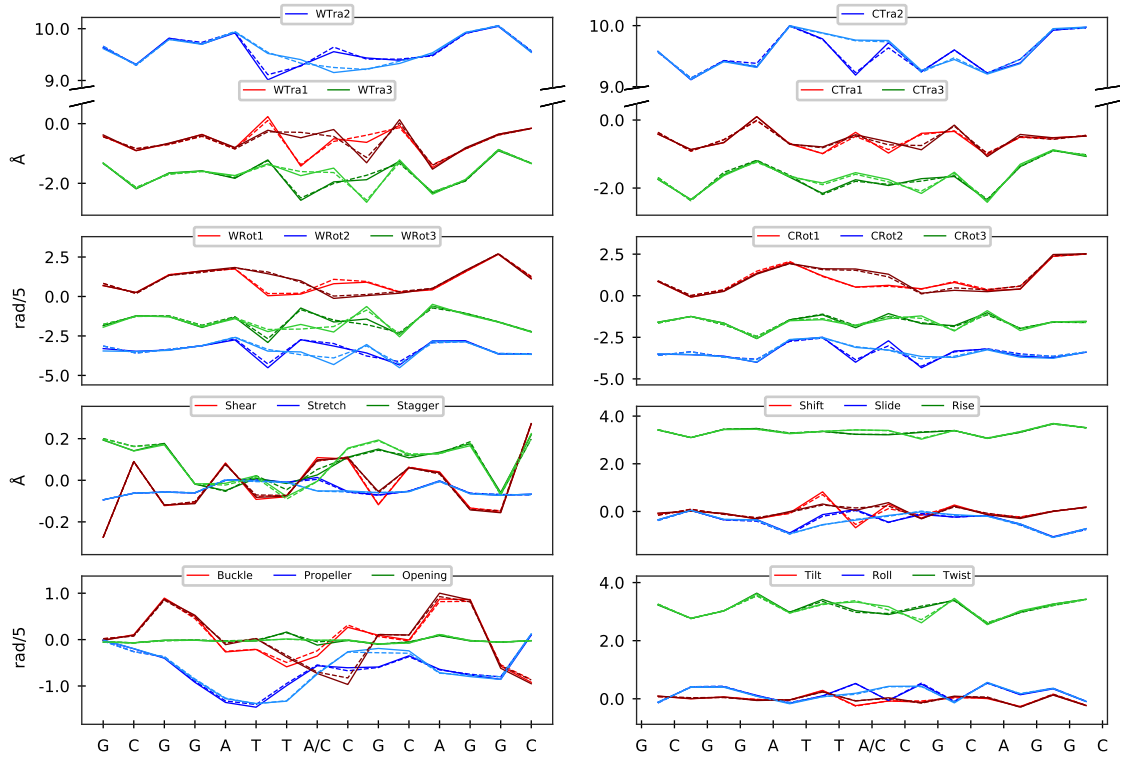
Moreover, in section 3.6, we have shown that the distributions for inter base-pair step and phosphate coordinates for dsDNA often deviate from Gaussian behavior (which also depend on the flanking sequence context). In contrast, the distributions for various internal coordinates in dsRNA are almost Gaussian. In DRH, we observed a mixed kind of behavior in the distribution of the internal coordinates. However, for modeling purposes, we have imposed Gaussianity to the underlying distributions for internal coordinates, leading to an inevitable modeling error.

We have quantified this modeling error, $\mathcal{E}_{\mathrm{KL}}^{\mathrm{Gauss}}$ by computing the KL divergence between the observed pdf and the best-fit Gaussian pdf to the observed pdf as described in section 2.5.2 and quantified in section 3.7. Except for a Wtra1 phosphate coordinate, $\mathcal{E}_{\mathrm{KL}}^{\mathrm{Gauss}}$ is less than *scale*.

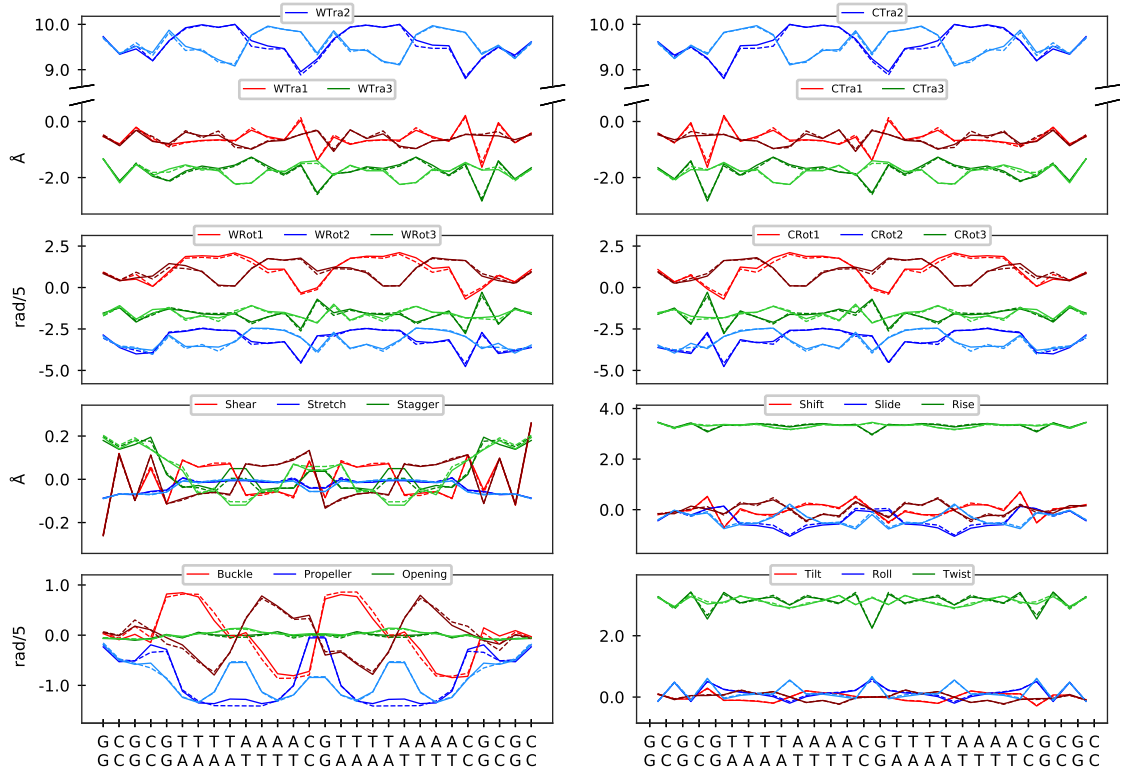
It should be noted that the reconstruction error is defined as the deviation of cgNA+ predicted Gaussian pdf with the stationary observed Gaussian pdf in MD simulations, i.e., observed MD Gaussian pdf is the ground truth for the cgNA+ model. Therefore, the palindromic and Gaussian approximation errors do not contribute to the aforementioned reconstruction error.

	$\mathrm{Lb}_{\mathrm{DNA}}$		$\mathrm{Lb}_{\mathrm{RNA}}$		$\mathrm{Lb}_{\mathrm{DRH}}$	
Training sequences						
Index	$\mathcal{E}_{\mathcal{M}}^{\mathrm{res}}$	$\mathcal{E}_{\mathrm{KL}}^{\mathrm{res}}$	$\mathcal{E}_{\mathcal{M}}^{\mathrm{res}}$	$\mathcal{E}_{\mathrm{KL}}^{\mathrm{res}}$	$\mathcal{E}_{\mathcal{M}}^{\mathrm{res}}$	$\mathcal{E}_{\mathrm{KL}}^{\mathrm{res}}$
1	0.0018	0.0240	0.0010	0.0058	0.0018	0.0239
2	0.0025	0.0439	0.0011	0.0064	0.0019	0.0254
3	0.0020	0.0302	0.0013	0.0070	0.0016	0.0227
4	0.0016	0.0267	0.0011	0.0083	0.0015	0.0165
5	0.0021	0.0289	0.0012	0.0081	0.0017	0.0226
6	0.0025	0.0368	0.0013	0.0063	0.0017	0.0213
7	0.0021	0.0353	0.0012	0.0070	0.0018	0.0209
8	0.0017	0.0266	0.0011	0.0071	0.0015	0.0247
9	0.0022	0.0328	0.0013	0.0080	0.0018	0.0215
10	0.0020	0.0276	0.0011	0.0074	0.0017	0.0209
11	0.0020	0.0342	0.0013	0.0095	0.0015	0.0167
12	0.0020	0.0322	0.0013	0.0067	0.0017	0.0432
13	0.0018	0.0297	0.0014	0.0101	0.0017	0.0214
14	0.0016	0.0282	0.0014	0.0092	0.0019	0.0218
15	0.0023	0.0344	0.0014	0.0101	0.0027	0.0395
16	0.0017	0.0296	0.0013	0.0076	0.0047	0.0532
Average	0.0020	0.0313	0.0012	0.0078	0.0019	0.0260
Test sequences						
Index	$\mathcal{E}_{\mathcal{M}}^{\mathrm{res}}$	$\mathcal{E}_{\mathrm{KL}}^{\mathrm{res}}$	$\mathcal{E}_{\mathcal{M}}^{\mathrm{res}}$	$\mathcal{E}_{\mathrm{KL}}^{\mathrm{res}}$	$\mathcal{E}_{\mathcal{M}}^{\mathrm{res}}$	$\mathcal{E}_{\mathrm{KL}}^{\mathrm{res}}$
17	0.0026	0.0357	0.0015	0.0087	0.0031	0.1032
18	0.0037	0.0291	0.0014	0.0095		
19	0.0034	0.0483	0.0021	0.0100		
20	0.0027	0.0306	0.0018	0.0101		
21	0.0026	0.0291	0.0015	0.0070		
22	0.0022	0.0285	0.0011	0.0092		
23	0.0019	0.0283	0.0010	0.0076		
24	0.0024	0.0254	0.0016	0.0061		
25	0.0027	0.0297				
26	0.0016	0.1449				
Average	0.0027	0.0316	0.0015	0.0085		
scale	0.0245	0.4395	0.0177	0.2185	0.0209	0.3273

Table 4.1 Model reconstruction error in terms of KL divergence ($\mathcal{E}_{\mathrm{KL}}^{\mathrm{res}}$) and Mahalanobis distance ($\mathcal{E}_{\mathcal{M}}^{\mathrm{res}}$) as defined in section 2.5.5. The list of sequences is provided in the table B.1 where the first 16 are training sequences, and the rest are test sequences. The *scale* (which quantifies variation over sequence) is obtained by computing the average pair-wise difference between all the training sequences.

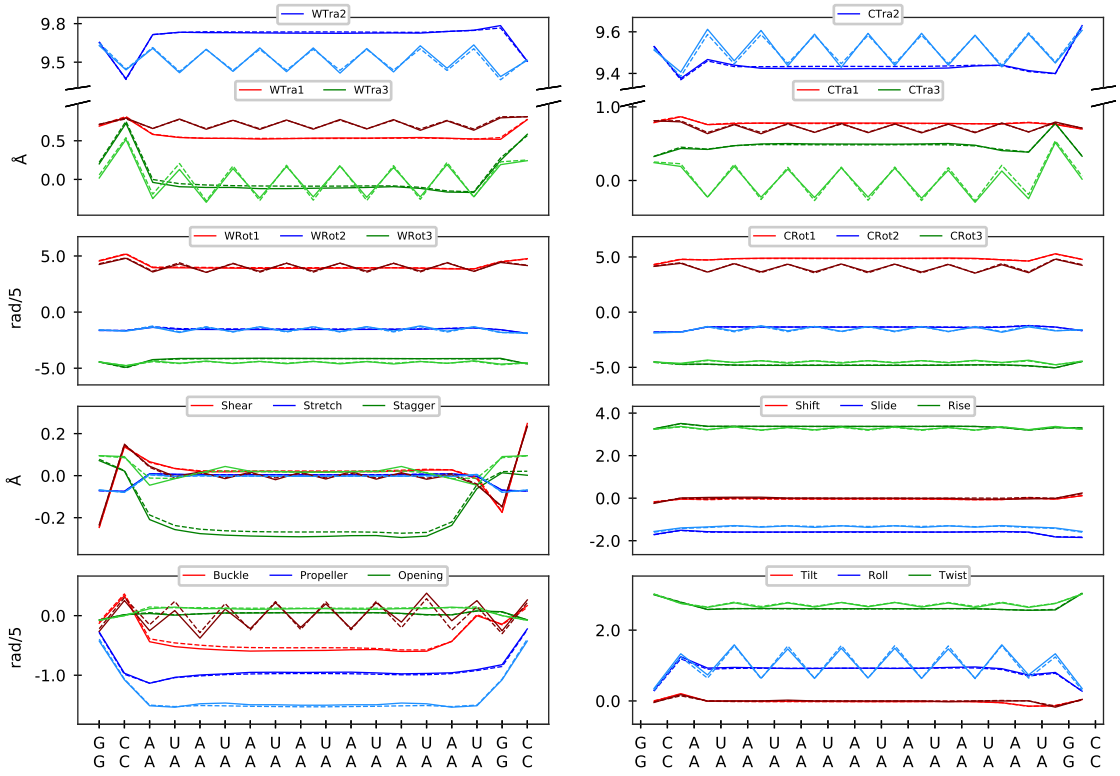


(a) Point mutation.

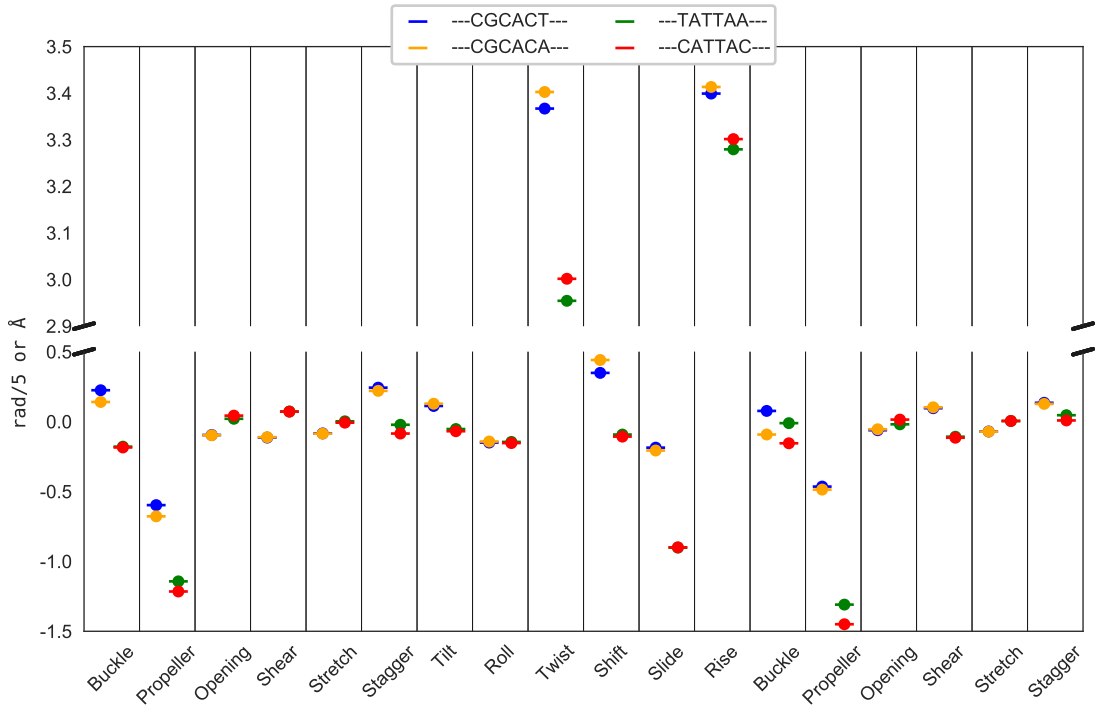


(b) Two A-tracts.

Fig. 4.1 Groundstate coordinates (elements of w) for (a) sequence indices 20 (in red, blue, and green as shown in legend) and 21 (in dark red, dark blue, dark green) and (b) sequence indices 22 (in red, blue, and green as shown in legend) and 23 (in dark red, dark blue, dark green) in $\mathrm{Lb}_{\mathrm{DNA}}$. The figure highlights the cgNA+ model accuracy in capturing (a) point mutation and (b) mechanically exceptional behavior of A-tracts. MD estimates are in solid lines while dashed lines are cgNA+ reconstructions.



(a) poly(A) and poly(AU).



(b) Hexamer context effect on CA and TT steps

Fig. 4.2 (a) Groundstate coordinates (elements of w) for sequence indices 18 (in red, blue, and green as shown in legend) and 19 (in dark red, dark blue, dark green) in Lb_{RNA}. MD estimates are in solid lines while dashed lines are cgNA+ reconstructions. (b) Internal coordinates of middle-junction dimer in different beyond tetramer context highlighting beyond tetramer flanking context influence on groundstate of the middle-junction dimer. The • is MD simulations data, and – is cgNA+ predictions, and the two data sets are indistinguishable. Note that beyond hexamer flanking sequence is also different but concisely denoted as `---`.

With these two approximations on MD time-series, we obtain a Gaussian pdf for each of the training sequences, which are used to compute the dimer-dependent parameter set based on two assumptions: a) the nearest-neighbor interactions assumption, i.e., the total energy of any given oligomer is the sum of local junction energies, and b) the local junction energy parameters depend only on the sequence of the corresponding junction dimer. Note that, in the updated cgNA+ training protocol, we directly computed the model parameters from the observed MD Gaussian pdf, unlike previously [149], in which the parameters were obtained in two steps, first, a banded Gaussian pdf (corresponding to the assumption of nearest-neighbor interactions) was obtained for all training sequences followed by the estimation of dimer-dependent parameters. As a consequence of this direct computation, we cannot precisely determine the error associated with these two assumptions. However, we can approximate the errors associated with these two assumptions a) by computing the banded stiffness matrix, which corresponds to nearest-neighbor interactions from the observed stiffness matrix in MD simulations, and then defining the truncation error as the KL divergence between banded and observed Gaussian pdfs and b) sequence locality error in the junction energy parameters by computing the KL divergence between banded and reconstructed Gaussian pdfs.

4.3.4 Contribution of nearest-neighbor interactions assumption in cgNA+ reconstruction error

Firstly, note that even though the nearest-neighbor interactions assumption simplifies the modeling, it is based on the observations made in the MD statistics. In figure 4.3, we have plotted the observed stiffness matrix in MD time-series for sequence index 1 in $\mathrm{Lb_{DNA}}$ along with the stencils corresponding to the nearest-neighbor interactions approximation. Note that the stiffness matrix is shown for only half of the sequence as the remaining entries are dependent due to the palindromic nature of the sequence. It can be observed from the plot that the stiffness matrix is highly banded, and there are very few entries outside the stencils, thus, justifying the nearest-neighbor interactions approximation. However, it must be noted that the non-zero entries outside the stencils are located very close to the stencils, which may suggest developing a model beyond nearest-neighbor interactions. The possibility of next-to-nearest-neighbor interactions is discussed in refs. [62, 149, 159]. These works concluded that extending the current nearest-neighbor to next-to-nearest-neighbor interactions approximation would lead to a significant increase in model parameters, while the gain in accuracy will be comparatively smaller. Moreover, with more model parameters, training the model and ensuring a positive-definite reconstruction for any sequence will be challenging. Therefore, in the cgNA+ model, we continued with the nearest-neighbor interactions approximation. We want to emphasize that the accuracy gained in the model from cgDNA to cgDNA+ is remarkable, as discussed in ref. [149]. Moreover, we found a similar sparsity pattern in the observed MD stiffness matrix for dsRNA and DRH as shown in figure 4.4 for sequence index 1 in $\mathrm{Lb_{RNA}}$ and $\mathrm{Lb_{DRH}}$. Note that the sequence in $\mathrm{Lb_{DRH}}$ is not palindrome, but we plotted half stiffness matrix for better visualization and comparison.

To quantify the error associated with this approximation, we have first computed the banded stiffness matrix corresponding to the nearest-neighbor interactions approximation using the

maximum entropy fit algorithm [60]. Then this approximation error (referred to as truncation error, $\mathcal{E}_{\mathrm{KL}}^{\mathrm{Trunc}}$) can be computed as the symmetric KL divergence between the observed stiffness and the corresponding banded stiffness as given in section 2.5.5. Note that the corresponding Mahalanobis contribution will be zero, since there is no change in the average shape of the oligomer when computing the banded stiffness (refer equation (C.11)). In table 4.2, we have listed the truncation errors, $\mathcal{E}_{\mathrm{KL}}^{\mathrm{Trunc}}$ for the training sequences (for brevity, we have not provided results for all sequences) in $\mathrm{Lb}_{\mathrm{DNA}}$, $\mathrm{Lb}_{\mathrm{RNA}}$, and $\mathrm{Lb}_{\mathrm{DRH}}$. It can be observed that for all sequences $\mathcal{E}_{\mathrm{KL}}^{\mathrm{Trunc}}$ is almost similar, with average values of 0.0046, 0.0026, and 0.0049 per dof for dsDNA, dsRNA, and DRH, respectively, which is approximately 100 times smaller than the corresponding *scale*.

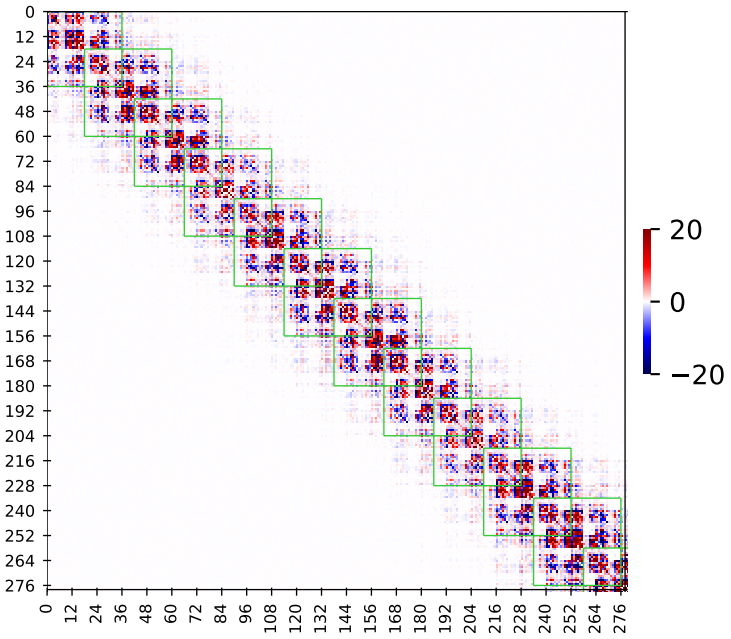
Lastly, the truncation error in dsRNA sequences is approximately half of the error in dsDNA sequences, which might be attributed to the larger conformational space of dsDNA/DRH, but can not be ascertained.

4.3.5 Contribution of sequence locality assumption in cgNA+ reconstruction error

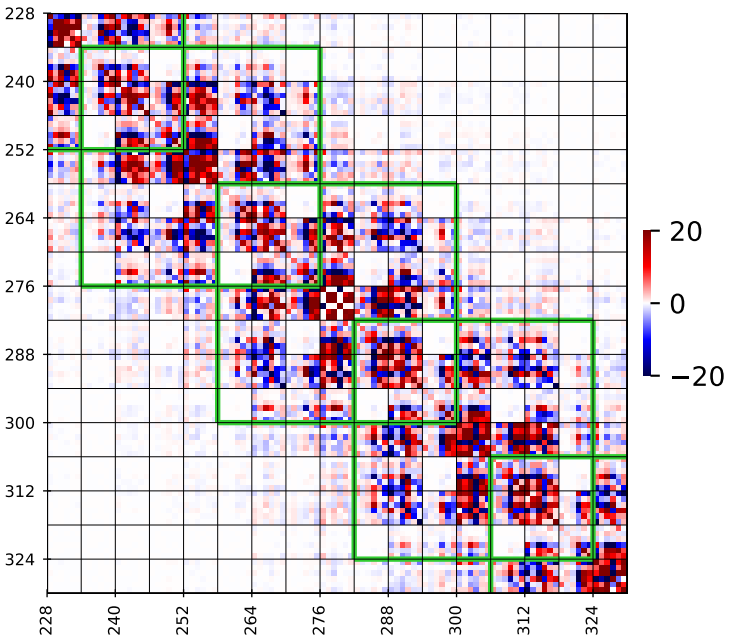
The final assumption in the cgNA+ model is the dependence of local junction energy parameters on the local dimer sequence. For instance, in $\mathcal{P}_{\mathrm{DNA}}$, σ^{XY} and $\mathcal{K}^{\mathrm{XY}}$ depend on the local dimer step XY. Note that the position of a given junction is crucial; therefore, we have different parameters for the interior and terminal junctions. The errors associated with this assumption, $\mathcal{E}_{\mathrm{KL}}^{\mathrm{local}}$ and $\mathcal{E}_{\mathcal{M}}^{\mathrm{local}}$ (described in section 2.5.4), are tabulated in table 4.2 for training sequences in $\mathrm{Lb}_{\mathrm{DNA}}$, $\mathrm{Lb}_{\mathrm{RNA}}$, and $\mathrm{Lb}_{\mathrm{DRH}}$. Remarkably, $\mathcal{E}_{\mathrm{KL}}^{\mathrm{local}}$ and $\mathcal{E}_{\mathcal{M}}^{\mathrm{local}}$ are at least one order of magnitude smaller than the corresponding *scale* for each training sequence in $\mathrm{Lb}_{\mathrm{DNA}}$, $\mathrm{Lb}_{\mathrm{RNA}}$, and $\mathrm{Lb}_{\mathrm{DRH}}$.

	$\mathrm{Lb}_{\mathrm{DNA}}$			$\mathrm{Lb}_{\mathrm{RNA}}$			$\mathrm{Lb}_{\mathrm{DRH}}$		
Index	$\mathcal{E}_{\mathrm{KL}}^{\mathrm{Trunc}}$	$\mathcal{E}_{\mathcal{M}}^{\mathrm{local}}$	$\mathcal{E}_{\mathrm{KL}}^{\mathrm{local}}$	$\mathcal{E}_{\mathrm{KL}}^{\mathrm{Trunc}}$	$\mathcal{E}_{\mathcal{M}}^{\mathrm{local}}$	$\mathcal{E}_{\mathrm{KL}}^{\mathrm{local}}$	$\mathcal{E}_{\mathrm{KL}}^{\mathrm{Trunc}}$	$\mathcal{E}_{\mathcal{M}}^{\mathrm{local}}$	$\mathcal{E}_{\mathrm{KL}}^{\mathrm{local}}$
1	0.0046	0.0018	0.0198	0.0026	0.0011	0.0034	0.0051	0.0019	0.0196
2	0.0048	0.0026	0.0401	0.0026	0.0012	0.0039	0.0050	0.0019	0.0213
3	0.0046	0.0021	0.0260	0.0025	0.0014	0.0047	0.0049	0.0017	0.0185
4	0.0046	0.0016	0.0228	0.0025	0.0012	0.0060	0.0047	0.0016	0.0125
5	0.0048	0.0021	0.0249	0.0025	0.0013	0.0058	0.0049	0.0017	0.0185
6	0.0045	0.0026	0.0328	0.0026	0.0013	0.0039	0.0047	0.0018	0.0172
7	0.0043	0.0022	0.0314	0.0026	0.0013	0.0046	0.0048	0.0019	0.0168
8	0.0047	0.0017	0.0224	0.0025	0.0011	0.0047	0.0049	0.0016	0.0206
9	0.0046	0.0022	0.0288	0.0026	0.0014	0.0057	0.0051	0.0019	0.0173
10	0.0045	0.0021	0.0236	0.0026	0.0012	0.0051	0.0053	0.0018	0.0166
11	0.0043	0.0021	0.0305	0.0026	0.0013	0.0071	0.0045	0.0015	0.0127
12	0.0046	0.0021	0.0284	0.0025	0.0014	0.0044	0.0050	0.0018	0.0393
13	0.0050	0.0018	0.0255	0.0026	0.0014	0.0077	0.0055	0.0017	0.0172
14	0.0043	0.0017	0.0242	0.0025	0.0015	0.0069	0.0049	0.0019	0.0176
15	0.0046	0.0023	0.0306	0.0027	0.0014	0.0077	0.0048	0.0028	0.0355
16	0.0046	0.0017	0.0255	0.0024	0.0013	0.0053	0.0049	0.0047	0.0491
Average	0.0046	0.0021	0.0273	0.0026	0.0013	0.0054	0.0049	0.0020	0.0219
scale	0.4395	0.0245	0.4395	0.2185	0.0177	0.2185	0.3273	0.0209	0.3273

Table 4.2 Truncation error due to nearest-neighbor interactions assumption in terms of symmetric KL divergence ($\mathcal{E}_{\mathrm{KL}}^{\mathrm{Trunc}}$) and locality error due to sequence locality assumption in the junction parameters in terms of KL divergence ($\mathcal{E}_{\mathrm{KL}}^{\mathrm{local}}$) and Mahalanobis distance ($\mathcal{E}_{\mathcal{M}}^{\mathrm{local}}$). The list of sequences are provided in the table B.1. The *scale* (which quantifies variation over sequence) is obtained by computing the average pair-wise difference between all the training sequences.

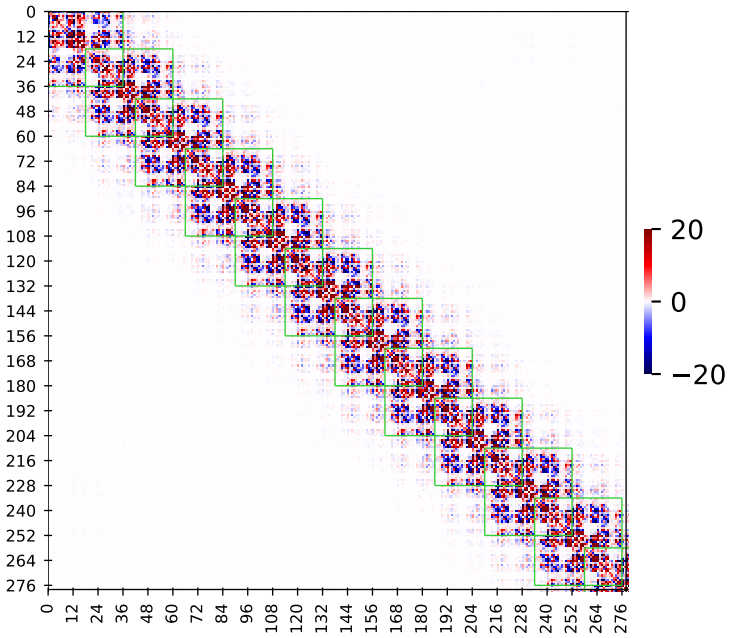


(a) Stiffness matrix observed in MD simulations for sequence 1 in Lb_{DNA}

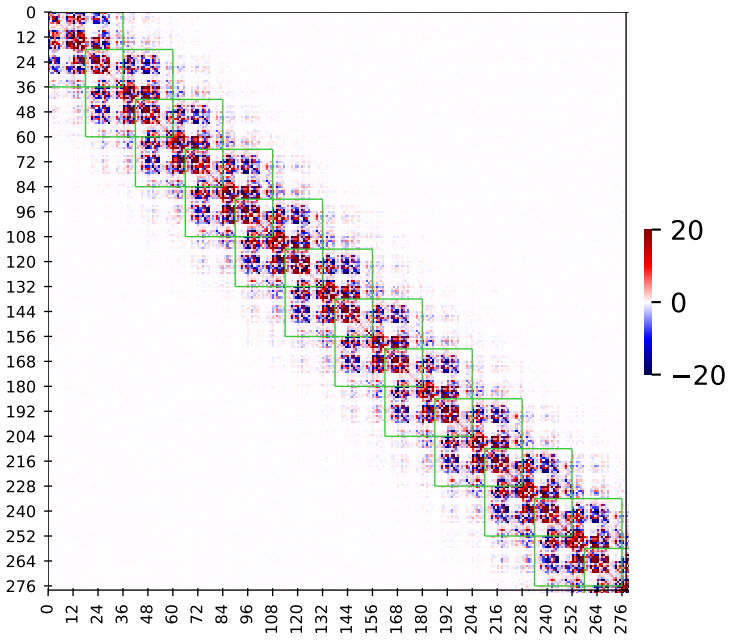


(b) Enlarged version of the above plot for central tetramer

Fig. 4.3 (a) Sparsity pattern in observed stiffness matrix in MD simulation for sequence index 1 in Lb_{DNA} (only half sequence is shown as the sequence is a palindrome), and (b) is a zoom-in image of the same matrix corresponding to central tetramer of the sequence. The green stencils correspond to the nearest-neighbor interactions approximation.



(a) Stiffness matrix observed in MD simulations for sequence 1 in $\mathrm{Lb}_{\mathrm{RNA}}$



(b) Stiffness matrix observed in MD simulations for sequence 1 in $\mathrm{Lb}_{\mathrm{DRH}}$

Fig. 4.4 Sparsity pattern in observed stiffness matrix in MD simulation for sequence index 1 (a) in $\mathrm{Lb}_{\mathrm{RNA}}$ and (b) in $\mathrm{Lb}_{\mathrm{DRH}}$ (only half sequence is shown). The green stencils correspond to the nearest-neighbor interactions approximation.

Now, of these two sources ($\mathcal{E}^{\text{Trunc}}$ and $\mathcal{E}^{\text{local}}$) in the total reconstruction error ($\mathcal{E}^{\text{res}}$), the locality assumption in sequence dependence of local junction energy parameters dominates. For instance, for training sequences in Lb_{DNA}, the average reconstruction error in terms of KL divergence, $\mathcal{E}^{\text{res}}_{\text{KL, avg}}$ is 0.0313, of which the contribution from the nearest-neighbor interactions assumption is 0.0046 and the locality assumption in the sequence dependence of junction energy parameters is 0.0273. This implies that the nearest-neighbor interactions assumption is reasonable and contributes negligibly to the modeling error. Whereas, the primary source of modeling error is the sequence locality assumption in junction energy parameters. Once again, this error, $\mathcal{E}^{\text{local}}$ is only a fraction of the *scale* set by computing the pair-wise difference between the training sequences in the respective libraries. Anyhow, it highlights the non-local sequence dependence in the local junction energy. We would like to remind the reader that $\mathcal{E}^{\text{res}}_{\text{KL, avg}}$ or $\mathcal{E}^{\text{local}}_{\text{KL, avg}}$ has two components, Mahalanobis (which quantifies the error in the groundstate) and stiffness component (which quantifies the error in the stiffness matrix). Further note that the Mahalanobis contribution in $\mathcal{E}^{\text{res}}_{\text{KL, avg}}$ or $\mathcal{E}^{\text{local}}_{\text{KL, avg}}$ is relatively negligible, thus implying that the major error is in the stiffness matrix, which has a dimer/trimer local sequence dependence. Remarkably, groundstate has a highly non-local sequence dependence due to the inversion of the stiffness matrix and the corresponding frustration energy associated with it. Lastly, we would like to recall that in the last two subsections, we have only approximated the contributions due to the nearest-neighbor interactions assumption and local sequence dependence in the junction energy parameters. cgNA+ model performs the computation for these two approximations in one step, and therefore, it is impossible to quantify the associated error individually. Truncation using different methods may lead to slightly different quantification of these errors [63, 158].

4.4 Comparison of dsDNA, dsRNA, and DNA:RNA hybrid

In the previous sections, we established that the cgNA+ model is extremely accurate in predicting groundstate and stiffness matrix for any given dsDNA/dsRNA/DRH sequence and quantified the associated errors. Moreover, the prediction is extremely fast, making possible the reconstruction of groundstate and stiffness matrix for millions of sequences and, thus, statistical estimation of various dsNA properties. For example, one can compute groove widths for all decamers and obtain statistical conclusions about sequence-dependence in groove widths. Such a computation is otherwise impossible to perform using traditional computational or experimental techniques. This section has rigorously investigated various such observables for dsDNA, dsRNA, and DRH for a large sequence space and compared the trends in these three kinds of dsNA. Some such comparisons can be found for various dsNAs in refs. [34, 117, 131, 132, 152, 163, 192, 196]; however, most of these studies are done for a minimal number of sequences (often less than 5) that question the generalizability of those results, especially when it is known that the properties/features of dsNAs are highly sequence-dependent (often non-local dependence) [9, 22, 50, 102, 147].

4.4.1 Comparison of average shape of dsDNA, dsRNA, and DNA:RNA hybrid

This subsection compares the average shape of base-pairs and base-pair steps in the average flanking context for dsDNA, dsRNA, and DRH. Moreover, we have highlighted the sensitivity of dimer average shape to the flanking context. For this comparison, we have used statistics obtained from MD simulations of Lb_{DNA}, Lb_{RNA}, and Lb_{DRH} along with the corresponding cgNA+ predictions (which also allowed comparing MD statistics with cgNA+ predictions). We have extracted average intra base-pairs coordinates and inter base-pair step and phosphate coordinates in average flanking contexts and tetramer flanking contexts and plotted intra base-pair coordinates in figure 4.5 and inter base-pair step and phosphate coordinates in figure 4.6.

It should be noted that such comparisons have been made previously in the literature [34, 131]; however, for only limited sequences. Therefore, some trends in the dimer sequence dependence, particularly for inter base-pair coordinates, have been observed before; notably, this is subjected to the source of data, such as experimental X-ray data or MD simulation data using various MD protocols. The objective of such comparison is that at length scales at which a few MD simulations can be performed, the cgNA+ model accurately captures the underlying trends in the average shape.

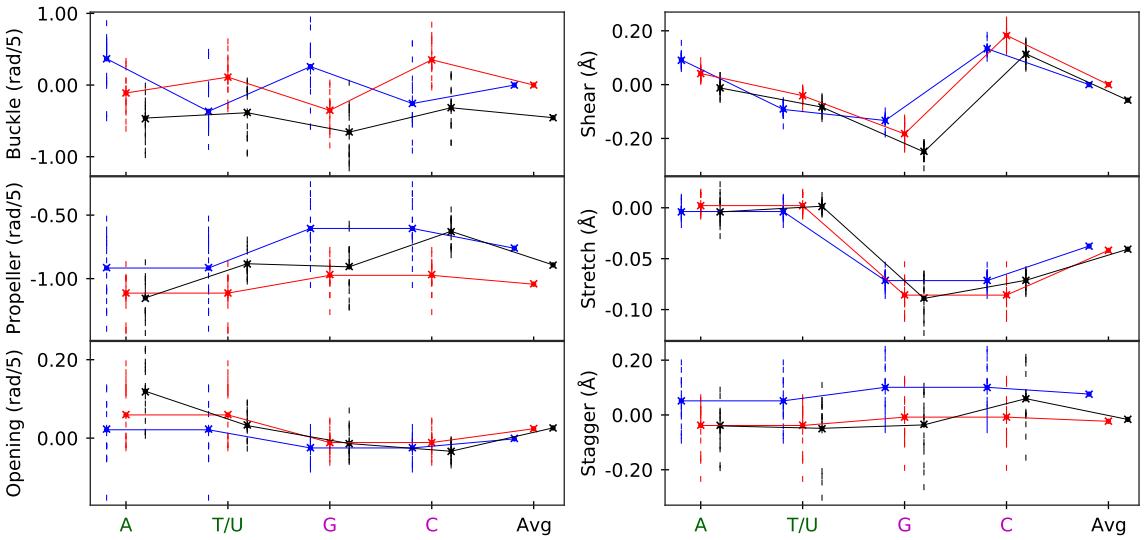


Fig. 4.5 Comparison of intra base-pair coordinates for dsDNA (in Blue), dsRNA (in Red), and DRH (in Black) at the X-axis. For each base-pair in average context, coordinates observed in MD simulations and cgNA+ predictions are plotted in • and ×, respectively, along with the coordinates in various flanking trimer contexts in vertical lines (|) to highlight the role of flanking sequence. A line plot is plotted along • for better visualization, and the data corresponding to dsDNA, dsRNA, and DRH is slightly shifted along the X-axis.

This work presents the most extensive comparisons of the average shape of base-pair and base-pair steps (along with phosphate coordinates) for dsDNA, dsRNA, and DRH. In figures 4.5 and 4.6, we have one panel for each of the internal coordinates, and the value of the given internal coordinate is plotted on the Y-axis while base-pair or base-pair step is plotted on the X-axis. The cgNA+ predictions are plotted as × in blue, red, and black for dsDNA, dsRNA, and DRH, respectively, together with the MD observations as •.

In figure 4.5, we have compared intra base-pair coordinates. Comparing MD observations (•) with the corresponding cgNA+ predictions (×), which are almost superimposed, again highlights the cgNA+ model accuracy. Note that for dsDNA and dsRNA, AT and TA base-pairs and GC and CG base-pairs represent the same physical base-pair except for reading the sequence from the different strand (Watson or Crick). Therefore, intra base-pair coordinates for these base-pairs are the same except for the sign convention determined by the change of reading strand transformation (refer to section 2.2.2). The same is not true for DRH as the GC base-pair has G on DNA strand and C on RNA strand in contrast to the CG base-pair, which has C on DNA strand and G on RNA strand, thus representing two chemically different molecules. Therefore, one can observe that the intra coordinates for A/T (or U) and G/C are identical (except for Buckle and Shear, in which coordinates are anti-symmetric due to reading strand transformation). For Buckle and Shear, the average values for A and G (the independent set of base-pairs) are in the order dsDNA > dsRNA > DRH, with sequence average values (up to 3 decimal points) equal to 0.000 (rad/5 and Å) for dsRNA and dsDNA and -0.456 rad/5 and -0.058 Å for DRH. Note that the difference in magnitude is much smaller for Shear than Buckle. Furthermore, it is interesting that in A and G, the direction of Buckle for dsDNA and DRH is opposite but with a similar magnitude. On the contrary, for dsRNA, it remains closer to zero (planarity). Propeller values observed in dsRNA are more negative than in dsDNA, whereas DRH adopts intermediate values of dsDNA and dsRNA. A similar trend is also observed in Stagger with DRH values relatively closer to dsRNA. The order is dsRNA > DRH > dsDNA for Opening. Lastly, for Stretch, the values for dsDNA, dsRNA, and DRH are incredibly close, with almost zero Stretch for the A-T base-pair and slightly negative Stretch for the C-G base-pair, which can be attributed to the three H-bonds in C-G as compared to two in A-T base-pair. Similar observations can be made for Propeller and Opening, where the average values for the C-G base-pair tend to stay close to zero, whereas, in the A-T base-pair, the deviation from zero is more. Lastly, from the spread of the vertical lines (|) around •, it can be concluded that flanking trimer contexts significantly impact intra-base-pair coordinates with the variation due to flanking contexts often being larger than the variation for different base-pairs. Moreover, similar trends are observed for all three dsNAs considered in this work.

Furthermore, for inter base-pair step and phosphate coordinates, out of the 16 dimers, six are dependent in the case of dsDNA and dsRNA, while for DRH, 16 dimers are independent as there is no CW symmetry. Therefore, figure 4.6 contains data corresponding to all the 16 dimers. To start with the discussion, we focused on helical inter coordinates (Shift, Slide, Rise, Tilt, Roll, and Twist). Shift and Tilt are odd parameters (i.e., change sign on reading strand transformation as described in section 2.2.2); therefore, the average value for all dimers and palindromic dimers (AT, GC, CG, and TA) of Shift and Tilt over various dimers in the average context is zero, while the rest six dependent pairs are in positive and negative pairs. Shift for dsRNA remains very close to zero, while it fluctuates between positive and negative values for dsDNA. Shift for DRH is always positive, with an average value of 0.189 Å. The same observations are true for Tilt, with values for dsRNA close to zero and relatively larger sequence-dependent fluctuations around zero for dsDNA. At the same time, Tilt values for DRH are always positive with clear dimer sequence dependence. The trends for Slide are in the order dsDNA > DRH > dsRNA with

an average Slide of around -0.397 Å for dsDNA, -1.666 Å for dsRNA, and in between dsDNA and dsRNA for DRH but slightly closer to dsRNA. The average Twist in dsDNA is considerably higher than in dsRNA due to the B-form and A-form of dsDNA and dsRNA, whereas DRH has a mixed A and B-form with A-form dominating and Twist, in general, close to dsRNA. In particular, Noy et al. [132] had similar findings, with Twist and Slide values being closer to the observed values in dsRNA, while the rest of the inter-coordinates are in between dsDNA and dsRNA. It is interesting that Rise is greater for dsRNA than dsDNA for YR steps, whereas it is reversed for the rest of the dimer steps. The Rise for DRH varies between dsDNA and dsRNA values, but is generally closer to the values observed for dsRNA. Lastly, it is evident from the spread of the vertical lines (|) around • that the tetramer context is crucial and highly influences the average shape of the dimer (in terms of inter-helical coordinates). In particular, for inter-translational coordinates, dsDNA dimers are much more sensitive to flanking tetramer context than dsRNA, whereas DRH is in between dsDNA and dsRNA. The trends are still very similar for rotational coordinates, but the difference is relatively smaller. Note that the dsDNA backbone fluctuates more and occupies a larger conformational space [131, 132] and is coupled with inter base-pair step coordinates leading to a higher sensitivity of dsDNA inter coordinates to flanking context compared to dsRNA and DRH.

Finally, such an analysis for phosphate coordinates is entirely novel. For DRH, we chose the DNA strand as the reading strand and the RNA strand as the complementary strand. First, one can observe that the phosphate coordinates for dsDNA and dsRNA are very different (except WTra2 and CTra2), which can be attributed to the A- and B-form geometry of the dsNA in dsRNA and dsDNA, respectively. Furthermore, in contrast to inter base-pair coordinates, the observed phosphate coordinates in DRH are not in between dsDNA and dsRNA; instead, Watson phosphate coordinates are closer to those observed in dsDNA, whereas Crick phosphates are closer to those observed in dsRNA. This implies that the backbone behavior of the DNA strand in DRH is closer to pure dsDNA, and the RNA strand is closer to pure dsRNA. Similar findings have been reported in [34, 108, 132, 133] but are characterized differently. Notably, dimer-step-dependent fluctuations in phosphate coordinates for dsDNA are considerably higher than those in dsRNA. Moreover, for a given dimer step, the differences in average phosphate coordinates for various flanking tetramer contexts are also much higher in dsDNA than in dsRNA. This behavior can be expected as the dsDNA backbone exhibits a larger conformational space (geometry can also change from B- to A-form depending on sequence) than dsRNA (mostly adhering to A-form geometry). In the case of DRH, once again, it can be observed that the DRH Crick strand behaves similar to pure dsRNA (with less variation due to the dimer step sequence and the flanking tetramer context), and the Watson strand behaves similar to pure dsDNA (i.e., sensitive to dimer step sequence and flanking tetramer context).

4.4.2 Comparison of persistence lengths of dsDNA, dsRNA, and DNA:RNA hybrid

One of the most popular and traditional measures to quantify the rigidity of the NAs is persistence length, which can be defined as the length scale over which correlations in the direction of tangent along a polymer centerline are lost [68]. In the context of DNA (and other NAs), the definition of persistence length has been traditionally and frequently used in the sequence-average sense, which has two crucial governing factors, stiffness and intrinsic shape [201]. However, it is well understood that both governing factors depend on the sequence of the given NA. Mitchell et al. [123] rigorously studied sequence-dependent persistence lengths of dsDNA (referred to as apparent persistence length, ℓ_p) using the cgDNA model and, for the first time, introduced the notion of sequence-dependent dynamic persistence length, ℓ_d by factoring out the contributions of the intrinsic shape from ℓ_p. This work has been summarized in section 2.6 in the context of the cgNA+ model, and more details can be found in refs. [123, 149].

We want to remind the readers that persistence length is also sensitive to experimental techniques and conditions, along with sequence dependence. For example, increasing the salt concentration decreases the persistence length of dsDNA from 57 to 43 nm. In the present literature, 150 bps or 50 nm is the agreed sequence-average persistence length of the dsDNA. The persistence length predicted by the cgDNA or cgDNA+ model is, in general, considerably higher than the experimental consensus of 150 bps. It has been thoroughly discussed in previous work [123, 149] where they demonstrated that although the model predictions are higher than experimental consensus, the trends in sequence-dependent persistence lengths of dsDNA are similar to those observed in the experiments. There can be several reasons for this discrepancy of persistence lengths given by cgNA+ tools. Firstly, in experiments, the salt concentration is relatively higher, and often divalent counter ions are used as opposed to mono-valent ions under physiological concentrations in the MD simulations (used to train the cgNA+ model), which have a significant effect on the persistence length of dsDNA (or dsNA) [27, 175, 211]. Moreover, the parameterization of DNA in various MD forcefields might be stiffer [43, 76]. Notably, it has been shown in ref. [149] that for shorter sequences (24mer), the tangent-tangent correlation observed in the MD simulations is incredibly close to cgDNA+ predictions implying that this discrepancy is not inherent to the model.

Such a rigorous study of the persistence lengths of dsRNA and DRH is entirely novel. A few experimental investigations have measured the persistence length of a few dsRNA sequences in various experimental setups, again providing very different values of the persistence length for dsRNA. However, in general, the dsRNA is considered to be stiffer than dsDNA. Experiments [1] performed using magnetic tweezers and atomic force microscopy (believed to be highly accurate experimental techniques to measure persistence length) found that the mean persistence length of dsRNA is equal to 63.8 and 62 nm, respectively. In the case of DRH, Zhang et al. [211] found the persistence length of DRH $\approx$ 63 nm at one mM NaCl. Moreover, magnetic tweezers experiments at various salt conditions have shown that the persistence length of DRH is generally in between dsDNA and dsRNA. Lastly, all such experiments are performed under specific conditions and for limited sequences, making it difficult to draw any statistical conclusions about the persistence length of dsNAs.

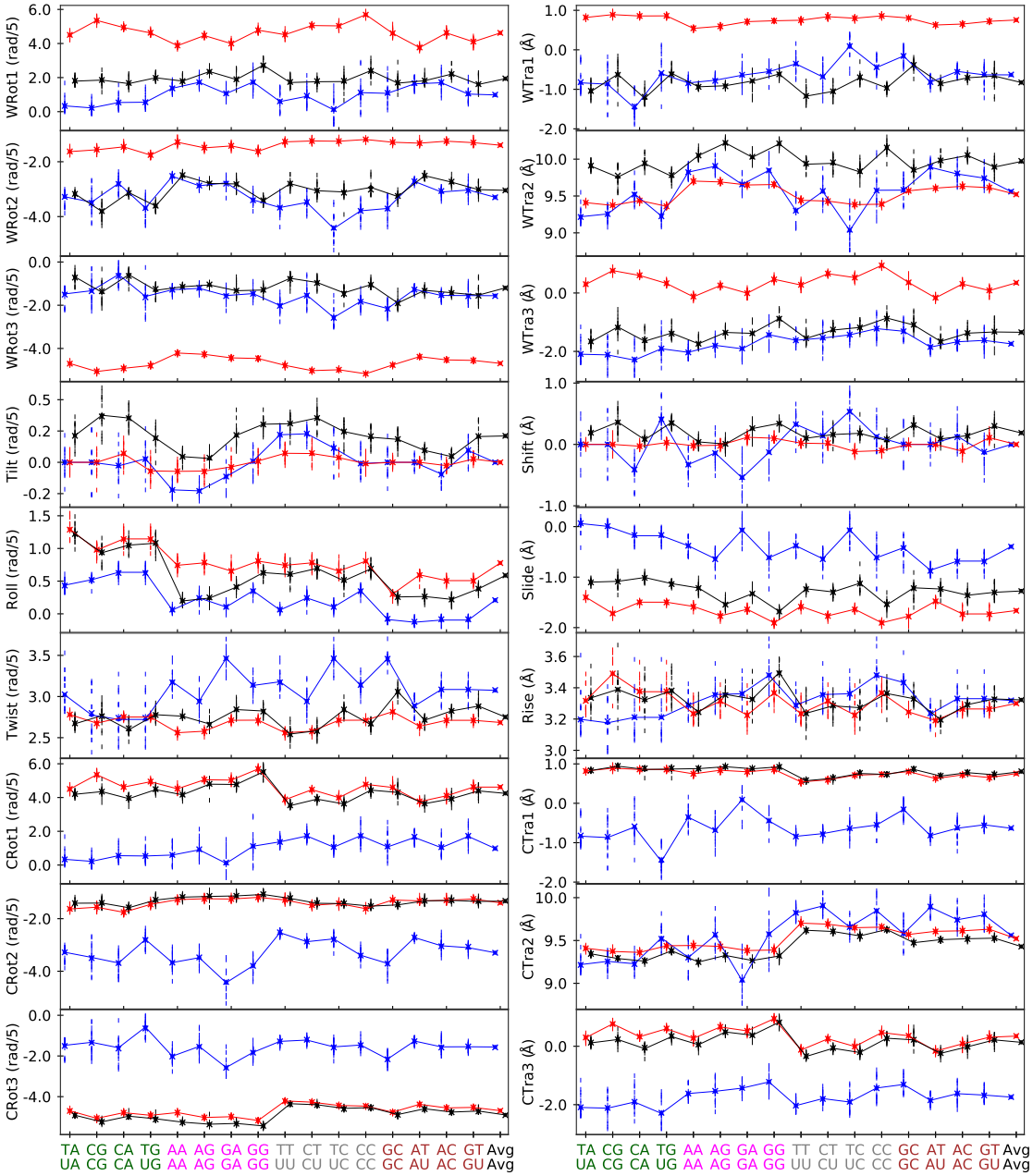


Fig. 4.6 Comparison of base-pair step and phosphate coordinates for dsDNA (in Blue), dsRNA (in Red), and DRH (in Black) at the X-axis. For each base-pair step in average context, coordinates observed in MD simulations and cgNA+ predictions are plotted in • and ×, respectively, along with the coordinates in various flanking tetramer contexts in vertical lines (|) to highlight the role of flanking sequence. For better visualization, a line plot is plotted along •, and the data corresponding to dsDNA, dsRNA, and DRH is slightly shifted along the X-axis.

In this work, we have generated an extensive database of persistence lengths for two million random sequences as well as all poly-dimers for dsDNA, dsRNA, and DRH, exploiting the capabilities of the cgNA+ model and the Monte Carlo code. All sequences were of length 220 bps embedded in GC ends, and 10^5 Monte Carlo samples were drawn for each sequence. It is worth highlighting that performing MD simulation for even a single sequence (of length > 200 bps) is almost impossible, whereas using cgNA+ tools to generate an ensemble of configurations only takes a few minutes on a standard laptop. The objective of this database is to draw a statistical conclusion about the sequence-dependent rigidity of dsNAs (defined in terms of ℓ_p and ℓ_d) and to compare the persistence lengths of dsDNA, dsRNA, and DRH. The results are plotted in figure 4.7 with the following observations:

- **Sequence of a given dsNA plays a crucial role in the determination of persistence length.** In the top panel of figure 4.7, we have plotted the histograms of ℓ_p and ℓ_d for dsDNA, dsRNA, and DRH. The first observation that can be made from the plot is that the persistence length is highly sequence-dependent. The range of ℓ_p observed in this database of 2 million random sequences is 138, 175, and 166 bps for dsDNA, dsRNA, and DRH, respectively, which are significant as the corresponding average ℓ_p are 211, 214, and 182 bps. In contrast, the range of ℓ_d observed in this database is relatively narrow (as ℓ_p has a combined effect of stiffness and intrinsic shape, whereas ℓ_d only accounts for the stiffness) with a range of 50, 76, and 106 bps for dsDNA, dsRNA, and DRH with average ℓ_d equals to 226, 270, and 236 bps. This range of ℓ_p and ℓ_d highlights the role of sequence in the persistence length of dsNAs. Noteworthy that the range of ℓ_p observed for dsDNA, dsRNA, and DRH is comparable, whereas, for ℓ_d, the range observed for DRH is significantly larger than that for dsDNA and dsRNA. The mixed type of DNA and RNA strands in DRH might make its behavior more complicated and sensitive to the sequence.
- The difference between ℓ_p and ℓ_d for a given sequence emphasizes the role of the intrinsic shape of the sequence. For bent sequences like A-tracts, the difference is larger, whereas, for sequences with straight groundstate, the difference is tiny.
- **On average, the observed persistence lengths are in the order:** $\ell_p^{\mathbf{RNA}} \gtrapprox \ell_p^{\mathbf{DNA}} \gtrapprox \ell_p^{\mathbf{DRH}}$ **and** $\ell_d^{\mathbf{RNA}} \gtrapprox \ell_d^{\mathbf{DRH}} \gtrapprox \ell_d^{\mathbf{DNA}}$. In the bottom panel of figure 4.7, we have plotted the histograms for sequence-wise difference of ℓ_p and ℓ_d for DRH and dsRNA from dsDNA.

Lastly, in figure 4.8, we have plotted ℓ_p and ℓ_d for all independent poly-dimers for dsDNA, dsRNA, and DRH. Note that poly(AA) and poly(TT) or poly(UU) represent the same physical dsDNA or dsRNA; however, these are two different DRH molecules. To comment on the rigidity of a sequence, ℓ_d is a better choice as it factors out the effect of the intrinsic shape of a sequence from ℓ_p. For dsDNA, the stiffest sequence (in terms of ℓ_d) we found is poly(A) or poly(T) with $\ell_p \approx 253$ bps and $\ell_d \approx 254$ bps whereas poly(AT) is the softest with $\ell_p \approx 194$ and $\ell_d \approx 195$ bps. Note that both sequences have a straight groundstate but a vast difference in the rigidity. Furthermore, all dsDNA poly-dimers are intrinsically almost straight in contrast to dsRNA and DRH, where the difference between ℓ_p and ℓ_d is significant. For dsRNA, the stiffest sequence is poly(C)/poly(G) with $\ell_d \approx 338$ bps. For DRH, it is very interesting that the stiffest sequence is poly(A) with $\ell_d \approx 312$ bps, whereas poly(T) is incredibly softer $\ell_d \approx 196$ bps. Similar behavior can be observed for other sequence pairs such as poly(C) and poly(G), poly(TG) and

poly(AC), and poly(TC) and poly(AG) (these pairs represent the same physical dsNA in the case of dsDNA and dsRNA, while the two are different for DRH), i.e., a significant difference in persistence length (both ℓ_d and ℓ_p) within the pairs. It highlights that the flexibility/rigidity behavior in DRH is not a simple average of dsRNA and dsDNA; instead, it shows a unique and complicated pattern.

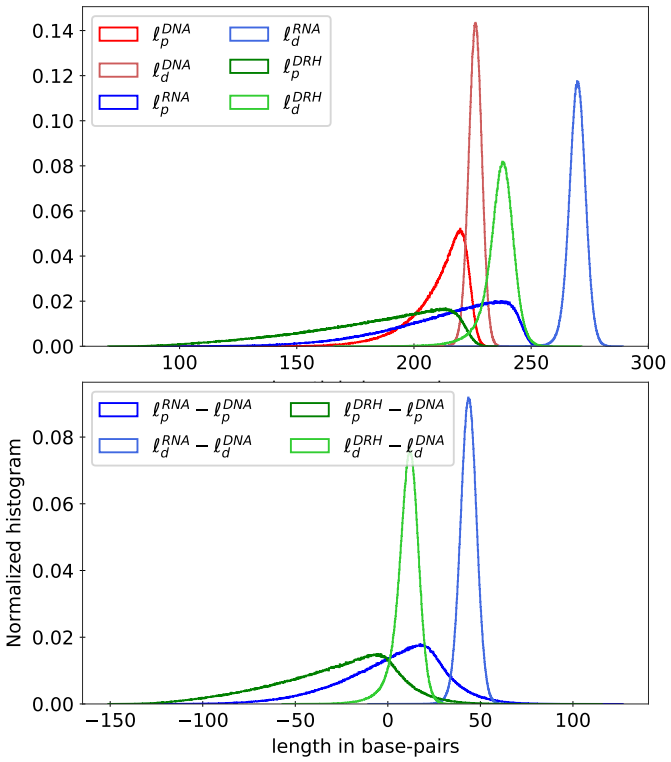


Fig. 4.7 Top: Histogram for dynamic (ℓ_{d}) and apparent (ℓ_{p}) persistence lengths for ≈ 2 million random sequences (of length 220 bp) and all poly-dimers (110 repeats) for dsDNA, dsRNA, and DRH. Bottom: Histogram for sequence-wise difference in persistence lengths of dsRNA and DRH from dsDNA.

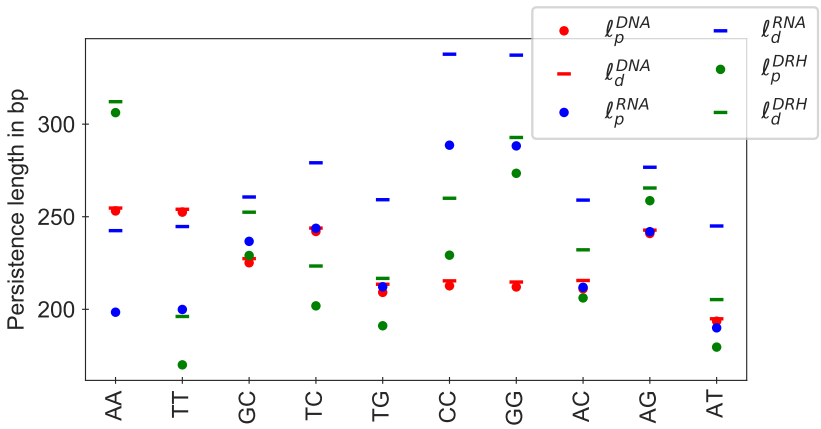


Fig. 4.8 Dynamic (ℓ_{d}) and apparent (ℓ_{p}) persistence lengths for all independent poly-dimers ($(\mathrm{XY})_{110}$ embedded in GC ends) for dsDNA, dsRNA, and DRH.

4.4.3 Comparison of groove widths of dsDNA, dsRNA, and DNA:RNA hybrid

The mechanical properties of DNA play a central role in protein binding. Proteins recognize specific DNA sequences [82, 129, 171, 172], which are often controlled by the ability of the DNA sequence to deform in a certain way to facilitate protein binding. Most proteins interact with DNA through grooves. One common mechanism involves protein interaction with DNA via major grooves by forming H-bonds between amino acids and bases in a sequence-specific manner. Furthermore, protein also recognizes DNA without interacting with the bases directly but recognizes the specific conformation/shape assumed by certain DNA sequences, called indirect readout. For instance, positive side-chains of proteins bind to narrow minor grooves having strong negative electrostatic potential [172]. It is worth noting that groove widths depend on the sequence, as observed in both experiment [134, 171, 195] and simulation.

This work studies the sequence dependence in the groove widths of dsDNA, dsRNA, and DRH and compares them. Basic details on grooves can be found in section 1.1. In this work, for sequence-dependent analysis of grooves, we reconstructed all possible decamers embedded in four random bases plus GC ends on both sides, i.e., 4^{10} ($\approx$ one million) sequences of length 22 bps. We have followed a similar protocol as that given in refs. [101, 195] for groove width computation. For a given sequence, we chose the central phosphate on the Watson strand for each sequence as the reference phosphate. Subsequently, we fit the cubic spline curves along the Crick strand backbone passing through P_i (phosphate group) and tangential to $P_{i+1} - P_{i-1}$. Then we mark nine equidistant points (finer mesh has negligible effects on the results) on the cubic spline fitted between each subsequent phosphate group and compute the distance of these points from the chosen reference phosphate on the Watson strand. The minor groove is in the $3'$ to $5'$ direction (or **upstream**) from the reference phosphate on the Watson strand, and the major groove is in the $5'$ to $3'$ direction (or **downstream**). Lastly, we subtract an offset of 5.8 Å (as recommended in CURVES+ [101]) from the observed groove widths to take into account the van der Waals radius of the phosphate.

In figure 4.9, we have plotted histograms for minor and major grooves for dsDNA, dsRNA, and DRH. For dsDNA, the major groove is, in general, seven Å wider than the minor grooves. Moreover, the variation in minor groove width in the sequence space of all decamers is larger than in major groove width. In contrast, for dsRNA, minor groove is approximately four Å wider than major groove. Furthermore, the variation in the major groove width is larger than in minor groove width. Lastly, DRH shows average behavior for both major and minor groove widths, with no clear distinction between the two. The minor groove width of DRH is between the observed values of dsDNA and dsRNA, and the same observation is also true for major groove width. These trends are comparable with previous observations in experiment [134, 171] and simulation [34, 117, 132], however, performed for only a few sequences.

From the histograms in figure 4.9, it is clear that the sequence strongly influences the groove widths. In particular, the range of minor groove width observed for dsDNA is approximately six Å. It raises a natural question: which sequences lead to extreme groove widths? We have addressed this question by plotting sequence logos [178] for the central decamer in outlier sequences, defined as sequences with groove widths outside three standard deviations from the mean (i.e., 0.15% on both sides). Sequence logos are excellent graphical tools for visualizing

and comprehending the underlying sequence pattern for any observable. Sequence logos are described in more details in section 1.2.1.

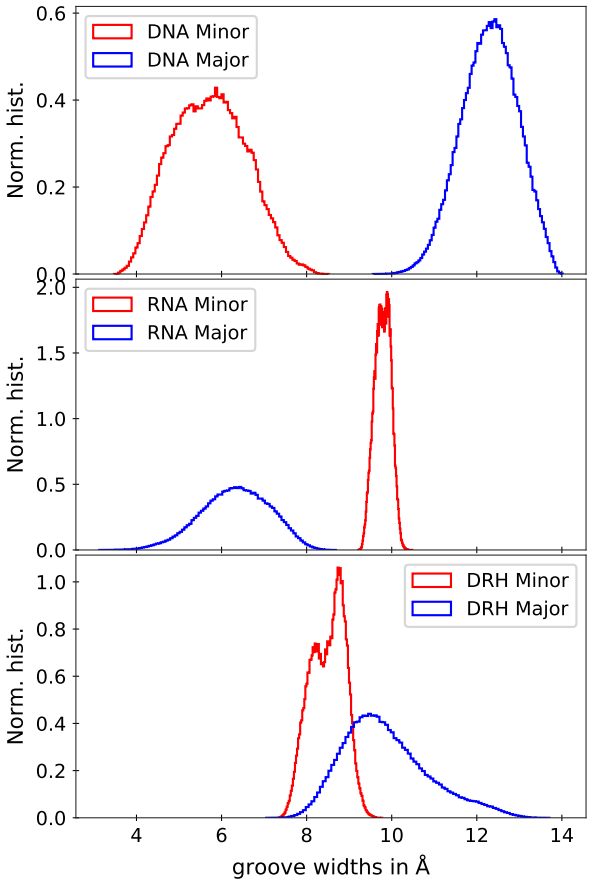
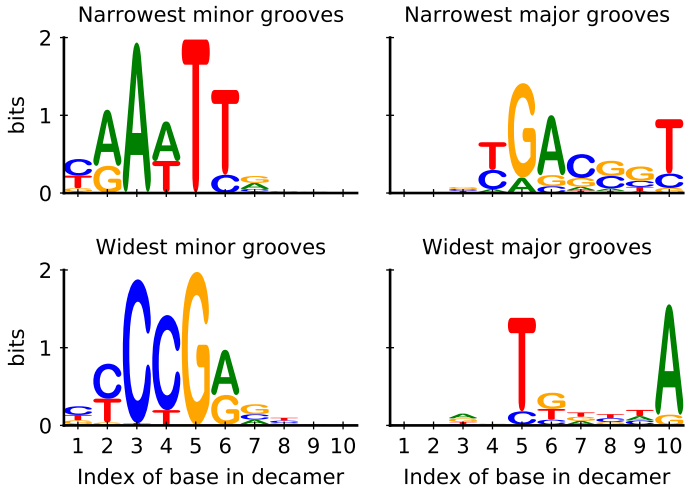


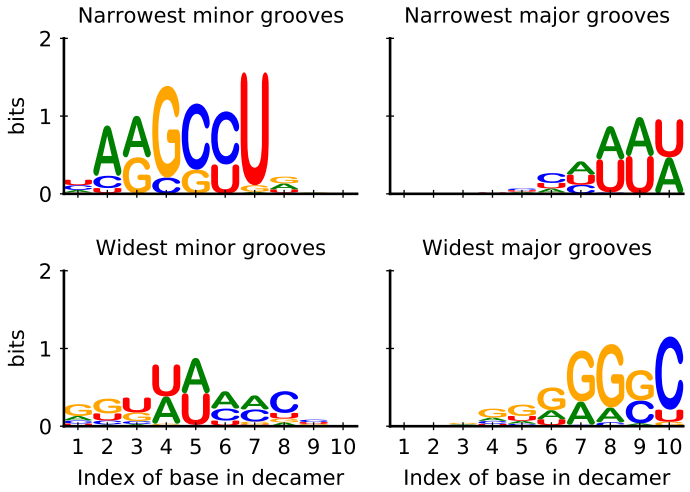
Fig. 4.9 Distribution of major and minor grooves in dsDNA, dsRNA, and DRH

Firstly, note that we have taken the central phosphate on the reading strand as the reference frame such that upstream to this reference phosphate is the minor groove and the major groove is downstream. Now, in figure 4.10(a), from the statistics obtained for dsDNA, we can observe that the presence of A and T leads to narrow minor grooves, whereas C and G lead to wider minor grooves. For minor grooves, the information is only present in 2 to 6 positions of the decamer and is zero for the rest because the minor groove is upstream of the reference phosphate (between indices 5 and 6). Furthermore, the narrow major grooves result from G/C rich sequences, and the wide major grooves result from A/T rich sequences (even though the information content is less). The findings that A/T-rich sequences have narrow minor grooves and C/G-rich sequences have wider minor grooves agree with the experimental crystallographic data for free DNA and protein-DNA complexes [134, 172] and NMR solution results [134]. Moreover, we observed that most sequences with narrow minor grooves are A-tracts and surprisingly do not have a single TA step (in position indices 2 to 6 of decamer). A similar conclusion has been reached in experiments where TA steps are found to be correlated with minor grooves widening [172].

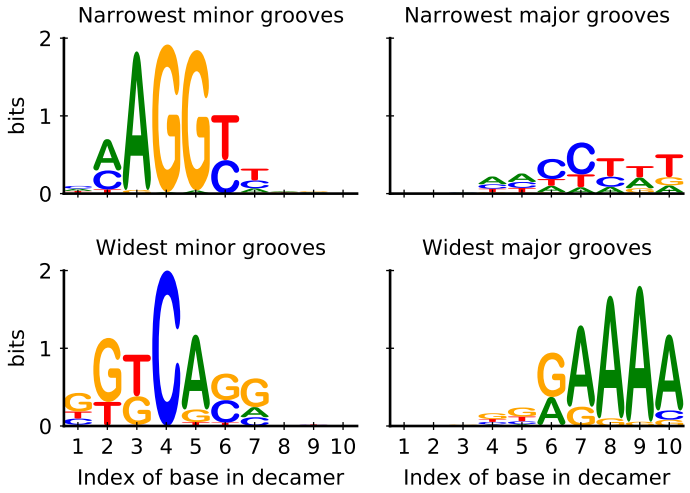
In contrast to dsDNA, the major groove of dsRNA shows much higher variation (range $\approx$ 6 Å) in sequence space (refer to figure 4.9). Moreover, from sequence logos in figure 4.10(b), it can be deduced that narrow major grooves are correlated with A/U content and wider major grooves with C/G content. Similar observations have been made in MD simulations for six



(a) Sequence logos for sequences with extreme major and minor groove widths in dsDNA



(b) Sequence logos for sequences with extreme major and minor groove widths in dsRNA



(c) Sequence logos for sequences with extreme major and minor groove widths in DRH

Fig. 4.10 Sequence logos for sequences with extreme major and minor groove widths in various dsNAs. The statistics are obtained for all decamers (≈ one million sequences) embedded in fixed flanking contexts. The x-axis is base index in the decamer with information content at that index on the y-axis. The Watson phosphate between 5^{th} and 6^{th} is taken as the reference phosphate.

dsRNA sequences [117] with $(AUAU)_n$ and $(CGCG)_n$ with the narrowest and widest major grooves. The distribution for minor grooves in dsRNA is relatively narrow. The range is less than two Å with sequences rich in G/C favoring narrow minor grooves and rich in A/U favoring wider minor grooves (although the information content is quite less).

RNAase H enzyme (crucial for genome stability and DNA replication of the mitochondrial genome [24, 185]) selectively recognizes DRH (among other dsNAs) and degrades the RNA strand without affecting the complementary DNA strand [193]. Experimental findings [54, 198] attribute this selectivity to intermediate minor groove width as the key structural characteristic. Initially, the activity of this enzyme was considered sequence-independent, while some recent experimental evidence [75, 174] suggested otherwise, with a higher purine content in RNA (or lower purine content in DNA) strand leads to resistance in RNAase H activity. Gorle et al. [196] attributed this inactivity of the RNAase H enzyme to the widening of minor grooves by showing that varying the pyrimidine content (on the DNA strand) from 0 to 100 % gradually widens the minor grooves comparable to pure dsRNA. In our results, in DRH, the distributions for major and minor grooves are between dsDNA and dsRNA. Moreover, there exists an overlap between the distributions of major and minor grooves, which implies that which groove is wider or narrower depends on the sequence. From the sequence logos in figure 4.10, it is clear that A/G rich sequences prefer narrow minor grooves and wider major grooves, whereas there is no equally clear sequence preference for wider minor grooves and narrow major grooves.

In summary, we have shown that sequence influences groove widths and specific sequences adopt extreme groove widths. We would like to emphasize that the sequence logos are plotted for a standard outliers definition (i.e., outside three standard deviations from the mean). However, similar choices, such as the extreme 1 or 2% sequences, also give similar results. Finally, the major and minor grooves have no clear correlation in sequence space.

4.5 Single nucleotide polymorphism

Single nucleotide polymorphisms (SNPs) [23] are the most common genetic variations in the human genome that include substitution variants (in which a nucleotide is replaced by other), insertion or deletion variants (in which a nucleotide is either inserted or deleted). SNPs are abundantly present in both coding and non-coding regions of the genome and are related to individuality, diversity in the population, susceptibility to diseases, and individual's response to chemicals/medicines/vaccines. SNPs are found to play a crucial role in several common medical conditions, such as migraine, diabetes, high blood pressure, cancer, and heart disease [184]. Therefore, understanding SNPs opens potential applications in personalized medicine, pharmacogenetics, forensics, and disease causation.

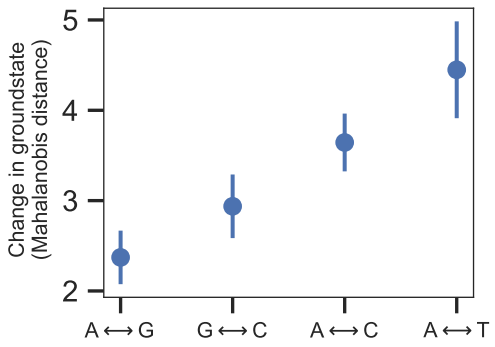
In this section, we have focused on substitution SNPs and have attempted to understand, along with the chemical changes, how SNPs influence dsDNA mechanics. To address this problem systematically, we have computed the Mahalanobis distance between all possible nonamers (9mers) with SNPs at the central nucleotide. The total possible substitution SNPs in the standard DNA alphabets (A, T, C, and G) are twelve. As we want to compute the distance between SNPs, 12 possible SNPs reduce to six (A$\longleftrightarrow$T, A$\longleftrightarrow$C, A$\longleftrightarrow$G, T$\longleftrightarrow$C, T$\longleftrightarrow$G, and G$\longleftrightarrow$C). Furthermore, in a systematic study of all nonamers, A$\longleftrightarrow$G and T$\longleftrightarrow$C are dependent; sim-

ilarly, A$\longleftrightarrow$C and T$\longleftrightarrow$G, thus, only four independent SNPs. Then, we have computed the change in groundstate in terms of symmetric Mahalanobis distance (see equation (C.12)) for each independent SNP embedded in all tetramers and a fixed flanking sequence to avoid end-effects. Note that the SNP modifies the NA on both strands and maintain CW pairs and should not be confused with DNA mismatches.

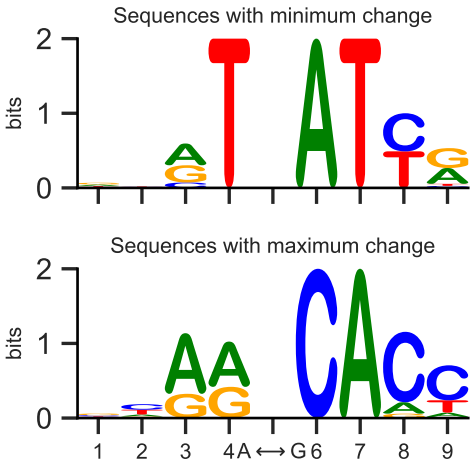
In figure 4.11(a), we have plotted the change in groundstate for each SNP, where the error bar shows the effect of flanking context. It can be expected that purine$\longleftrightarrow$purine substitution should result in less change in groundstate as compared to purine$\longleftrightarrow$pyrimidine. The observations in figure 4.11(a) align with this anticipation, i.e., A$\longleftrightarrow$G substitution lead to the minimum change in groundstate whereas A$\longleftrightarrow$T lead to maximum change in groundstate, which is approximately twice the change by A$\longleftrightarrow$G. However, the order obtained for change in groundstate for various purine$\longleftrightarrow$pyrimidine SNPs, i.e., A$\longleftrightarrow$T > A$\longleftrightarrow$C > G$\longleftrightarrow$C can not be explained easily.

It is noteworthy that the flanking sequence strongly influences the change in groundstate on point mutation, as can be deduced from the errorbars in figure 4.11(a) and therefore, in figure 4.11(b-e), we have plotted sequence logos (refer section 1.2.1) for flanking contexts in which a given point mutation leads to most or least change in groundstate. In other words, the sequence logos are plotted for outlier sequences that have a minimum/maximum change in groundstate (outside three standard deviations from the mean) on SNP at the central nucleotide. Firstly in figure 4.11(b-e), one can notice that along with the immediate flanking context to the SNP location, the next or next-nearest flanking context also plays a considerable role. Moreover, we point out that the findings here are not very sensitive to the definition of outliers or the measure of difference in groundstate.

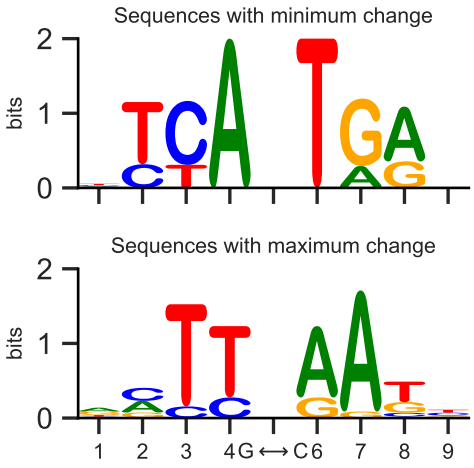
Lastly, to visualize the effect of point mutation on groundstate of a given sequence, we have plotted groundstate of two sequences in figure 4.12, one has a slight change in groundstate on A$\longrightarrow$G substitution and the other has a highly non-local change in groundstate on A$\longrightarrow$T substitution. The non-local change is up to four base-pairs on either side of the A$\longrightarrow$T substitution.



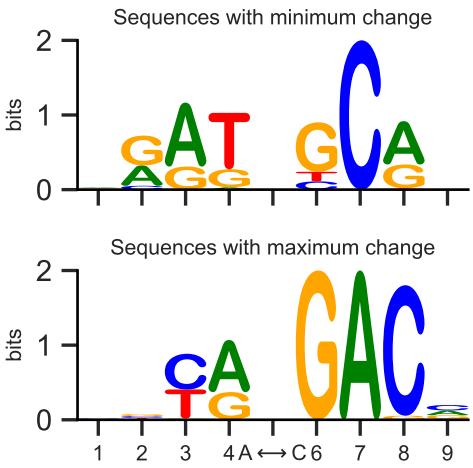
(a) Total change in groundstate due to SNPs



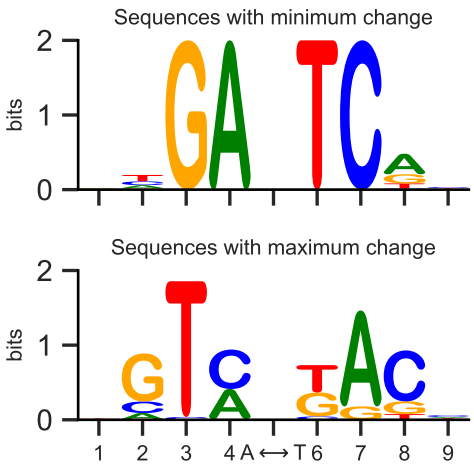
(b) A ⟷ G SNPs



(c) G ⟷ C SNP

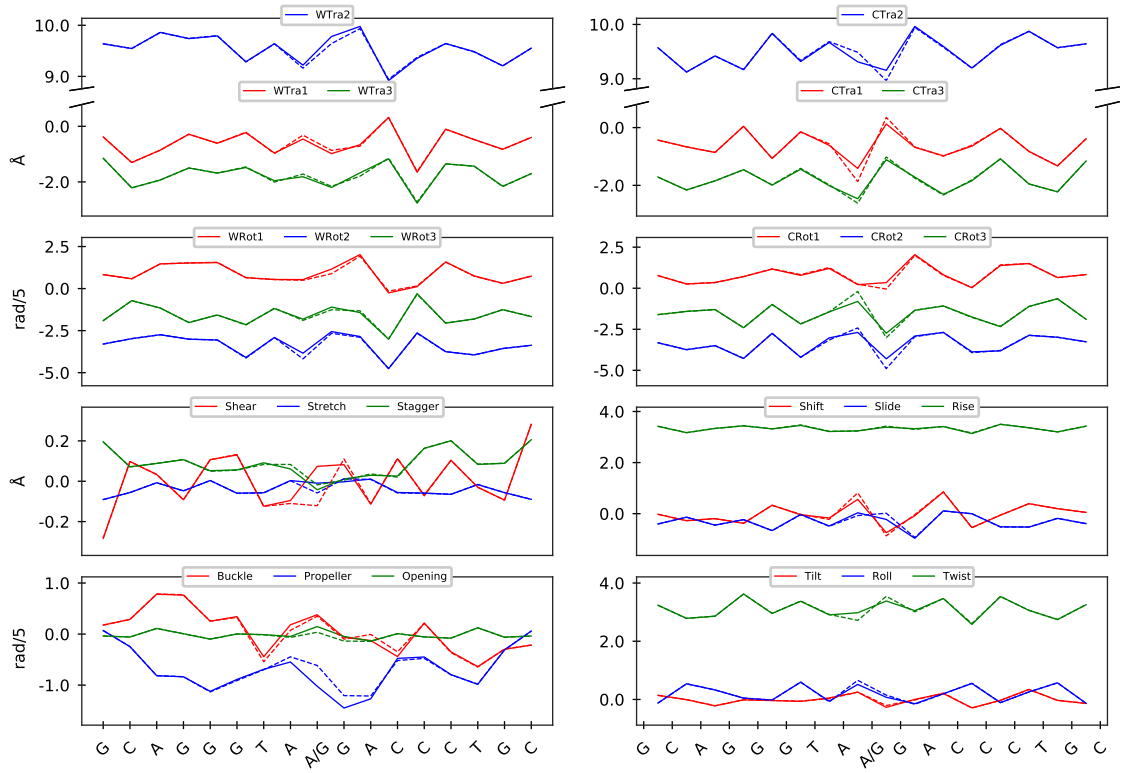


(d) A ⟷ C SNP

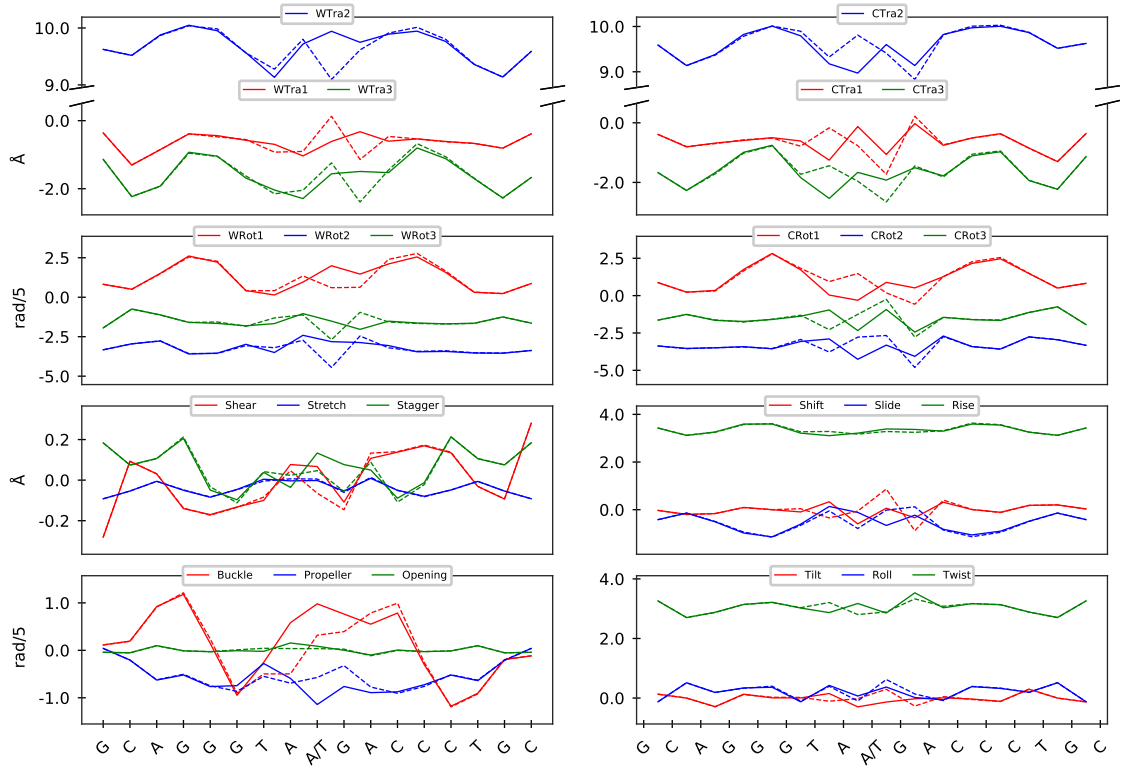


(e) A ⟷ T SNPs

Fig. 4.11 (a) Change in groundstate in terms of symmetric Mahalanobis distance on single nucleotide polymorphism (SNP) at central base-pair with error bars showing the influence on the flanking context. (b)-(e) sequence logos for flanking contexts that least and most change the groundstate on various SNPs at 5^{th} position.



(a) A ⟶ G SNPs



(b) A ⟶ T SNPs

Fig. 4.12 cgNA+ predicted groundstate coordinates (elements of w) for sequences with (a) A ⟶ G SNPs and (b) A ⟶ T SNPs at the middle base-pair. The figure highlights change in groundstate on SNPs as predicted by the cgNA+ model. cgNA+ groundstate for a given sequence is in solid lines and with the same sequence after point mutation in dashed lines.

CHAPTER 5

Comparison of non-local sequence-dependent mechanics of double-stranded DNA in protein-DNA crystal structures ensemble with the cgNA+ model

The work in this chapter is done in collaboration with Prof. Wilma K. Olson, Dr. Luke Czapla, and Dr. Helen Lindsay. In particular, Prof. Olson and Dr. Czapla helped obtain and curate the protein-DNA X-ray data set and Dr. Lindsay helped with the initial analysis.

How well double-stranded DNA (dsDNA) conformations in the protein-DNA crystal structure ensemble relate to the thermodynamic fluctuations of dsDNA in the simulated solution and exhibit non-local sequence-dependence is unclear. In this chapter, we have made a detailed comparison between the groundstate and covariance of dsDNA dimer predicted by the cgNA+ model, a non-local sequence-dependent coarse-grained model trained on atomistic MD simulations with an ensemble of protein-DNA crystal X-ray structures. For the first time, we have compared all independent dimers in tetramer contexts in both intra- and inter-base-pair coordinates. For the groundstate of the middle junction dimer in a tetramer context, we have shown that a) to study the mechanics of dsDNA at the dimer level, the tetramer context plays a crucial role and thus, can not be ignored, b) the cgNA+ model, which is indistinguishable from the corresponding MD simulation statistics provides an efficient alternative to repeating atomistic MD simulations for different flanking sequences, and c) the groundstate of dimers in their tetramer context predicted by the cgNA+ model is in agreement with the corresponding X-ray crystal data, and the direction of variation of groundstate in sequence space aligns very closely in the two data sets. Furthermore, the directions of dsDNA deformations as given by the eigenvectors of sequence-average configuration covariance are very close in the two data sets, as well as an excellent correlation between the non-local sequence-dependent configurational volume (a measure of dsDNA deformability) in terms of inter variables, and principal components of intra and inter variables both. Lastly and most interestingly, we found that the directions of maximum variation in groundstate over sequence space align with the softest modes in the configuration space in both the data sets. It justifies the nearest-neighbor assumption in the model with the observation that the minimum energy configuration (i.e., groundstate) can only be achieved by compromising more on the softest modes. Thus, in this chapter, we demonstrate that the cgNA+ model explains the dsDNA mechanics observed in the protein-DNA crystal structures ensemble and is an extensive model (almost indistinguishable from atomistic MD) that could not be trained otherwise on limited experimental data.

5.1 Introduction

Sequence-dependent mechanics of DNA plays a crucial role in several biological processes such as nucleosome positioning [179, 180], indirect readout [36, 126], and DNA looping [3, 176]. For example, specific proteins recognize the groundstate and deformability of DNA, which are often highly sequence-dependent [36, 85, 126, 146, 171]. Such direct evidence piqued significant interest in understanding the sequence-dependent mechanics of DNA. It has been established that the groundstate and stiffness of DNA can be modeled as a combination or the overall effect of local dimer contributions [49, 103, 136, 137]. However, there has been growing evidence both in experiments [110, 208] as well as in atomistic simulations [50, 97, 147] that along with the base composition of the dimer-step, non-local sequence context is also an important factor in the mechanical behavior of DNA and thus, can not be ignored.

Due to the immense sequence space of DNA, it is not feasible to investigate all sequences (even for DNA dodecamers) either experimentally or using atomistic molecular dynamics (MD) simulations. For example, the most extensive analysis using atomistic MD simulations published so far is only for the 136 independent tetramers [50, 147] by the ABC consortium. Due to these limitations, coarse-grained modeling provides an excellent alternative. By choosing the right degrees of freedom to efficiently model sequence-dependent mechanics of DNA, such coarse-grained models allow statistical sampling in this vast sequence space to better understand the mechanics of DNA.

There have been several attempts to model DNA, starting from worm-like chain models [92, 186], but one of the first and widely applied models for sequence-dependence in coarse-grained models of dsDNA was a base-pair model developed by the Olson group [139]. In that model, dimer-dependent parameters were obtained from available X-ray crystal data of protein-DNA complexes which have been a great source of information for studying groundstate and flexibility of DNA [128]. The model holds under the assumption that in protein-DNA complexes, proteins distort the DNA structure in different random directions but in specific ways that are compatible with sequences' intrinsic deformability, thus, the available X-ray protein-DNA crystal data can be used to study the conformational space of dsDNA and the propensity for different sequences to naturally distort in different ways. Similar base-pair models were also obtained using atomistic MD simulations as the training data [61, 97]. One of the major drawbacks of base-pair models is local sequence-dependence. It has been shown several times that sequence dependence limited to the dimer level is not always sufficient to explain all the properties of specific DNA sequences, and non-local sequence dependence often plays a pivotal role in DNA mechanics [6, 56, 102, 145, 155, 208]. The only way to obtain a base-pair model that captures sequence dependence beyond the dimer level is to train parameters for all possible tetramers or even beyond, which is unfeasible with the limited experimental data.

The Maddocks group provided a novel approach to overcome the limitations with the iteration over indefinitely longer DNA contexts by developing finer coarse-grain models of dsDNA trained on atomistic MD simulations with state-of-the-art force fields optimized explicitly for nucleic acid conformational flexibility and structure. In such finer coarse-grained models, cgDNA [47, 62, 159] (rigid-base level) and cgDNA+ [149] (rigid base and rigid phos-

phate level), individual base-pair steps cannot achieve their local minima, and frustration energy arises in the nearest-neighbors; thus, this approach naturally captures the non-local sequence-dependence of dsDNA but only uses dimer-dependent parameters. The latest development in this direction, the cgDNA+ model, predicts non-local sequence-dependent Gaussian pdfs for a given sequence instantly and almost indistinguishably from the corresponding atomistic MD statistics [149]. There also exist other coarse-grained models for DNA [71, 190, 203] that have been developed using a mixture of sources for training data.

Irrespective of the modeling approach, there has been a debate about training data. It is not clear how well the average structure and conformational flexibility of dsDNA in protein-DNA crystal structure ensembles reflect the average structure and thermodynamic fluctuations of dsDNA in the solvent, for example, modeled in atomistic MD simulations. In general, the deformability of dsDNA in X-ray crystal structures is significantly less than in MD simulations, since the effective temperature in X-ray crystal structures is lower [14, 97]. There have been several studies comparing X-ray crystal structure data with atomistic MD simulations [41, 56, 111, 141, 152, 155] and they have been shown to agree reasonably well for both groundstate and stiffness (ignoring the magnitude of stiffness). For example, in ref. [155] a general agreement was found between various MD force-fields (AMBER parambsc0 and CHARMM27) and the X-ray crystal database, with subtle differences seen in the force fields. Along with the equilibrium values of the helicoidal parameters of dsDNA, Dans et al. [41] explored the distributions of these parameters for both X-ray structures in the RCSB database and MD simulations, and found that the bimodality in helical coordinates has similar trends (although the study was limited due to the scarcity of X-ray data). Moreover, in ref. [97], authors argued that at the dimer level, X-ray crystal structures data and MD simulations are in good agreement but demonstrated that the dimer level model is not sufficient for either the groundstate or to study the deformability of DNA. However, the available literature comparing MD with X-ray data is limited to inter base-pair coordinates and primarily at the dimer level.

Furthermore, the base-pair resolution is insufficient to understand dsDNA mechanics and requires an explicit treatment of intra-base-pair interactions. In ref. [148], large-scale atomic force microscopy measurements found that a high Propeller (or propeller twist) is linked to higher DNA flexibility as well as higher surface accessibility, allowing propeller changes to act in regulatory elements. Also, the propeller plays a crucial role in discriminating ChIP-seq (Chromatin immunoprecipitation sequencing) bound sites from background genomic regions [118]. Furthermore, Buckle plays a crucial role in the intercalation of amino acid side groups [89, 90]. Therefore, a finer comparison, including both intra and inter base-pair coordinates is required, and with the increase in the number of protein-DNA crystal structures in the RCSB database, we believe that now there are sufficient X-ray data available for a finer comparison and to make comparisons for all independent dimers in tetramer contexts.

In this chapter, for the first time, we have carried out a systematic comparison of the groundstate and stiffness of dsDNA in protein-DNA X-ray crystal structure ensemble (say X-ray data set) with the cgNA+ model data set (which is indistinguishable from atomistic MD data) for dimers in specific tetramer context, in terms of both intra base-pair and inter base-pair step coordinates. First, we have justified why the cgNA+ model is a better choice than atomistic

MD simulations for such a comparison. Then, we have shown how crucial flanking context is to studying sequence-dependent mechanics of dsDNA and why dimer level sequence-dependence is insufficient for such purposes. Furthermore, we found an excellent agreement between sequence-independent groundstate of dimer and directions of variation of groundstate over sequence space. Subsequently, we compared the groundstate of dimers (both in average context and specific tetramer context) in the two data sets. Lastly, we have compared the stiffness in terms of configurational volume. We have found that tetramer context influences the stiffness (in inter base-pair coordinates) of flexible dimer steps more significantly than the rigid ones and have found an excellent agreement between the two data sets.

Details of all the codes and data used in this chapter are provided appendix F.

5.2 Methodology

We compare the groundstate (equilibrium shape) and stiffness (inverse configuration covariance) of dsDNA dimers (in average flanking sequence) and middle junction dimer step in a specific tetramer context (for 136 independent tetramers) in X-ray and cgNA+ model data set. In the following subsections, we describe various choices, assumptions, and methods used in this work.

5.2.1 Choices in dimers and tetramers

For dsDNA, the total number of possible dimers and tetramers are 16 and 256, of which only 10 and 136 are independent, respectively. As X-ray data are limited, we have carried out our analysis only for independent dimers and tetramers to enhance the statistics. For dimers, we have chosen RR, RY, and YR steps (a non-unique but deliberate choice) where R and Y denote purine and pyrimidine bases, respectively, and then opted for a set of 10 independent dimers. We have again chosen a non-unique set of 136 independent tetramers, but kept the same dimer steps as the central junction dimer steps. The chosen 136 independent tetramers are depicted concisely in figure 5.1.

5.2.2 Database definition

Protein-DNA X-ray crystal structures have been taken from the RCSB Protein Data Bank [20] (www.rcsb.org) (last updated here in August 2020) using Biojava [73, 95] for retrieval and caching, and for implementing the criteria to identify redundant structures. Since some of the protein-DNA crystal entries in this database include identical sequences bound to the same protein or nearly the same protein with small chemical changes (examples include nucleosome structures with an identical 147 bp DNA sequence bound to histone core protein, such as structures with PDB accession codes 1KX5, 5AV5, 5AV6, 5AV8, 5AVA, 5AVB, and 5AV9), which can bias the statistics, we developed and implemented a method to identify redundant structures.

The different sequences bound to the same protein do provide new information about the compatible flexibility as the induced bending is different for different sequences due to aspects like different hydrogen-bonding patterns with the protein; however, the inclusion of the same sequence several times would bias the statistics toward certain protein-induced bending modes.

Therefore, such redundant structures must be discarded to explore the unbiased dsDNA conformational space. To identify redundant information, which would involve the same sequence or nearly an identical sequence bound to the same protein, ECOD (evolutionary classifications of domains) [37] classification groups are used to identify the protein category, and then sequence matching (having a longest common sub-sequence in two structures that is greater than 70 % of the total DNA length) is used as the second criterion to determine if a structure is redundant compared to one previously chosen if it is bound to the same protein. For structures that meet both criteria, the one with the better X-ray resolution is chosen, and the other structures that are nearly identical in sequence and bound to the same protein are discarded.

Subsequently, we have coarse-grained those selected atomistic structures by fitting a rigid-body frame [191] using standard DNA atomic coordinates defined in the Tsukuba convention [140]. From these frames, we obtained CURVES+ internal coordinates [101]. Lastly, from a sequence of any length, we have extracted coordinates of all dimers in their tetramer contexts with the constraint that the tetramer should be at least two base-pairs away from ends.

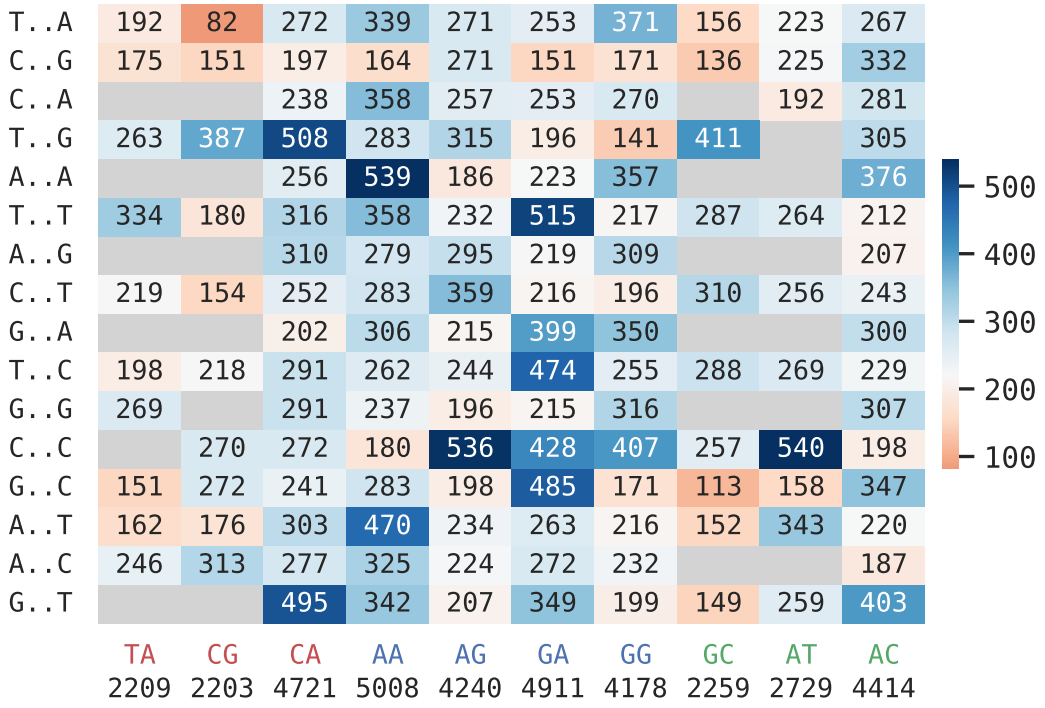


Fig. 5.1 Number of appearances of 136 independent tetramers in X-ray data set (case-I). Abscissa is middle junction dimer-step and ordinate is flanking tetramer context. The blank entries in the plot represent the dependent tetramer. Note that palindromic steps (self-complementary) are only read from the reading strand here. Further note that while computing the sequence-independent average and covariance, we consider all 256 tetramers and for palindromic steps, we have used double of their corresponding weights (details in section 5.2.3.1).

Furthermore, the crude data contain some entries with highly distorted DNA, broken H-bonds, and nicks which requires filtering. One of the consequences of these distortions is that the rotation angles may go very close to π. In cgNA+ coordinates, rotation angles are defined

as the norm of Cayley vector [149, 158] which tends to infinity when the angle approaches π (details in section 2.2.1). It allows us to efficiently filter such cases as the corresponding Cayley components become very high (of the order $\sim 10^6$). This step removed $\approx 9\%$ of the data. Following this, to ensure that each parameter follows a quasi-normal distribution and does not have long asymmetric tails, we adopted a variant of the 3σ method used originally in ref. [139]. We remove a snapshot if any parameters are outside three standard deviations from the mean. This method has been often used to curate X-ray data [139, 152], but only while using inter parameters and at the dimer level, which allows the algorithm to converge in 5-6 steps. However, in our case, we are using 18 parameters and analyzing 136 tetramers, which rejects $\approx 50\%$ of the data and convergence requires 10-20 cycles. An alternative approach might be to reject the dimer steps that have broken H-bonds [56], but with a risk of long tails in the distribution of some of the internal coordinates for some of the tetramers. Therefore, in this chapter, we decided to use the 3σ approach for just one cycle (to eliminate long asymmetric tails, if any) and ensure that the accepted data contains no broken H-bonds keeping $\approx 70\%$ of the crude data. Moreover, we also analyzed the data obtained after two and three cycles of the above approach and found a negligible impact on the conclusions as it only removes the data from the tails without influencing the mean.

Lastly, we performed our analysis for two sets of data I) No resolution cut-off on PDB structures and II) PDB structures with at least 3.0 Å resolution or better. In the chapter, we have presented results for case-I and for case-II, corresponding results are provided in appendix E. The precise number of instances for each tetramer in the X-ray data set for case-I and case-II are provided in figures 5.1 and E.7, respectively.

5.2.3 Methods to compare X-ray statistics with cgNA+ statistics

5.2.3.1 Computation of sequence-independent groundstate and covariance

To compute the sequence-independent (or sequence-average) groundstate and covariance, we have considered all 256 tetramers with double the corresponding weights for palindromic steps. It is because, for non-palindromic steps, we are counting the same physical dimer step twice while reading from both the strands; therefore, to balance the statistics, we are taking double weights for the palindromic steps. In this chapter, we have defined two covariance matrices, namely, shape covariance (C_s) and configuration covariance (C) defined as:

$$C_s = \frac{\sum_{i=1}^{N} w_i \left(x_i - \mu^*\right)\left(x_i - \mu^*\right)^T}{\sum_{i=1}^{N} w_i} \text{ where } \mu^* = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i} \tag{5.1}$$

$$C = \frac{\sum_{i=1}^{N} w_i \left(C_i + x_i x_i^T\right)}{\sum_{i=1}^{N} w_i} - \mu^* (\mu^*)^T \tag{5.2}$$

where w_i, x_i, C_i, and μ^* are the weight (or number of instances), average shape, covariance for a given tetramer, and weighted sequence-independent (i.e., sequence-average) average shape, respectively. Shape covariance (C_s) is computed over the groundstate of dimer in tetramer context and can be described as the directions in which groundstate varies over sequence space. For the X-ray data set, configuration covariance (C) is defined as the deformation of DNA in

the configuration space computed over all entries in the X-ray data set. For cgNA+ model data set, it is the Gaussian average of the covariance matrices for DNA dimer in tetramer contexts. We would like to emphasize that the groundstate and covariance computed using 256 tetramers will respect the palindromic properties as discussed in section 5.2.3.2. Note that we have used the subscript X and M to highlight the statistics from the X-ray and cgNA+ model data sets, respectively. For instance, C_s^X and C_s^M are shape covariance for X-ray and cgNA+ model data sets, respectively.

5.2.3.2 Eigenvector parity

There is an inherent CW symmetry in the groundstate of dsDNA due to the reading strand choice. E (= E^T) matrix (reading strand transformation matrix defined in section 2.2.2) maps the groundstate of dsDNA read from one strand to another, i.e., $\mu(\bar{S}) = E\mu(S)$. This inherent CW symmetry is also reflected in the configuration covariance matrices and thus, follows the CW symmetry condition $C(\bar{S}) = EC(S)E$. For a palindromic or self-complementary sequence (invariant of reading strand) or the average sequence (which is also invariant of reading strand), $S = \bar{S}$, the relation becomes $\mu(\bar{S}) = \mu(S)$ and $C(S) = EC(S)E$. For such palindromic cases,

$$D = P^T CP = P^T ECEP = (EP)^T C(EP) \tag{5.3}$$

where D and P are the eigenvalue (eigenvalues on the diagonal) and eigenvector matrix (with columns as eigenvector), respectively. Due to CW symmetry, if P_i is an eigenvector, then EP_i is also an eigenvector with the relation $P_i = \pm EP_i$ where positive or negative signs define a parity of the eigenvector. Furthermore, we used cosine similarity to compare eigenvectors in the two covariance matrices, defined as the dot product $(P_i \cdot P_j)$ between the corresponding eigenvectors. Note that we have continued to use the subscript X and M to highlight the statistics from the X-ray and cgNA+ model data sets, respectively.

5.2.3.3 Hierarchical Clustering

Clustering is an unsupervised machine learning method that finds patterns in the data sets consisting of input data without labels. It finds meaningful structure, features, and groupings inherent in the input data. In particular, we have performed hierarchical clustering (which group data into a tree of clusters) on the groundstate of dsDNA dimers in tetramer context using the square root of symmetric Mahalanobis distance, $\mathcal{M}$ (defined in equation (C.12)) as the metric and average linkage as the linkage algorithm [124]. The standard python or Matlab linkage algorithm is used in which the distance $\mathcal{D}(p, q)$ between two clusters is computed. The algorithm starts by treating every data point as an individual cluster and then combining two nearest clusters, say p and q, into one cluster, and then removing p and q. It iterates until only one cluster is left, which becomes the root. The average linkage algorithm defines the distance between two clusters as

$$\mathcal{D}(p, q) = \frac{1}{|p| * |q|} \sum_{i=0,j=0}^{i=|p|,j=|q|} \sqrt{\mathcal{M}(\rho_i, \rho_j)} \tag{5.4}$$

where $|p|$ and $|q|$ are cardinalities of clusters.

5.2.3.4 Configurational volume

For X-ray crystal data or atomistic simulations, the deformability of DNA can be quantified in terms of fluctuations of internal coordinates in the configuration space. For fluctuations $\in \mathbb{R}^N$, configurational volume [139, 141] can be defined as:

$$S = \sqrt{\lambda_1 \lambda_2 \cdots \lambda_N} \tag{5.5}$$

where λ_i are the eigenvalues of the covariance matrix, C $\in \mathbb{R}^{N \times N}$. The unit of S is equivalent to $\text{Å}^{N/2} \cdot (rad/5)^{N/2}$.

5.2.4 Assumptions in this study

- The primary assumption in this comparison is that in protein-DNA X-ray crystal ensemble, distortions of dsDNA resulting from proteins effectively balance out for a sufficiently large ensemble exposing the intrinsic mechanical behavior of dsDNA. It leads to the following assumption that we have sufficient data for each dimer in the tetramer context. The exact number of instances for each dimer in the tetramer context is provided in figure 5.1. For most tetramer contexts, we have at least 150 instances (after filtering), and the distribution of internal coordinates is peaked around a particular value (sometimes bi-modal) providing confidence in the statistics, at least for the groundstate. Moreover, as described in section 5.2.3.2, groundstate for the palindromic dimer should be invariant of the reading strand, which allows us to define the palindromic error (refer section 2.5.1) to quantify the convergence in X-ray statistics. For X-ray data set, palindromic error for dimer in tetramer context is 0.0197 while for dimer in average tetramer context is 0.0042 (details in figure E.4). Notably, the corresponding palindromic error in MD simulations training data for the cgNA+ model is 0.0025 and 0.00037, respectively, which contains $\sim 10^7$ snapshots. Even though the palindromic error is the norm of a scaled vector with mixed rotational and translational entries, Å or rad/5 can be treated as the units of palindromic error. The palindromic error obtained in the X-ray data set is still reasonable because, in the X-ray data set, flanking sequence beyond tetramer context is different for most entries.
- Furthermore, in this comparison, we have only considered the first and second moments of the distribution of helical internal coordinates for each dimer step (either in average context or tetramer context). However, it is well known that for some of the internal coordinates, there exists an inherent bi-modality in helical internal coordinates [41]. Further investigation of bimodality is outside the scope of this chapter (also limited by the scarcity of experimental data) and was previously discussed at dimer level in ref. [41, 138].
- Lastly, it is not clear how physical conditions in crystallization experiments (which might be different for each protein-DNA complex) influence the mechanics of dsDNA. For example, the effective temperature of protein-DNA crystal ensemble is significantly less than in atomistic MD simulations and is unknown and not easy to determine [14, 97]. Furthermore, other factors such as divalent cations, salt concentration, buffer-type, and packing forces are poorly understood.

5.3 Results and Discussion

5.3.1 cgNA+ model over atomistic MD simulations

In the X-ray data set, each instance of a dimer in the tetramer context is most likely flanked by a different sequence. However, it is not computationally feasible to perform atomistic MD simulations for tetramers in all the possible flanking sequences, even up to two base-pair steps on both sides. In contrast, the cgNA+ model provides an excellent alternative whose predictions are almost indistinguishable from MD and can efficiently compute statistics for millions of sequences. In figure 4.1, we have exemplified that the cgNA+ model prediction is almost indistinguishable from the corresponding MD statistics for a sequence outside the cgNA+ model training library. Furthermore, in figure 4.2(b), we have also compared internal coordinates of middle junction dimers embedded in different beyond tetramer contexts for MD simulations and demonstrated that beyond tetramer context could influence the groundstate of junction dimer, and the cgNA+ model can capture such highly non-local changes due to change in hexamer or beyond context. A detailed assessment of the cgNA+ model prediction accuracy is in chapter 4.

Thus, in this chapter, we have used the cgNA+ model over MD simulations and reconstructed the groundstate and stiffness matrix of all possible sequences with a length of 10 base-pair steps (plus GC ends) using the cgNA+ model. Subsequently, we extracted the marginal of the middle junction dimer in each tetramer context to obtain the statistics in the average flanking sequence context beyond tetramer context.

Lastly, we would like to remind that the cgNA+ model is a rigid-base and rigid-phosphate model which predicts the groundstate of a given sequence in base and phosphate internal coordinates. However, in this chapter, we have only compared the cgNA+ model predictions with X-ray data set in base-coordinates (both intra- and inter-base coordinates) by taking marginals over the phosphate coordinates. It is well known that phosphate coordinates are highly multimodal (see section 3.6) and is related to BI–BII backbone conformations [44, 69] which is again found to be dependent on the sequence [44]. Such comparison of backbone conformations in X-ray and MD data set was carried out previously in ref. [111]. Therefore, we believe that there are not enough experimental data to compare phosphates coordinates, particularly at the tetramer level, and is, therefore, left for further detailed investigations in the future.

5.3.2 Comparison of groundstate

5.3.2.1 Comparison of dsDNA dimer groundstate and the directions of variations in groundstate over sequence space

In figure 5.2(a), firstly, we have plotted the sequence-independent (or sequence-average) average shape of the dimer predicted by the model along with the corresponding observations in the X-ray data set. The average shapes of dimers in the two data sets are very close. These results agree well with previous findings [41, 56, 111, 141, 152, 155] limited to inter variables.

Furthermore, it is interesting to understand the directions in which groundstate of dsDNA varies over sequence space and whether these directions are consistent in the two data sets. We have computed "shape covariance matrices" denoted as C_s^X and $C_s^M \in \mathbb{R}^{18\times18}$ for X-ray and

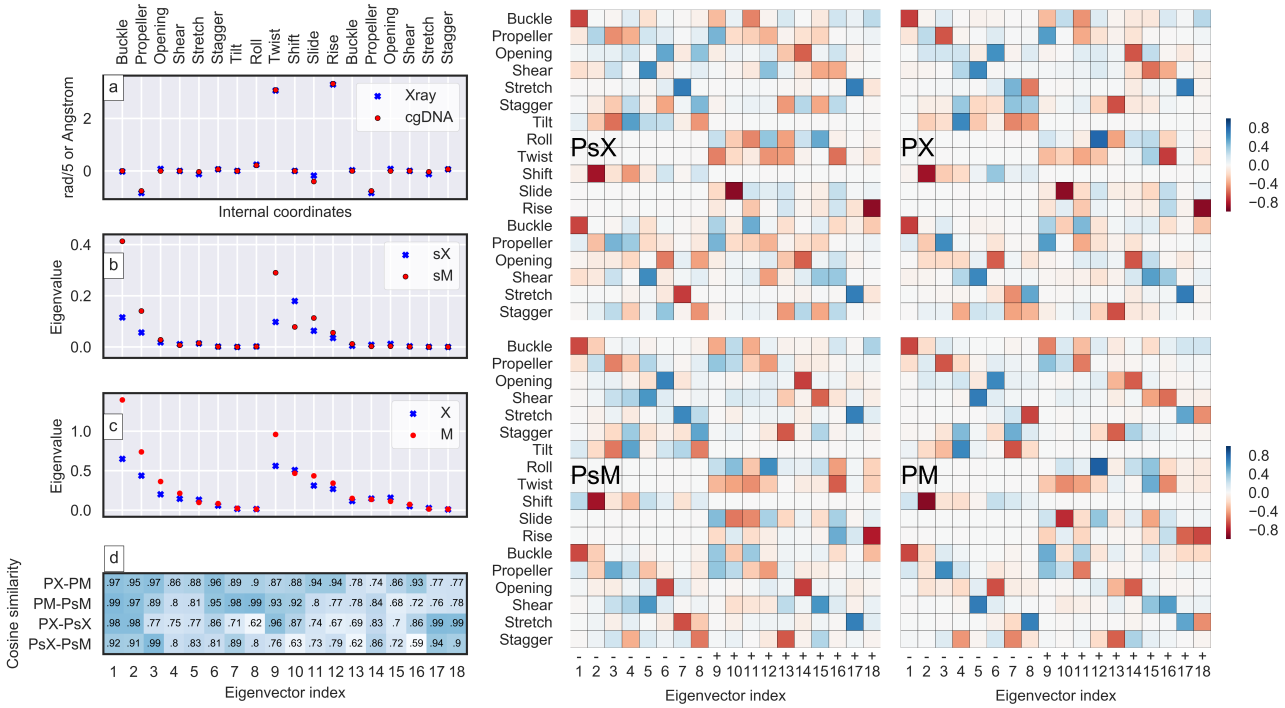


Fig. 5.2 a) Plot comparing sequence-independent groundstate (average shape) of dimer coordinates in X-ray and cgNA+ model data set. On right, P_s^X and P_s^M are the associated eigenvector matrices for the shape covariance matrix (denoted by subscript s) describing the directions of variation in groundstate over sequence space for X-ray (denoted by superscript X) and cgNA+ model (denoted by superscript M) data sets, respectively and D_s^X and D_s^M are corresponding eigenvalues in b). While P^X and P^M are the eigenvectors of average configuration covariance describing the direction of deformation of dsDNA in configuration space and D^X and D^M are corresponding eigenvalues in c). In d), there is cosine similarity index for corresponding eigenvectors in (C^X, C^M), (C_s^M, C^M), (C_s^X, C^X), and (C_s^X, C_s^M).

cgNA+ model data sets, respectively, and plotted the corresponding eigenvectors (P_s^X and P_s^M) and eigenvalues (D_s^X and D_s^M) in figure 5.2. The shape covariance matrices (defined in section 5.2.3.1) describe the directions in which groundstate varies over sequence space. C_s^X and C_s^M are shown in figure E.1 and ignoring the scale, the two matrices look very close and therefore, explored further by looking into the eigenvalues and eigenvectors. For both the data sets, observed eigenvectors are quite sparse, in particular, follow a unique sparsity pattern with decoupling of inter variables with intra1 and intra2 variables (intra1 and intra2 are intra base-pair coordinates for the first and second base-pair of the dimer). This decoupling originates from the inherent CW symmetry in groundstate of dimer and is algebraically explained in appendix D. There are 8 negative parity eigenvectors ($P_i = -EP_i$) with no fluctuations in the direction of Roll, Twist, Slide, and Rise, while for 10 positive parity eigenvectors ($P_i = +EP_i$), there are no fluctuations in the directions of Shift and Tilt where P_i and E are eigenvectors and reading strand transformation matrix. More details on reading strand transformation and eigenvector parity are provided in section 5.2.3.2. The number of positive and negative parity eigenvectors and this sparsity pattern in eigenvectors can be explained by inherent CW symmetry in groundstate of dsDNA. Now, the question arises how much eigendirections in the two data sets align?

To quantify this, we have computed the cosine similarity (details in section 5.2.3.2) between the best aligned eigenvectors in the two data sets which is plotted in figure 5.2(d). The average cosine similarity between the eigendirections of two data sets is 0.81 ± 0.11. More importantly, in the two data sets, eigenvectors align approximately according to the magnitude of the corresponding eigenvalues, i.e., eigenvectors with comparable eigenvalues in both the data sets align with each other. However, note that the magnitude of eigenvalues for the model shape covariance C_s^M is larger than the corresponding aligned eigenvalues for X-ray data set C_s^X, particularly the larger ones. It indicates that even though the direction of variation in the average shape over sequence space is similar, the magnitude of variation in the X-ray data set is less, which can be attributed to lower effective temperature in the X-ray data set.

Another interesting observation is to identify the directions (in CURVES+ coordinates) from which eigenvectors are composed of. For example, eigenvectors with large eigenvalues are dominated by the Buckle, Propeller, and Shear (intra variables). While in inter variables, Rise is associated with the eigenvector that has the smallest eigenvalue and the rest of the CURVES+ inter base-pair coordinates seems important. For the CURVES+ coordinates associated with eigenvectors with the lowest eigenvalues, the variation over the sequence space is so low that it is almost impossible to distinguish the sequence effect from the underlying noise. This can be quantified by looking at the variance of internal coordinates for groundstates over sequence space as listed in table E.1. For example, the variance in Stretch and Rise over tetramer sequence space is 0.007 and 0.006 Å^2 in X-ray data set. Notably, in the cgNA+ model data set, Stretch and Rise have relatively low variance than others.

5.3.2.2 How crucial is tetramer context?

In this section, we have investigated how crucial tetramer contexts are by comparing the dimers in specific tetramer contexts with those in the average context. In figure 5.3, we have directly plotted the internal coordinates of a dimer in tetramer and average context as small and large horizontal lines, respectively, using a different color for each specific tetramer context and the last column as the sequence-average groundstate in each panel. As can be seen in the plots, the internal coordinates of a dimer are, in general, sensitive to its context, and this sensitivity varies for different coordinates. In some cases, such as Buckle or Rise, the variation in groundstate due to flanking tetramer context change is larger than the variation in groundstate due to change in central dimer step in the average context. Furthermore, the variation in some internal coordinates over sequence space is negligible, for instance, Stretch and Rise. Figure 5.3 contains all the data condensed into one plot to provide an overall picture of various aspects of the two data sets and for identifying exciting questions. For instance, is there any pattern in dimer sequences (in average context) for which a given internal coordinate is far from the corresponding sequence-averaged internal coordinate or for a given dimer, which tetramer contexts result a dimer groundstate far from its groundstate in average context? More standard questions include the correlation between the various internal coordinates in two data sets.

Firstly, we have plotted the shifted (by the sequence-average shape) average shape for all the dimers in figure 5.4(a) to visualize which dimers have the shape farthest or nearest to the sequence-average shape. Once again, it can be observed that the spread of various internal

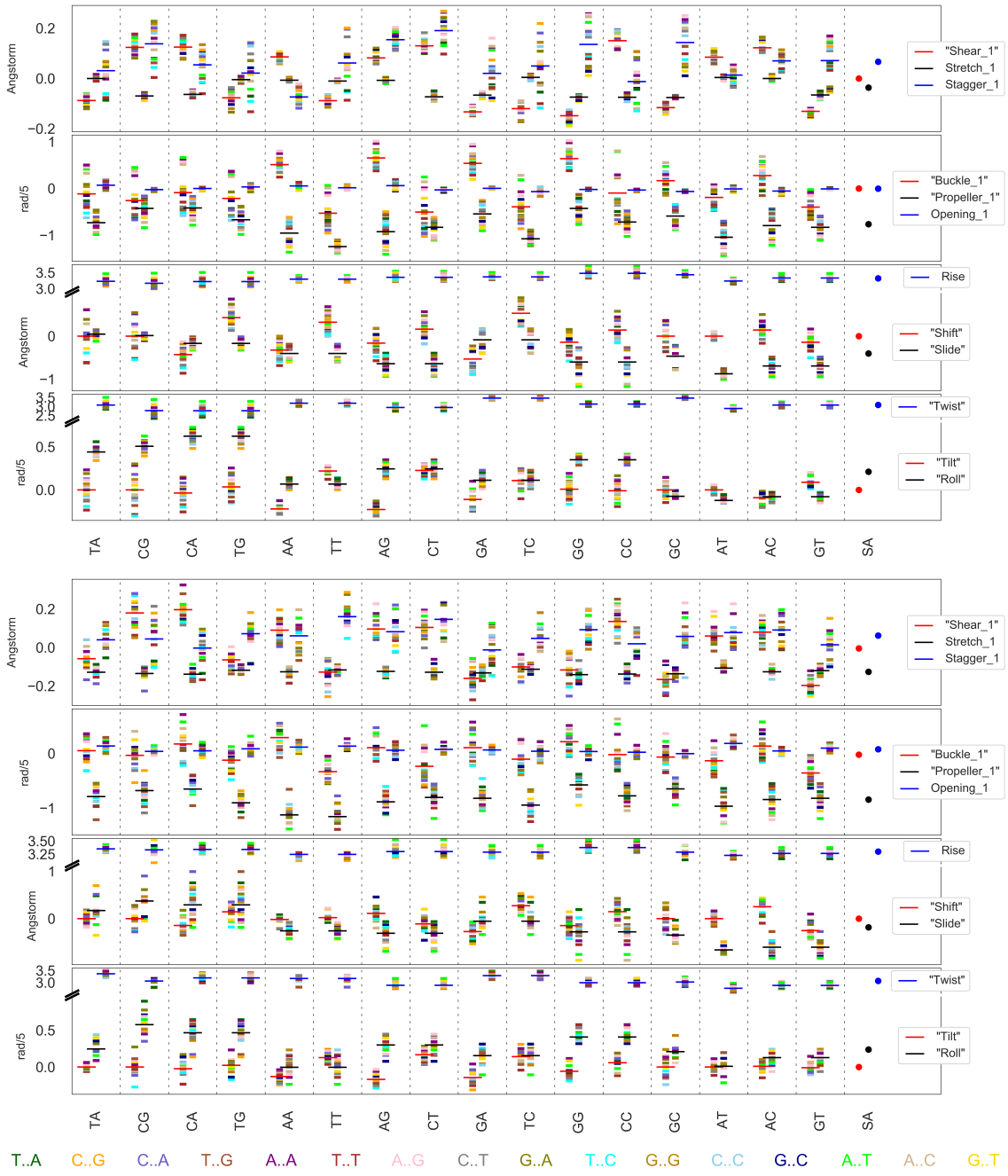


Fig. 5.3 Plot of Intras and Inter for X-ray (bottom) and cgNA+ model (top) data set in which large dash lines depict internal coordinates of a dimer (in average context) while the other smaller dash lines are the internal coordinates for that dimer in a specific tetramer context. For better and concise visual representation, in each subplot, the three internal coordinates are slightly shifted on the X-axis. Also, separate flanking contexts are plotted in different colors as described at the bottom of the plot. SA is sequence-average groundstate.

coordinates is smaller in X-ray data than in the model data. Furthermore, different dimer steps adopt extreme values for different internal coordinates. For instance, RR and YY steps adopt the extreme values for Tilt and Shift, whereas RY and YR for Roll and Slide, and this trend is generally present in both data sets. In contrast, there is no clear pattern for Twist and Rise.

Moreover, from figure 5.3, it is clear that the flanking tetramer context significantly influences the average shape adopted by a dimer. In figure 5.4(b), we have attempted to answer which flanking contexts influence the shape of dimers the most. For each internal coordinate (IC), we have defined $\gamma_{\mathrm{XUVZ}} = \mathrm{IC}_{\mathrm{XUVZ}} - \mathrm{IC}_{\mathrm{X_{avg}UVZ_{avg}}}$ as the difference of the internal coordinate of a dimer (UV) in tetramer context (X - - Z) with the same dimer in average context, where X, U, V, Z $\in$ [A, T, C, G]. Then, for each internal coordinate, we have defined positive and negative outliers as, $\gamma_{\mathrm{XUVZ}} < -\sigma$ and $\gamma_{\mathrm{XUVZ}} > +\sigma$, where σ is standard deviation of γ_{XUVZ}. In figure 5.4(b), we have sequence logos plot with the information content in the tetramer flanking context (X - - Z) for which γ_{XUVZ} are negative or positive outliers. Sequence logos are described in more detail in section 1.2.1. From the sequence logos, it can be observed that dimer adopts extreme values only in specific flanking contexts. For instance, in the presence of Y - - Y flanking contexts, Tilt and Shift adopt a lower value than the average context, and the converse is true for R - - R flanking contexts. Similarly, R - - Y and Y - - R flanking contexts tend to decrease and increase the equilibrium Slide values adopted by dimers, whereas the same contexts increase and decrease the equilibrium Twist and Rise values. Lastly, C/G flanking contexts decrease the equilibrium Roll values preferred by dimers, whereas A/T contexts have the opposite effect. These observations are made for the cgNA+ model data set, but similar conclusions can also be made for X-ray data sets. In general, less information is present for the X-ray data set, which implies that the preference for particular flanking contexts is not equally strong. Moreover, for Roll and Shift, no information in flanking contexts. Note that the statistics are obtained for a limited X-ray data set, which might be noisy, so such an agreement is still impressive. Thus, it can be concluded from the two subplots in figure 5.4 that different sequences prefer different equilibrium shape, and some conclusions can be made about their preference based on the pyrimidine or purine nature of the sequence.

An alternate way of exploring the role of tetramer contexts can be dendrograms using hierarchical clustering in the two data sets as plotted in figure 5.5. For clustering, we have used the square root of symmetric Mahalanobis distance [113] as the distance metric, and "average linkage" algorithm to compute the distance between clusters, which is defined in terms of the square root of Mahalanobis distance. More technical details on the dendrograms are provided in section 5.2.3.3. In figure 5.5, we observed similar clustering in both data sets,

- there are three main clusters corresponding to YR, RR, and RY dimer steps with an exception in RR and RY clusters, where R is Purine and Y is Pyrimidine base. This classification highlights the importance of the middle-junction dimer step;
- in each cluster, there are sub-clusters that correspond to the same middle junction dimer-steps in different tetramer contexts. However, sub-clusters in YR cluster are not well-resolved;
- YR step is farthest from all the clusters (interestingly, YR dimer steps are found to be exceptionally flexible [139]).

At the same time, there are some differences in the dendrograms for the two data sets; notably, the distance between clusters is not the same in the two data sets, which can be attributed to the fact that the magnitude of shape covariance and configurational covariance in the two data sets are not identical due to lower effective temperature in X-ray data set. With these observa-

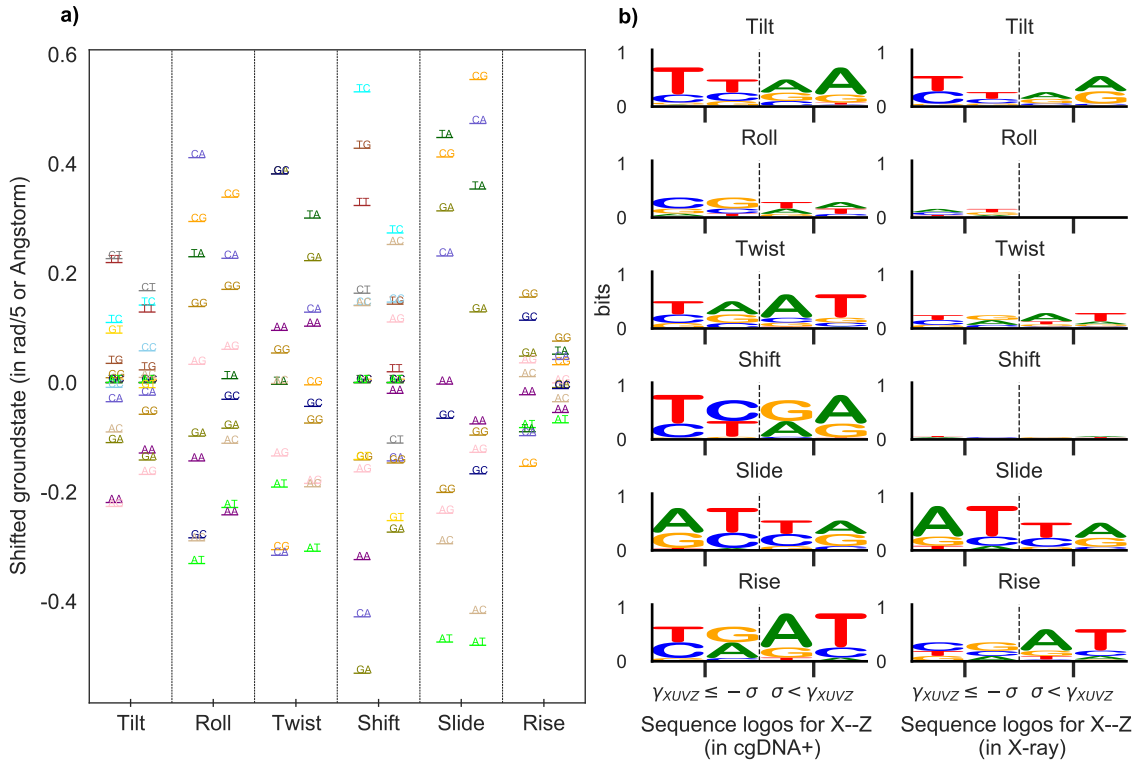


Fig. 5.4 a) Inter coordinates (shifted with respect to sequence-average groundstate) are plotted for dimers in average flanking context to identify which dimers assume distant values from sequence-average groundstate for a given variable and whether that signal is consistent in the two data sets. For each IC, the left column is for the cgNA+ model data set and the right column is for the X-ray data set. b) Sequence logos plot to statistically quantify the role of tetramer context on the groundstate (in inter variables) of a given dimer. For each internal coordinate (IC), we have defined $\gamma_{\mathrm{XUVZ}} = \mathrm{IC}_{\mathrm{XUVZ}} - \mathrm{IC}_{\mathrm{X_{avg}UVZ_{avg}}}$ as the difference of the internal coordinate of a dimer (UV) in tetramer context (X - - Z) with the same dimer in average context, where X, U, V, Z $\in$ [A, T, C, G]. Then, for each internal coordinate, we have defined positive and negative outliers as, $\gamma_{\mathrm{XUVZ}} < -\sigma$ and $\gamma_{\mathrm{XUVZ}} > +\sigma$, where σ is standard deviation of γ_{XUVZ}. In the sequence-logos plot, we have plotted the information content in the tetramer flanking context (X - - Z) for which γ_{XUVZ} are negative or positive outliers.

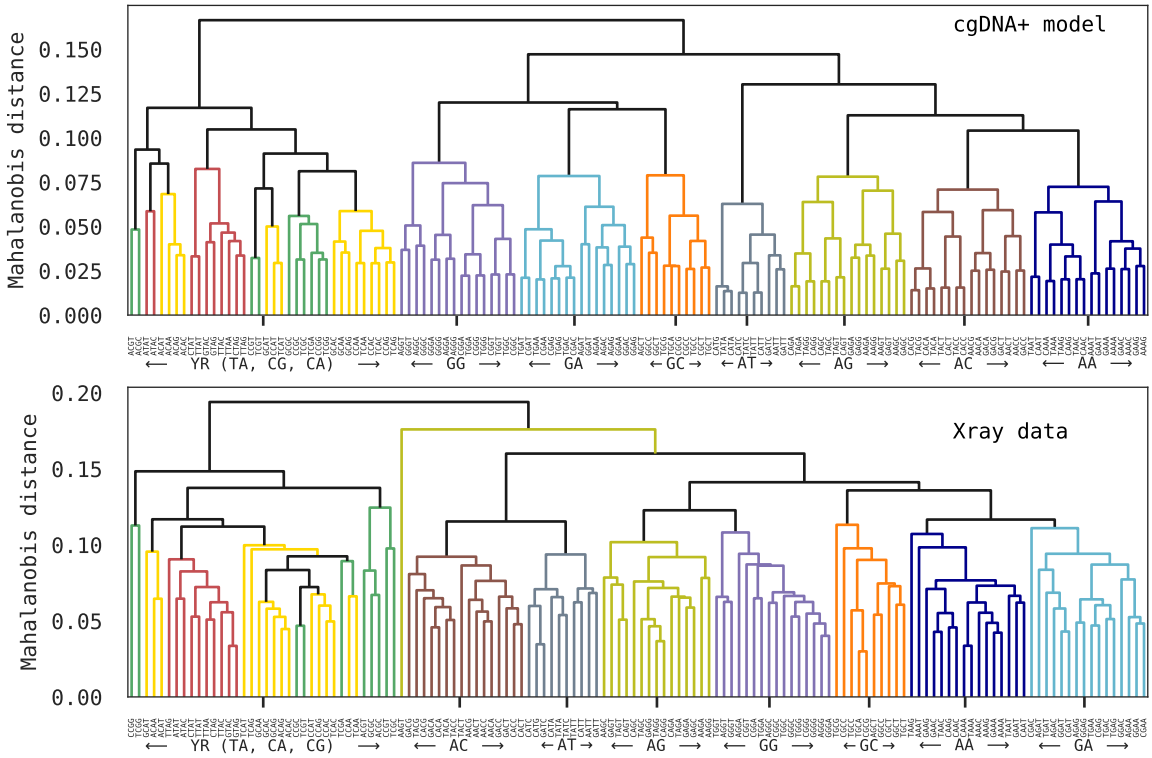


Fig. 5.5 Dendrograms using hierarchical clustering on independent tetramers using square root of symmetric Mahalanobis distance (taking inverse of sequence-dependent configuration covariance as the weight matrix) as metric and average linkage algorithm section 5.2.3.3.

tions, we can conclude that the tetramer context plays a crucial role in the dimer groundstate and thus, dimer models are not sufficient for a complete description of sequence-dependent mechanical properties of dsDNA. Therefore, in the next section, we have performed a further rigorous analysis comparing groundstate of dsDNA dimer in all independent tetramer contexts.

5.3.2.3 Comparison at tetramer level

In this sub-section, we have first compared the groundstate of dimers in specific tetramer contexts by computing Pearson correlation (PC) between each internal coordinate for the two data sets as plotted in figure 5.6(a). One can observe that PC for some of the internal coordinates (such as Buckle, Propeller, Shear, Tilt, Roll, Slide) is excellent, while for others (such as Stretch, Stagger, Rise) PC is quite low. We observed that the internal coordinates with the lowest PC (except Twist) are the internal coordinates that are the stiffest modes in shape covariance or, say, varies the least in sequence space for groundstate as listed in table E.1. Note that the X-ray data have an inherent noise, and for internal coordinates, which have very low variation over the sequence space, it is almost impossible to distinguish the sequence effect from the underlying noise. For example, the variance in the Stretch and Rise over tetramer sequence space is 0.007 and 0.006 Å^2 in the X-ray data set. Notably, in the cgNA+ model data set, the same internal coordinates also have relatively lower variance than others. Moreover, we have also shown a corresponding correlation in two data sets for dimers in the average context in the same figure.

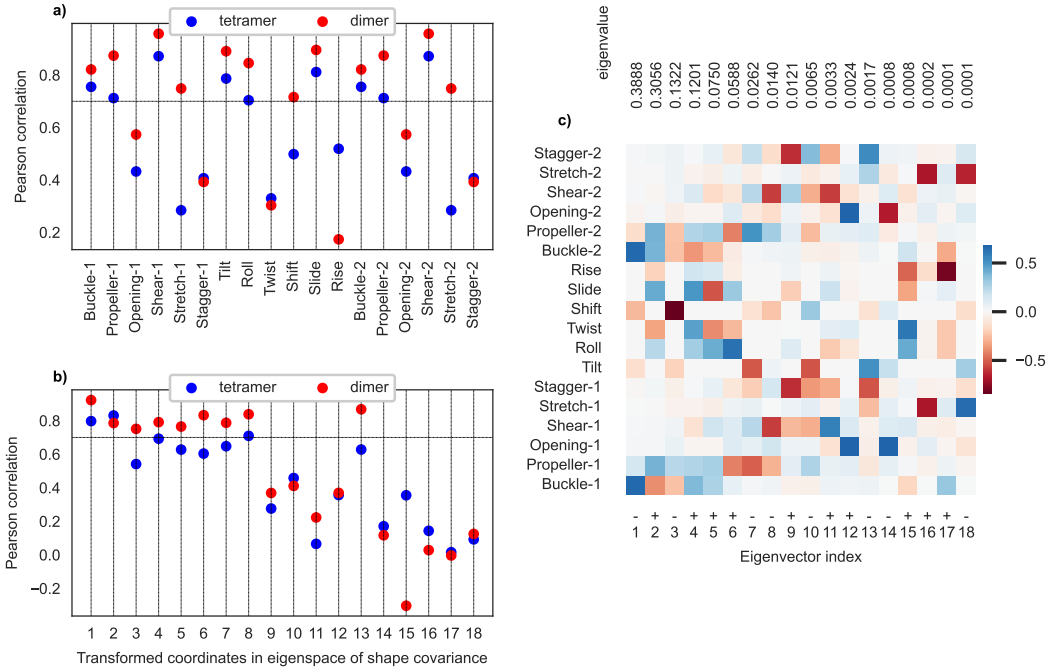


Fig. 5.6 Pearson correlation between X-ray and cgNA+ data set a) in standard CURVES+ coordinates and b) in transformed coordinates in the eigenspace of cgNA+ shape covariance and the corresponding eigenvectors shown in c) with the $+/-$ parity as defined in section 5.2.3.2.

The agreement between the two data sets at the dimer level is generally better.

To further justify this hypothesis, we have transformed CURVES+ internal coordinates into the eigenspace of cgNA+ shape covariance matrix. The eigenvectors and eigenvalues are plotted in figure 5.6(c) (note this is the same matrix plotted earlier in figure 5.2 but differently). In this plot, one can observe that the higher modes are populated mainly by Buckle, Propeller, Shift, and inter coordinates except for Rise. It roughly fits the hypothesis that the PC is lower for variables with less variation in average shape over sequence space. So, we have computed the PC in the transformed coordinates of the eight principal modes for both dimer and tetramers, and it can be observed in figure 5.6(b) that the PC between the two data sets is excellent for both dimers in average and specific tetramer contexts. In contrast, the PC for transformed coordinates in lower modes is very low (except index 13). Thus, the outcomes agree with our hypothesis that in the directions with least variation in the sequence space, inherent noise might have a dominating effect, and thus, in those directions, it is not sensible to compare the two data sets directly.

Therefore, lastly, we have compared 136 independent tetramers in terms of symmetric Mahalanobis distance (defined in section 2.5.5) but only in eight principal components of cgNA+ shape covariance matrix. The cgNA+ shape covariance matrix is shown in figure 5.6(c), and the eight principal modes explain $\approx 97.6\%$ variability in the data computed as the sum of eigenvalues of the eight principal modes divided by the sum of all eigenvalues. Thus, in this way, we have removed the directions which are believed to be dominated by the inherent noise.

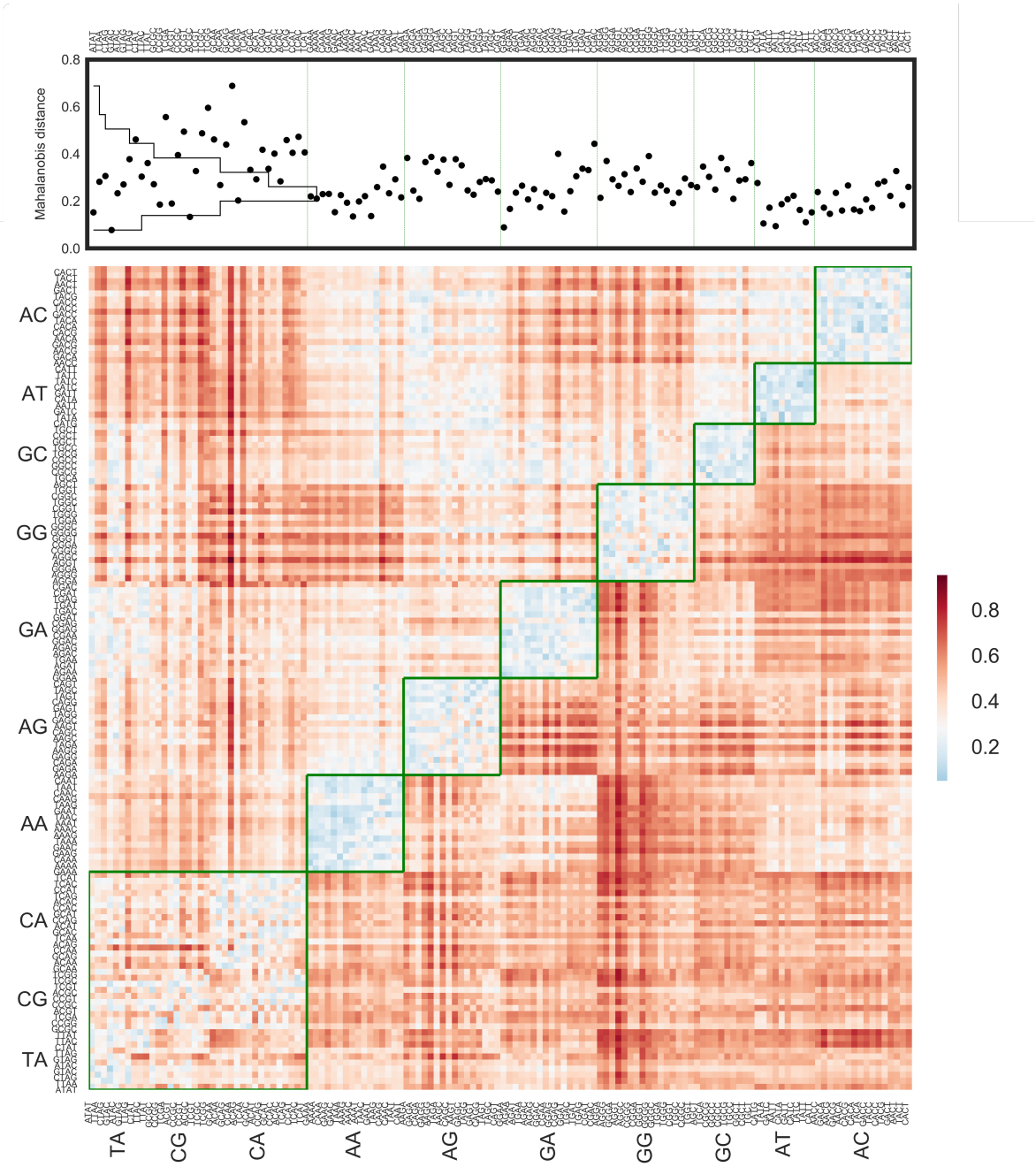


Fig. 5.7 In the heat map (bottom), the diagonal entries are Mahalanobis distance between the groundstate of dimers (in 136 independent tetramer contexts) in the X-ray and cgNA+ model data set. Whereas lower and upper off-diagonal entries are Mahalanobis distance between different dimers (in specific tetramer context) within the cgNA+ model and X-ray data set, respectively. The diagonal entries of the heat-map are again plotted in the scatter plot (top) along with the histogram in the same plot. Note that the Mahalanobis distance (defined in section 2.5.5) is computed in the transformed coordinates in the eight principal modes of cgNA+ shape covariance and using cgNA+ shape covariance matrix (in transformed coordinates) as the weight matrix. The equivalent plot using all 18 CURVES+ coordinates is shown in figure E.13.

In figure 5.7, we have plotted a heat map with the diagonal entries as Mahalanobis distance between the groundstate of dimers (in 136 independent tetramer contexts) in the X-ray and cgNA+ model data set and the lower and the upper off-diagonal entries as Mahalanobis distance between different dimers (in specific tetramer context) within cgNA+ model and X-ray data set, respectively. The diagonal entries of the heat-map are again plotted in the scatter plot (top) along with the histogram in the same plot. Note that in Mahalanobis distance computation, we have used cgNA+ shape covariance matrix (in transformed coordinates) as the weight matrix. Firstly, it can be observed in the heat-map that a given dimer step in various tetramer flanking contexts are closer to each other than the rest of the dimer steps (with some exceptions in YR steps). YR dimer steps in various tetramer contexts are farther from each other than other dimer steps, indicating a stronger influence of tetramer contexts on YR steps. Moreover, the pattern observed above and below the diagonal in the heat-map is similar. Similar conclusions were also drawn from the dendrograms in figure 5.5. Along the diagonal in the heat-map are Mahalanobis distance between the average shape of dimer in specific tetramer context in X-ray data set with the corresponding dimer in cgNA+ data set, which is approximately 0.2 for most cases (as can also be seen in the scatter plot above the heat-map). The Mahalanobis distance is reasonably small, differentiating a given dimer step from the other. However, for a given central dimer step, the change in average shape due to various flanking contexts is sometimes less than 0.2 implying that it is not always possible to resolve the influence on average shape due to variation in tetramer context. Alternatively, in the two data sets, there is no one-to-one mapping between various dimers (in tetramer context) with the least Mahalanobis distance. It could also be inferred from the Pearson correlation (which is ≈ 0.7) in transformed coordinates between the two data sets. Thus, there is no perfect agreement between the two data sets, but the data sets are very close given the scarcity of the X-ray data set and the two very different data sources.

5.3.3 Comparison of sequence-independent deformability of dsDNA in configurational space

It is not very clear how to best compare the dsDNA deformation in X-ray data which come from an ensemble of different dsDNA conformations in different protein-DNA crystals, in contrast, to dsDNA simulations in a solvent under particular physical conditions and is a result of thermal fluctuations. Furthermore, it is well known that the magnitude of dsDNA deformations in MD simulations is quite large (thus, also reflected in the cgNA+ model) as compared to X-ray data set due to unknown effective temperature for the X-ray crystal data and finding such T is also non-trivial [14, 97]. Ignoring the magnitude of deformations, in this section, we have compared the directions of deformations in the two data sets in a sequence-independent manner. We have computed the average covariance matrix in the configuration space (say configuration covariance) C^X and C^M for X-ray data and cgNA+ model data, respectively (see details in section 5.2.3.1 and plotted in figure E.1). To obtain uncorrelated directions, we have computed eigenvectors matrices, P^X and P^M corresponding to C^X and C^M configuration covariances and are shown in figure 5.2. Both the P^X and P^M matrices are quite sparse and similar in eyeball metrics. Once again, we observed the decoupling of inter coordinates with intra1 and intra2 coordinates as well as the ratio of positive to negative parity eigenvectors to be 10:8. Such

behavior of these eigenvectors originates from the inherent CW symmetry in the groundstate of dsDNA also reflected in configuration covariance (see section section 5.2.3.2 for more details). To best compare P^X and P^M, we have computed the cosine similarity index (defined in section 5.2.3.2) between the corresponding columns of P^X and P^M and found an excellent match with average cosine similarity 0.88 ± 0.07. It shows a remarkable similarity in the directions of dsDNA deformations in the two data sets. As expected, the corresponding eigenvalues, D^X and D^M have a significant difference in magnitude because of effective temperature. However, more importantly, eigenvectors with large eigenvalues in one data set align with eigenvectors with large eigenvalues in another data set, and the same is true for eigenvectors corresponding to smaller eigenvalues. This trend can be observed almost perfectly for P^X and P^M further highlighting that the direction as well as the trends in magnitude of deformations (ignoring the scaling due to the effective temperature) in those directions are similar in the two data sets.

Lastly, we observed that both data sets have similar eigenvectors corresponding to shape covariance and average configuration covariance. The average cosine similarity between the eigenvectors of C^X and C_s^X is 0.82 ± 0.11 and between C^M and C_s^M is 0.85 ± 0.1. Such a similarity between the eigenvectors of shape and configuration covariance is unexpected if we perform a similar analysis for a set of random pdfs with some mean and positive definite covariance. It is a remarkable similarity in both data sets with the observation that the largest sequence variation of groundstate (eigenvector of P_s with largest eigenvalue) is highly aligned with softest configuration dependent modes (eigenvector of P_s with largest eigenvalue). Similarly, smaller sequence variations of groundstate are aligned with the stiffest modes in the configuration space. It justifies the nearest-neighbor assumption in which all base-pair steps can not simultaneously achieve their individual local minima, and frustration energy arises in nearest-neighbors, and base-pair steps compromise from their local minima to attain a minimum energy configuration (which is not zero energy). For this minimum energy configuration, the consecutive base-pair steps have to negotiate the deformations in various directions such that it minimizes the sum of nearest-neighbor junction energies, and the findings suggest that the deformations are more in the directions of the soft modes of configuration space (as these cost the least). In other words, for various sequences/flanking contexts, the dimer adopts groundstate by compromising more in the soft modes of configuration space.

5.3.4 Comparison of Co-variance or sequence-dependent deformability of dsDNA

One of the methods to quantify the deformability of DNA is to compute the configurational volume or entropy of DNA as defined in section 5.2.3.4. Once again, it must be noted that the magnitude of S is larger for the cgNA+ model data set than for the X-ray data set, and therefore, we have compared the two data sets ignoring this scaling due to effective temperature.

Moreover, computing configuration covariance matrix for dimers in tetramer context with fewer representations in X-ray data set is questionable and thus, should be treated with caution. We observed that the range and product of eigenvalues for a dimer in tetramer context and dimer in average context (which have more than 2000 instances) are comparable, which provides confidence in this computation.

We have carried out this comparison computing S_{inter} (for inter coordinates, taking marginal

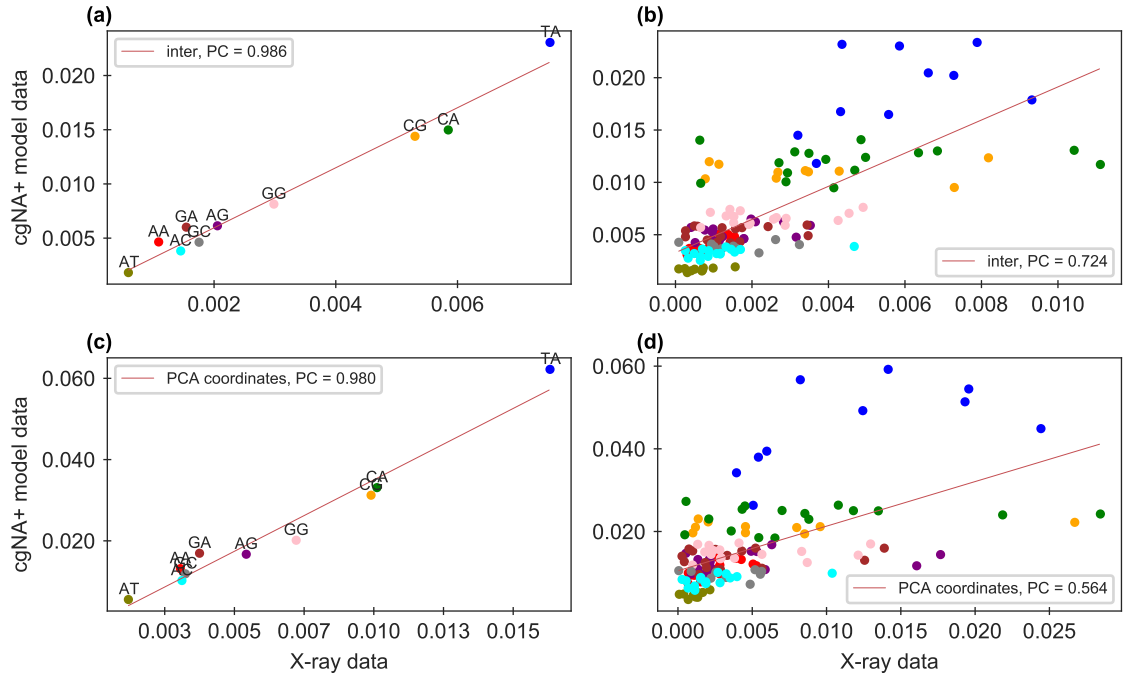


Fig. 5.8 Comparison of configurational volume for cgNA+ model covariance vs X-ray data set covariance a) in inter coordinates for independent dimer steps in average context, b) in inter coordinates for dimers in independent tetramer contexts, c) in PCA coordinates (in eight principal modes of cgNA+ model shape covariance $\in \mathbb{R}^{18}$) for independent dimer steps in average context, d) in PCA coordinates (in eight principal modes of cgNA+ model shape covariance $\in \mathbb{R}^{18}$) for dimers in independent tetramer contexts. The red line is best-fit line between the two data sets using linear regression.

over intra coordinates) and S_{PCA8} (for the transformed coordinates in the eight principal modes of cgNA+ shape covariance matrix). For S_{inter}, first we compared the two data sets at dimer level (i.e., average flanking context) in figure 5.8a and found an excellent correlation (PC = 0.98) between the two data sets. We also observed that the YR step is significantly more flexible in inter variables than other dimer steps. TA being the most flexible and AT most stiff (for inters parameters), which was also observed previously [53, 97, 141]. In figure 5.8b), we have compared S_{inter} for independent dimer in tetramer context and found good agreement (with PC = 0.72). In the plot, each dimer is color-coded differently with the label described in the figure 5.8b. S highly depends on the tetramer context for some dimers, while for others, the flanking context has a negligible effect. In general, dimer steps that are easily deformable than the stiff ones are more sensitive to the flanking tetramer context. For example, the most flexible YR steps (TA, CA, and CG) show a higher variation in configuration volume than RR and RY steps over the tetramer context in both data sets.

For S_{PCA8}, PC in the two data sets is 0.98 for dimers in average context, while for dimer in tetramer context, the PC in S_{PCA8} for the two data sets is 0.564 as shown in figure 5.8. Again, we observed that soft dimers are more affected by the change in the tetramer context. Note that we obtained similar results on carrying out comparison in 8 ± 2 principal components. We would

like to highlight that the comparison of S_{PCA8} or S_{inter} for dimer in tetramer context is limited by the scarcity of experimental data, which might be a reason for a poor correlation.

Lastly, in figure E.2, we have put an analogous comparison for the cgNA+ model data set and MD data set. For dimer in the average context, we found PC of 0.998 for S_{inter} and 0.998 for S_{PCA8}. For dimer in tetramer context, the observed PC for S_{inter} is 0.918 and for S_{PCA8} is 0.922.

5.4 Conclusions

In this chapter, we have shown that cgNA+ model prediction is in reasonable agreement with the available protein-DNA X-ray structure database for all dimers in various tetramer contexts. As dimer steps in the X-ray data set are often present in various flanking sequence contexts and as shown previously (in figure 4.2) that beyond tetramer flanking context could have a considerable effect on the average shape, we argued that the cgNA+ model is a better alternative over the MD simulations for such comparisons due to its accuracy and efficiency, which allowed computing the average shape of dimer in various tetramer contexts by averaging over all possible beyond flanking tetramer contexts. Moreover, we have compared both intra base-pair and inter base-pair step coordinates in the two data sets. Notably, the comparison of the intra base-pair coordinates and for all dimers in all tetramer contexts are complete novelties.

Firstly, we have shown that the sequence-independent (or sequence-average) average shape in the two data sets is extremely close. Moreover, defining "shape covariance" as the variation of groundstate in sequence space, we have shown that the direction of variation of groundstate in sequence space, i.e., the eigenvectors of shape covariance align closely with a cosine similarity of 0.81 ± 0.11. Then, we have demonstrated that, in the X-ray data, along with the central dimer step immediate tetramer flanking context is crucial in determining the average shape of dimer and in some cases, change in average shape can be larger due to change in tetramer contexts than change in central dimer step. Also, we found that certain flanking contexts systematically influence the average shape of dimer more than others. Furthermore, we have also emphasized the role of sequence by performing hierarchical clustering on the average shape of dimer in tetramer contexts, resulting in four main clusters based on pyrimidine-purine steps of the central base-pair step further sub-clusters based on the specific base-pair steps. Notably, the results are reasonably similar in the two data sets.

The next part of this chapter is dedicated to directly comparing internal coordinates in the two data sets. Firstly, we found that the Pearson correlation between the two data sets for some internal coordinates is excellent; whereas the correlation is poor for a few other coordinates, such as Rise or Stretch. We observed that coordinates with poor correlation are the ones that change the least in the sequence space, which leads to the hypothesis that for such coordinates, inherent noise in the X-ray data might have a dominating effect and, thus, it is not possible to resolve the sequence effect, in particular, the role of tetramer flanking context. We justified this hypothesis by demonstrating that correlation for the transformed coordinates in the principal modes of shape covariance is excellent, in contrast, to the correlation in the lower modes. Furthermore, we have also shown that Mahalanobis distance between the average shape for dimers in tetramer contexts in the two data sets is also close, however, it is not always possible to resolve dimers in

various tetramer contexts.

In the final part, we have defined sequence-average "configuration covariance" which tells dsDNA deformation in the configuration space and found as excellent alignment (cosine similarity of 0.88 ± 0.07) in the eigenvectors of the configuration covariance, i.e., the direction of deformation of the two data sets. However, the magnitude of the deformation is significantly less in the X-ray data set, which can be attributed to its lower effective temperature. More interestingly, the directions of deformation in configuration space align well to the direction of variation in the average shape in sequence space, implying that a given dimer attains minimum total nearest-neighbor junction energies, i.e., equilibrium shape by negotiating more in the direction of soft modes. Furthermore, for sequence-dependent analysis of dsDNA deformability, we have used the configurational volume as a metric to quantify deformability and found an excellent correlation of 0.98 at dimer level for both inter-variables and PCA coordinates (transformed on eight principal modes). Notably, at the tetramer level, the correlation in the data set is not equally good with a Pearson correlation of 0.72 for inter-coordinates and 0.56 for PCA coordinates but are still reasonable given the scarcity of the X-ray data for dimers in various tetramer flanking contexts. Moreover, we found that the sensitivity in deformability to flanking contexts is maximum in flexible dimer steps (YR).

Thus, we have demonstrated that the flanking tetramer contexts are crucial for dsDNA mechanics in X-ray data, and cgNA+ model predictions are in reasonable agreement for both the average shape and deformability; thus, it presents itself as an excellent tool for routine investigation of non-local sequence-dependent dsDNA mechanics for various applications.

CHAPTER 6

Extension of cgNA+ parameter sets for epigenetically modified DNA

The primary focus of this chapter is DNA base modifications, in particular, methylation or hydroxymethylation at the 5-position of cytosine in CpG steps. DNA methylation, regulated by the DNA methyltransferase enzyme, plays a pivotal role in several biological processes, such as X-chromosome inactivation and genomic imprinting, while aberrations in methylation patterns are often associated with diseases such as cancer [93, 161]. Around 70-80% of the CpG steps are methylated in mammalian cells [77] except for the CpG islands (dominantly present in the gene promoter regions). In general, methylation of CpG steps in promoter regions is anti-correlated with gene expression [32, 151]. It is believed that CpG methylation reduces the flexibility of DNA [38, 156, 162] and thus, reduces the ability of DNA to interact with transcription factors, modulates DNA accessibility, and makes them less prone to wrap around nucleosomes. However, recent works have shown contrasting findings with hypermethylation related to increased DNA flexibility [107, 115, 160, 187]. It highlights that the influence of base modifications on DNA mechanics is complex and depends on the extent and position of base modifications. Moreover, Rausch et al. [166] showed *in vivo* and *in vitro* experiments that cytosine methylation stabilizes double-stranded DNA (dsDNA) helix by increasing its melting temperature and resisting enzymatic activities toward dsDNA, therefore, suggesting its crucial role in regulating dsDNA access and genomic processes.

In comparison to DNA methylation, DNA hydroxymethylation has received much less attention because of its relatively low abundance in genomes and the lack of experimental techniques to resolve hydroxymethylated C from methylated C. Hydroxymethylated C is the result of oxidation of methylated cytosine in CpG steps catalyzed by ten-eleven translocation proteins [199]. DNA hydroxymethylation has been observed in the genomic regions of many organisms, in particular, prevalent in mammalian brain cells [125] and embryonic stem cells. The disturbed hydroxymethylation pattern of DNA cytosine may result in disordered cell function and, thus, in different types of cancers, e.g., myeloid cancers [91].

There are both experimental and theoretical evidence that these modifications bring about changes at the structural level [12, 28, 104, 164, 165] as well as modulate overall mechanics [28, 59, 81, 130, 181, 182]. However, these studies lack consensus on their findings. It is probably because alterations in dsDNA properties on base modifications are highly dependent on the modification level and the flanking sequence. Thus, a systematic investigation of how epigenetic base modifications influence dsDNA mechanics is required. A coarse-grained model such as cgNA+, which has been demonstrated to be indistinguishably accurate in predicting the mechanics (equilibrium shape and stiffness) of dsDNA, dsRNA, and DRH, will be highly

beneficial to obtain insights into the role of epigenetic base modifications in dsDNA mechanics and, thus, better understand its function in biology. This chapter discusses the extension of the cgNA+ model for epigenetically modified dsDNA, in particular, for sequences containing methylated and hydroxymethylated CpG steps.

Details of all the codes and data used in this chapter are provided appendix F.

6.1 cgNA+ for epigenetically modified dsDNA

This section describes the extension of cgNA+ to predict the Gaussian pdfs for sequences containing methylated and hydroxymethylated CpG steps. First, we describe the epigenetic base modifications in dsDNA and introduce a notation for modified cytosine. We then describe the training sequences used to train the cgNA+ parameters for modified base-pair steps. Lastly, we discuss the training of the cgNA+ parameter set $\mathcal{P}_{\text{Met/Hmet}}$ which in combination with existing $\mathcal{P}_{\text{DNA}}$ will allow predicting groundstate and stiffness matrix for any sequence containing methylated/hydroxymethylated CpG steps.

6.1.1 Epigenetic modifications in DNA bases

The most common epigenetic modifications in DNA bases are methylation and hydroxymethylation of cytosine at the 5–position. The chemical structures of these modified cytosines are shown in figure 1.1. Other base modifications, such as 5-formyl-C, 5-carboxyl-C, and N6-methyl-A, are comparatively rare in biology. Furthermore, most often, cytosine methylation or hydroxymethylation occurs at CpG dinucleotide steps, which can be di-substituted if both strands are symmetrically modified or hemi-substituted if only one of the strands is asymmetrically modified. This work focuses only on cytosine modification in CpG steps.

6.1.2 Alphabets for epigenetically modified cytosine

In this work, to describe the DNA sequence, we have used the standard alphabets A, T, C, and G for bases. However, the notation for hydroxymethylated or methylated cytosines is not standardized. This thesis uses the letter M for 5-methylated-cytosine, and N for Guanine when complementary cytosine is methylated. Similarly, letters H and K are used for 5-hydroxymethylated-cytosine and Guanine complementary to 5-hydroxymethylated-cytosine, respectively. For example, MN represents symmetrically methylated cytosine on both strands, CN denotes asymmetrically methylated cytosine on the Crick (complementary) strand, and MG denotes asymmetrically methylated cytosine on the Watson (reading) strand. Following this notation, M and H are complementary bases to N and K, respectively, and dimer steps such as MN, NM, HK, and KH are palindromes. Thus, for modified DNA, any sequence can be described using alphabets $X_i \in \{\text{A, T, C, G, M, N, H, K}\}$.

6.1.3 Training library

As discussed in chapter 2, cgNA+ is a coarse-grained model trained on MD simulations for a set of sequences called the training library. To train the cgNA+ parameter set that allows

cytosine base modifications in CpG steps, we have used an extensive library of 12 sequences provided in table B.2. The libraries are denoted as $\mathrm{Lb_{Met}}$ and $\mathrm{Lb_{Hmet}}$ containing sequences with methylated and hydroxymethylated CpG steps. The key features of these libraries include: (a) all the training sequences are palindromes, which allows quantifying the convergence of the MD simulations (refer to chapter 3 for details), (b) contains both di-substituted and hemi-substituted CpG steps in diverse sequence contexts, and (c) contains various combinations on modified CpG steps (for instance, MNMN or MGCN or CNMG in $\mathrm{Lb_{Met}}$). Thus, training sequences are designed optimally to have minimal sequences in the library with various combinations of modified base-pair steps in diverse sequence contexts. It should be noted that library design is a crucial step in the cgNA+ model. The training data must have sufficient diversity for any data-driven model to ensure accurate/reasonable predictions for unseen samples. In the case of the cgNA+ model, as discussed in section 4.1, lack of diversity in the training sequence may lead to a non-positive reconstruction of the stiffness matrix as observed in some dsDNA sequences with non-GC ends [149]. This problem was solved using a comprehensive library with diverse contexts for non-GC ends (refer section 4.1 and table B.3). A similar problem was also encountered during experimentation of modified parameter set; in particular, a parameter set trained on sequences without various combinations of modified CpG steps (i.e., without using sequence indices 9 to 12) predicts a non-positive definite stiffness matrix for sequences with adjacent repeats of modified CpG steps (for instance, GC···ATMNCNMG···GC). It again highlights how crucial library design is for the cgNA+ model.

Lastly, MD simulations for all sequences in $\mathrm{Lb_{Met}}$ and $\mathrm{Lb_{Hmet}}$ are performed using the same MD protocol used for dsDNA with additional force-field parameters for modified cytosine. Details of MD protocol and post-processing are provided in sections 3.2 and 3.4, respectively.

6.1.4 Training of cgNA+ parameter set to allow epigenetically modified cytosine

The cgNA+ model requires a parameter set containing dimer-dependent blocks for stiffness matrix and stress vector to predict a Gaussian pdf for a given sequence. In this work, the aim is to extend $\mathcal{P}_{\mathrm{DNA}}$ (defined in equation (4.2)) that allows prediction of Gaussian pdf for any sequence containing methylated or hydroxymethylated CpG steps. We started with the approximation that the parameter blocks for the unmodified base-pair steps remain the same.It implies that we need to estimate parameters only for {MN, NM, MG, CN, AM, TM, CM, GM, NA, NG, NT, NC} dimer steps out of which {MN, NM, MG, AM, TM, CM, GM} dimer steps are independent. Similarly, to allow hydroxymethylated CpG steps, parameters for {HK, KH, HG, AH, TH, CH, GH} dimer steps are required, while for the dependent dimers, the parameters can be obtained using the CW symmetry relation defined in equation (2.20). Thus, for any sequence containing modified CpG steps, the groundstate and stiffness matrix can be predicted using cgNA+ model with a combination of $\mathcal{P}_{\mathrm{DNA}}$ and $\mathcal{P}_{\mathrm{Met/Hmet}}$ given as

$$\mathcal{P}_{\mathrm{Met/Hmet}} = \{\sigma^{\mathrm{XY}}, \mathcal{K}^{\mathrm{XY}}\} \in \left[\mathbb{R}^{42}\right]^7 \times \left[\mathbb{R}^{42\times 42}\right]^7, \tag{6.1}$$

where XY $\in$ {MN, NM, MG, AM, TM, CM, GM} for $\mathcal{P}_{\mathrm{Met}}$ and XY $\in$ {HK, KH, HG, AH, TH, CH, GH} for $\mathcal{P}_{\mathrm{Hmet}}$.

Once again, it must be noted that $\mathcal{P}_{\text{Met/Hmet}}$ only allows modified CpG steps, not modified GpC steps. Taking the example of dsDNA methylation, NM and GM (or NC) steps are not allowed. It is contradictory and confusing, as $\mathcal{P}_{\text{Met}}$ already contains parameters for these dimer steps. This is because NM and GM (or NC) steps naturally arise in various combinations of methylated CpG steps which are allowed. For example, NM and GM steps are present in repeated combinations of methylated CpG steps such as MNMN/CNMN/CNMG and MGMN/MGMG, respectively. Therefore, to write precisely, XNMZ and XGMZ (or XNCZ) steps are not allowed where X $\in$ {A, T, G} and Z $\in$ {A, T, C} as these steps represent methylated GpC step. Moreover, in the current version of cgNA+, a sequence containing both hydroxymethylated and methylated CpG steps is only allowed as input when they are not present adjacent to each other.

Lastly, since we already have parameters for NM and GM, one can contemplate using these parameters to predict Gaussian pdf for a sequence containing methylated GpC steps. Such an exercise might lead to a non-positive definite stiffness matrix for that sequence. The only explanation justifying the non-positive definite reconstruction of such sequences is that the training library does not contain any such example cases. It again highlights the importance of training library design. However, it leads to a technical issue in $\mathcal{P}_{\text{Met/Hmet}}$ parameter sets. As once the cgNA+ parameter set is computed, the next step is to search the block elements in the null space to check whether dimer stiffness blocks in the parameter set are positive-definite (refer to sections 2.4 and 2.4.5) which ensure a positive-definite stiffness matrix for any given sequence. However, in the case of $\mathcal{P}_{\text{Met/Hmet}}$, we already know that it is not possible to find such block elements as sequences containing XNMZ and XGMZ (or XNCZ) steps where X $\in$ {A, T, G} and Z $\in$ {A, T, C} give non-positive definite stiffness matrix. The next best check to confirm positive-definite reconstruction for any sequence is to examine the reconstructed stiffness matrix for a large ensemble of sequences and hope that for any other sequences, not in this ensemble; the cgNA+ model will predict positive-definite Gaussian pdf. We checked the definiteness of the predicted stiffness matrix for (a) all 16mers in GC ends and (b) 10^8 random sequences of length varying from 16 to 300 bps containing at least one modified CpG steps. This test was performed individually for both $\mathcal{P}_{\text{Met/Hmet}}$, and positive-definite stiffness matrices were obtained for all sequences. Note that for dsDNA/dsRNA/DRH, before searching for block elements in null-space, we reconstruct all sequences of length 4-10 and check the definiteness of the stiffness matrix. In all the cases (RNA/DNA/DRH), even if stiffness matrices for all hexamers for a given parameter set are positive definite, we were always able to find positive-definite dimer stiffness parameter blocks. It further establishes trust in $\mathcal{P}_{\text{Met/Hmet}}$ that for any given sequence containing modified CpG steps, cgNA+ reconstruction will be positive-definite, however, it can not be guaranteed as in the case of $\mathcal{P}_{\text{DNA/RNA/DRH}}$.

6.2 cgNA+ reconstructions and associated modeling errors

This section is similar to the previous discussion in section 4.3 and investigates the performance of the cgNA+ model in predicting groundstate and stiffness for dsDNA sequences with epigenetic base modifications. In particular, we have tested the cgNA+ model on sequences that are not part of the training sequences for $\mathcal{P}_{\text{Met/Hmet}}$.

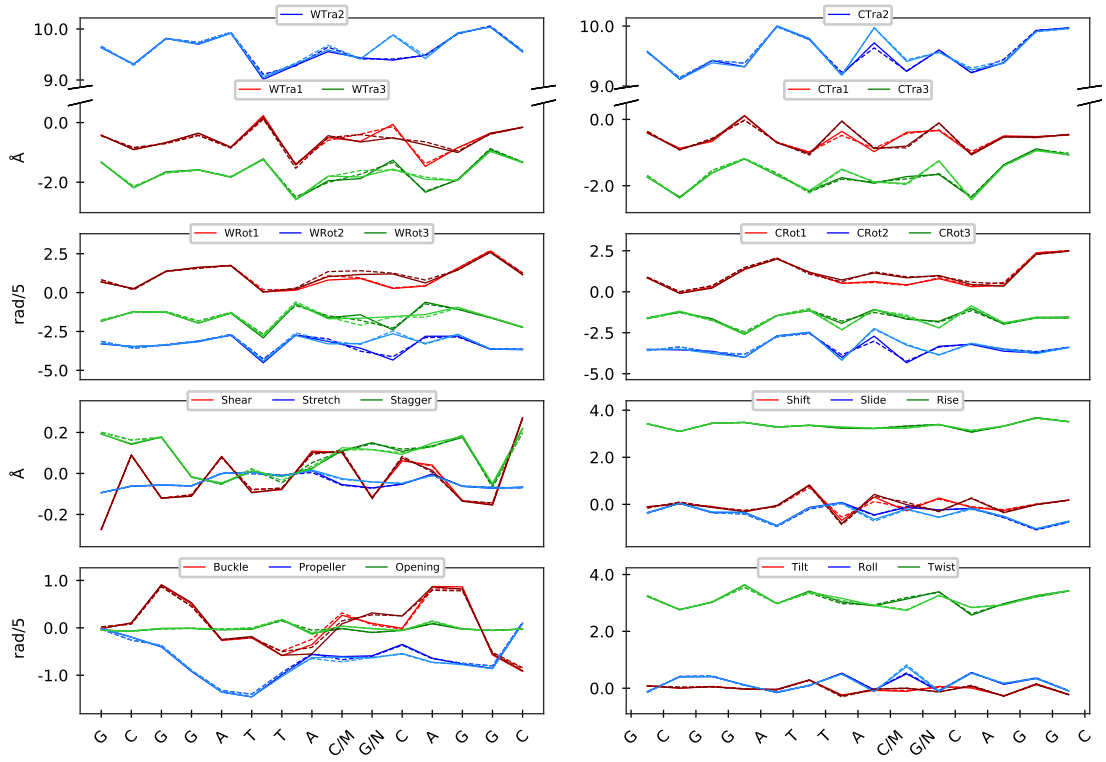
6.2.1 Test library

To assess the cgNA+ model, we have simulated several test sequences (listed in table B.2) using the same MD protocol as used for the training sequences. These test sequences are designed carefully to critically examine various aspects of the cgNA+ model's predictive capability. For instance, sequence index 18 in Lb_{Met} or Lb_{Hmet} is the symmetric modified (methylated or hydroxymethylated) version of sequence index 20 in Lb_{DNA} at the interior CpG step. This sequence allows to check how well cgNA+ captures the change in groundstate of this sequence "GCGGATTA**CG**CAGGC" upon symmetric modification of the CpG step (highlighted in bold). Furthermore, sequence index 24 in Lb_{DNA} is a typical CpG island, and to check the effect of CpG methylation/hydroxymethylation, we have simulated sequence indices 19 to 21 (in Lb_{Hmet} or Lb_{Met}) which are differently modified variants of the same CpG islands.

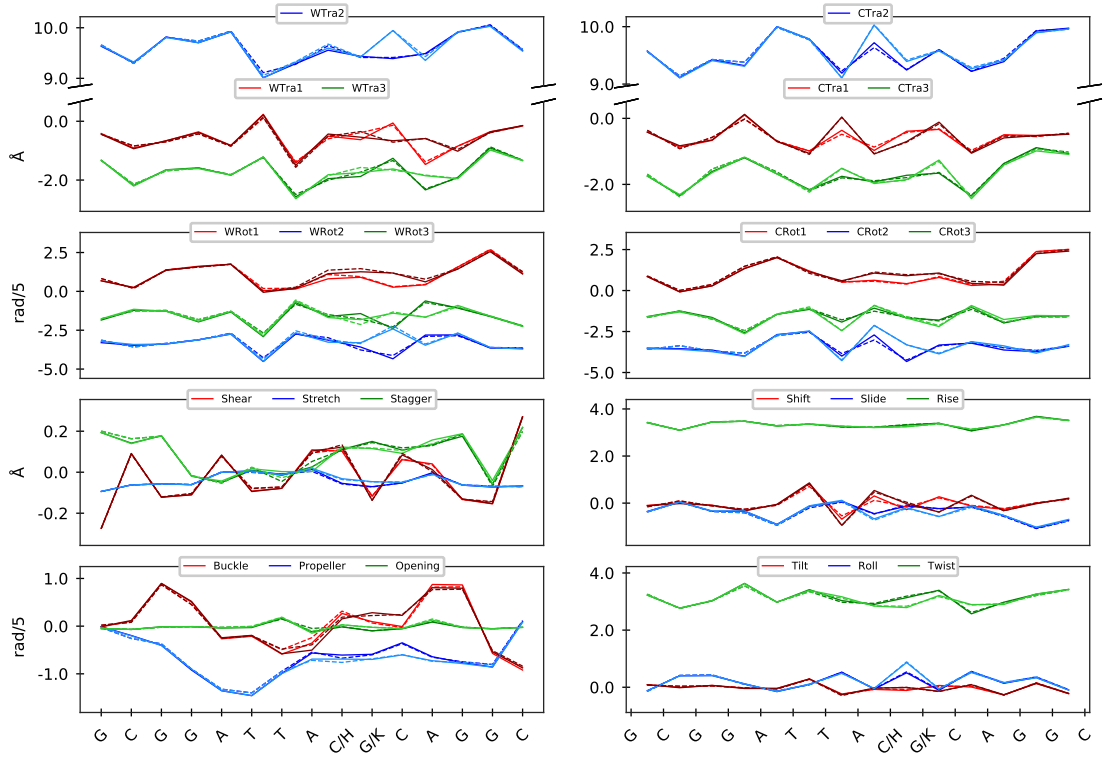
6.2.2 Reconstruction error in cgNA+ model

In this subsection, first, we have plotted the groundstate for a few selected sequences along with the observed MD estimates to visualize the model accuracy and highlighted that the cgNA+ model captures non-local changes (i.e., the change in groundstate is propagated to the neighboring base-pair steps) in groundstate due to base modifications in the sequence (notably base modification can be considered as a smaller change than point mutation). In figure 6.1(a), we have compared the groundstate of unmethylated and methylated versions of "GCGGATTA**CG**CAGGC" (symmetric methylation at the highlighted CpG step). First, note that the change in groundstate due to methylation of one CpG step is highly non-local. Furthermore, it can be observed in the plot that the average shape in MD simulations and groundstate predicted by the cgNA+ model are indistinguishable and, thus, the cgNA+ model accurately captures non-local changes in groundstate on methylation of CpG steps. This is further quantified as the reconstruction error in terms of the Mahalanobis distance and is equal to ≈ 0.0027 Å^2 or $(\text{rad}/5)^2$ for test sequences (refer table 6.1(a)). An analogous plot for hydroxymethylation of the highlighted CpG step in "GCGGATTA**CG**CAGGC" is provided in figure 6.1(b), and similar conclusions can be drawn about the accuracy of the model and the impact of CpG hydroxymethylation on groundstate.

Furthermore, in figure 6.2(a), we have compared the groundstate for monomethylated (asymmetric) and dimethylated (symmetric) typical CpG islands to highlight the difference in groundstate for two differently substituted CpG steps. It can be observed that asymmetric and symmetric methylation of CpG steps leads to a significantly different groundstate and the model accurately captures those changes. Similar results were also observed for asymmetric and symmetric hydroxymethylation of CpG steps in CpG islands (not shown in the thesis for brevity). Lastly, we would like to emphasize that internal coordinates for CpG steps are often multimodal (refer to figure 3.7) and highly sensitive to flanking sequence context. So, one can expect a larger reconstruction error in the prediction of groundstate for CpG steps compared to other dimer steps for dsDNA or, in general, for dsRNA (where internal coordinate distributions are close to Gaussian) due to much complicated and larger conformational space. Therefore, it again highlights how impressive the cgNA+ model is in predicting groundstate for such modified sequences containing CpG steps.

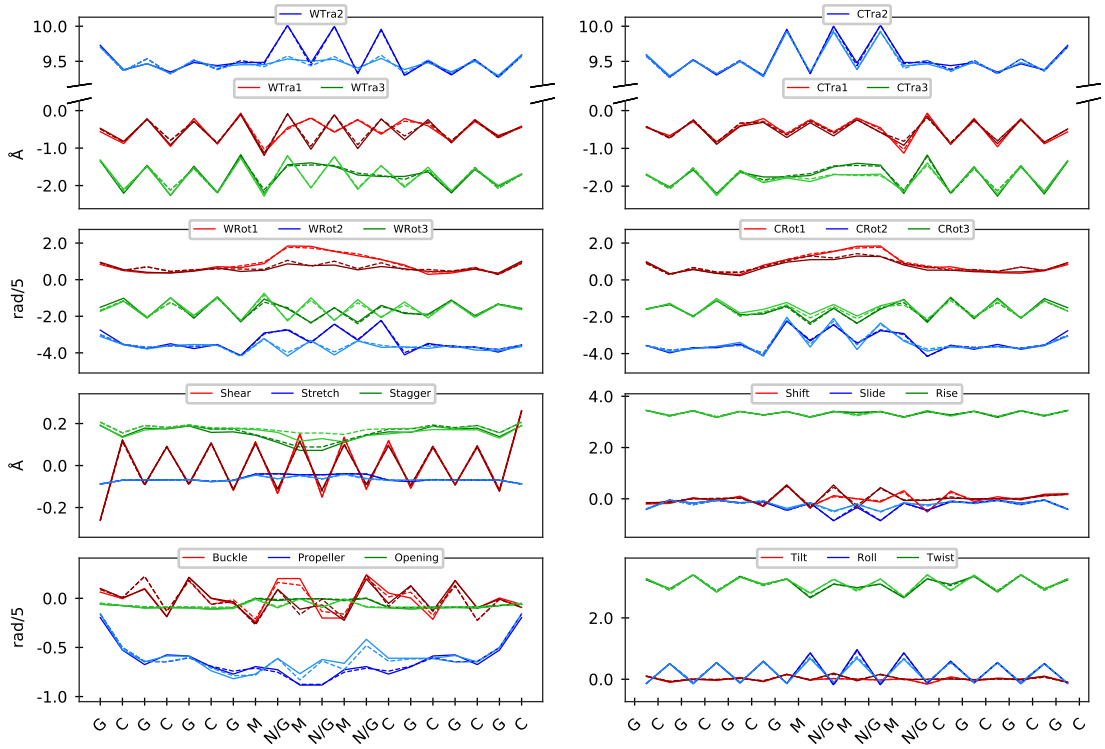


(a) Methylation of CpG step

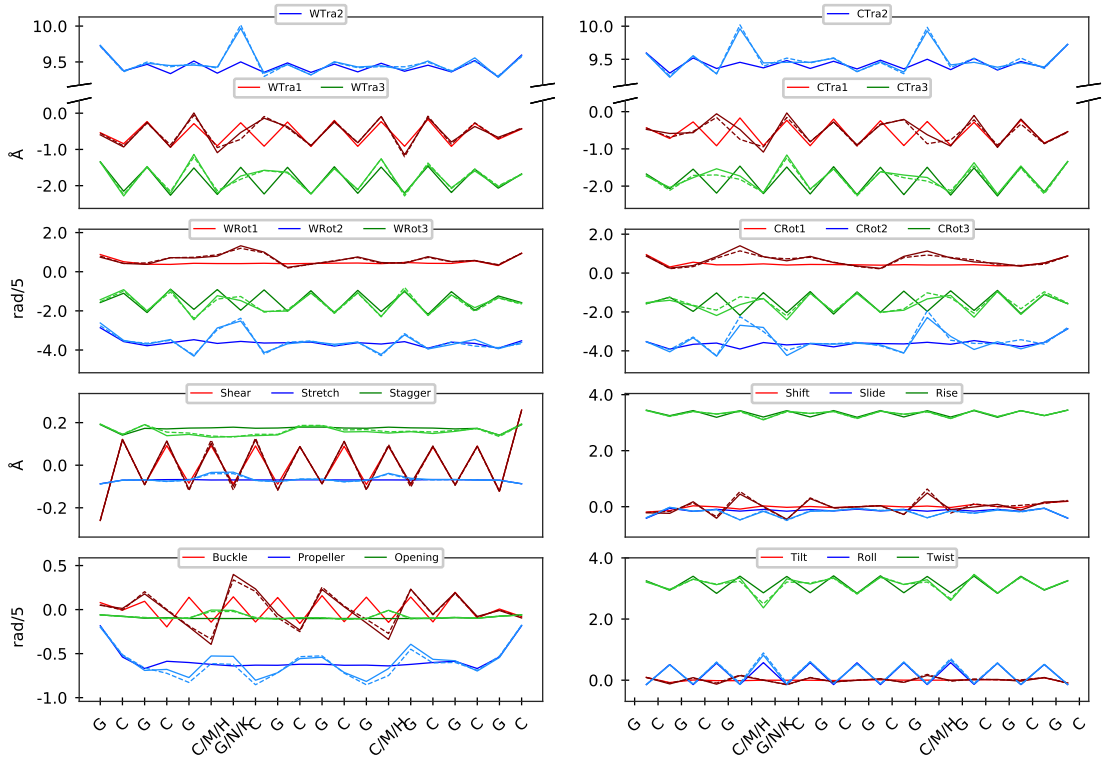


(b) Hydroxymethylation of CpG step

Fig. 6.1 Groundstate coordinates (elements of w) for (a) sequence index 20 in $\mathrm{Lb}_{\mathrm{DNA}}$ (red, blue, and green as shown in legend) and 18 in $\mathrm{Lb}_{\mathrm{Met}}$ (in dark red, dark blue, dark green) and (b) sequence index 20 in $\mathrm{Lb}_{\mathrm{DNA}}$ (red, blue, and green as shown in legend) and 18 in $\mathrm{Lb}_{\mathrm{Hmet}}$ (in dark red, dark blue, dark green). The figure highlights the cgNA+ model accuracy in predicting the non-local change in groundstate due to (a) methylation and (b) hydroxymethylation of CpG step. MD estimates are in solid lines while dashed lines are cgNA+ reconstructions.



(a) Symmetric vs asymmetric methylation of CpG islands



(b) Symmetric and asymmetric methylation and hydroxymethylation of CpG islands

Fig. 6.2 Groundstate coordinates (elements of w) for (a) sequence index 19 (red, blue, and green as shown in legend) and 20 in Lb_{Met} (in dark red, dark blue, dark green) where MD estimates are in solid lines while dashed lines are cgNA+ reconstructions, and (b) sequence index 20 in Lb_{DNA} (red, blue, and green as shown in legend) and 21 in Lb_{Met} (in solid lines) and Lb_{Hmet} (in dashed lines) in dark red, dark blue, dark green. The figure highlights (a) the cgNA+ model accuracy in predicting change in groundstate due to symmetric and asymmetric methylation of CpG islands and (b) impact of methylation and hydroxymethylation on groundstate of CpG islands.

In the second part, we have quantified the total reconstruction error in the cgNA+ model for sequences with epigenetic modifications. The reconstruction error $\mathcal{E}^{\text{res}}$ (refer to section 2.5.5) for the cgNA+ model is defined as the deviation of the predicted Gaussian pdf from the corresponding observed Gaussian pdf in MD simulations and is computed in terms of symmetric KL divergence and symmetric Mahalanobis distance. Note that $\mathcal{E}^{\text{res}}_{\text{KL}}$ (reconstruction error in terms of KL divergence) describes the total reconstruction error (both in groundstate and stiffness matrix) in the cgNA+ model, while the $\mathcal{E}^{\text{res}}_{\mathcal{M}}$ (reconstruction error in terms of Mahalanobis distance) highlights the difference in the predicted groundstate and MD average shape scaled by stiffness. In table 6.1(a), we have tabulated the reconstruction errors per degree of freedom, dof (which is $24N - 18$, i.e., the number of internal coordinates required to describe a given sequence of length N bp) in the training and test sequences for Lb_{Met} and Lb_{Hmet}. Firstly, the average reconstruction error in the training sequences in Lb_{Met} is 0.0023 and 0.0304 in terms of $\mathcal{E}^{\text{res}}_{\mathcal{M}}$ and $\mathcal{E}^{\text{res}}_{\text{KL}}$, respectively, which is roughly one order of magnitude smaller than the corresponding *scale*, 0.0211 and 0.3378, computed as average pair-wise difference of training sequences (see section 2.5.6) and is a quantification of variation over sequence. It highlights the precision of the cgNA+ model in accurately capturing the non-local sequence-dependent mechanics of dsDNA. Moreover, the analogous reconstruction errors, $\mathcal{E}^{\text{res}}_{\text{KL}}$ and $\mathcal{E}^{\text{res}}_{\mathcal{M}}$ for the test sequences in Lb_{Met} are 0.0025 and 0.0311 which are almost equal to the errors in the reconstruction of the training sequences. It comments on the generalizability of the cgNA+ model. Similar observations can be made for training and test sequences in Lb_{Hmet}. It should be noted that the reconstruction errors obtained here are comparable to the corresponding results for Lb_{DNA} (in table 4.1). However, the *scale* obtained here for Lb_{Hmet} and Lb_{Met} is noticeably smaller than *scale* for Lb_{DNA} because the training sequences in Lb_{Hmet} or Lb_{Met} are similar to each other than in Lb_{DNA}.

This total reconstruction error in the cgNA+ model results from several modeling assumptions in our model, as listed in section 2.3 and the error associated with each assumption can be quantified as described in section 2.5. We have discussed the contributions of various modeling assumptions to the reconstruction error in the following subsections.

6.2.3 Approximation error in the training data

The first modeling assumption in the cgNA+ model is that the MD time series is fully converged. The associated convergence error (referred to as palindromic error) is discussed in section 2.5.1, and details of this error quantification are provided in section 3.5 and table 3.4. For the training sequences in Lb_{Hmet} and Lb_{Met}, the average palindromic error in terms of KL divergence $\mathcal{E}^{\text{palin}}_{\text{KL, avg}}$ and Mahalanobis distance $\mathcal{E}^{\text{palin}}_{\mathcal{M}\text{, avg}}$ are of the order 10^{-4} and 10^{-3}, respectively, which are approximately two orders smaller than the corresponding *scales*.

Moreover, in section 3.6, we have discussed the distribution of internal coordinates in MD time-series and shown that the distributions for inter base-pair step and phosphate coordinates for dsDNA often deviate from Gaussian behavior and depend on the flanking sequence context. However, for modeling purposes, we have imposed Gaussianity to the underlying distribution for internal coordinates, leading to an inevitable modeling error. We have quantified this modeling error by computing the KL divergence between observed pdf and best-fit Gaussian pdf to the observed pdf and found that except for Wtra1 phosphate coordinate, $\mathcal{E}^{\text{Gauss}}_{\text{KL}}$ is less than *scale*.

	Lb_{Met}		Lb_{Hmet}	
Training sequences				
Index	$\mathcal{E}^{res}_{\mathcal{M}}$	$\mathcal{E}^{res}_{KL}$	$\mathcal{E}^{res}_{\mathcal{M}}$	$\mathcal{E}^{res}_{KL}$
1	0.0021	0.0329	0.0020	0.0322
2	0.0020	0.0266	0.0019	0.0268
3	0.0021	0.0261	0.0020	0.0245
4	0.0016	0.0211	0.0017	0.0357
5	0.0019	0.0260	0.0019	0.0256
6	0.0017	0.0290	0.0018	0.0291
7	0.0028	0.0329	0.0028	0.0322
8	0.0026	0.0281	0.0027	0.0304
9	0.0024	0.0342	0.0025	0.0352
10	0.0033	0.0402	0.0031	0.0387
11	0.0026	0.0329	0.0025	0.0307
12	0.0028	0.0350	0.0029	0.0360
Average	0.0023	0.0304	0.0023	0.0314
Test sequences				
Index	$\mathcal{E}^{res}_{\mathcal{M}}$	$\mathcal{E}^{res}_{KL}$	$\mathcal{E}^{res}_{\mathcal{M}}$	$\mathcal{E}^{res}_{KL}$
13	0.0025	0.0339	0.0026	0.0349
14	0.0033	0.0389	0.0032	0.0377
15	0.0022	0.0295	0.0024	0.0297
16	0.0023	0.0306	0.0022	0.0299
17	0.0025	0.0356	0.0023	0.0333
18	0.0027	0.0324	0.0027	0.0319
19	0.0019	0.0244	0.0018	0.0226
20	0.0025	0.0281	0.0028	0.0316
21	0.0024	0.0262	0.0023	0.0266
Average	0.0025	0.0311	0.0025	0.0309
scale	0.0211	0.3378	0.0214	0.3449

(a) Model reconstruction error

	Lb_{Met}	Lb_{Hmet}	Lb_{Met}		Lb_{Hmet}	
Index	$\mathcal{E}^{Trunc}_{KL}$	$\mathcal{E}^{Trunc}_{KL}$	$\mathcal{E}^{local}_{\mathcal{M}}$	$\mathcal{E}^{local}_{KL}$	$\mathcal{E}^{local}_{\mathcal{M}}$	$\mathcal{E}^{local}_{KL}$
1	0.0049	0.0051	0.0021	0.0288	0.0021	0.0280
2	0.0044	0.0047	0.0021	0.0223	0.0020	0.0223
3	0.0045	0.0046	0.0021	0.0221	0.0020	0.0204
4	0.0047	0.0049	0.0016	0.0169	0.0017	0.0314
5	0.0049	0.0050	0.0019	0.0218	0.0019	0.0212
6	0.0045	0.0050	0.0018	0.0250	0.0019	0.0247
7	0.0041	0.0045	0.0029	0.0290	0.0028	0.0280
8	0.0047	0.0051	0.0027	0.0240	0.0027	0.0261
9	0.0047	0.0048	0.0025	0.0302	0.0026	0.0311
10	0.0041	0.0043	0.0034	0.0362	0.0032	0.0345
11	0.0043	0.0044	0.0027	0.0290	0.0025	0.0267
12	0.0048	0.0047	0.0029	0.0307	0.0030	0.0316
Average	0.0046	0.0048	0.0024	0.0263	0.0024	0.0272
scale	0.3378	0.3449	0.0211	0.3378	0.0214	0.3449

(b) Truncation and locality error

Table 6.1 (a) Model reconstruction error in terms of KL divergence ($\mathcal{E}^{res}_{KL}$) and Mahalanobis distance ($\mathcal{E}^{res}_{\mathcal{M}}$) defined in section 2.5.5, and (b) truncation error due to nearest-neighbor interactions assumption ($\mathcal{E}^{Trunc}_{KL}$) and sequence locality error ($\mathcal{E}^{local}_{KL}$ and $\mathcal{E}^{local}_{\mathcal{M}}$). The list of sequences is provided in the table B.2. The first 12 sequences are training sequences, while the rest are test sequences in Lb_{Met} or Lb_{Hmet}. The *scale* (quantifies variation over sequence) is obtained by computing the average pair-wise difference between all training sequences.

It must be noted that the reconstruction error is defined as the deviation of cgNA+ predicted Gaussian pdf with the stationary observed Gaussian pdf in MD simulations, i.e., observed MD Gaussian pdf is the ground truth for the cgNA+ model. Therefore, the palindromic and Gaussian approximation errors do not contribute to the aforementioned reconstruction error.

With these two assumptions on MD time-series, we obtain Gaussian pdf for each training sequence, which are used to compute the dimer-dependent parameter set based on two assumptions: a) the nearest-neighbor interactions assumption, i.e., the total energy of any given oligomer is the sum of local junction energies, and b) the local junction energy parameters depend only on the sequence of corresponding junction dimer. We have approximated the error associated with these two assumptions in the following subsections.

6.2.4 Contribution of nearest-neighbor interactions assumption in cgNA+ reconstruction error

As discussed previously in section 4.3.4, the nearest-neighbor interactions assumption is a modeling choice inspired by the observations in the MD time series. One can refer back to figure 4.3 where we have plotted the observed stiffness matrix in the MD time series along with the stencils corresponding to the nearest-neighbor interactions assumption. Note that similar conclusions can be made from the corresponding plots for modified sequences (not shown for brevity). To quantify the error associated with this approximation, we have first computed the banded stiffness matrix corresponding to the nearest-neighbor interactions approximation using the maximum entropy fit algorithm [60] to the observed stiffness matrix. Then this approximation error (referred to as truncation error) can be computed as the symmetric KL divergence between the observed stiffness and the corresponding banded stiffness as defined in section 2.5.5 and denoted as $\mathcal{E}_{\mathrm{KL}}^{\mathrm{Trunc}}$. Notably, the corresponding Mahalanobis contribution will be zero, since there is no change in the average shape of the oligomer while computing banded stiffness. In table 6.1(b), we have listed the truncation errors, $\mathcal{E}_{\mathrm{KL}}^{\mathrm{Trunc}}$ for the training sequences (for brevity, we have not provided results for the test sequences) in $\mathrm{Lb}_{\mathrm{Met}}$ and $\mathrm{Lb}_{\mathrm{Hmet}}$. It can be observed that for all training sequences in $\mathrm{Lb}_{\mathrm{Met}}$ and $\mathrm{Lb}_{\mathrm{Hmet}}$, $\mathcal{E}_{\mathrm{KL}}^{\mathrm{Trunc}}$ is comparable with average values of 0.0046 and 0.0048 per dof, respectively, which is approximately two orders smaller than the corresponding *scale*.

6.2.5 Contribution of sequence locality assumption in cgNA+ reconstruction error

The last assumption in the cgNA+ model is that the local junction energies, which sum up to make the total oligomer level energies, depend on the local dimer sequence. The error associated with this assumption (described in section 2.5.4) is tabulated in table 6.1(b) with average $\mathcal{E}_{\mathcal{M}}^{\mathrm{local}}$ and $\mathcal{E}_{\mathrm{KL}}^{\mathrm{local}}$ equal to 0.0024 and 0.0263, and 0.0024 and 0.0272 for training sequences in $\mathrm{Lb}_{\mathrm{Met}}$ and $\mathrm{Lb}_{\mathrm{Hmet}}$, respectively. The error associated with the locality in sequence dependence is one order smaller than the corresponding *scale* for $\mathrm{Lb}_{\mathrm{Met}}$ and $\mathrm{Lb}_{\mathrm{Hmet}}$.

When comparing the two sources of errors ($\mathcal{E}^{\mathrm{Trunc}}$ and $\mathcal{E}^{\mathrm{local}}$) in the total reconstruction error in the cgNA+ model ($\mathcal{E}^{\mathrm{res}}$), the sequence locality assumption for junction energy parameters dominates. For instance, for training sequences in $\mathrm{Lb}_{\mathrm{Met}}$, the average reconstruction error

in terms of KL divergence, $\mathcal{E}^{\text{res}}_{\text{KL, avg}}$ is 0.0304, out of which the contribution from the nearest-neighbor interactions assumption in interaction energies is 0.0046 while from the locality assumption in the sequence dependence of junction energy parameters is 0.0263. It implies that the nearest-neighbor interactions assumption in the interaction energy is reasonable and contributes negligible to the modeling error, whereas the primary source of modeling error is the locality assumption in sequence dependence of junction energy parameters. Once again, this error, $\mathcal{E}^{\text{local}}$ is only a fraction of the *scale* set by computing the pair-wise difference between the training sequences in the respective libraries. Anyhow it highlights the non-local sequence dependence in the local junction energy. It should be noted that even though the stiffness matrix in the cgNA+ model has dimer/trimer local sequence dependence, the groundstate has a highly non-local sequence dependence due to the inversion of the stiffness matrix and the corresponding frustration energy associated with it.

6.3 Effect of cytosine substitution on dsDNA mechanics

In the previous section, we have demonstrated that the cgNA+ model is highly accurate in predicting the groundstate and stiffness matrix for any modified dsDNA sequence. Moreover, the prediction is extremely fast, making possible the prediction of groundstate and stiffness matrix for millions of sequences and, thus, statistical estimation of various dsDNA properties. Such a computation is impossible to perform using traditional computational or experimental techniques. This section has rigorously investigated various such observables for modified dsDNA for a large sequence space and the impact of CpG modifications in dsDNA. Notably, most of the prior studies [12, 28, 59, 81, 104, 130, 156, 164, 165, 181, 182] are done for a minimal number of sequences, which questions the generalizability of those results, primarily when it is known that the properties of dsNAs are highly sequence-dependent (often non-local dependence) [9, 22, 50, 102, 147].

6.3.1 Effect of cytosine substitution on the groundstate of dsDNA

In figure 6.3, we have compared base-pair step coordinates of dimers containing unmodified, methylated, and hydroxymethylated bases by plotting the MD observations (as •) in the training sequences along with corresponding cgNA+ model predictions (as ×). Firstly, for all dimers and various internal coordinates, • and × are indistinguishable, highlighting the accuracy of the cgNA+ model. Secondly, in general, both hydroxymethylation and methylation have a similar effect on the average shape of dimers, and their magnitude depend on the base-pair step and internal coordinate. The observations in figure 6.3 can be summarized as:

- Intra coordinates: The intra-coordinates of CG base-pair have a negligible effect due to methylation/hydroxymethylation substitution. The results are not shown for brevity.
- Inter coordinates: cytosine modification decreases Twist for all dimer steps, whereas increases Roll for modified CpG steps and decreases for other steps. Rise and Slide reduce, in general, upon cytosine modification. Lastly, Tilt and Shift either decrease or increase for a set of independent dimers following the CW symmetry conditions, while are zero for palindromes (CG and GC).

- Phosphate coordinates: Only the Crick phosphate coordinates are shown as the Watson phosphate coordinates are linearly dependent (refer equation (2.20)). In general, on cytosine modification, rotational coordinates increase, whereas translational coordinates decrease except CTra2. Interestingly, phosphate coordinates of base-pair steps adjacent to CpG step are more affected by CpG modification.

Furthermore, in figure 6.4, we have plotted the analogous plot to figure 6.3 but only for CpG step in various tetramer contexts to highlight that the flanking sequence context influences the effect of cytosine modification. It can be observed that, in general, a) variations in internal coordinates in different flanking contexts are larger than due to epigenetic modifications, and b) the magnitude of change in a given coordinate due to CpG step modification is highly sensitive to flanking sequence context. Some of these results have been observed in earlier works [12, 28] for inter-coordinates, such as a decrease in Twist and an increase in Roll upon cytosine modification or tetramer context induces larger changes in dimer shape than epigenetic modifications.

6.3.2 Role of flanking sequence context in epigenetic base modifications

In figures 6.1 to 6.3, we have shown that epigenetic modifications in CpG step lead to a significant change in groundstate of a given sequence. Moreover, this change in groundstate is also influenced by the flanking context of the modified CpG step (figure 6.4). It raises a natural question: which flanking sequence context to CpG step lead to a minimum or maximum change in groundstate of a given sequence upon epigenetic modification of that CpG step.

To address the question formally, we have considered all sequences of length 10 bps with central CpG step, i.e., S_i = GCGTCG$\mathrm{X_4X_3X_2X_1}$**CG**$\mathrm{Y_1Y_2Y_3Y_4}$GTCGGC embedded in random but fixed flanking context on both sides and X_j and $\mathrm{Y}_j \in$ {A, T, C, G} $\forall j \in \{1, 2, 3, 4\}$. It leads to 4^8 (60,000) sequences of 22 bps length. It can be expected that modifying the highlighted central CpG step will change the groundstate of a given sequence. Now, the question is for which X_j and Y_j, the central CpG step modification will lead to a minimum and maximum change in the groundstate where the change is defined as the symmetric Mahalanobis distance (refer equation (C.12)) in groundstate before and after the central CpG step modification.

Firstly, we observed that CpG step modification results in significant changes in the groundstate and the change is highly sensitive to the flanking contexts. For instance, the Mahalanobis distance between the groundstate before and after central CpG methylation and hydroxymethylation (symmetric) ranges from 1.57 to 3.36 and 1.69 to 3.63 Å^2 or rad/5^2, respectively. In figure 6.5, we have presented the sequence logos (described in section 1.2.1) to highlight which flanking contexts (X_j and Y_j) lead to a minimum and maximum change in groundstate of a given sequence upon epigenetic modification of the central CpG step. We have plotted sequence logos (detail in section 1.2.1) for outlier sequences defined as 0.5% sequences with the least and most change in groundstate on the central CpG step modification.

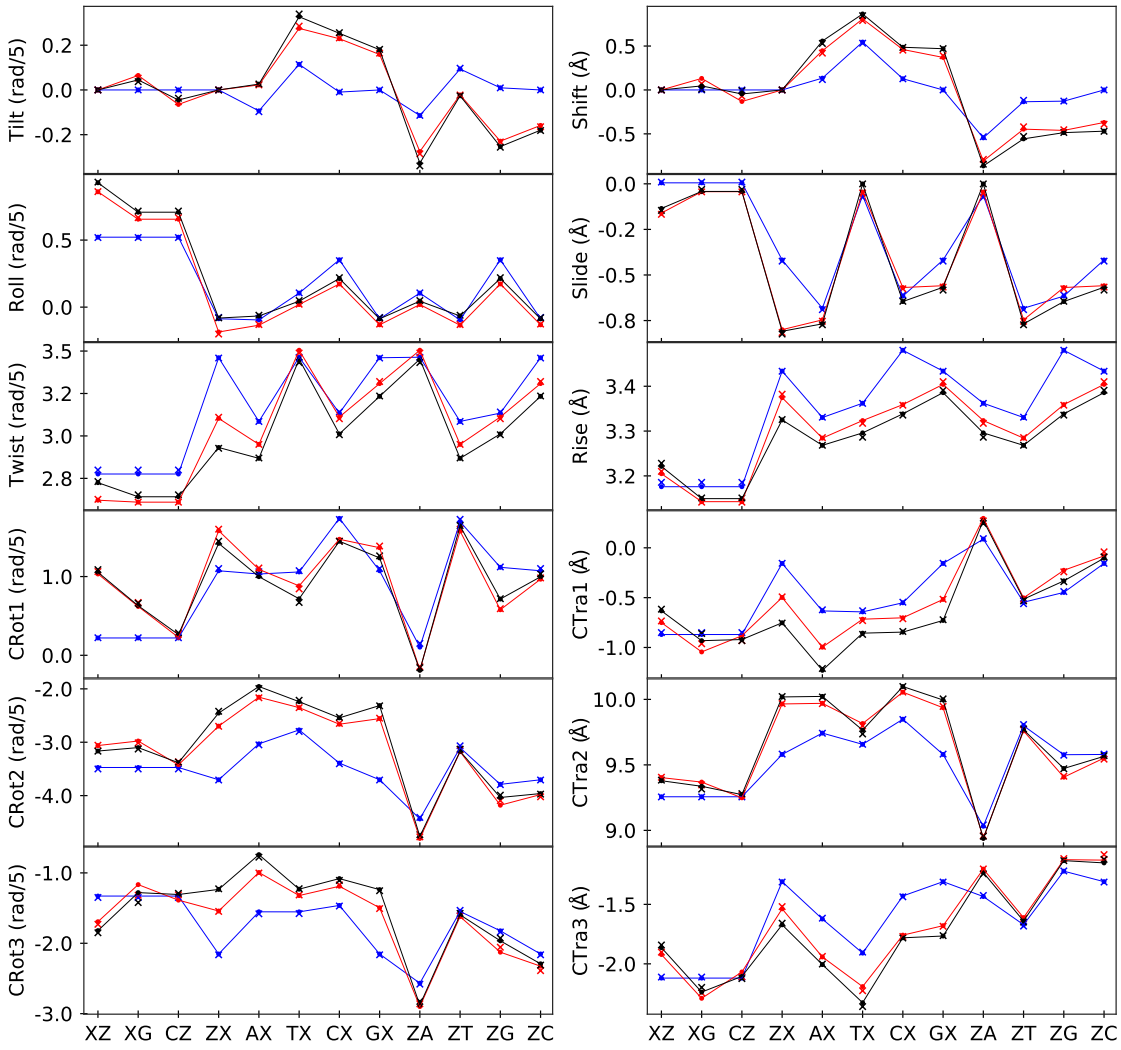


Fig. 6.3 Comparison of base-pair step coordinates for dsDNA where unmodified steps (X = C and Z = G) are in Blue, methylated steps (X = M and Z = N) are in Red, and hydroxymethylated steps (X = H and Z = K) are in Black. For each base-pair step, average coordinates observed in MD simulations and corresponding cgNA+ predictions are plotted in • and ×, respectively. For better visualization, a line plot is plotted along •.

In figure 6.5(a), we have plotted the findings when the central CpG step was substituted by MN. It can be seen that sequences with the least change in groundstate have X_1,X_2 = A, G and Y_1,Y_2 = T, C alphabets to the upstream (left) and the downstream (right) of CpG step, respectively, with information content ≈ 2. If we read the sequence from both strands in $5'-$ to $3'-$ direction, there is a GA before C/M, implying a strong correlation between the least change in groundstate of the modified sequence and the presence of purines upstream to C. Lastly, there is no information content for X_3, X_4, Y_3, and Y_4, indicating that beyond hexamer sequence does not influence the groundstate on epigenetic modifications. Moreover, in the bottom plot, we have presented the information content in the sequences that are most influenced by the methylation of the central CpG step. The findings indicate that the maximum change in groundstate is correlated with the presence of C or G base-pairs adjacent to the central CpG steps and A or T base-pairs at hexamer context. Figure 6.5(b) is an analogous plot for the asymmetric methylation

(MG) to the central CpG step. The sequences with the least change in groundstate when CpG step is substituted by either MG or MN are almost the same. The similarities in the sequence logos include X_1 = A, X_2 = A or G, and Y_1 = T, whereas differences include no information at Y_2 (possibly because as Y_2 is away from M) and X_3 is more likely to be C or T (pyrimidine). In contrast, the sequences with maximum changes in groundstate due to asymmetric methylation are slightly different from those with symmetric methylation. The sequence logos show a strong preference for T and G at X_2 and X_1, respectively, and A at Y_1, which can be explained by the fact that in asymmetric methylation, Y_1 position is, in fact, two base-pairs away from methylated C and can be considered as Y_2 which agrees with the sequence logos for symmetric methylation. Lastly, the corresponding plots for CpG step hydroxymethylation are almost identical as shown in figure 6.5(c) and (d).

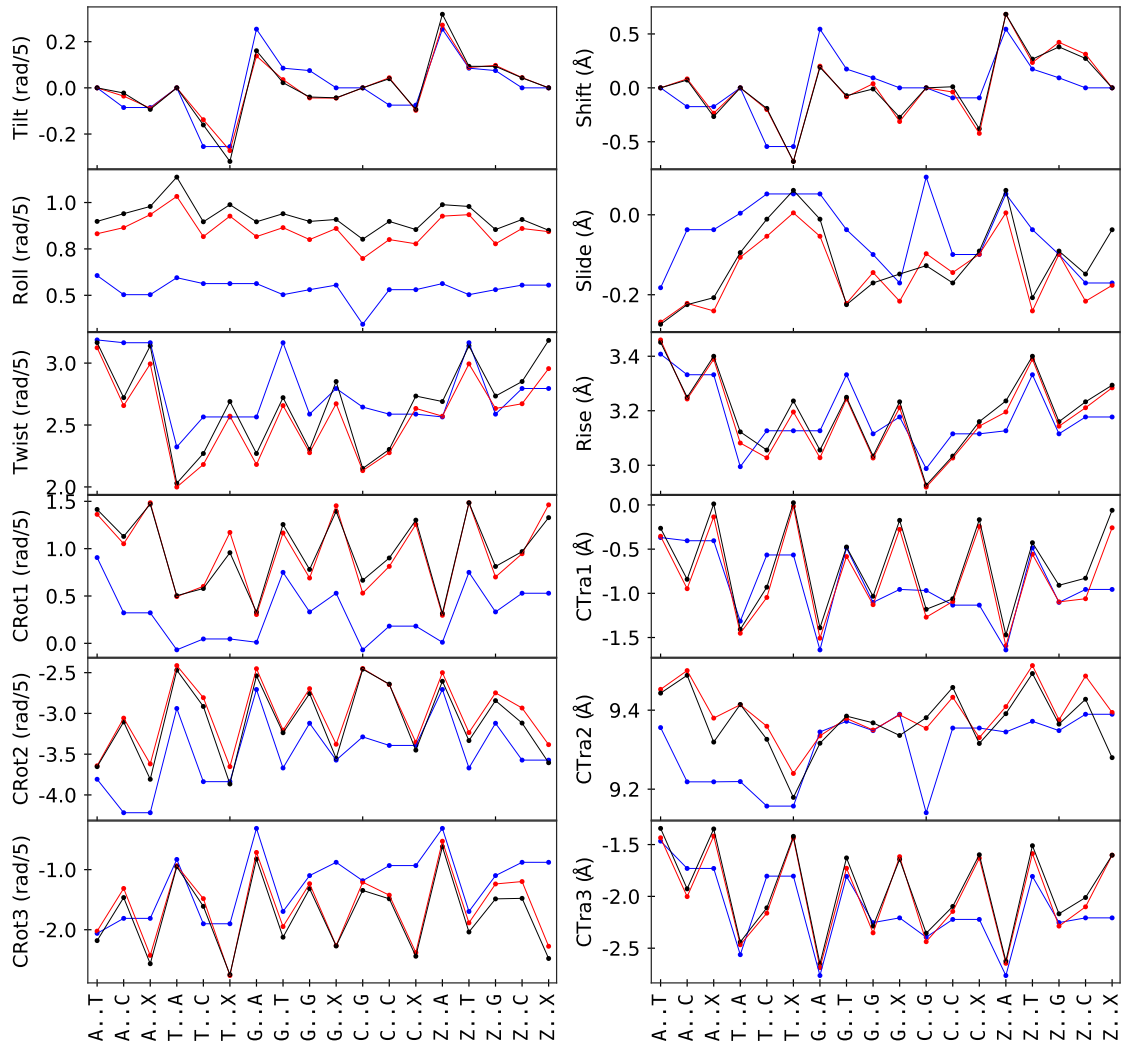
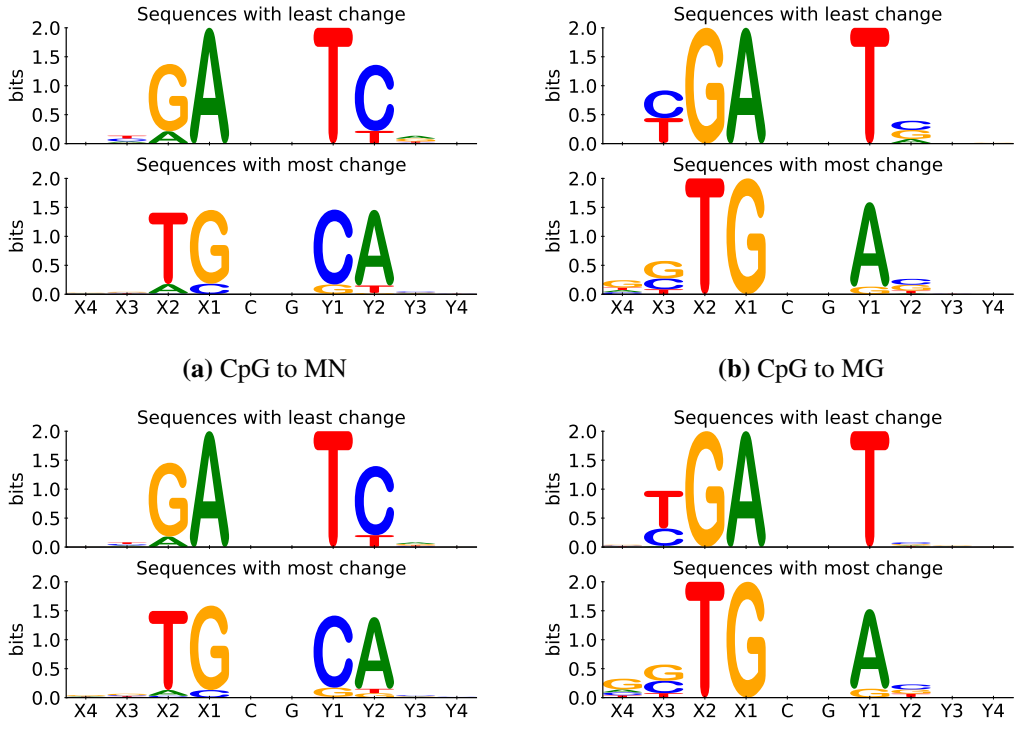


Fig. 6.4 Comparison of CpG step coordinates in various flanking contexts where coordinates for unmodified, methylated, and hydroxymethylated CpG steps are in blue, red, and black, respectively. X-axis are the flanking contexts where X = C and Z = G for unmodified CpG steps, X = M and Z = N for methylated CpG steps, X = H and Z = K for hydroxymethylated CpG steps. For better visualization, a line plot is plotted along •.

In summary, sequences with flanking A or T before modified cytosine (in CpG step) are relatively inert to epigenetic modifications, whereas sequences with C or G before modified cytosine are highly sensitive. In the plots, we have shown statistics obtained from extreme 0.5% sequences and using symmetric Mahalanobis distance as a metric; however, we would like to emphasize that the results are not sensitive to these choices. Any other choices for metric (such as L^1-norm of the difference in groundstate) and outlier cut-off lead to similar conclusions. Moreover, in figure 6.6, we have plotted the groundstate for two sequences which only differ by immediate flanking context (underlined) to CpG step, GCGTCGGAA**CG**TTTTGTCGGC and GCGTCGGTG**CG**CTTTGTCGGC to visualize the change in groundstate on the symmetric methylation of the central CpG step (in bold). The change in groundstate for the latter sequence on CpG methylation is much larger and non-local compared to the former, particularly for the phosphate coordinates.



(a) CpG to MN

(b) CpG to MG

(c) CpG to HK

(d) CpG to HG

Fig. 6.5 Sequence logos to highlight flanking contexts that least and most influence the change in groundstate upon epigenetic modification of central CpG step. Statistics are obtained from all decamers with central CpG steps embedded in a 22mer, i.e., GCGTCG$X_4X_3X_2X_1$**CG**$Y_1Y_2Y_3Y_4$GTCGGC and information content in X_j and Y_j for most (top 0.5%) and least change (bottom 0.5%) in the groundstate are plotted on the ordinate.

We want to highlight that the probability of occurrence of C or G base-pairs adjacent to CpG step is much more likely in CpG islands than in non-CpG islands. Traditionally, increased dsDNA stiffness due to epigenetic modifications is considered the controlling factor to act as a gene silencer [38, 156, 162]. It is believed that methylation of CpG steps reduces dsDNA flexibility and therefore, reduces its ability to interact with transcription factors, modulates dsDNA

accessibility, and makes them less prone to wrap around nucleosomes. Here, we have shown that the change in the equilibrium shape, i.e., groundstate is also a contributing factor to the differential behavior of modified dsDNA, in particular, for CpG islands.

6.4 Impact of epigenetic base modifications on groove widths

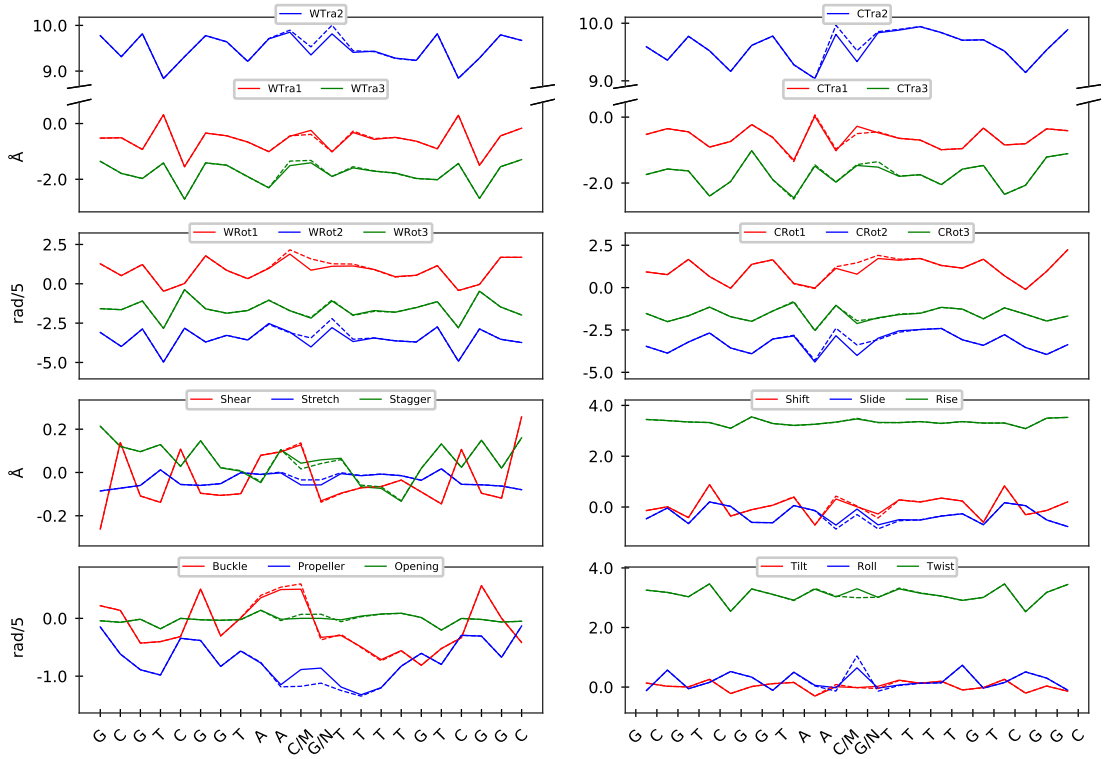
As discussed earlier in section 2.4.3, dsDNA displays a wide range of groove widths as a function of sequence; for example, minor groove widths range from 3 to 9 Å. The groove widths were computed for all decamers (one million) embedded in a flanking sequence of length six bps on both sides and taking central Watson phosphate (index 6 in decamer) as the reference phosphate. Notably, we concluded from the sequence logos that the sequence alphabets at positions 2-6 in decamers influence the minor groove, and 4-10 influence the major groove.

In limited prior studies exploring the influence of CpG modification on groove widths, the observations are inconsistent. For instance, the general belief is that CpG methylation narrows and widens the minor and major grooves, respectively [45, 106]. However, a thorough analysis of available X-ray crystallographic and NMR data [164] has shown that the methylation of CpG step may reduce or widen the minor groove depending on the sequence and location of the modified CpG step.

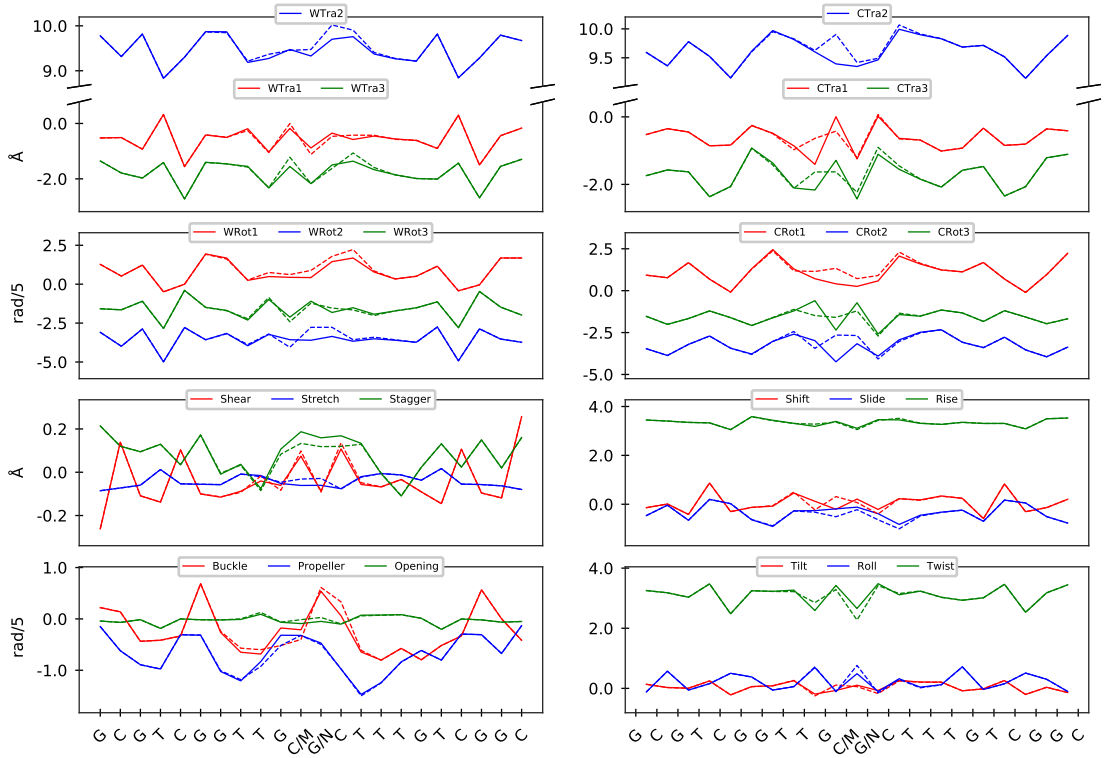
In this section, we have investigated how epigenetic modifications of the CpG step influence groove widths. Firstly, in figure 6.7(a), we have plotted a schematic diagram of the CG base-pair showing that the methyl or hydroxymethyl (shown as X) is present in the major groove, thus, explicitly change the major groove chemical environment, while the minor groove remains the same. For a systematic analysis of how this change affects the groove widths, we have performed two studies to explore the effect on groove widths due to (a) location of CpG modification, and (b) extent of CpG modification in the sequence. Note that we have used the same protocol described in section 2.4.3 for both studies.

For the first study, we have considered sequences of length 22 bps, i.e., GCTGTG$X_1\mathbf{X}_2\mathbf{X}_3$-$\mathbf{X}_4\mathbf{X}_5\mathbf{X}_6X_7X_8X_9X_{10}$CATGGC and varied the position of CpG step in $\mathbf{X}_2\mathbf{X}_3\mathbf{X}_4\mathbf{X}_5\mathbf{X}_6$ and computed the difference in groove widths before and after the modification of that CpG step. The results are plotted in figure 6.7(b), showing that the minor grooves generally widen on CpG step modification and depend on the position of the CpG step. Moreover, the widening of the minor groove on CpG modification is more for symmetric modifications than for asymmetric ones, and hydroxymethylation of CpG step widens the minor groove more than methylation. Lastly, the error bars in the plot are for various possible sequences by changing X_is.

Furthermore, to investigate the influence on groove widths due to the extent of CpG modifications, we have considered a sequence GCTGTG**CGCGCGCGCG**CATGGC of length 22 bp such that the central decamer is $(CG)_5$. Then, we have iteratively modified (symmetric/asymmetric methylation/hydroxymethylation) this sequence by replacing one, two, three, four, and all five CpG steps and computed the groove widths, and presented the findings in figure 6.7. We have plotted minor groove widths with the percentage of CpG modification, which shows a positive correlation between the minor groove widths and % CpG modification. The minor groove width for a fully symmetrically methylated sequence is approximately one Å larger than the corresponding unmodified sequence.

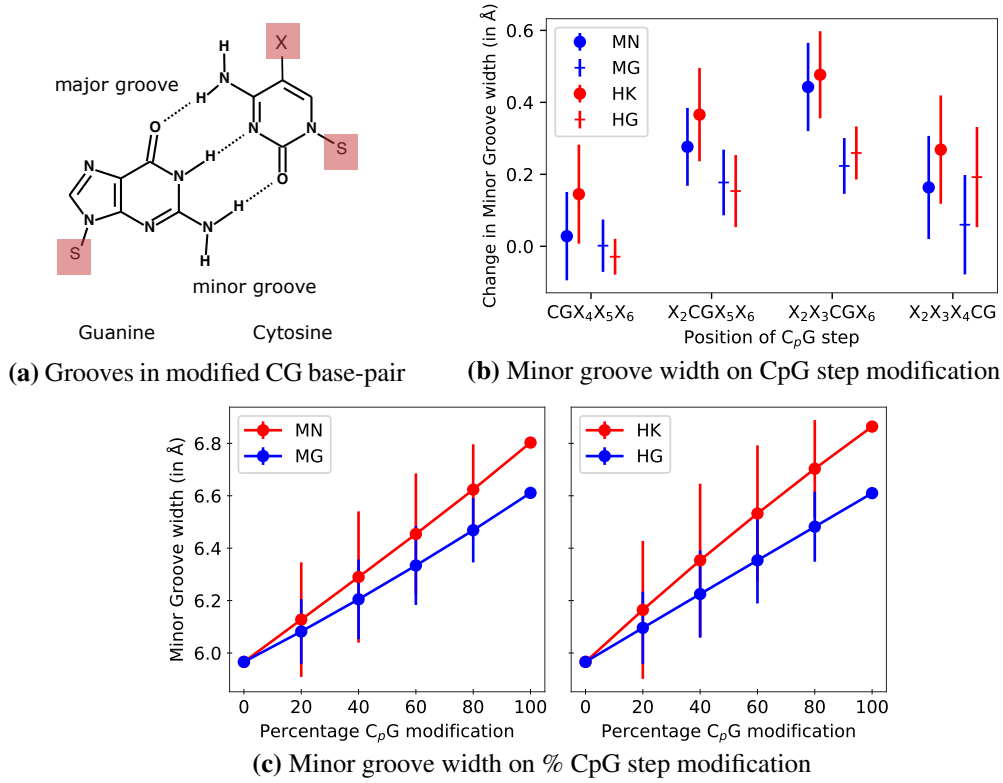


(a) Change in groundstate on symmetric methylation of central CpG step



(b) Change in groundstate on symmetric methylation of central CpG step

Fig. 6.6 Groundstate coordinates (elements of w) for (a) GCGTCGGAA**CG**TTTTGTCGGC (red, blue, and green as shown in legend) and same sequence with symmetric methylation on central CpG step in dashed lines, and (b) groundstate coordinates for GCGTCGGTG**CG**CTTTGTCGGC (red, blue, and green as shown in legend) and same sequence with symmetric methylation on central CpG step in dashed lines. The two sequences differ only in the immediate flanking sequence context (underlined) of the central CpG step (in bold), and the figure highlights the role of flanking sequence context in the change of dsDNA groundstate upon CpG methylation.



(a) Grooves in modified CG base-pair

(b) Minor groove width on CpG step modification

(c) Minor groove width on % CpG step modification

Fig. 6.7 (a) Schematic diagram for grooves in modified CG base-pair where methyl/hydroxymethyl group (X) is in major groove. Change in minor groove widths due to (b) CpG modification at different positions in the highlighted sub-sequence of GCTGTGX$_1$**X**$_2$**X**$_3$**X**$_4$**X**$_5$**X**$_6$X$_7$X$_8$X$_9$X$_{10}$CATGGC, and (c) various extent of CpG modification in the highlighted sub-sequence of GCTGTG**CGCGCGCGCG**CATGGC.

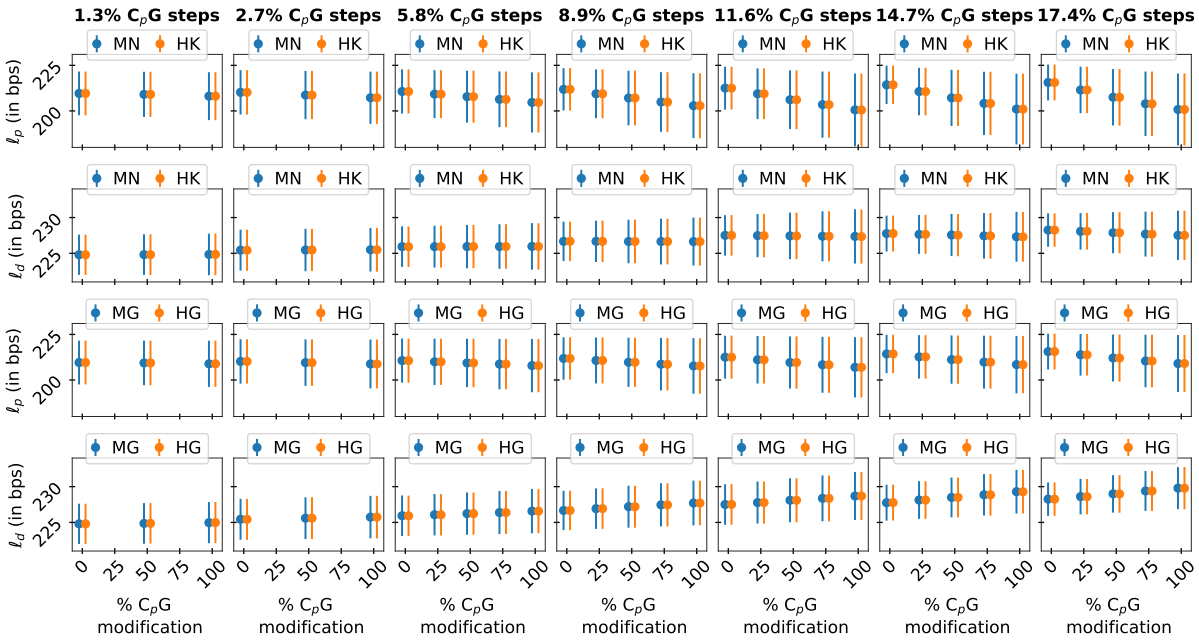


Fig. 6.8 Each subplot plots apparent (ℓ_p) or dynamic (ℓ_d) persistence lengths for sequences containing x% CpG steps (shown in title) for increasing % randomly modified CpG steps (shown in legend). • and error bar are the mean and standard deviation for 20,000 random sequences.

Similar trends are observed for asymmetric methylation of CpG steps, but the range of variation is slightly smaller (≈ 0.8 Å). Moreover, hydroxymethylation of CpG steps has an almost identical effect on the minor groove widths. Also, we would like to highlight that the position of the modified CpG step is crucial, as can be deduced from the error bar in the plot, in which sequences with the same percentage of CpG step modification have different minor groove widths.

Lastly, CpG modification slightly widens the major grooves. The plots are not shown here for brevity. However, it must be noted that even though the major groove width does not change significantly on CpG modification, it significantly changes the chemical environment inside the major grooves (with methyl group being hydrophobic while hydroxymethyl being hydrophilic) and therefore, has implications in protein-DNA interactions [164].

6.5 Effect of CpG modification on the persistence lengths of dsDNA

In this work, we have computed sequence-dependent dynamic persistence length, ℓ_d by factoring out the contributions of the intrinsic shape from apparent persistence length ℓ_p as described in ref. [123] and summarized in section 2.6 in the context of the cgNA+ model. Moreover, in section 4.4.2, we have shown that both the apparent and dynamic persistence lengths strongly depend on the sequence.

A general consensus is that methylation leads to an increase in persistence length [10, 68, 156, 181]. However, recent experimental studies [160, 187] revealed that hypermethylation could increase dsDNA flexibility. Therefore, to investigate the influence of CpG modification on the persistence length of dsDNA, we conducted a systematic study for 0.5 million sequences of length 220 bps for each type of CpG modification. In particular, we have generated several lists of 20,000 sequences with different percentages of CpG steps, namely, 1.3%, 2.7%, 5.8%, 8.9%, 11.6%, 14.7%, and 17.4% and then randomly modified 0%, 25%, 50%, 75%, and 100% of the CpG steps. In figure 6.8, we have plotted the dynamic and apparent persistence of all sequences. Each subplot, from left to right, plots persistence length for sequences with increasing %CpG steps, and in each subplot, we have shown persistence length for sequences with different %CpG modification, i.e., 0% and 100% CpG modification means unmodified and completely modified (CpG steps) sequences. Each data point is the mean and standard deviation of persistence length for 20,000 random sequences with a particular percentage of CpG steps and a certain number of those CpG steps are randomly modified. The following observations can be made from figure 6.8:

- With an increase in %CpG steps in the sequence (which may also correspond to GC content) both ℓ_p and ℓ_d increase.
- The effects on persistence length for both modifications, methylation and hydroxymethylation, are almost similar.
- According to the definition, $\ell_d - \ell_p > 0$, but this difference increases with increasing % modified CpG steps. It implies that the modified CpG steps make the intrinsic or groundstate shape of the sequence more bent, which might be attributed to the increased Roll of the CpG steps (refer to figure 6.4).
- The symmetric and asymmetric modifications show a similar pattern for ℓ_p, i.e, ℓ_p decreases when % modification increases. Moreover, the decrease in ℓ_p also depends on the

number of CpG steps in the sequence, since more modified CpG steps lead to a further decrease in ℓ_p.

- Whereas, ℓ_d remains constant on symmetric modification and ℓ_d increases when % asymmetric modification increases.

Thus, by rigorous computations, we found that the apparent persistence length of a given sequence decreases with the CpG step modification, while the dynamic persistence length remains almost similar for symmetric modification and increases for asymmetric modification.

CHAPTER 7

Neural networks to predict the location of sugar atoms in cgNA+ configurations

DNA consists of three elementary units: base, phosphate, and deoxyribose sugar, and the latter two form DNA backbone (refer figure 1.3). The main focus of this chapter is on the sugar and DNA backbone. The deoxyribose sugar is an inherently non-planar ring molecule that primarily stays in either C2′−endo or C3′−endo conformations. These conformations are strongly linked to two common geometries adopted by dsDNA (or any dsNA), i.e., A-form (C3′−endo) and B-form (C2′−endo), having distinct structural and mechanical properties. Moreover, the sugar-phosphate backbone is quite flexible and is characterized by the six dihedral parameters, but these dihedral angles are not entirely free to rotate due to steric constraints. Therefore, the ds-DNA backbone prefers some conformations over others; in particular, the two typical backbone conformations include BI and BII identified based on the difference between the dihedral angle $\epsilon - \zeta$, which is negative for BI and positive for BII conformation (refer section 1.1 for details). These backbone conformations directly related to the structural properties of dsDNA, for example, groove widths [70, 134] and base-pair step coordinates [51, 52, 144, 147] and are found to be important in protein-DNA recognition [51, 65]. Furthermore, it has been observed that the sugar pucker modes directly influence dsDNA backbone conformations with BII conformation is strongly constrained by the C2′−endo pucker, while BI conformations are much less affected by various sugar pucker modes [177].

Moreover, dsDNA backbone conformations are found to have specific sequence preferences in X-ray and NMR structures [51, 70, 144, 197, 208] with BI as the dominant conformation. In computational studies, it has been observed that along with the associated dimer step, the backbone conformations are also affected by the flanking sequence contexts [21, 39, 42, 69, 147]. A rigorous analysis by the ABC consortium [147] using MD simulations found that for most RR and RY steps, the backbone conformation is highly sensitive to flanking contexts with 5′−flanking Y and R favoring higher BII and BI %, respectively. In contrast, YR steps strongly prefer the BI state, irrespective of the flanking contexts. Thus, the dsDNA backbone and sugar conformations have specific sequence (often non-local) preferences and are crucial in determining the structural and mechanical properties of dsDNA and its functioning in biology.

The cgNA+ model explicitly treats bases and phosphates as rigid bodies, while sugar is treated implicitly; thus, it does not provide direct information on the dsDNA backbone and sugar conformations. For a given sequence, the cgNA+ model predicts the groundstate and the stiffness matrix in the internal coordinates of the base and the phosphate. Furthermore, using Monte Carlo sampling, one can obtain an ensemble of configurations corresponding to this

groundstate and stiffness matrix. For any such configuration (including groundstate) described in internal coordinates, the absolute position and orientation of each rigid base and each rigid phosphate can be reconstructed (refer section 2.2.3), and subsequently, an atomistic representation can be obtained by embedding localized ideal atoms coordinates (listed in table A.1) in bases and phosphates using equation (2.17). However, this atomistic representation of the configuration is missing the sugar group in the dsDNA backbone as the model does not consider sugar explicitly. This chapter presents a machine learning tool to completely fine-grain any cgNA+ configuration by additionally predicting sugar atoms. In particular, we have used feed-forward Neural Network (NN) to predict the location of sugar atoms in cgNA+ configuration using the positions of the phosphates and bases atoms. Note that the sugar atoms are the only heavy atoms missing in cgNA+ configurations.

With the same goal of finding missing heavy atoms, an alternative approach based on minimizing the force field potential (taken from the same force field used for MD simulations) was proposed in ref. [149]. In this approach, the total potential energy of the sub-molecule (given sugar and the covalently attached base and phosphates) is minimized to obtain the most stable sugar configuration while freezing the position of the attached phosphate and base as predicted by the cgNA+ model. Even though the approach worked reasonably well for cgNA+ groundstate, this approach had limitations. The primary constraint is that the minimization is slow and the algorithm is not guaranteed to find the minima. Thus, reconstructing the sugar ring for an ensemble of configurations for a given groundstate and stiffness (to perform any statistical analysis of the backbone configurations) is computationally expensive. Moreover, for any configuration that is not groundstate, the sugar ring should not adopt the local minima state; thus, the approach is unsuitable.

In the following section, we have first described the elementary details of the sugar ring and NNs. Then, a detailed mathematical framework to approach the problem and the necessary details of the NN training are provided. Once neural networks are trained, we have evaluated the accuracy of the cgNA+ sugar module to predict the location of sugar atoms and dihedral angles for the sugar ring and the backbone. Then we have shown an application of the model in the sequence-dependent analysis of pucker modes and backbone angles and discussed the potential application in obtaining an equilibrium structure that can be used to start MD simulations, particularly useful for dsDNA mini-circles. Finally, we showed that, despite some limitations, the module's predictions are impressively good; furthermore, we argued that the module has a scope for improvement and discussed the possible directions for future work.

Details of all the codes and data used in this chapter are provided appendix F.

7.1 Elementary details and implementation of the Neural Networks

In this section, we have discussed the implementation of NNs to predict the location of sugar atoms from bases and phosphates coordinates in cgNA+ configurations. First, we have described modeling assumptions, followed by a brief description of NN, and finally, the implementation and performance of the NNs in predicting the location of sugar atoms.

7.1.1 Sugar ring in DNA

Deoxyribose sugar in DNA is a five-member ring (C1′, C2′, C3′, C4′, and O4′) with C substitution (C5′) at C4′ as shown in figures 1.4 and 7.1. The sugar ring is typically characterized in terms of dihedral angles ($\theta_i \ \forall \ i \in [0,1,2,3,4]$) and pseudo-rotation phase angle (**P**) as defined in figure 1.4 and equation (1.1), respectively. The pseudo-rotation phase angle (**P**) and the maximum degree of pucker (θ_{max}) are defined as

$$tan(\mathbf{P}) = \frac{(\theta_4 + \theta_1) - (\theta_3 + \theta_0)}{2\theta_2\,(sin(36^\circ) + sin(72^\circ))} \text{ and } \theta_{max} = \frac{\theta_2}{cos(\mathbf{P})} \tag{7.1}$$

where **P** can be anything between $0 - 360^\circ$ and if $\theta_2 < 0$ then $\mathbf{P} = \mathbf{P} + 180^\circ$. Sugar ring is highly flexible, non-planar, and exists in various puckered conformations (characterized by **P**), which are inter-convertible into each other. A figure explaining various pucker conformations and respective definitions is provided in figure 1.4. Due to steric constraints, not all conformations are equally accessible, and the two most common puckered conformations are C2′− and C3′−endo.

7.1.2 Assumptions and mathematical formulation

The primary goal of this work is to predict the position of sugar atoms given the atomistic coordinates of adjacent phosphate and base atoms. It is reasonable to assume that the position of the surrounding bases and phosphates atoms can determine the location of the missing sugar atoms. In particular, it has been observed in previous works that the sugar modes and backbone orientations depend non-locally on the sequence [21, 39, 42, 69, 147].

To formulate a mathematical framework, we first introduce the notation. Let us assume that a configuration $w(S) \in \mathbb{R}^{24N-18}$ is provided in cgNA+ internal variables where N is the length of the sequence S in the number of base-pairs (bps). Using transformation $T_{I\to F} : \mathbb{R}^{24N-18} \to SE(3)^{4N-2}$ as defined in section 2.2.3, one can rewrite

$$T_{I\to F}(w) = \{F^n_{B^+}, F^n_{B^-}, F^{n'}_{P^+}, F^{n''}_{P^-}\}_{n=1\cdots N,\ n'=2\cdots N,\ n''=1\cdots N-1} \tag{7.2}$$

where $F \in SE(3)$ represents the frame (with orientation $R \in SO(3)$ and translation $r \in \mathbb{R}^3$) with subscript B and P representing the base and phosphate frame, respectively, and + and − denote the Watson (chosen as reading) and the Crick (complementary) strands. Note that the first $5'$-phosphate on both strands is not considered. Subsequently, in these frames, ideal atoms can be embedded using the transformation $T_{F\to C} : SE(3) \to \mathbb{R}^{3\times K}$ defined as

$$\mathcal{C}_{X^k} = T_{F\to C}(F_X) =: R\mathcal{A}_{X^k} + r, \ \forall \ k = 1\cdots K \tag{7.3}$$

where K is the number of atoms in base or phosphate, X is the kind of rigid body (base or phosphate), $\mathcal{A}_{X^k} \in \mathbb{R}^{3\times 1}$ is the coordinate of the k^{th} ideal atom in X type rigid body (listed in table A.1), and $\mathcal{C}_{X^k} \in \mathbb{R}^{3\times 1}$ is the coordinate of the k^{th} atom embedded in frame F_X. Thus, equation (7.2) can be written as

$$T_{F\to C}(T_{I\to F}(w)) = \{\mathcal{C}^n_{B^+}, \mathcal{C}^n_{B^-}, \mathcal{C}^{n'}_{P^+}, \mathcal{C}^{n''}_{P^-}\}_{n=1\cdots N,\ n'=2\cdots N,\ n''=1\cdots N-1} \tag{7.4}$$

where $\mathcal{C}_X \in \mathbb{R}^{3\times K(X)}$ and K is a function of the rigid body type (X). Lastly, we have denoted the sugar ring as $\{\mathrm{S}^n_+, \mathrm{S}^n_-\}_{n=1\cdots N}$ where + and − denote the reading and the complementary strands and n denotes the index of the base to which the sugar is attached.

In this work, we have assumed that the atomic coordinates of any sugar in a DNA strand can be determined from the location of the covalently attached base and phosphates and the two nearest bases (which are not directly bonded to the sugar) on the same strand. Figure 7.1 depicts a typical atomistic structure of a DNA strand where the sugar atoms for the n^{th} base-pair level (S^n_+) are highlighted in red, and the assumption is that the location of these atoms can be predicted from the atomic coordinates of the three nearest bases (i.e., bases at the n-1, n, and n+1 base-pair level) and two phosphates (between these three base-pair levels) on the same strand. In other words, $\mathrm{S}^n_\pm$ can be determined from the positions of $\{\mathcal{C}^{n+1}_{B^\pm}, \mathcal{C}^{n+1}_{P^\pm}, \mathcal{C}^{n}_{B^\pm}, \mathcal{C}^{n}_{P^\pm}, \mathcal{C}^{n-1}_{B^\pm}\}$, i.e.,

$$\mathrm{S}^n_\pm = \mathrm{f}\left(\{\mathcal{C}^{n+1}_{B^\pm}, \mathcal{C}^{n+1}_{P^\pm}, \mathcal{C}^{n}_{B^\pm}, \mathcal{C}^{n}_{P^\pm}, \mathcal{C}^{n-1}_{B^\pm}\}\right) \tag{7.5}$$

We have used NNs to approximate the underlying function $(\mathrm{f}(\cdot))$ that predicts the position of sugar atoms from the position of neighboring bases and phosphates. In the following subsection, we have briefly discussed NN and its training. More details about NNs can be found in an excellent book by Goodfellow et al. [64].

Notably, we assumed that location of sugar atoms depends on neighboring three bases and two phosphates (on the same strand), however, the positions of these bases and phosphates are non-locally dependent on further neighbors on both strands; thus, the predicted location of sugar atoms will have non-local sequence dependence. We have tested the NN predictions for various choices and concluded that this approximation is optimal.

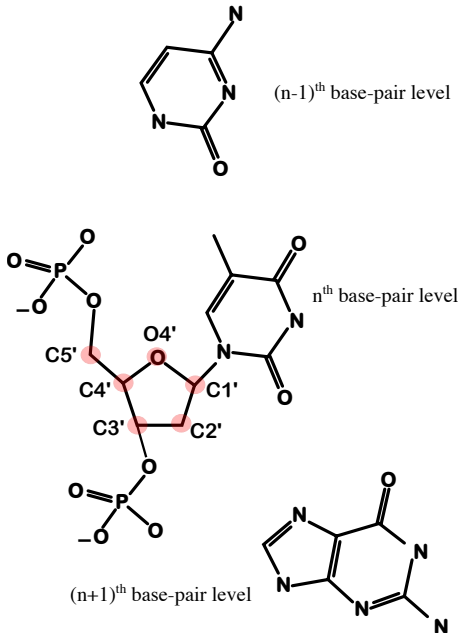


Fig. 7.1 A schematic diagram of a DNA strand with bases and phosphates (which can be obtained from the cgNA+ model) along with the missing sugar atoms highlighted in light red color. The figure only focuses on one middle sugar ring; the rest of the sugar rings and the complementary strand are not shown.

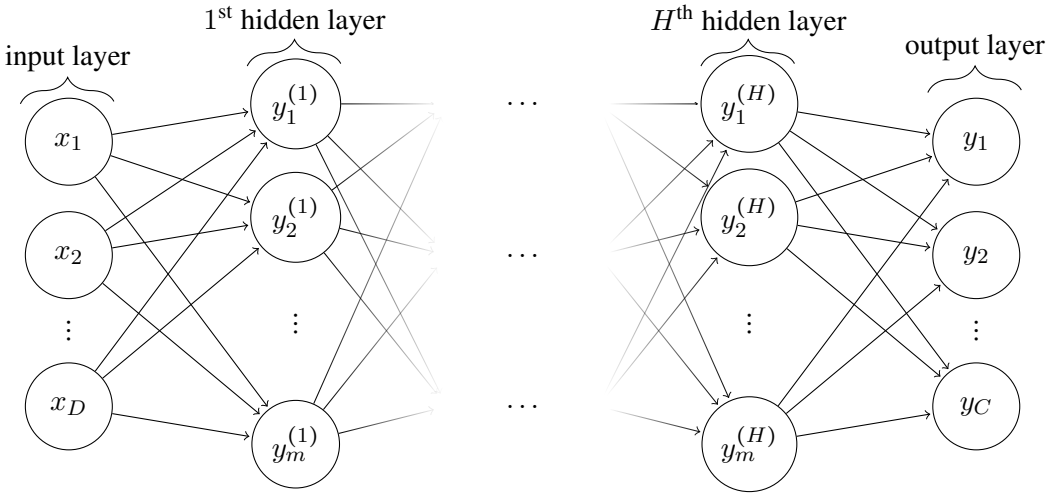


Fig. 7.2 Typical schematic diagram for a feed-forward Neural Network with D input units, C output units, and H hidden layers each containing m neurons. The input and output layers are considered as 0^{th} and $(H+1)^{\text{th}}$ layers.

7.1.3 Feed-forward Neural Network

Feed-forward Neural Networks (NNs), also known as multilayer perceptrons, are an important class of machine learning algorithms whose structure and name are inspired by neurons in the human brain. The goal of an NN is to approximate the true underlying function (f^*) for the given input (X) and output (Y^*) as $Y = f(X;\Theta)$ by learning the NN parameters Θ. A typical NN diagram is shown in figure 7.2 with the zeroth layer as input layer (containing D input units), H hidden layers (containing m neurons each), and the final layer as output layer (containing C outputs). The inputs $X(x_1, x_2, ..., x_D)$ can also be denoted as $(y_1^{(0)}, y_2^{(0)}, ..., y_D^{(0)})$ and similarly, the outputs $Y(y_1, y_2, ..., y_C)$ as $(y_1^{(H+1)}, y_2^{(H+1)}, ..., y_C^{(H+1)})$. $y_j^{(h)}$ is the output of the j^{th} neuron in layer h that is given as a function of the output of neurons in layer $h-1$ as,

$$y_j^{(h)} = f^{(h)}(y^{(h-1)}) = \Phi\left(\sum_i w_{i,j}^{(h)} y_i^{(h-1)} + b_j^{(h)}\right) \tag{7.6}$$

where i denotes the index of neuron in $(h-1)^{\text{th}}$ layer, $w_{i,j}^{(h)}$ is an NN parameter called weight that connects i^{th} neuron in $(h-1)^{\text{th}}$ layer in j^{th} neuron of the h^{th} layer and $b_j^{(h)}$ is the bias term at h^{th} layer for the j^{th} neuron. Lastly, $\Phi(\cdot)$ is the activation function, which is often non-linear. Popular choices for $\Phi(\cdot)$ include sigmoid function, inverse tangent function, and rectified linear unit (ReLU). The final output, Y, is given as the composition of functions applied at every layer as

$$Y = f(X) = f^{(H+1)} \circ \cdots \circ f^{(2)} \circ f^{(1)}(X). \tag{7.7}$$

This step is also known as the forward pass.

Given the training data with P samples, $S_{Train} = (X_p, Y_p^*)_{p=1..P}$, the training of NN involves finding NN parameters, w and b, such that it minimizes the cost function defined as

$$\mathcal{L} = \frac{1}{P}\sum_{p=1}^{P}\left(Y_p^* - Y_p\right)^2. \tag{7.8}$$

The above loss function is the most popular choice called mean square error. Other popular choices include the mean absolute error, a combination of mean square error and mean absolute error, or some custom choices. The NN training step is called the back-propagation step, and there exist several algorithms to minimize the loss function and, thus, optimize w and b starting from some random initialization. Popular algorithms for the back-propagation step are gradient descent, stochastic gradient descent, and Adam optimizer.

Before training NN, one must make various choices, including the number of hidden layers (H), neurons in each layer (m), activation function ($\Phi(\cdot)$), optimizer, network initialization, and loss function. These choices are collectively called hyperparameters, and various options should be explored to find optimal hyperparameters for the given training data. One of the standard techniques for searching in hyperparameter space is the k-fold cross-validation (CV) technique[64]. In the CV technique, a subset of training data is separated (randomly chosen), termed the validation set, used to test the performance of the NN trained on the remaining training data. This performance of the NN for a given choice of hyperparameters is called validation accuracy. The k-fold CV repeats the train-validation split k times so that all the samples in the training data become part of the validation data exactly once. The most common choice for k is 5, i.e., in every split, 20% of the data are taken as validation data, and the model is trained on the remaining 80%. Thus, using k-fold CV, one can compute the average validation accuracy for a given choice of hyperparameters, and based on this average validation accuracy, one can choose the optimal hyperparameters while searching in hyperparameter space.

7.1.4 Training data

As described in chapter 3, we have performed 10 μs long MD simulations for 16 palindromic sequences (of length 24 bps) referred to as the training sequences in Lb_{DNA} (see table B.1) to train the $\mathcal{P}_{\text{DNA}}$ in cgNA+ model. To train the NN, we have used snapshots from the same training data. Note that the training sequences in Lb_{DNA} contain all possible trimers with almost the same frequency. We have used the following steps to obtain the training data for NN:

i) In the MD time-series of each training sequence, we sub-sampled snapshots such that the configurations are uncorrelated (100 picoseconds apart from each other). It leads to $16 \cdot 10^5$ snapshots in the training data, since we have 10μs of MD simulations for 16 sequences. Subsequently, we removed snapshots with broken H-bond (as done in training the cgNA+ model), which discarded $\approx 15\%$ of the initial training data; thus, $0.85 \cdot 16 \cdot 10^5$ snapshots.

ii) Each MD configuration is split into trimers centered around the middle sugar ring, as shown in figure 7.1 while reading separately from both strands. It led to 22 trimers per strand for each configuration (as sugar rings associated with terminal base-pairs are ignored). Thus, the total samples to train NN $\approx 0.85 \cdot 16 \cdot 10^5 \cdot 2 \cdot \ 22$ where the atomistic coordinates of the phosphates and bases are inputs, while the sugar atoms are outputs.

iii) Before training, we have aligned all the trimers about its central base, i.e., the frame associated with the central base is taken as $\{I, \mathbf{0}\}$ where $I \in \mathbb{R}^{3\times3}$ is an identity matrix and $\mathbf{0} \in R^{3\times1}$ is a zero vector.

Model	Weight initialization	Optimizer	Batch size	Epoch	Activation function	Number of nodes	Hidden layers	Learning rate
			(a) Hyperparameter space explored					
	Xavier normal Random normal	Adam SGD	32 64 128	50 100 200	ReLU tanh sigmoid	200 to 1800 in steps of 200	2, 4, 6, 8	0.001 0.005
			(b) Optimal hyperparameters					
RRR	Xavier normal	Adam	32	200	ReLU	1200	4	0.001
RRY	Xavier normal	Adam	32	200	ReLU	1200	4	0.001
RYR	Xavier normal	Adam	32	200	ReLU	1800	4	0.001
YRR	Xavier normal	Adam	32	200	ReLU	1200	4	0.001
YYR	Xavier normal	Adam	32	200	ReLU	1200	4	0.001
YRY	Xavier normal	Adam	32	200	sigmoid	400	4	0.0005
RYY	Xavier normal	Adam	32	200	ReLU	1800	6	0.001
YYY	Xavier normal	Adam	32	200	ReLU	400	4	0.0005

Table 7.1 (a) Hyperparameters space explored and (b) the optimal hyperparameters found for neural networks trained for each trimer.

Since the overarching aim of this work is to fit the sugar ring in cgNA+ predicted coarse-grained configurations, which have rigid phosphates and rigid bases, i.e., fixed position of atoms within a given rigid body. Therefore, the model should also be trained on similar data rather than crude MD snapshots, which have non-rigid phosphates and bases. Thus, the input data for the NN are the best-fit ideal coordinates in the phosphate and base units, which are obtained by first fitting the frames in the MD snapshots (refer section 2.1) and then re-embed ideal atoms as described in equation (7.4).

7.1.5 Implementation

In the previous sections, we have described the typical architecture of the NN and the training data. In this section, we have discussed the implementation of the NN to predict the location of sugar atoms from the positions of neighboring bases and phosphates. In other words, we have trained the NN for which the input is the atomic positions of the three nearest bases and the two nearest phosphates to the sugar ring on the same strand, and the output is the coordinate of the sugar atoms. It is mathematically described in equation (7.5).

One of the crucial aspects of the NN architecture pertinent to this implementation is that the NN has a fixed number of input features (once chosen). However, the number of atoms in the neighboring bases depends on the type of base (A/T/C/G). A/G are purines with more atoms, while C/T are pyrimidines with fewer atoms. Therefore, we have trained eight different NNs for eight possible trimers in the pyrimidine and purine alphabets.

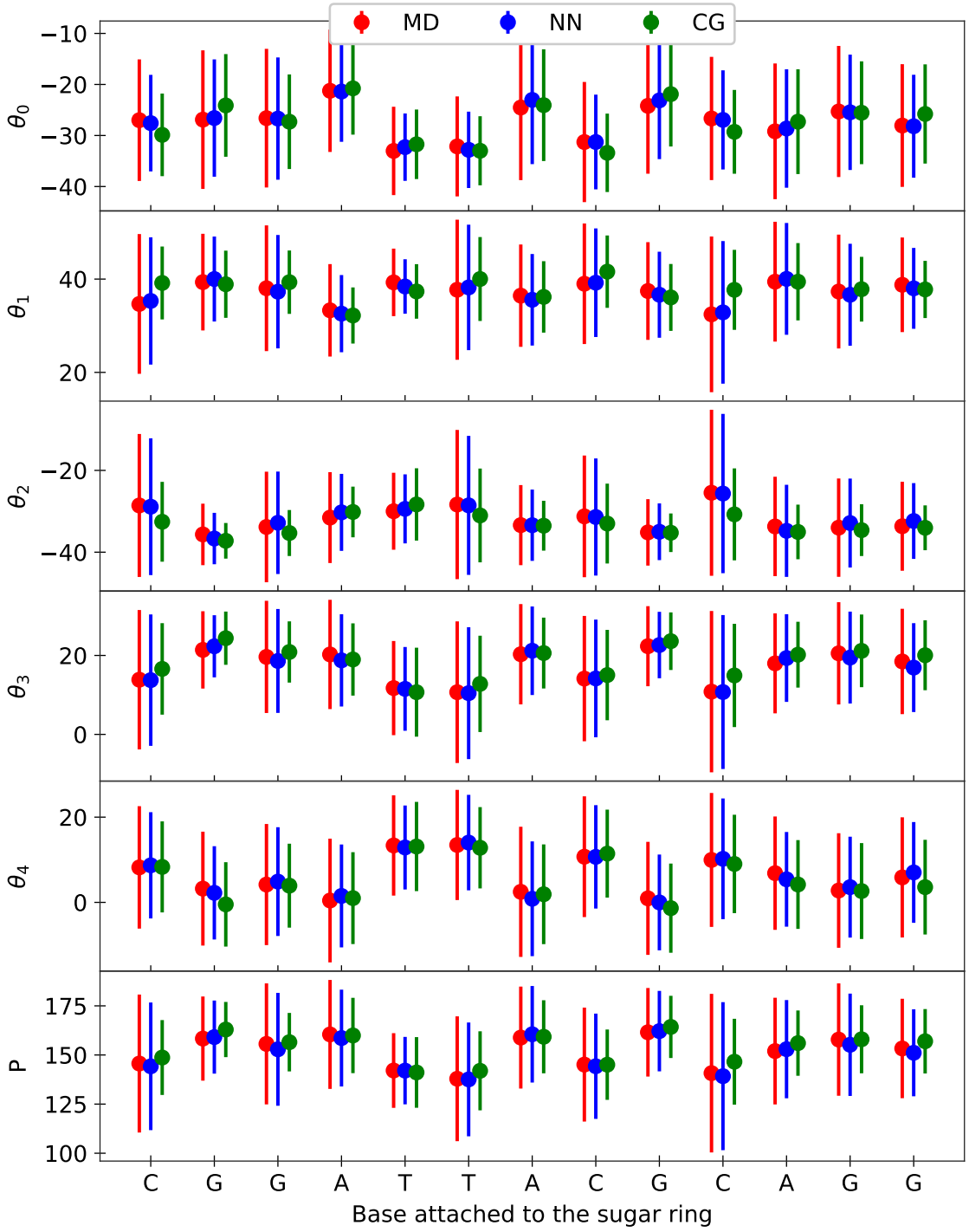


Fig. 7.3 Sugar pucker angles on Watson strand of sequence index 20 in Lb_{DNA} (GCGGATTACGCAGGC). The parameters observed in MD simulations (labeled as MD) are in red, obtained by re-fitting sugar in coarse-grained MD snapshots (labeled as NN) are in blue, and obtained by fitting sugar in an ensemble of coarse-grained configurations generated by the cgNA+ Monte Carlo (labeled as CG) are in green. The ensemble mean and standard deviation for a given parameter are plotted as • and vertical line, respectively.

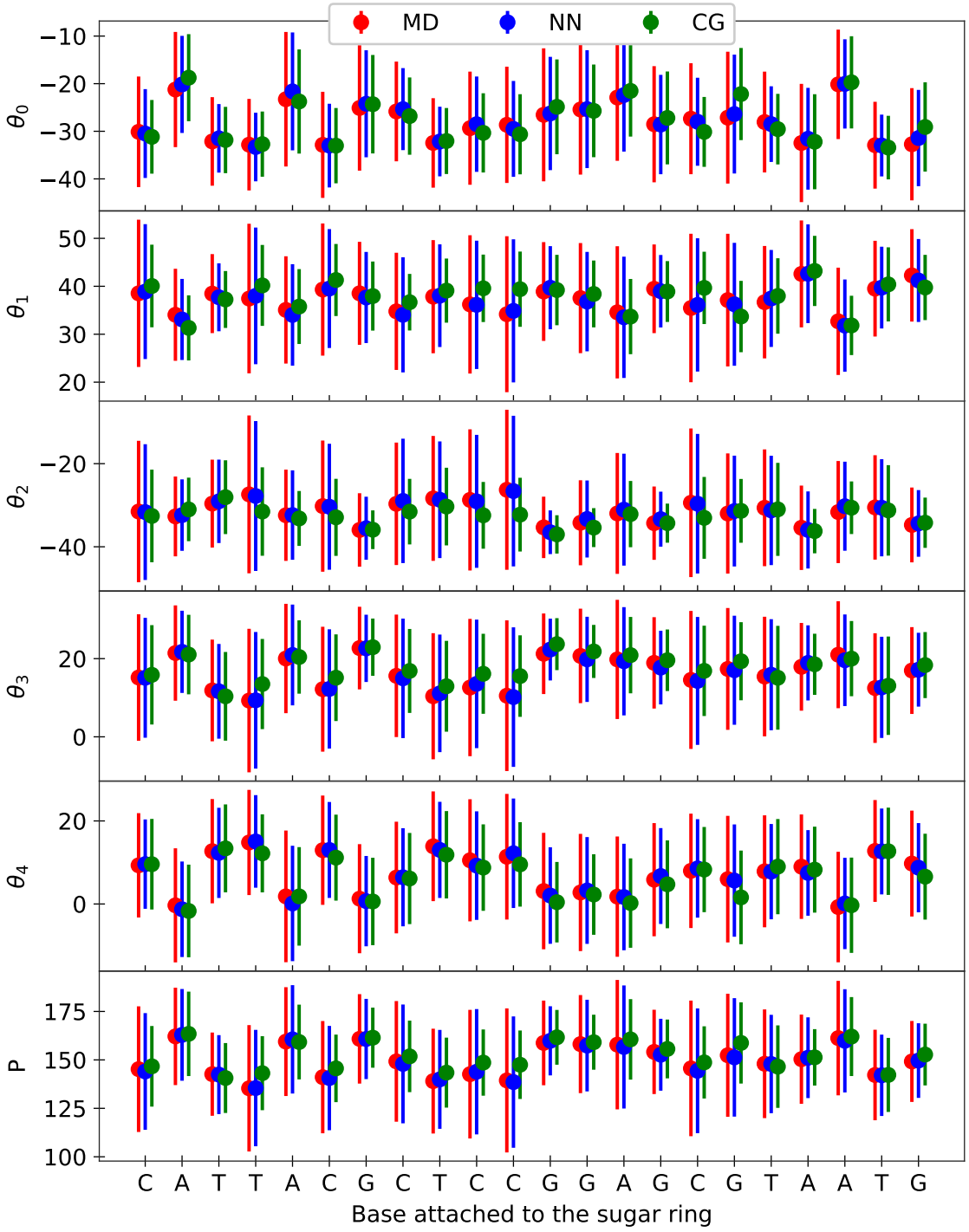


Fig. 7.4 Sugar pucker angles on Watson strand of sequence index 17 in Lb_{DNA} (GCATTACGCTCCGGAGCGTAATGC). The parameters observed in MD simulations (labeled as MD) are in red, obtained by fitting sugar in coarse-grained MD snapshots (labeled as NN) are in blue, and obtained by fitting sugar in an ensemble of coarse-grained configurations generated by the cgNA+ Monte Carlo (labeled as CG) are in green. The ensemble mean and standard deviation for a given parameter are plotted as • and vertical line, respectively.

7.1.5.1 Hyperparameters selection

Before training the final (best-fit) NN model, one must choose the hyperparameters for NN. We have chosen the hyperparameters for each NN based on the average validation accuracy as described in section 7.1.3. We have used a five-fold CV, i.e., randomly split the total training data (for each trimer) into five equal sets, of which four sets were used as training data, while the remaining one was used as validation data. This process is repeated five times such that each set becomes the validation data exactly once. Based on this average validation accuracy, we have selected the hyperparameters for the best-fit model. In table 7.1, we have listed the hyperparameter space explored to find the optimal choice for the best-fit model. In total, we have explored $7,776$ combinations, of which many choices gave a comparable average validation accuracy. Finally, for each trimer model, the set of hyperparameters with the highest average validation accuracy was selected as the optimal hyperparameters as tabulated in table 7.1. In particular, we found that the Adam optimizer performed best with 200 epochs (the number of times the training data pass through the algorithm), batch-size (number of samples in one iteration of training) of 32, and Xavier normal initialization of the neural network parameters. The ReLU activation function generally works best, and the optimal choices for the number of neurons and hidden layers are different for different trimer models.

7.1.5.2 Best-fit model

Once the optimal hyperparameters are found (listed in table 7.1), the best-fit model is trained for each trimer using the complete training data. The training of each model took approximately 12 hours on one CPU.

7.1.6 How accurate is the model?

In this section, we have discussed the NN accuracy in predicting sugar atom coordinates. First, we have described the test data and then quantified NN performance.

7.1.6.1 Test data for model

Test data should not be involved in model training or hyperparameter selection. Recall that NNs are trained on MD snapshots that are 100 picoseconds apart in the MD times-series (which has snapshots at two picosecond intervals) of the training sequences in $\mathrm{Lb_{DNA}}$. Therefore, in principle, the remaining MD snapshots not used in NN training can be used as test data. A more severe test would be to check the model prediction accuracy for sequences not used in the training data, i.e., test sequences in $\mathrm{Lb_{DNA}}$. In the following section, we have compared the predictions with observations in MD simulations for two sequences that are not part of the training library.

7.1.6.2 Model accuracy in the prediction of sugar atoms position

The cgNA+ sugar module predicts the location of sugar atoms in any cgNA+ coarse-grained configuration. Then, various dihedral angles (commonly used to characterize the DNA back-

bone and sugar ring) can be computed from these atomic positions. We have assessed the quality of the model predictions for a) atomistic coordinates of sugar and b) various backbone dihedrals and sugar pucker angles for two test sequences (indices 17 and 20 in Lb_{DNA}).

To evaluate the accuracy of the cgNA+ sugar module, we first have coarse-grained MD snapshots (by fitting phosphate and base frames) and then fine-grain those coarse-grained snapshots using the cgNA+ sugar module. The mean square error per degree of freedom (dof) in predicting the location of sugar atoms for the test data is approximately 0.005 Å^2 or ≈ 0.07 Å as the root mean square error per dof. The statistics are obtained from 10^5 snapshots (minus snapshots with broken H bond) for each of two test sequences (indices 17 and 20 in Lb_{DNA}). Thus, in absolute terms, the mean error in the predictions of the cgNA+ sugar module for test sequences is negligible. Compared to the reconstruction error in the cgNA+ model, which is ≈ 0.003 Å^2 or $(\text{rad}/5)^2$ per dof in terms of the Mahalanobis distance (see table 4.1 for more details), the prediction error in the cgNA+ sugar module is only slightly larger. It highlights that the cgNA+ sugar module can accurately fine-grain any cgNA+ coarse-grained configuration.

The standard parameters for analyzing the DNA backbone and sugar conformations are dihedral angles. In figures 7.3 and 7.4, we have plotted the sugar ring dihedral angles and the pseudo-rotation phase angle (**P**) (defined in section 1.1.1) for two test sequences with MD observations in red and the corresponding NN predictions (for the same MD configurations) in blue. The following observations can be made from the figures:

- Both the mean and standard deviation for various parameters are highly sequence-dependent.
- NNs capture the distribution of various parameters exceptionally well, with mean values almost identical and standard deviation slightly smaller in the predictions.
- In a one-to-one comparison of the configurations in the ensemble, the Pearson correlation for any parameter is greater than 0.9.

Furthermore, in figures 7.5 and 7.6, we have plotted various backbone dihedral angles (defined in section 1.1.1). First, note that the sugar pucker angles are defined using the sugar atoms predicted by the NNs; in contrast, backbone dihedral angles involve some atoms that are part of either rigid base or rigid phosphate. For instance, α_n is defined as the dihedral angle between $\text{O3}'_{n-1}$–P_n–$\text{O5}'_n$–$\text{C5}'_n$ (see figure 1.3) involving frozen atoms $\text{O3}'_{n-1}$, P_n, and $\text{O5}'_n$. Therefore, a larger prediction error can be expected in the backbone dihedral angles because of the rigid base/phosphate assumption, which can not be attributed to the NNs performance. As can be observed in figures 7.5 and 7.6, NNs capture the dihedral angles of the backbone; their mean and standard deviation are almost similar and highly dependent on the underlying sequence. However, one can also notice that the various backbone dihedrals are consistently underestimated or overestimated (only a few degrees) irrespective of the sequence, e.g., α and ζ are overestimated, while the rest are underestimated. Lastly, in figure 7.7, we have compared the backbone conformations in the two test sequences by plotting % BII population identified as $\epsilon - \zeta > 0$. NNs predictions systematically overestimate % BII conformations compared to those observed in the corresponding MD simulations, which can be expected as a slight underestimation and overestimation of ϵ and ζ, respectively, have a pronounced effect on % BII or BI conformations.

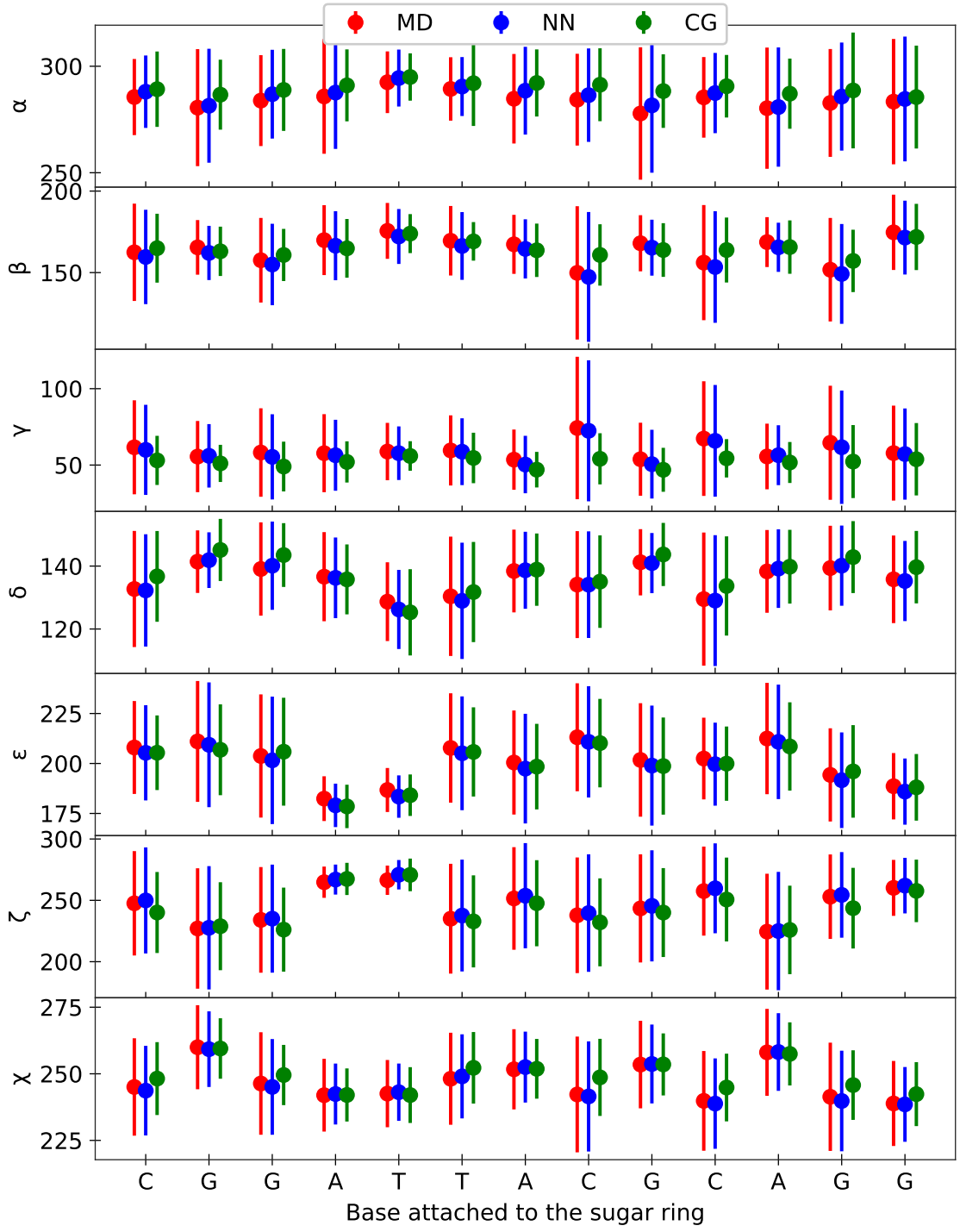


Fig. 7.5 Backbone dihedrals (on the Watson strand) for sequence index 20 in Lb_{DNA} (GCGGATTACGCAGGC). The parameters observed in MD simulations (labeled as MD) are in red, obtained by fitting sugar in coarse-grained MD snapshots (labeled as NN) are in blue, and obtained by fitting sugar in an ensemble of coarse-grained configurations generated by the cgNA+ Monte Carlo (labeled as CG) are in green. The ensemble mean and standard deviation for a given parameter are plotted as • and vertical line, respectively.

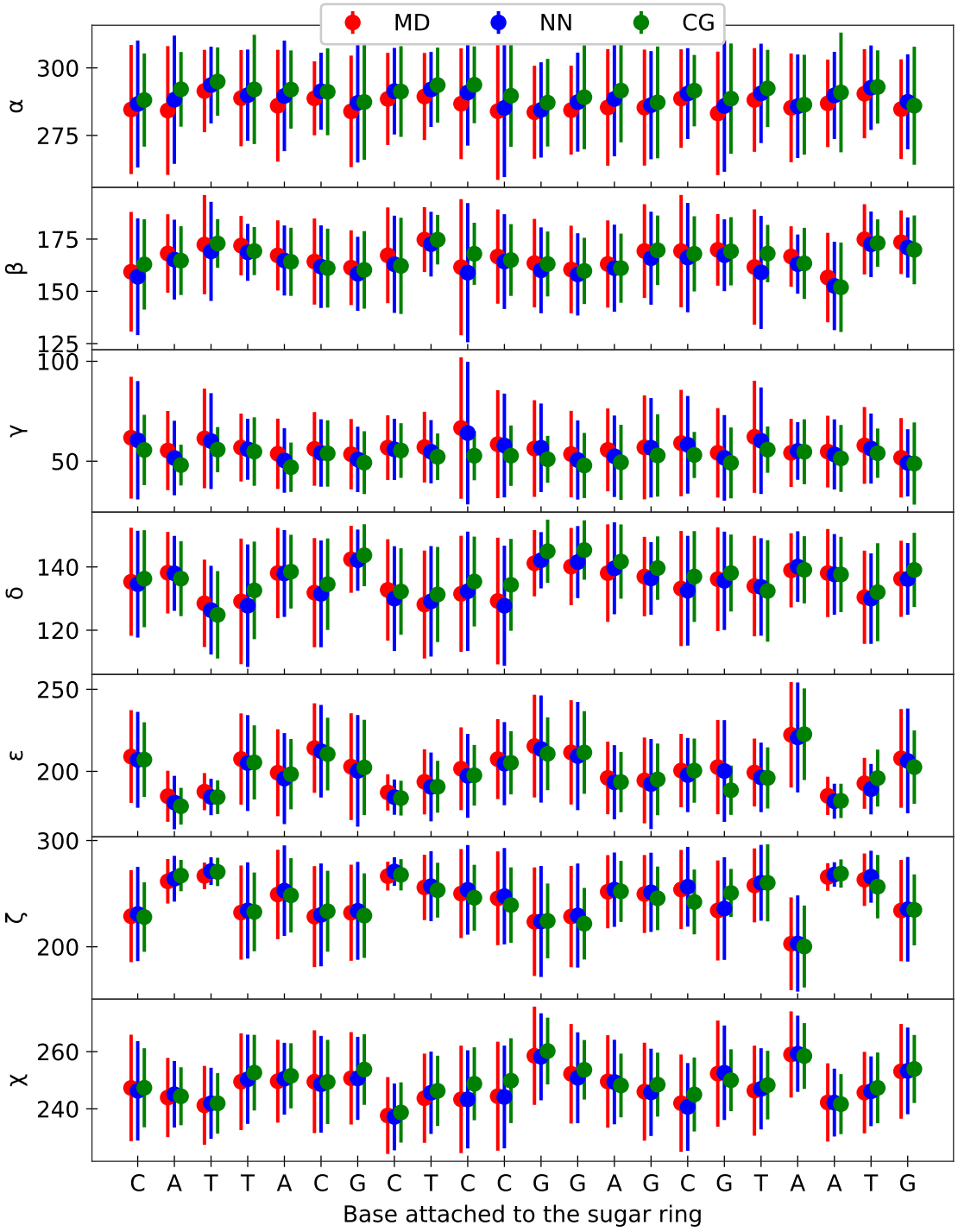
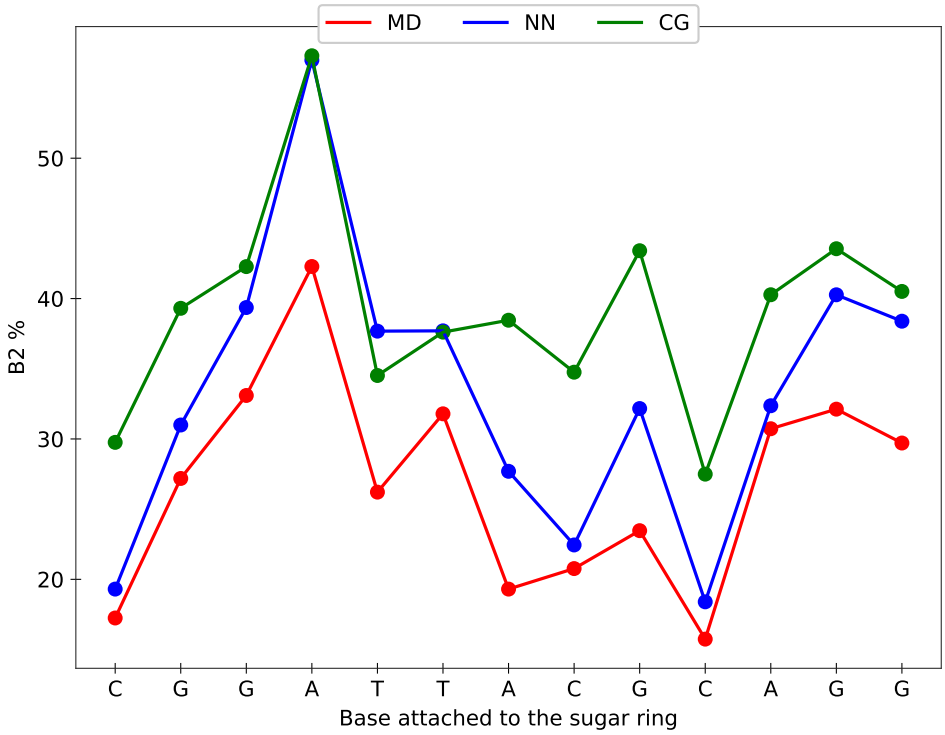
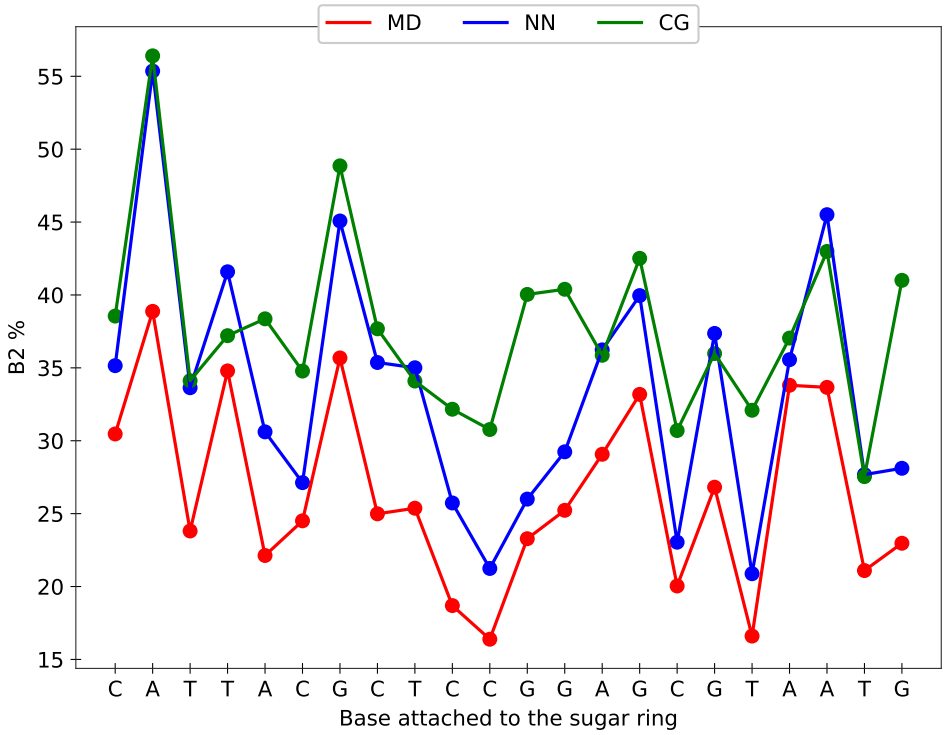


Fig. 7.6 Backbone dihedrals (on the Watson strand) for sequence index 17 in Lb_{DNA} (GCATTACGCTCCGGAGCGTAATGC). The parameters observed in MD simulations (labeled as MD) are in red, obtained by fitting sugar in coarse-grained MD snapshots (labeled as NN) are in blue, and obtained by fitting sugar in an ensemble of coarse-grained configurations generated by the cgNA+ Monte Carlo (labeled as CG) are in green. The ensemble mean and standard deviation for a given parameter are plotted as • and vertical line, respectively.



(a) BII % on the Watson strand for sequence index 20 in Lb_{DNA} (GCGGATTACGCAGGC)



(b) BII % on the Watson strand for sequence index 17 in Lb_{DNA} (GCATTACGCTCCGGAGCGTAATGC)

Fig. 7.7 BII % on the Watson strand for sequence indices 20 and 17 in Lb_{DNA}. The parameters observed in MD simulations (labeled as MD) are in red, obtained by fitting sugar in coarse-grained MD snapshots (labeled as NN) are in blue, and obtained by fitting sugar in an ensemble of coarse-grained configurations generated by the cgNA+ Monte Carlo (labeled as CG) are in green. The ensemble mean and standard deviation for a given parameter are plotted as • and vertical line, respectively.

7.2 Applications of the cgNA+ sugar module

The primary goal of the cgNA+ sugar module is to fine-grain any cgNA+ coarse-grained configuration. It allows to predict a complete atomistic equilibrium structure for an arbitrary sequence. Moreover, using the cgNA+ Monte Carlo code (refer section 2.6), one can obtain an ensemble of configurations (in cgNA+ coarse-grained coordinates) and then using the cgNA+ sugar module, an ensemble of atomistic configurations. This ensemble allows for studying backbone and sugar-pucker conformations for any sequence. In particular, we have predicted the ensemble of atomistic configurations (using the cgNA+ Monte Carlo code and sugar module) for sequence indices 17 and 20 (of $\mathrm{Lb}_{\mathrm{DNA}}$ listed in table B.1) and plotted the various pucker angles, dihedral angles and backbone conformations in figures 7.3 to 7.7 in green. For the sugar pucker angles shown in figures 7.3 and 7.4, the two data sets, a) MD observations and b) ensemble of atomistic configurations obtained using the cgNA+ Monte Carlo code, are close in terms of mean and standard deviation (standard deviation is greater in MD observations). However, there are some noticeable exceptions, for example, C at the 11^{th} position of index 20 as shown in figure 7.3. Similar observations can also be made for the dihedral angles and conformations of the backbone in figures 7.5 to 7.7. We would like to highlight that the provided MD statistics are for 10 μs of simulation data which took approximately two months (for a sequence of length 24 base-pairs) on a highly efficient GPU node (containing 2 Xeon-Gold processors and 2 NVIDIA V100 PCIe 32 GB GPUs); in contrast, the ensemble of 10^5 atomistic configurations obtained using the cgNA+ tools merely took an hour. Thus, this approach provided an accurate and highly efficient alternative to obtain a sampling of configurations for any sequence. Such an analysis can be easily performed for a large number of sequences. It will be beneficial to generate such an ensemble for some mechanically exceptional sequences discovered using the cgNA+ model, for instance, sequences with extreme groove widths or persistence lengths, as discussed in chapter 4. This analysis is not performed in this thesis, as the authors believe that there is scope for improvement in the current cgNA+ sugar module, as discussed in the next section; therefore, an extensive analysis of interesting sequences will be undertaken in the future.

Another application of this module is to obtain a sequence-dependent atomistic equilibrium structure for an arbitrary sequence, which can then be used as a starting point for the MD simulations. Note that the starting structure in an MD simulation is crucial and desirable that it is close to the equilibrium structure under the given physical conditions. A starting structure far from the equilibrium structure may take a prohibitively long simulation time to equilibrate. There are several reliable sources to obtain an initial structure for short linear NA fragments, such as the nucleic acid builder (NAB) in AMBERTOOLS 18 [29]. However, for large systems, in particular, dsDNA mini-circles (typically are of length 60-500 base-pairs), obtaining a good initial structure is non-trivial and crucial as full atomistic MD simulation is computationally expensive. Previous studies [98] have used JUMNA software [100] to obtain the initial structures, and then the energy of that structure is minimized using some force fields. In particular, Glowacki et al. [13, 60] developed an algorithm to compute the equilibrium structure of the dsDNA mini-circle (in cgDNA/cgDNA+ internal coordinates) for a given sequence and linking number from a linear groundstate predicted by the cgDNA/cgDNA+ model. This al-

gorithm, combined with the cgNA+ sugar module, presents an excellent method to obtain a sequence-dependent equilibrium atomistic structure that can be used to start MD simulations. Such an accurate sequence-dependent equilibrium structure for dsDNA mini-circles should take significantly less simulation time to equilibrate in MD simulations. Note that the location of sugar atoms predicted by the cgNA+ sugar module for dsDNA mini-circles might not be highly accurate compared to linear fragments (as the NNs are trained on linear fragments data), in particular, for highly overwound or underwound dsDNA mini-circles, but it is still an excellent tool to obtain a good initial structure for dsDNA mini-circles. At the time of the thesis writing, we do not have any MD simulation results for dsDNA mini-circles to compare the predicted dsDNA mini-circle equilibrium structure with the MD observations, but in the future, such a comparison will help validate and further improve the cgNA+ sugar module for dsDNA mini-circles.

7.3 Limitations of the cgNA+ sugar module and improvement directions

We would like to emphasize that the current version of the module presented in this chapter has a further scope of improvement in several directions. In particular, we have used only the feed-forward architecture of the NNs, which works reasonably well. However, other popularly used NN architectures, such as recurrent and long-short-term memory NN, might improve the predictions.

Calculating the dihedral angles and backbone conformational state involves phosphate or base atoms that are assumed to be frozen in the cgNA+ configurations. This approximation for phosphate and base as rigid bodies is reasonable; however, it might lead to erroneous backbone dihedral angles. In particular, the atoms within a phosphate group fluctuate more and are involved in the computation of most dihedral angles. Therefore, a possible solution to improve the prediction of the backbone dihedral angles is to allow perturbations in the phosphate atoms. This can be implemented by predicting perturbations in the phosphate atoms (or perturbed phosphate position) using the NNs as additional outputs.

Finally, in the current module, we do not have parameters for the terminal sugar as it does not have the same number of neighbors as the interior sugar. It requires training new networks with a different number of input features. Moreover, the current module is limited to dsDNA only, and it would be particularly beneficial to extend the module for dsDNA with epigenetic base modifications that significantly impact backbone conformation and sugar puckering modes [10, 107]. The module should also be extended for dsRNA and DRH for completeness.

CHAPTER 8

Conclusions and future work

8.1 Summary and conclusions

In this work, we have extended the cgDNA+ model, which predicts a non-local sequence-dependent Gaussian pdf for any arbitrary DNA sequence, to the cgNA+ model by estimating parameter sets for a wider variety of double-stranded nucleic acids (dsNAs), including dsDNA with epigenetic base modifications, dsRNA, and DNA:RNA hybrid (DRH). Just as in its precursor cgDNA+ model, the cgNA+ model explicitly treats bases and phosphates as rigid bodies $\in SE(3)$ and uses helicoidal CURVES+ coordinates to parameterize the configuration of dsNAs. For a sequence S of length N base-pairs, any configuration can be described using internal coordinates in $24N - 18$ dimensions with $6N$ intra base-pair, $6(N-1)$ inter base-pair step, $6(N-1)$ Crick phosphate and $6(N-1)$ Watson phosphate coordinates. Thus, given a sequence S and a parameter set $\mathcal{P}_{\text{NA}}$, the cgNA+ model predicts a Gaussian pdf in configuration space by reconstructing a ground-state $\hat{w}(S, \mathcal{P}_{\text{NA}}) \in \mathbb{R}^{24N-18}$, and a positive-definite stiffness matrix $\mathcal{K}(S, \mathcal{P}_{\text{NA}}) \in \mathbb{R}^{24N-18 \times 24N-18}$:

$$\rho(w; S, \mathcal{P}_{\text{NA}}) = \frac{1}{Z} exp\{-\frac{1}{2}(w - \hat{w}) \cdot \mathcal{K}(w - \hat{w})\}. \tag{8.1}$$

The parameter set $\mathcal{P}_{\text{NA}}$ contains dinucleotide-step dependent stiffness blocks $\in \mathbb{R}^{42\times42}$ and stress vectors $\in \mathbb{R}^{42}$ for interior dinucleotide steps and stiffness blocks $\in \mathbb{R}^{36\times36}$ and stress vectors $\in \mathbb{R}^{36}$ for terminal dinucleotide steps. From these dinucleotide-step dependent parameters, oligomer level stiffness matrix $\mathcal{K}(S)$ and stress vector $\sigma(S)$ are constructed by overlaying the blocks with 18×18 overlap in $\mathcal{K}(S)$ and 18×1 overlap in $\sigma(S)$. Notably, the parameter set is dinucleotide-step dependent; thus, both the stiffness matrix $\mathcal{K}(S)$ and stress vector $\sigma(S)$ have local sequence-dependence, but the groundstate $\hat{w}(S)$ constructed as,

$$\hat{w}(S) = \mathcal{K}^{-1}(S)\sigma(S), \tag{8.2}$$

and so has a non-local (often strongly non-local) sequence dependence due to the inversion of the banded stiffness matrix $\mathcal{K}$. It reflects the physical phenomenon of frustration that originates because each base-pair level (complementary bases along with their $5'$-phosphates) participates in two base-pair level junctions which can not simultaneously minimize their energy. This phenomenon of frustration energy and, thus, non-local sequence dependence is only possible in rigid base models or higher hierarchy models such as cgDNA+ and cgNA+ (details in chapter 2).

The cgNA+ model is trained on atomistic molecular dynamics (MD) simulations of a comprehensive set of diverse sequences (sixteen 24mers for each dsDNA, dsRNA, and DRH) using state-of-the-art MD simulation protocols. In chapter 3, we have discussed the training sequences whose palindromic nature (for dsDNA and dsRNA) allows quantifying the convergence error in the MD time-series. By defining a *scale* (which quantifies variation over sequence) as the average pair-wise distance/divergence between pdfs for all training sequences for a given dsNA, we have demonstrated that 10 μs of MD time-series for each sequence is sufficient with convergence error almost two orders smaller than *scale*. Moreover, we found that the distributions of internal coordinates for dsDNA are often non-Gaussian, particularly for phosphates and some inter base-pair coordinates. In contrast, for dsRNA, the corresponding distributions of internal coordinates are close to Gaussian. Most interestingly, DRH behavior is between dsDNA and dsRNA, with the DNA strand similar to pure dsDNA and the RNA strand similar to pure dsRNA. Notably, the cgNA+ model assumes that the internal coordinates follow a Gaussian behavior, i.e., the cgNA+ model energy is quadratic. Lastly, we have shown that the corresponding Gaussian approximation error is negligible, with a few exceptions in the phosphate coordinates.

In chapter 4, we have introduced the cgNA+ parameter sets and illustrated that the model predictions are almost indistinguishable from the corresponding MD statistics (first and second moments) by assessing the model for diverse test sequences, including sequences with exceptional mechanical behavior (e.g., A-tracts) and showing that the model accurately captures changes in groundstate due to change in the hexamer context or beyond (highly nonlocal change). We have quantified the error due to various modeling assumptions and showed that the largest error in the model originates from the sequence locality assumption (dinucleotide dependence) in the parameter set; however, the model is highly accurate with a prediction/reconstruction error one order smaller than *scale* (which quantifies variation over sequence). Furthermore, we have presented a systematic and rigorous comparison of various observables, including average shape, persistence length, and groove widths for dsDNA, dsRNA, and DRH. It is worth highlighting that some of these observables, such as average shape and groove widths, can be obtained using MD simulations but only for few short sequences, while it is almost infeasible to estimate persistence length even for a single sequence (of length greater than 100 bps) using atomistic MD simulations.

Firstly at length scales at which a few MD simulations can be performed, we confirmed that the model predictions are extremely close to the corresponding observations in MD simulations and agree well with the findings in prior literature. For instance, the average shape for various dimers in dsDNA and dsRNA are considerably different; Twist and Slide in dsDNA are higher than in dsRNA, whereas the trend is opposite for Roll. Moreover, the phosphate coordinates are dramatically different for the two dsNAs. The difference, in general, can be interpreted with the A-form and B-form geometry adopted by dsRNA and dsDNA, respectively. Notably, DRH adopts a mixed geometry (slightly closer to A-form) with base coordinates slightly closer to pure dsRNA and phosphate coordinates on the DNA strand closer to pure dsDNA and on the RNA strand closer to pure dsRNA. For persistence length, the general trends computed for approximately two million random sequences (of length 220 bps) are in order $\ell_p^{\text{RNA}} \gtrsim \ell_p^{\text{DNA}} \gtrsim \ell_p^{\text{DRH}}$ and $\ell_d^{\text{RNA}} \gtrsim \ell_d^{\text{DRH}} \gtrsim \ell_d^{\text{DNA}}$, where ℓ_p and ℓ_d are apparent and dynamic (obtained by factoring out

shape contributions from ℓ_p) persistence length. It is worth noting that both dsDNA and dsRNA exhibit a wide range of persistence lengths over sequence space, which is further complicated for DRH with much larger variation, for instance, ℓ_d is approximately 312 bps for poly(A) while 196 bps for poly(T). In general, the trends for persistence lengths agree fairly well with the limited experimental observations, except that the predictions of the cgNA+ model are larger. It should be noted that the tangent-tangent correlations obtained from the cgNA+ Monte Carlo code and MD simulations for shorter sequences (24mers) are incredibly close; thus, the discrepancy in persistence lengths between model predictions and the experimental consensus is not inherent to the cgNA+ model and might be due to the MD protocol, which differs strongly from the experimental setups. Regardless, this work provided a detailed insight into the persistence length spectrum of various dsNAs in the sequence space, and notably, such computation for even a single sequence of length greater than 200 bps is almost unfeasible using MD simulations.

Moreover, we found that groove widths, which play a significant role in indirect readout and other protein-DNA interactions, are highly sensitive to the sequence and showed that A/T rich sequences tend to have very narrow minor grooves (with the exception that the TA step is never present), whereas C/G rich sequences have minor groove widths almost twice as compared to A/T rich sequences. The major groove (wider than the minor groove) for dsDNA does not exhibit strong sequence preferences for extreme widths. In contrast, the minor groove is wider than the major groove in dsRNA and comparable to the major groove for DRH. Similar to dsDNA, extreme groove widths are also adopted by specific sequences in dsRNA and DRH as well. Lastly, we demonstrated that single nucleotide polymorphisms (SNPs) have a strongly non-local impact on the groundstate of dsDNA which depends on the kind of mutation as well as the flanking contexts. We revealed that the impact on groundstate due to various SNPs is in the order A$\longleftrightarrow$G < C$\longleftrightarrow$G < A$\longleftrightarrow$C < A$\longleftrightarrow$T and all of these SNPs are highly sensitive to flanking contexts. Notably, these statistics are obtained over millions of sequences in only a few hours on a standard laptop. Thus, in this chapter, we have illustrated the efficiency and potential of the cgNA+ model to explore the sequence-dependent structural and mechanical properties of various dsNAs.

The next chapter systematically compared the predictions of the cgNA+ model with the available protein-DNA X-ray crystal structure data for dimers in all flanking tetramer contexts. First, we have shown that the flanking tetramer context strongly influences the average shape of a given dimer in the X-ray data set and thus can not be ignored. Moreover, using hierarchical clustering, we have demonstrated that the trends in the sequence space for the dimer's average shape are similar in the two data sets. In a direct comparison, we found a reasonable agreement between the average shape and a close alignment in the direction of variation of the average shape over the sequence space. Furthermore, the directions of dsDNA deformations in configuration space are very close in the two data sets, with an excellent correlation between the non-local sequence-dependent configurational volume (a measure of DNA deformability). Lastly and most interestingly, we found a striking alignment between the direction of variation of groundstate in sequence space and the direction of dsDNA deformation in configuration space, implying that the dimer adopts minimum energy configurations for various sequences/flanking contexts by compromising more in the soft modes of configuration space.

In chapter 6, we have extended the cgNA+ model to include alphabets for epigenetically modified bases, in particular, for 5-methylated or 5-hydroxymethylated cytosine in CpG steps, by training parameters for additional dinucleotide steps (to standard dsDNA) using MD statistics obtained from a palindromic library containing twelve 24mers. We have demonstrated that the model accuracy is similar to that obtained in predicting a Gaussian pdf for the standard dsDNA sequence. The model can capture the non-local change in groundstate due to epigenetic base modifications (arguably a smaller change than a point mutation). The change in groundstate as a result of either methylation or hydroxymethylation is similar and highly sensitive to the flanking sequence contexts. Moreover, a rigorous analysis of change in groundstate upon methylation or hydroxymethylated revealed that the minimum change in groundstate is when the C to be modified is preceded by A and the maximum change is when CpG step is present in the C/G context. This implies that the CpG modifications are likely to have a much larger influence on the groundstate in CpG islands. Furthermore, we found that increasing base modifications lead to a widening of the minor groove and depend on the modification position. Lastly, contrary to general belief, we found that apparent persistence lengths decrease upon symmetric modification of CpG steps, whereas dynamic persistence remains almost the same. However, asymmetric modifications of the CpG steps lead to an increase in persistence length. Also, in general, the influence of both types of modification (methylation or hydroxymethylation) on persistence length is similar.

In the final chapter of the thesis, we have introduced a neural network module to predict the atomistic position of the sugar atoms from the knowledge of adjacent phosphate and base atoms. The module is trained on the same MD simulation data used for training the cgNA+ model. It allows fine-graining any cgNA+ coarse-grained configuration or generating an ensemble of atomistic configurations for any sequence comparable to MD simulations but in a very short time. In particular, we have shown that this module can generate an ensemble of 10^5 configurations for a 24mer within an hour with statistics comparable to 10 μs of MD simulations which take approximately two months on a highly efficient GPU. It enables an accurate analysis of sequence-dependent sugar-pucker modes and backbone configuration for any sequence in negligible time. Furthermore, a fine-grain sequence-dependent equilibrium structure can be used to start MD simulations, particularly useful for dsDNA mini-circles.

Thus, with the overarching goal of widening the impact and applicability of the cgDNA+ model, we have extended it to the cgNA+ model, which allows predicting a non-local sequence-dependent Gaussian pdf for any dsDNA (with epigenetic base modifications), dsRNA, or DRH sequence and an additional machine learning tool has been developed to predict the positions of sugar atoms in any cgNA+ configuration. Moreover, we have demonstrated that the error in the model prediction for mechanically diverse test sequences is negligible, and for dsDNA, we have also shown that the model predictions are in reasonable agreement with the available protein-DNA X-ray structure data for both the average shape and stiffness. Lastly, we have illustrated the model efficiency by exploring applications such as persistence length, groove widths, the impact of SNPs and the role of flanking contexts, and the impact of base-modifications and the role of flanking contexts for millions of sequences, which is otherwise unfeasible.

8.2 Future work

There are many possible further extensions of the cgNA+ model that could be of practical interest. For example,

1. In terms of implementation, the current machine learning module to predict the location of sugar atoms is only developed for dsDNA and will be particularly useful to extend it for dsDNA with epigenetically modified bases which have a significant impact on the backbone conformation and sugar puckering modes [10, 107]. Similarly, it could also be extended for dsRNA and DRH for completeness.

2. In this work, we have only compared model prediction with the available protein-DNA X-ray data in base coordinates, and it would be interesting to extend the comparison for phosphate coordinates which are currently ignored in this work due to the scarcity of the experimental data and multi-modal behavior of phosphate coordinates. Similarly, there are not enough experimental data for dsRNA, DRH, or epigenetically modified dsDNA to make such a comparison.

3. One particular direction in which the cgNA+ parameter sets should be extended is to allow both methylated and hydroxymethylated CpG steps in the same sequence (currently allowed but not adjacent to each other) that are frequent in biology. It will require computing additional parameters for dimer steps such as NH and KM, which is relatively simple, but ensuring a positive-definite reconstruction of stiffness matrix for any sequence is a challenging task and is in progress. Furthermore, parameters for other base modifications can also be added for further generalisability of the cgNA+ model. In particular, methylation or hydroxymethylation of GpC steps is relatively rare but has a potential role in regulating mitochondrial gene expression. However, note that any such extension leads to considerable expansion of the parameter set, which requires comprehensive MD simulation data to train those parameters, and lastly, considerable effort to ensure that the predicted stiffness for any sequence will be positive-definite. Moreover, the extension of the model parameter sets for other rare base modifications such as 5-formyl-C, 5-carboxyl-C, and N6-carboxymethyl-A is also limited by the availability of reliable MD forcefields.

4. Another interesting extension of the cgNA+ model is to include parameters for DNA mismatches (when two non-complementary bases align in the same base-pair, e.g., A aligns with C or A or G). Such mismatches are frequent in biology and can occur during DNA replication and due to ionizing radiation, mutagenic chemicals, or spontaneous deamination. The inclusion of parameters for DNA mismatches would help better understand how DNA mismatches influence the local mechanics of the DNA and provide insights into DNA mismatch repair.

5. Another avenues for future research lie in the development of tools for various applications of the model. For instance, T. Zwahlen, in his thesis, developed cgDNA+loc to scan the whole genome, and thus, identified exceptional sequences in various genomes. Similar efforts are required in several other directions (some of them are ongoing in the

LCVMM or collaboration) at the time of thesis writing. One particular ongoing project is the easy implementation of the method that predicts sequence-dependent equilibria for dsDNA (or for any dsNAs) mini-circles of various lengths using the groundstate predicted by the cgNA+ model. Another interesting problem is to compute the energy required for a linear dsDNA fragment to wrap around the nucleosome core particle and to understand the role of sequence in dsDNA wrapping energy and the changes induced by the epigenetic modifications (which are often related to gene silencing). Moreover, the deformation of dsNAs on the application of external loads such as pulling and twisting is another exciting application actively pursued in the LCVMM group. Other potential applications include modeling protein-DNA interactions.

6. Lastly, to further expand the impact of the cgNA+ model, one of the ongoing projects includes the incorporation of sequence-dependent mechanics of dsDNA (using cgNA+) in the oxDNA model [143, 190] which is a successful model for studying the mechanical [183] and thermodynamic properties [143] of large DNA nanostructures. This will allow fine-tuning of DNA nanostructures or origamis, which have potential applications in molecular machine and drug delivery, catalysis, and biophysics.

REFERENCES

[1] J. Abels, F. Moreno-Herrero, T. Van der Heijden, C. Dekker, and N. H. Dekker. Single-molecule measurements of the persistence length of double-stranded RNA. *Biophysical Journal*, 88(4):2737–2744, 2005.

[2] R. L. Adams. *The biochemistry of the nucleic acids*. Springer Science and Business Media, 2012.

[3] S. Adhya. Multipartite genetic control elements: communication by DNA loop. *Annual Review of Genetics*, 23(1):227–250, 1989.

[4] C. T. Altona and M. Sundaralingam. Conformational analysis of the sugar ring in nucleosides and nucleotides. New description using the concept of pseudorotation. *Journal of the American Chemical Society*, 94(23):8205–8212, 1972.

[5] H. C. Andersen. Rattle: A "velocity" version of the shake algorithm for molecular dynamics calculations. *Journal of Computational Physics*, 52(1):24–34, 1983.

[6] M. J. Araúzo-Bravo, S. Fujii, H. Kono, S. Ahmad, and A. Sarai. Sequence-dependent conformational energy of DNA derived from molecular dynamics simulations: Toward understanding the indirect readout mechanism in protein- DNA recognition. *Journal of the American Chemical Society*, 127(46):16074–16089, 2005.

[7] S. Arnott and D. Hukins. Optimised parameters for A-DNA and B-DNA. *Biochemical and Biophysical Research Communications*, 47(6):1504–1509, 1972.

[8] S. Arnott, D. Hukins, S. Dover, W. Fuller, and A. Hodgson. Structures of synthetic polynucleotides in the A-RNA and A'-RNA conformations: X-ray diffraction analyses of the molecular conformations of polyadenylic acid· polyuridylic acid and polyinosinic acid· polycytidylic acid. *Journal of Molecular Biology*, 81(2):107–122, 1973.

[9] A. Balaceanu, D. Buitrago, J. Walther, A. Hospital, P. D. Dans, and M. Orozco. Modulation of the helical properties of DNA : next-to-nearest neighbour effects and beyond. *Nucleic Acids Research*, 47(9):4418–4430, 2019.

[10] M. Banyay and A. Gräslund. Structural effects of cytosine methylation on DNA sugar pucker studied by FTIR. *Journal of Molecular Biology*, 324(4):667–676, 2002.

[11] A. Basu, D. G. Bobrovnikov, Z. Qureshi, T. Kayikcioglu, T. Ngo, A. Ranjan, S. Eustermann, B. Cieza, M. T. Morgan, M. Hejna, et al. Measuring DNA mechanics on the genome scale. *Nature*, 589(7842):462–467, 2021.

[12] F. Battistini, P. D. Dans, M. Terrazas, C. L. Castellazzi, G. Portella, M. Labrador, N. Villegas, I. Brun-Heath, C. González, and M. Orozco. The impact of the hydroxymethylcytosine epigenetic signature on DNA structure and function. *PLoS Computational Biology*, 17(11):e1009547, 2021.

[13] M. Beaud, R. Singh, and J. H. Maddocks. Using the cgDNA+ model to compute sequence-dependent shapes of DNA minicircles. Master's thesis, EPFL, 2021.

[14] N. B. Becker, L. Wolff, and R. Everaers. Indirect readout: detection of optimized subsequences and calculation of relative binding affinities using different DNA elastic potentials. *Nucleic Acids Research*, 34(19):5638–5649, 2006.

[15] J. Bednar, P. Furrer, V. Katritch, A. Stasiak, J. Dubochet, and A. Stasiak. Determination of DNA persistence length by cryo-electron microscopy. Separation of the static and dynamic contributions to the apparent persistence length of DNA. *Journal of Molecular Biology*, 254(4):579–594, 1995.

[16] A. Bendandi, A. S. Patelli, A. Diaspro, and W. Rocchia. The role of histone tails in nucleosome stability: An electrostatic perspective. *Computational and Structural Biotechnology Journal*, 18:2799–2809, 2020.

[17] H. Berendsen, J. Grigera, and T. Straatsma. The missing term in effective pair potentials. *Journal of Physical Chemistry*, 91(24):6269–6271, 1987.

[18] H. J. Berendsen, J. V. Postma, W. F. Van Gunsteren, A. DiNola, and J. R. Haak. Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics*, 81(8): 3684–3690, 1984.

[19] S. L. Berger, T. Kouzarides, R. Shiekhattar, and A. Shilatifard. An operational definition of epigenetics. *Genes and Development*, 23(7):781–783, 2009.

[20] H. M. Berman, T. Battistuz, T. N. Bhat, W. F. Bluhm, P. E. Bourne, K. Burkhardt, Z. Feng, G. L. Gilliland, L. Iype, S. Jain, et al. The protein data bank. *Acta Crystallographica Section D: Biological Crystallography*, 58(6):899–907, 2002.

[21] H.-O. Bertrand, S. Fermandjian, T. Ha-Duong, and B. Hartmann. Flexibility of the B-DNA backbone: effects of local and neighbouring sequences on pyrimidine-purine steps. *Nucleic Acids Research*, 26(5):1261–1267, 1998.

[22] D. L. Beveridge, G. Barreiro, K. S. Byun, D. A. Case, T. E. Cheatham III, S. B. Dixit, E. Giudice, F. Lankas, R. Lavery, J. H. Maddocks, et al. Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA oligonucleotides. I. Research design and results on d(CpG) steps. *Biophysical Journal*, 87(6):3799–3813, 2004.

[23] D. Botstein, R. L. White, M. Skolnick, and R. W. Davis. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American journal of Human Genetics*, 32(3):314, 1980.

[24] A. Brambati, L. Zardoni, E. Nardini, A. Pellicioli, and G. Liberi. The dark side of RNA:DNA hybrids. *Mutation Research/Reviews in Mutation Research*, 784:108300, 2020.

[25] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. a. Swaminathan, and M. Karplus. CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry*, 4(2):187–217, 1983.

[26] B. R. Brooks, C. L. Brooks III, A. D. Mackerell Jr, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, et al. CHARMM: the biomolecular simulation program. *Journal of Computational Chemistry*, 30(10):1545–1614, 2009.

[27] A. Brunet, C. Tardin, L. Salomé, P. Rousseau, N. Destainville, and M. Manghi. Dependence of DNA persistence length on ionic strength of solutions with monovalent and divalent salts: a joint theory–experiment study. *Macromolecules*, 48(11):3641–3652, 2015.

[28] A. T. Carvalho, L. Gouveia, C. R. Kanna, S. K. Wärmländer, J. A. Platts, and S. C. L. Kamerlin. Understanding the structural and dynamic consequences of DNA epigenetic modifications: Computational insights into cytosine methylation and hydroxymethylation. *Epigenetics*, 9(12):1604–1612, 2014.

[29] D. Case, I. Ben-Shalom, S. Brozell, D. Cerutti, T. Cheatham III, V. Cruzeiro, T. Darden, R. Duke, D. Ghoreishi, M. Gilson, et al. AMBER 2018; 2018. *University of California, San Francisco*.

[30] D. A. Case, T. E. Cheatham III, T. Darden, H. Gohlke, R. Luo, K. M. Merz Jr, A. Onufriev, C. Simmerling, B. Wang, and R. J. Woods. The Amber biomolecular simulation programs. *Journal of Computational Chemistry*, 26(16):1668–1688, 2005.

[31] D. A. Case et al. Amber 2021: Reference manual. Workingpaper, University of California Press, United States, June 2021.

[32] H. Cedar and Y. Bergman. Linking DNA methylation and histone modification: patterns and paradigms. *Nature Reviews Genetics*, 10(5):295–304, 2009.

[33] D. Chakraborty, N. Hori, and D. Thirumalai. Sequence-dependent three interaction site model for single-and double-stranded DNA. *Journal of Chemical Theory and Computation*, 14(7):3763–3779, 2018.

[34] T. E. Cheatham and P. A. Kollman. Molecular dynamics simulations highlight the structural differences among DNA:DNA , RNA:RNA, and DNA:RNA hybrid duplexes. *Journal of the American Chemical Society*, 119(21):4805–4825, 1997.

[35] T. E. Cheatham III, P. Cieplak, and P. A. Kollman. A modified version of the cornell et al. force field with improved sugar pucker phases and helical repeat. *Journal of Biomolecular Structure and Dynamics*, 16(4):845–862, 1999.

[36] S. Chen, A. Gunasekera, X. Zhang, T. A. Kunkel, R. H. Ebright, and H. M. Berman. Indirect readout of DNA sequence at the primary-kink site in the CAP-DNA complex: alteration of DNA binding specificity through alteration of DNA kinking. *Journal of Molecular Biology*, 314(1):75–82, 2001.

[37] H. Cheng, R. D. Schaeffer, Y. Liao, L. N. Kinch, J. Pei, S. Shi, B.-H. Kim, and N. V. Grishin. ECOD: an evolutionary classification of protein domains. *PLoS Computational Biology*, 10(12):e1003926, 2014.

[38] R. Cortini, M. Barbi, B. R. Caré, C. Lavelle, A. Lesne, J. Mozziconacci, and J.-M. Victor. The physics of epigenetics. *Reviews of Modern Physics*, 88(2):025002, 2016.

[39] G. Da Rosa, L. Grille, V. Calzada, K. Ahmad, J. P. Arcon, F. Battistini, G. Bayarri, T. Bishop, P. Carloni, T. Cheatham III, et al. Sequence-dependent structural properties of B-DNA: what have we learned in 40 years? *Biophysical Reviews*, pages 1–11, 2021.

[40] L. X. Dang. Mechanism and thermodynamics of ion selectivity in aqueous solutions of 18-crown-6 ether: a molecular dynamics study. *Journal of the American Chemical Society*, 117(26):6954–6960, 1995.

[41] P. D. Dans, A. Pérez, I. Faustino, R. Lavery, and M. Orozco. Exploring polymorphisms in B-DNA helical conformations. *Nucleic Acids Research*, 40(21):10668–10678, 2012.

[42] P. D. Dans, I. Faustino, F. Battistini, K. Zakrzewska, R. Lavery, and M. Orozco. Unraveling the sequence-dependent polymorphic behavior of d(CpG) steps in B-DNA. *Nucleic Acids Research*, 42(18):11304–11320, 2014.

[43] P. D. Dans, I. Ivani, A. Hospital, G. Portella, C. González, and M. Orozco. How accurate are accurate force-fields for B-DNA? *Nucleic Acids Research*, 45(7):4217–4230, 2017.

[44] P. D. Dans, A. Balaceanu, M. Pasi, A. S. Patelli, D. Petkeviciūtė, J. Walther, A. Hospital, G. Bayarri, R. Lavery, J. H. Maddocks, et al. The static and dynamic structural heterogeneities of B-DNA: extending Calladine–Dickerson rules. *Nucleic Acids Research*, 47 (21):11090–11102, 2019.

[45] A. C. Dantas Machado, T. Zhou, S. Rao, P. Goel, C. Rastogi, A. Lazarovici, H. J. Bussemaker, and R. Rohs. Evolving insights on how cytosine methylation affects protein–DNA binding. *Briefings in Functional Genomics*, 14(1):61–73, 2015.

[46] T. Darden, D. York, and L. Pedersen. Particle mesh Ewald: An N.log(N) method for Ewald sums in large systems. *The Journal of Chemical Physics*, 98(12):10089–10092, 1993.

[47] L. De Bruin and J. H. Maddocks. cgDNAweb: a web interface to the cgDNA sequence-dependent coarse-grain model of double-stranded DNA. *Nucleic Acids Research*, 46 (W1):W5–W10, 2018.

[48] D. Demurtas, A. Amzallag, E. J. Rawdon, J. H. Maddocks, J. Dubochet, and A. Stasiak. Bending modes of DNA directly addressed by cryo-electron microscopy of DNA minicircles. *Nucleic Acids Research*, 37(9):2882–2893, 2009.

[49] R. E. Dickerson. DNA bending: the prevalence of kinkiness and the virtues of normality. *Nucleic Acids Research*, 26(8):1906–1926, 1998.

[50] S. B. Dixit, D. L. Beveridge, D. A. Case, T. E. Cheatham 3rd, E. Giudice, F. Lankas, R. Lavery, J. H. Maddocks, R. Osman, H. Sklenar, 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. *Biophysical Journal*, 89(6):3721–3740, 2005.

[51] D. Djuranovic and B. Hartmann. DNA fine structure and dynamics in crystals and in solution: the impact of BI/BII backbone conformations. *Biopolymers: Original Research on Biomolecules*, 73(3):356–368, 2004.

[52] T. Drsata, A. Pérez, M. Orozco, A. V. Morozov, J. Sponer, and F. Lankas. Structure, stiffness and substates of the Dickerson-Drew dodecamer. *Journal of Chemical Theory and Computation*, 9(1):707–721, 2013.

[53] M. El Hassan and C. Calladine. Conformational characteristics of DNA : empirical classifications and a hypothesis for the conformational behaviour of dinucleotide steps. *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 355(1722):43–100, 1997.

[54] O. Y. Fedoroff, Y. Ge, and B. R. Reid. Solution structure sf r(gaggacug):d(CAGTCCTC) hybrid: implications for the initiation of HIV–1(+)–strand synthesis. *Journal of Molecular Biology*, 269(2):225–239, 1997.

[55] A. Fire, S. Xu, M. K. Montgomery, S. A. Kostas, S. E. Driver, and C. C. Mello. Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature*, 391(6669):806–811, 1998.

[56] S. Fujii, H. Kono, S. Takenaka, N. Go, and A. Sarai. Sequence-dependent DNA deformability studied using molecular dynamics simulations. *Nucleic Acids Research*, 35 (18):6063–6074, 2007.

[57] R. Galindo-Murillo, D. R. Roe, and T. E. Cheatham III. Convergence and reproducibility in molecular dynamics simulations of the DNA duplex d(GCACGAACGAACGAACGC). *Biochimica et Biophysica Acta (BBA)-General Subjects*, 1850(5):1041–1058, 2015.

[58] X. Gao and P. W. Jeffs. Sequence-dependent conformational heterogeneity of a hybrid DNA.RNA dodecamer duplex. *Journal of Biomolecular NMR*, 4(3):367–384, 1994.

[59] K. B. Geahigan, G. A. Meints, M. E. Hatcher, J. Orban, and G. P. Drobny. The dynamic impact of CpG methylation in DNA. *Biochemistry*, 39(16):4939–4946, 2000.

[60] J. Glowacki. *Computation and Visualization in Multiscale Modelling of DNA Mechanics*. PhD thesis, EPFL, 2016. Thesis number 7062.

[61] O. Gonzalez and J. H. Maddocks. Extracting parameters for base-pair level models of DNA from molecular dynamics simulations. *Theoretical Chemistry Accounts*, 106(1-2): 76–82, 2001.

[62] O. Gonzalez, D. Petkeviciute, and J. Maddocks. A sequence-dependent rigid-base model of DNA. *The Journal of Chemical Physics*, 138(5):02B604, 2013.

[63] O. Gonzalez, M. Pasi, D. Petkeviciute, J. Glowacki, and J. H. Maddocks. Absolute versus relative parameter estimation in a coarse-grain model of DNA. *Multiscale Modeling and Simulation*, 15(3):1073–1107, 2017.

[64] I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT Press, 2016.

[65] D. G. Gorenstein. Conformation and dynamics of DNA and protein-DNA complexes by 31P NMR. *Chemical Reviews*, 94(5):1315–1338, 1994.

[66] T. Goto and M. Monk. Regulation of X-chromosome inactivation in development in mice and humans. *Microbiology and Molecular Biology Reviews*, 62(2):362–378, 1998.

[67] L. Greengard and V. Rokhlin. A fast algorithm for particle simulations. *Journal of Computational Physics*, 73(2):325–348, 1987.

[68] P. J. Hagerman. Flexibility of DNA. *Annual Review of Biophysics and Biophysical Chemistry*, 17(1):265–286, 1988.

[69] B. Hartmann, D. Piazzola, and R. Lavery. BI-BII transitions in B-DNA. *Nucleic Acids Research*, 21(3):561–568, 1993.

[70] B. Heddi, C. Oguey, C. Lavelle, N. Foloppe, and B. Hartmann. Intrinsic flexibility of B-DNA: the experimental TRX scale. *Nucleic Acids Research*, 38(3):1034–1047, 2010.

[71] D. M. Hinckley, G. S. Freeman, J. K. Whitmer, and J. J. De Pablo. An experimentally-informed coarse-grained 3-site-per-nucleotide model of DNA: Structure, thermodynamics, and dynamics of hybridization. *The Journal of Chemical Physics*, 139(14):10B604_1, 2013.

[72] C. Hognon, V. Besancenot, A. Gruez, S. Grandemange, and A. Monari. Cooperative effects of cytosine methylation on DNA structure and dynamics. *The Journal of Physical Chemistry B*, 123(34):7365–7371, 2019.

[73] R. C. Holland, T. A. Down, M. Pocock, A. Prlić, D. Huen, K. James, S. Foisy, A. Dräger, A. Yates, M. Heuer, et al. BioJava: an open-source framework for bioinformatics. *Bioinformatics*, 24(18):2096–2097, 2008.

[74] A. Hospital, J. R. Goñi, M. Orozco, and J. L. Gelpí. Molecular dynamics simulations: advances and applications. *Advances and Applications in Bioinformatics and Chemistry*, 8:37, 2015.

[75] H. Huang, R. Chopra, G. L. Verdine, and S. C. Harrison. Structure of a covalently trapped catalytic complex of HIV-1 reverse transcriptase: implications for drug resistance. *Science*, 282(5394):1669–1675, 1998.

[76] I. Ivani, P. D. Dans, A. Noy, A. Pérez, I. Faustino, A. Hospital, J. Walther, P. Andrio, R. Goñi, A. Balaceanu, et al. Parmbsc1: a refined force field for DNA simulations. *Nature Methods*, 13(1):55, 2016.

[77] K. Jabbari and G. Bernardi. Cytosine methylation and CpG, TpG (CpA) and TpA frequencies. *Gene*, 333:143–149, 2004.

[78] E. T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 106(4):620, 1957.

[79] E. T. Jaynes. Information theory and statistical mechanics. II. *Physical Review*, 108(2): 171, 1957.

[80] H. Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186 (1007):453–461, 1946.

[81] I. Jimenez-Useche, D. Shim, J. Yu, and C. Yuan. Unmethylated and methylated CpG dinucleotides distinctively regulate the physical properties of DNA. *Biopolymers*, 101 (5):517–524, 2014.

[82] S. Jones, P. Van Heyningen, H. M. Berman, and J. M. Thornton. Protein-DNA interactions: a structural analysis. *Journal of Molecular Biology*, 287(5):877–896, 1999.

[83] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, 79(2):926–935, 1983.

[84] I. S. Joung and T. E. Cheatham III. Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *The Journal of Physical Chemistry B*, 112(30):9020–9041, 2008.

[85] Z. S. Juo, T. K. Chiu, P. M. Leiberman, I. Baikalov, A. J. Berk, and R. E. Dickerson. How proteins recognize the TATA box. *Journal of Molecular Biology*, 261(2):239–254, 1996.

[86] S. U. Kass, D. Pruss, and A. P. Wolffe. How does DNA methylation repress transcription? *Trends in Genetics*, 13(11):444–449, 1997.

[87] T. K. Kelly, D. D. De Carvalho, and P. A. Jones. Epigenetic modifications as therapeutic targets. *Nature Biotechnology*, 28(10):1069–1078, 2010.

[88] C. L. Kielkopf, S. Ding, P. Kuhn, and D. C. Rees. Conformational flexibility of B-DNA at 0.74 å resolution: d(CCAGTACTGG)2. *Journal of Molecular Biology*, 296(3):787–801, 2000.

[89] J. L. Kim, D. B. Nikolov, and S. K. Burley. Co-crystal structure of TBP recognizing the minor groove of a TATA element. *Nature*, 365(6446):520–527, 1993.

[90] Y. Kim, J. H. Geiger, S. Hahn, and P. B. Sigler. Crystal structure of a yeast TBP/TATA-box complex. *Nature*, 365(6446):512–520, 1993.

[91] M. Ko, Y. Huang, A. M. Jankowska, U. J. Pape, M. Tahiliani, H. S. Bandukwala, J. An, E. D. Lamperti, K. P. Koh, R. Ganetzky, et al. Impaired hydroxylation of 5-methylcytosine in myeloid cancers with mutant TET2. *Nature*, 468(7325):839–843, 2010.

[92] O. Kratky and G. Porod. Röntgenuntersuchung gelöster fadenmoleküle. *Recueil des Travaux Chimiques des Pays-Bas*, 68(12):1106–1122, 1949.

[93] M. Kulis and M. Esteller. DNA methylation and cancer. *Advances in Genetics*, 70:27–56, 2010.

[94] S. Kullback. Information Theory and Statistics. John Riley and Sons. *Inc. New York*, 1959.

[95] A. Lafita, S. Bliven, A. Prlić, D. Guzenko, P. W. Rose, A. Bradley, P. Pavan, D. Myers-Turnbull, Y. Valasatava, M. Heuer, et al. BioJava 5: A community driven open-source bioinformatics library. *PLoS Computational Biology*, 15(2):e1006791, 2019.

[96] A. N. Lane, S. Ebel, and T. Brown. NMR assignments and solution conformation of the DNA.RNA hybrid duplex d(GTGAACTT).r(AAGUUCAC). *European Journal of Biochemistry*, 215(2):297–306, 1993.

[97] F. Lankas, J. Sponer, J. Langowski, and T. E. Cheatham III. DNA basepair step deformability inferred from molecular dynamics simulations. *Biophysical Journal*, 85(5): 2872–2883, 2003.

[98] F. Lankas, R. Lavery, and J. H. Maddocks. Kinking occurs during molecular dynamics simulations of small DNA minicircles. *Structure*, 14(10):1527–1534, 2006.

[99] F. Lankas, O. Gonzalez, L. Heffler, G. Stoll, M. Moakher, and J. H. Maddocks. On the parameterization of rigid base and basepair models of DNA from molecular dynamics simulations. *Physical Chemistry Chemical Physics*, 11(45):10565–10588, 2009.

[100] R. Lavery, K. Zakrzewska, and H. Sklenar. JUMNA (junction minimisation of nucleic acids). *Computer Physics Communications*, 91(1-3):135–158, 1995.

[101] R. Lavery, M. J. H. P. D. Moakher, M., and K. Zakrzewska. Conformational analysis of nucleic acids revisited: Curves+. *Nucleic Acids Research*, 37(17):5917–5929, 2009.

[102] R. Lavery, K. Zakrzewska, D. Beveridge, T. C. Bishop, D. A. Case, T. Cheatham III, S. Dixit, B. Jayaram, F. Lankas, C. Laughton, 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 Research*, 38(1):299–313, 2010.

[103] A. Lebrun, Z. Shakked, and R. Lavery. Local DNA stretching mimics the distortion caused by the TATA box-binding protein. *Proceedings of the National Academy of Sciences*, 94(7):2993–2998, 1997.

[104] A. Lefebvre, O. Mauffret, E. Lescot, B. Hartmann, and S. Fermandjian. Solution structure of the CpG containing d(CTTCGAAG)2 oligonucleotide: NMR data and energy calculations are compatible with a BI/BII equilibrium at CpG. *Biochemistry*, 35(38): 12560–12569, 1996.

[105] M. Levitt. Computer simulation of DNA double-helix dynamics. In *Cold Spring Harbor symposia on quantitative biology*, volume 47, pages 251–262. Cold Spring Harbor Laboratory Press, 1983.

[106] S. Li, Y. Peng, D. Landsman, and A. R. Panchenko. DNA methylation cues in nucleosome geometry, stability and unwrapping. *Nucleic Acids Research*, 50(4):1864–1874, 2022.

[107] K. Liebl and M. Zacharias. How methyl–sugar interactions determine DNA structure and flexibility. *Nucleic Acids Research*, 47(3):1132–1140, 2019.

[108] J.-H. Liu, K. Xi, X. Zhang, L. Bao, X. Zhang, and Z.-J. Tan. Structural flexibility of DNA-RNA hybrid duplex: stretching and twist-stretch coupling. *Biophysical Journal*, 117(1):74–86, 2019.

[109] X.-J. Lu and W. K. Olson. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Research*, 31 (17):5108–5121, 2003.

[110] D. R. Mack, T. K. Chiu, and R. E. Dickerson. Intrinsic bending and deformability at the TA step of CCTTTAAAGG: a comparative analysis of TA and AT steps within A-tracts. *Journal of Molecular Biology*, 312(5):1037–1049, 2001.

[111] A. Madhumalar and M. Bansal. Sequence preference for BI/BII conformations in DNA: MD and crystal structure data analysis. *Journal of Biomolecular Structure and Dynamics*, 23(1):13–27, 2005.

[112] T. Maehigashi, C. Hsiao, K. Kruger Woods, T. Moulaei, N. V. Hud, and L. Dean Williams. B-DNA structure is intrinsically polymorphic: even at the level of base pair positions. *Nucleic Acids Research*, 40(8):3714–3722, 2012.

[113] P. C. Mahalanobis. On the generalized distance in statistics. National Institute of Science of India, 1936.

[114] M. Mandal and R. R. Breaker. Gene regulation by riboswitches. *Nature Reviews Molecular Cell Biology*, 5(6):451–463, 2004.

[115] A. Marin-Gonzalez, J. Vilhena, F. Moreno-Herrero, and R. Perez. DNA crookedness regulates DNA mechanical properties at short length scales. *Physical Review Letters*, 122(4):048102, 2019.

[116] A. Marin-Gonzalez, J. Vilhena, F. Moreno-Herrero, and R. Perez. Sequence-dependent mechanical properties of double-stranded RNA. *Nanoscale*, 11(44):21471–21478, 2019.

[117] A. Marin-Gonzalez, C. Aicart-Ramos, M. Marin-Baquero, A. Martín-González, M. Suomalainen, A. Kannan, J. Vilhena, U. F. Greber, F. Moreno-Herrero, and R. Pérez. Double-stranded RNA bending by AU-tract sequences. *Nucleic Acids Research*, 48(22):12917–12928, 2020.

[118] A. Mathelier, B. Xin, T.-P. Chiu, L. Yang, R. Rohs, and W. W. Wasserman. DNA shape features improve transcription factor binding site predictions in vivo. *Cell systems*, 3(3): 278–286, 2016.

[119] A. W. Mauney, J. M. Tokuda, L. M. Gloss, O. Gonzalez, and L. Pollack. Local DNA sequence controls asymmetry of DNA unwrapping from nucleosome core particles. *Biophysical Journal*, 115(5):773–781, 2018.

[120] S. P. Meisburger, J. L. Sutton, H. Chen, S. A. Pabit, S. Kirmizialtin, R. Elber, and L. Pollack. Polyelectrolyte properties of single stranded DNA measured using SAXS and single-molecule FRET: Beyond the wormlike chain model. *Biopolymers*, 99(12): 1032–1045, 2013.

[121] M. Meselson and F. W. Stahl. The replication of DNA in Escherichia coli. *Proceedings of the National Academy of Sciences*, 44(7):671–682, 1958.

[122] F. Miescher-Rüsch. *Ueber die chemische Zusammensetzung der Eiterzellen*. 1871.

[123] J. S. Mitchell, J. Glowacki, A. E. Grandchamp, R. S. Manning, and J. H. Maddocks. Sequence-dependent persistence lengths of DNA. *Journal of Chemical Theory and Computation*, 13(4):1539–1555, 2017.

[124] D. Müllner. Modern hierarchical, agglomerative clustering algorithms. *arXiv preprint arXiv:1109.2378*, 2011.

[125] M. Münzel, D. Globisch, T. Brückl, M. Wagner, V. Welzmiller, S. Michalakis, M. Müller, M. Biel, and T. Carell. Quantification of the sixth DNA base hydroxymethylcytosine in the brain. *Angewandte Chemie International Edition*, 49(31):5375–5377, 2010.

[126] A. A. Napoli, C. L. Lawson, R. H. Ebright, and H. M. Berman. Indirect readout of DNA sequence at the primary-kink site in the CAP–DNA complex: Recognition of pyrimidine-purine and purine-purine steps. *Journal of Molecular Biology*, 357(1):173–183, 2006.

[127] S. Neidle. *Principles of Nucleic Acid Structure*. Elsevier, 2018.

[128] S. Neidle. Beyond the double helix: DNA structural diversity and the PDB. *Journal of Biological Chemistry*, page 100553, 2021.

[129] L. Nekludova and C. O. Pabo. Distinctive DNA conformation with enlarged major groove is found in Zn-finger-DNA and other protein-DNA complexes. *Proceedings of the National Academy of Sciences*, 91(15):6948–6952, 1994.

[130] T. T. Ngo, J. Yoo, Q. Dai, Q. Zhang, C. He, A. Aksimentiev, and T. Ha. Effects of cytosine modifications on DNA flexibility and nucleosome mechanical stability. *Nature Communications*, 7:10813, 2016.

[131] A. Noy, A. Pérez, F. Lankas, F. J. Luque, and M. Orozco. Relative flexibility of DNA and RNA: a molecular dynamics study. *Journal of Molecular Biology*, 343(3):627–638, 2004.

[132] A. Noy, A. Pérez, M. Márquez, F. J. Luque, and M. Orozco. Structure, recognition properties, and flexibility of the DNA.RNA hybrid. *Journal of the American Chemical Society*, 127(13):4910–4920, 2005.

[133] A. Noy, F. J. Luque, and M. Orozco. Theoretical analysis of antisense duplexes: determinants of the RNase H susceptibility. *Journal of the American Chemical Society*, 130(11): 3486–3496, 2008.

[134] C. Oguey, N. Foloppe, and B. Hartmann. Understanding the sequence-dependence of DNA groove dimensions: implications for DNA interactions. *PLoS One*, 5(12):e15931, 2010.

[135] C. Ohle, R. Tesorero, G. Schermann, N. Dobrev, I. Sinning, and T. Fischer. Transient RNA-DNA hybrids are required for efficient double-strand break repair. *Cell*, 167(4): 1001–1013, 2016.

[136] T. Okonogi, S. Alley, A. Reese, P. Hopkins, and B. Robinson. Sequence-dependent dynamics in duplex DNA. *Biophysical Journal*, 78(5):2560–2571, 2000.

[137] T. Okonogi, S. Alley, A. Reese, P. Hopkins, and B. Robinson. Sequence-dependent dynamics of duplex DNA: the applicability of a dinucleotide model. *Biophysical Journal*, 83(6):3446–3459, 2002.

[138] W. K. Olson and V. B. Zhurkin. Working the kinks out of nucleosomal DNA. *Current Opinion in Structural Biology*, 21(3):348–357, 2011.

[139] W. K. Olson, A. A. Gorin, X. J. Lu, L. M. Hock, and V. B. Zhurkin. DNA sequence-dependent deformability deduced from protein–DNA crystal complexes. *Proceedings of the National Academy of Sciences*, 95(19):11163–11168, 1998.

[140] W. K. Olson, M. Bansal, S. K. Burley, R. E. Dickerson, M. Gerstein, S. C. Harvey, U. Heinemann, X. Lu, S. Neidle, Z. Shakked, et al. A standard reference frame for the description of nucleic acid base-pair geometry. *Journal of Molecular Biology*, 313(1): 229–237, 2001.

[141] W. K. Olson, A. V. Colasanti, Y. Li, W. Ge, G. Zheng, and V. B. Zhurkin. DNA simulation benchmarks as revealed by X-ray structures. In *Computational Studies of RNA and DNA*, pages 235–257. Springer, 2006.

[142] M. Orozco, A. Pérez, A. Noy, and F. J. Luque. Theoretical methods for the simulation of nucleic acids. *Chemical Society Reviews*, 32(6):350–364, 2003.

[143] T. E. Ouldridge, A. A. Louis, and J. P. Doye. Structural, mechanical, and thermodynamic properties of a coarse-grained DNA model. *The Journal of Chemical Physics*, 134(8): 02B627, 2011.

[144] M. J. Packer and C. A. Hunter. Sequence-dependent DNA structure: the role of the sugar-phosphate backbone. *Journal of Molecular Biology*, 280(3):407–420, 1998.

[145] M. J. Packer, M. P. Dauncey, and C. A. Hunter. Sequence-dependent DNA structure: tetranucleotide conformational maps. *Journal of Molecular Biology*, 295(1):85–103, 2000.

[146] G. Paillard and R. Lavery. Analyzing protein-DNA recognition mechanisms. *Structure*, 12(1):113–122, 2004.

[147] M. Pasi, J. H. Maddocks, D. Beveridge, T. C. Bishop, D. A. Case, T. Cheatham III, P. D. Dans, B. Jayaram, F. Lankas, C. Laughton, et al. μABC: a systematic microsecond molecular dynamics study of tetranucleotide sequence effects in B-DNA. *Nucleic Acids Research*, 42(19):12272–12283, 2014.

[148] A. Pataskar, W. Vanderlinden, J. Emmerig, A. Singh, J. Lipfert, and V. K. Tiwari. Deciphering the gene regulatory landscape encoded in DNA biophysical features. *iScience*, 21:638–649, 2019.

[149] A. Patelli. *A sequence-dependent coarse-grain model of B-DNA with explicit description of bases and phosphate groups parametrised from large scale Molecular Dynamics simulations*. PhD thesis, EPFL, 2019. Thesis number 9552.

[150] D. A. Pearlman, D. A. Case, J. W. Caldwell, W. S. Ross, T. E. Cheatham III, S. DeBolt, D. Ferguson, G. Seibel, and P. Kollman. AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Computer Physics Communications*, 91(1-3):1–41, 1995.

[151] S. Pennings, J. Allan, and C. S. Davey. DNA methylation, nucleosome formation and positioning. *Briefings in Functional Genomics*, 3(4):351–361, 2005.

[152] A. Pérez, A. Noy, F. Lankas, F. J. Luque, and M. Orozco. The relative flexibility of B-DNA and A-RNA duplexes: database analysis. *Nucleic Acids Research*, 32(20):6144–6151, 2004.

[153] A. Pérez, F. J. Luque, and M. Orozco. Dynamics of B-DNA on the microsecond time scale. *Journal of the American Chemical Society*, 129(47):14739–14745, 2007.

[154] A. Pérez, I. Marchán, D. Svozil, J. Sponer, T. E. Cheatham III, C. A. Laughton, and M. Orozco. Refinement of the AMBER force field for nucleic acids: improving the description of α/γ conformers. *Biophysical Journal*, 92(11):3817–3829, 2007.

[155] A. Pérez, F. Lankas, F. J. Luque, and M. Orozco. Towards a molecular dynamics consensus view of B-DNA flexibility. *Nucleic Acids Research*, 36(7):2379–2394, 2008.

[156] A. Pérez, C. L. Castellazzi, F. Battistini, K. Collinet, O. Flores, O. Deniz, M. L. Ruiz, D. Torrents, R. Eritja, M. Soler-López, et al. Impact of methylation on the physical properties of DNA. *Biophysical Journal*, 102(9):2140–2148, 2012.

[157] J. P. Peters and L. J. Maher. DNA curvature and flexibility in vitro and in vivo. *Quarterly Reviews of Biophysics*, 43(1):23–63, 2010.

[158] D. Petkeviciute. *A DNA coarse-grain rigid base model and parameter estimation from molecular dynamics simulations*. PhD thesis, EPFL, 2012. Thesis number 5520.

[159] D. Petkeviciute, M. Pasi, O. Gonzalez, and J. H. Maddocks. cgDNA: a software package for the prediction of sequence-dependent coarse-grain free energies of B-form DNA. *Nucleic Acids Research*, 42(20):e153–e153, 2014.

[160] C. I. Pongor, P. Bianco, G. Ferenczy, R. Kellermayer, and M. Kellermayer. Optical trapping nanometry of hypermethylated CPG-island DNA. *Biophysical Journal*, 112(3):512–522, 2017.

[161] A. Portela and M. Esteller. Epigenetic modifications and human disease. *Nature Biotechnology*, 28(10):1057–1068, 2010.

[162] G. Portella, F. Battistini, and M. Orozco. Understanding the connection between epigenetic DNA methylation and nucleosome positioning from computer simulations. *PLoS Computational Biology*, 9(11):e1003354, 2013.

[163] U. D. Priyakumar and A. D. MacKerell. Atomic detail investigation of the structure and dynamics of DNA.RNA hybrids: A molecular dynamics study. *The Journal of Physical Chemistry B*, 112(5):1515–1524, 2008.

[164] S. Rao, T. P. Chiu, J. F. Kribelbauer, R. S. Mann, H. J. Bussemaker, and R. Rohs. Systematic prediction of DNA shape changes due to CpG methylation explains epigenetic effects on protein–DNA binding. *Epigenetics and Chromatin*, 11(1):1–11, 2018.

[165] C. Rauch, M. Trieb, B. Wellenzohn, M. Loferer, A. Voegele, F. R. Wibowo, and K. R. Liedl. C5-methylation of cytosine in B-DNA thermodynamically and kinetically stabilizes BI. *Journal of the American Chemical Society*, 125(49):14990–14991, 2003.

[166] C. Rausch, P. Zhang, C. S. Casas-Delucchi, J. L. Daiß, C. Engel, G. Coster, F. D. Hastert, P. Weber, and M. C. Cardoso. Cytosine base modifications regulate DNA duplex stability and metabolism. *Nucleic Acids Research*, 49(22):12870–12894, 2021.

[167] A. Rich. A hybrid helix containing both deoxyribose and ribose polynucleotides and its relation to the transfer of information between the nucleic acids. In *The Excitement of Discovery: Selected Papers of Alexander Rich: A Tribute to Alexander Rich*, pages 63–72. World Scientific, 1960.

[168] R. W. Roberts and D. M. Crothers. Stability and properties of double and triple helices: dramatic effects of RNA or DNA backbone composition. *Science*, 258(5087):1463–1466, 1992.

[169] D. R. Roe and T. E. Cheatham III. PTRAJ and CPPTRAJ: software for processing and analysis of molecular dynamics trajectory data. *Journal of Chemical Theory and Computation*, 9(7):3084–3095, 2013.

[170] D. R. Roe and T. E. Cheatham III. Parallelization of CPPTRAJ enables large scale analysis of molecular dynamics trajectory data. *Journal of Computational Chemistry*, 39(25): 2110–2117, 2018.

[171] R. Rohs, S. M. West, A. Sosinsky, P. Liu, R. S. Mann, and B. Honig. The role of DNA shape in protein–DNA recognition. *Nature*, 461(7268):1248–1253, 2009.

[172] R. Rohs, X. Jin, S. M. West, R. Joshi, B. Honig, and R. S. Mann. Origins of specificity in protein-DNA recognition. *Annual Review of Biochemistry*, 79:233–269, 2010.

[173] J.-P. Ryckaert, G. Ciccotti, and H. J. Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics*, 23(3):327–341, 1977.

[174] S. G. Sarafianos, K. Das, C. Tantillo, A. D. Clark, J. Ding, J. M. Whitcomb, P. L. Boyer, S. H. Hughes, and E. Arnold. Crystal structure of HIV-1 reverse transcriptase in complex with a polypurine tract RNA:DNA. *The EMBO Journal*, 20(6):1449–1461, 2001.

[175] A. Savelyev. Do monovalent mobile ions affect DNA's flexibility at high salt content? *Physical Chemistry Chemical Physics*, 14(7):2250–2254, 2012.

[176] R. Schleif. DNA looping. *Annual Review of Biochemistry*, 61(1):199–223, 1992.

[177] B. Schneider, S. Neidle, and H. M. Berman. Conformations of the sugar-phosphate backbone in helical DNA crystal structures. *Biopolymers: Original Research on Biomolecules*, 42(1):113–124, 1997.

[178] T. D. Schneider and R. M. Stephens. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Research*, 18(20):6097–6100, 1990.

[179] E. Segal and J. Widom. What controls nucleosome positions? *Trends in Genetics*, 25(8): 335–343, 2009.

[180] E. Segal, Y. Fondufe-Mittendorf, L. Chen, A. Thåström, Y. Field, I. K. Moore, J.-P. Z. Wang, and J. Widom. A genomic code for nucleosome positioning. *Nature*, 442(7104): 772–778, 2006.

[181] P. M. Severin, X. Zou, H. E. Gaub, and K. Schulten. Cytosine methylation alters DNA mechanical properties. *Nucleic Acids Research*, 39(20):8740–8751, 2011.

[182] P. M. Severin, X. Zou, K. Schulten, and H. E. Gaub. Effects of cytosine hydroxymethylation on DNA strand separation. *Biophysical Journal*, 104(1):208–215, 2013.

[183] R. Sharma, J. S. Schreck, F. Romano, A. A. Louis, and J. P. K. Doye. Characterizing the motion of jointed DNA nanostructures using a coarse-grained model. *ACS Nano*, 11(12): 12426–12435, 2017.

[184] B. S. Shastry. SNP alleles in human disease and evolution. *Journal of Human Genetics*, 47(11):561–566, 2002.

[185] N. N. Shaw and D. P. Arya. Recognition of the unique structure of DNA:RNA hybrids. *Biochimie*, 90(7):1026–1039, 2008.

[186] J. Shimada and H. Yamakawa. Statistical mechanics of helical wormlike chains. XII. Multivariate distribution functions. *The Journal of Chemical Physics*, 73(8):4037–4044, 1980.

[187] M. J. Shon, S.-H. Rah, and T.-Y. Yoon. Submicrometer elasticity of double-stranded DNA revealed by precision force-extension measurements with magnetic tweezers. *Science Advances*, 5(6):eaav1697, 2019.

[188] D. Shore, J. Langowski, and R. Baldwin. DNA flexibility studied by covalent closure of short fragments into circles. *Proceedings of the National Academy of Sciences*, 78(8): 4833–4837, 1981.

[189] S. Smith, L. Finzi, and C. Bustamante. Direct mechanical measurements of the elasticity of single DNA molecules by using magnetic beads. *Science*, 258(5085):1122–1126, 1992.

[190] B. E. K. Snodin, F. Randisi, M. Mosayebi, P. Sulc, J. S. Schreck, F. Romano, T. E. Ouldridge, R. Tsukanov, E. Nir, A. A. Louis, and J. P. K. Doye. Introducing improved structural properties and salt dependence into a coarse-grained model of DNA. *The Journal of Chemical Physics*, 142:234901, 2015.

[191] O. Sorkine-Hornung and M. Rabinovich. Least-squares rigid motion using SVD. *Computing*, 1(1), 2017.

[192] J. Sponer and F. Lankas. *Computational Studies of RNA and DNA*, volume 2. Springer Science and Business Media, 2006.

[193] H. Stein and P. Hausen. Enzyme from calf thymus degrading the RNA moiety of DNA-RNA hybrids: effect on DNA-dependent RNA polymerase. *Science*, 166(3903):393–395, 1969.

[194] S. H. Sternberg, B. LaFrance, M. Kaplan, and J. A. Doudna. Conformational control of DNA target cleavage by CRISPR–Cas9. *Nature*, 527(7576):110–113, 2015.

[195] E. Stofer and R. Lavery. Measuring the geometry of DNA grooves. *Biopolymers: Original Research on Biomolecules*, 34(3):337–346, 1994.

[196] G. Suresh and U. D. Priyakumar. DNA–RNA hybrid duplexes with decreasing pyrimidine content in the DNA strand provide structural snapshots for the A-to B-form conformational transition of nucleic acids. *Physical Chemistry Chemical Physics*, 16(34): 18148–18155, 2014.

[197] D. Svozil, J. Kalina, M. Omelka, and B. Schneider. DNA conformations and their sequence preferences. *Nucleic Acids Research*, 36(11):3690–3706, 2008.

[198] T. Szyperski, M. Götte, M. Billeter, E. Perola, L. Cellai, H. Heumann, and K. Wüthrich. NMR structure of the chimeric hybrid duplex r(gcaguggc)r(gcca)d(CTGC) comprising the tRNA-DNA junction formed during initiation of HIV-1 reverse transcription. *Journal of Biomolecular NMR*, 13(4):343–355, 1999.

[199] M. Tahiliani, K. P. Koh, Y. Shen, W. A. Pastor, H. Bandukwala, Y. Brudno, S. Agarwal, L. M. Iyer, D. R. Liu, L. Aravind, et al. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science*, 324(5929): 930–935, 2009.

[200] M. Tisdale, T. Schulze, B. A. Larder, and K. Moelling. Mutations within the RNase H domain of human immunodeficiency virus type 1 reverse transcriptase abolish virus infectivity. *Journal of General Virology*, 72(1):59–66, 1991.

[201] E. Trifonov, R. Tan, S. Harvey, et al. DNA bending and curvature. *Structure and Expression*, pages 243–253, 1987.

[202] D. W. Ussery. DNA Structure: A-, B- and Z-DNA Helix Families. *Encyclopedia of Life Sciences*, pages 1–7, 2002.

[203] J. J. Uusitalo, H. I. Ingolfsson, P. Akhshi, D. P. Tieleman, and S. J. Marrink. Martini coarse-grained force field: extension to DNA. *Journal of Chemical Theory and Computation*, 11(8):3932–3945, 2015.

[204] F. Vella. *Molecular biology of the cell*. Wiley Online Library, 1994.

[205] C. H. Waddington. The epigenotype. *Endeavour*, 1:18–20, 1942.

[206] A. H.-J. Wang, S. Fujii, J. H. van Boom, G. A. van der Marel, S. A. van Boeckel, and A. Rich. Molecular structure of r(GCG)d(TATACGC): a DNA–RNA hybrid helix joined to double helical DNA. *Nature*, 299(5884):601–604, 1982.

[207] J. D. Watson and F. H. C. Crick. Molecular Structure of Nucleic Acids. *Nature*, 171 (4356):737–738, 1953.

[208] K. Yanagi, G. G. Privé, and R. E. Dickerson. Analysis of local helix geometry in three B-DNA decamers and eight dodecamers. *Journal of Molecular Biology*, 217(1):201–214, 1991.

[209] J. D. Yesselman, S. K. Denny, N. Bisaria, D. Herschlag, W. J. Greenleaf, and R. Das. Sequence-dependent RNA helix conformational preferences predictably impact tertiary structure formation. *Proceedings of the National Academy of Sciences*, 116(34):16847–16855, 2019.

[210] M. Zgarbová, M. Otyepka, J. Sponer, A. Mladek, P. Banas, T. E. Cheatham III, and P. Jurecka. Refinement of the cornell et al. nucleic acids force field based on reference quantum chemical calculations of glycosidic torsion profiles. *Journal of Chemical Theory and Computation*, 7(9):2886–2902, 2011.

[211] C. Zhang, H. Fu, Y. Yang, E. Zhou, Z. Tan, H. You, and X. Zhang. The mechanical properties of RNA-DNA hybrid duplex stretched by magnetic tweezers. *Biophysical Journal*, 116(2):196–204, 2019.

[212] S. B. Zimmerman and B. H. Pheiffer. A RNA.DNA hybrid that can adopt two conformations: an x-ray diffraction study of poly(rA). poly(dT) in concentrated solution or in fibers. *Proceedings of the National Academy of Sciences*, 78(1):78–82, 1981.

[213] T. Zwahlen. *Landscapes of DNA mechanics and Genomes*. PhD thesis, EPFL, 2022.

Appendices

Appendix A

Ideal atoms coordinates in Tsukuba convention

This chapter lists the ideal coordinates for bases and phosphates used to fit the frames to coarse-grain NAs. The ideal coordinates for standard bases are as per Tsukuba convention [140] and phosphate is approximated as tetrahedron as provided in table A.1 For non-standard bases, 5-methylated Cytosine and 5-hydroxylmethylated Cytosine, Cytosine coordinates are taken.

Atom	**Adenine**			**Guanine**			**Phosphate**		
	x (Å)	y (Å)	z (Å)	x (Å)	y (Å)	z (Å)	x (Å)	y (Å)	z (Å)
C1′	-2.479	5.346	0.000	-2.477	5.399	0.000	—	—	—
N9	-1.291	4.498	0.000	-1.289	4.551	0.000	—	—	—
C8	0.024	4.897	0.000	0.023	4.962	0.000	—	—	—
N7	0.877	3.902	0.000	0.870	3.969	0.000	—	—	—
C5	0.071	2.771	0.000	0.071	2.883	0.000	—	—	—
C6	0.369	1.398	0.000	0.424	1.460	0.000	—	—	—
N6	1.611	0.909	0.000	—	—	—	—	—	—
O6	—	—	—	1.554	0.955	0.000	—	—	—
N1	-0.668	0.532	0.000	-0.700	0.641	0.000	—	—	—
C2	-1.912	1.023	0.000	-1.999	1.087	0.000	—	—	—
N2	—	—	—	-2.949	0.139	-0.001	—	—	—
N3	-2.320	2.290	0.000	-2.342	2.364	0.001	—	—	—
C4	-1.267	3.124	0.000	-1.265	3.177	0.000	—	—	—
P	—	—	—	—	—	—	0.000	0.000	0.000
O3′	—	—	—	—	—	—	1.518	0.000	-0.537
O5′	—	—	—	—	—	—	-0.759	-1.315	-0.537
OP1	—	—	—	—	—	—	-0.698	1.208	-0.493
OP2	—	—	—	—	—	—	0.000	0.000	1.480
Atom	**Thymine**			**Cytosine**			**Uracil**		
C1′	-2.481	5.354	0.000	-2.477	5.402	0.000	-2.481	5.354	0.000
N1	-1.284	4.500	0.000	-1.285	4.542	0.000	-1.284	4.500	0.000
C2	-1.462	3.135	0.000	-1.472	3.158	0.000	-1.462	3.131	0.000
O2	-2.562	2.608	0.000	-2.628	2.709	0.001	-2.563	2.608	0.000
N3	-0.298	2.407	0.000	-0.391	2.344	0.000	-0.302	2.397	0.000
C4	0.994	2.897	0.000	0.837	2.868	0.000	0.989	2.884	0.000
O4	1.944	2.119	0.000	—	—	—	1.935	2.094	-0.001
N4	—	—	—	1.875	2.027	0.001	—	—	—
C5	1.106	4.338	0.000	1.056	4.275	0.000	1.089	4.311	0.000
C5M	2.466	4.961	0.001	—	—	—	—	—	—
C6	-0.024	5.057	0.000	-0.023	5.068	0.000	-0.024	5.053	0.000

Table A.1 Cartesian coordinates defined for non-Hydrogen atoms of standard bases (A, G, T, C, and U) in Tsukuba convention [140] and phosphate coordinates used in this work.

Appendix B

MD libraries

For MD simulations of DNA, RNA, and DNA:RNA hybrid (DRH), we have used a palindromic library (given in table B.1) introduced in ref. [149]. This library contains all 256 tetramers on the reading strand, and the palindromic nature of the library allowed us to check the convergence of MD time-series and enhance the statistics. All the sequences in the palindromic library have GC ends to minimize fraying. Furthermore, we have imposed this palindromic property in the library of sequences with epigenetic modifications (table B.2). However, it is impossible to design a palindromic library for hybrid DNA-RNA (HDR). So, we used the same library as given in table B.1 as it provides comparable statistics for all the monomer, dimer, and trimer, as well as allows us a systematic comparison between DNA, RNA, and DRH at atomistic levels. All the libraries discussed above have GC ends. In order to obtain parameter set blocks for other end blocks, we have used the library described in table B.3. For each non-GC end, we have four sequences of length 12-nt of the form XYUV-(hex)-GC where XY $\in$ {15 non-GC ends}, UV is randomly chosen YY, YR, RR, and RY steps (to provide a rich context for XY), hex is randomly chosen hexamer, and the other end is fixed to be GC (as both non-GC ends lead to very low acceptance of the MD time-series after HB filtering).

B.1 Total number of monomers, dimers, monomers in trimer contexts, and dimers in tetramer contexts containing at least one modified base in monomers and dimers

We have only considered methylation and hydroxymethylation of CpG steps in this work. We have used the letter M for 5-methylated-Cytosine, and N for Guanine when the complementary Cytosine is methylated. Similarly, the letters H and K are used for 5- hydroxymethylated-Cytosine and Guanine complementary to 5-hydroxymethylated-Cytosine, respectively. This section discusses the total possible monomers, dimers, monomers in trimer contexts, and dimers in tetramer contexts containing at least one modified base in monomers and dimers by taking the example of methylated Cytosine.

Index	Sequence
	Training sequences
1	GCTTAGTTCAAATTTGAACTAAGC
2	GCTCTCTGTATTAATACAGAGAGC
3	GCCCTTGGCGATATCGCCAAGGGC
4	GCTAAAGCCTTATAAGGCTTTAGC
5	GCGGTAGAAAACGTTTTCTACCGC
6	GCCAAGACATTGCAATGTCTTGGC
7	GCAGATGGTCAGCTGACCATCTGC
8	GCCTCACCGCTCGAGCGGTGAGGC
9	GCAGTGGAATCATGATTCCACTGC
10	GCTTTACTTCGTACGAAGTAAAGC
11	GCTACCTATGCTAGCATAGGTAGC
12	GCGCACTGGGGATCCCCAGTGCGC
13	GCTGAGGAGTCCGGACTCCTCAGC
14	GCTGCCGTCGGGCCCGACGGCAGC
15	GCGCACAACACGCGTGTTGTGCGC
16	GCCTAACCCTGCGCAGGGTTAGGC
	Test sequences
17	GCATTACGCTCCGGAGCGTAATGC
18	GCAAAAAAAAAAAAAAGC
19	GCATATATATATATATGC
20	GCGGATTACGCAGGC
21	GCGGATTCCGCAGGC
22	GCGCGAAAATTTTCGAAAATTTTCGCGC
23	GCGCGTTTTAAAACGTTTTAAAACGCGC
24	GCGCGCGCGCGCGCGCGCGC
25	CGGCGCACGTGACCGCG
26	GCATCGCCACTGAAGTTGGTTATAACCAACTTCAGTGGCGATGC

Table B.1 Palindromic library in standard A, T, C, and G alphabets. For DNA this library has been referred as Lb_{DNA}. For RNA, we have used the same library except the T is replaced by U and referred as Lb_{RNA}. For HDR, we have intentionally chosen the DNA strand as the reading strand and thus keeping the same library which is called Lb_{DRH}.

Index	Sequence		
	Training library		Test library
1	GCTAMNTGTAMNMNTACAMNTAGC	13	GCTAMGTGTCMNMNGACACNTAGC
2	GCATMNACGAMNMNTCGTMNATGC	14	GCATMGACGTMNMNACGTCNATGC
3	GCGCMNGGAGMNMNCTCCMNGCGC	15	GCTGMGTTCGMNMNCGAACNCAGC
4	GCTCMNCTAAMNMNTTAGMNGAGC	16	GCCTMGCGTTMNMNAACGCNAGGC
5	GCTGMNTTCCMNMNGGAAMNCAGC	17	GCCTGAGTAMGMNCNTACTCAGGC
6	GCCTMNCGTGMNMNCACGMNAGGC	18	GCGGATTAMNCAGGC
7	GCGCMGGGATMNMNATCCCNGCGC	19	GCGCGCGMNMNMNCGCGCGC
8	GCTCMGCTACMNMNGTAGCNGAGC	20	GCGCGCGMGMGMGCGCGCGC
9	GCTAMGTGTCCNMGGACACNTAGC	21	GCGCGMNCGCGCGMGCGCGC
10	GCATMGACGTMGCNACGTCNATGC		
11	GCAGMGMGATAATTATCNCNCTGC		
12	GCCACAAGTCNMNMGACTTGTGGC		

Table B.2 Methylated or Hydroxymethylated libraries. The first 12 sequences are in the training library, and the rest of the sequences are in the test library. The Methylated and Hydroxymethylated libraries have been referred to as Lb_{Met} and Lb_{Hmet}, respectively.

Index	Sequence	Index	Sequence	Index	Sequence
1	AAGACCACTTGC	21	TGAGGCCACCGC	41	GTAAGATTACGC
2	AAGTTTAGGGGC	22	TGATCAAGTAGC	42	GTGCGACGCTGC
3	AATGCGTATCGC	23	TGTGCCGAGAGC	43	GTCAGGATAAGC
4	AATCACTTAGGC	24	TGCTTGATTTGC	44	GTTTCTAATAGC
5	ATAGACCCAAGC	25	TCAATTCGACGC	45	CGGACTACTCGC
6	ATGTATCACAGC	26	TCACAGCCATGC	46	CGGTGCTGCTGC
7	ATCAGGATAGGC	27	TCTGTGCAAAGC	47	CGTGGTGGAGGC
8	ATTTCTAGTGGC	28	TCTTGCGTTGGC	48	CGTCCTATTGGC
9	AGAAACTCGTGC	29	TTGATACCGCGC	49	CCGGCCCGCCGC
10	AGATAACACTGC	30	TTATCATGCAGC	50	CCACCCCGTCGC
11	AGCGCTCGTCGC	31	TTTGAATTATGC	51	CCTAAGTCTAGC
12	AGCCATGAAAGC	32	TTCTGGTTACGC	52	CCTTGCCTACGC
13	ACGGACGAATGC	33	GGGGCTCTTCGC	53	CTAGAGCGTGGC
14	ACGTTCAGTGGC	34	GGGTCGGACCGC	54	CTGCAACCCAGC
15	ACCGCGGTGAGC	35	GGTATCGACGGC	55	CTCATCCAACGC
16	ACCCAAAGCTGC	36	GGCCTATTATGC	56	CTCTGAGGTGGC
17	TAGACACTGTGC	37	GAGAGATGTCGC	57	CAAAGTCGACGC
18	TAATCCTCGCGC	38	GAATTATTACGC	58	CAACCCATTCGC
19	TATAGTGAGCGC	39	GACAGATCACGC	59	CACGGAAAGCGC
20	TATCGGGAATGC	40	GACTATGGTAGC	60	CATTAACGCCGC

Table B.3 Library for end-block parameters (Lb_{End})

Nmers	cases	total possibilities
X	M, N	2
WXY	AMN, TMN, GMN, CMN, NMN AMG, TMG, GMG, CMG, NMG - - - - - - - - - - - - - - - - - - MNA, MNT, MNG, MNC, MNM CNA, CNT, CNG, CNC, CNM	20
XY	**MN**, **NM**, MG, CN, AM, TM, CM, GM, NA, NT, NC, NG	12
WXYZ	AMNA, **TMNA**, CMNA, GMNA, NMNA **AMNT**, TMNT, CMNT, GMNT, NMNT AMNC, TMNC, CMNC, **GMNC**, NMNC AMNG, TMNG, **CMNG**, GMNG, NMNG AMNM, TMNM, CMNM, GMNM, **NMNM** - **CNMG**, MNMG, CNMN, **MNMN** - AMGA, TMGA, CMGA, GMGA, NMGA AMGT, TMGT, CMGT, GMGT, NMGT AMGC, TMGC, CMGC, GMGC, NMGC AMGG, TMGG, CMGG, GMGG, NMGG AMGM, TMGM, CMGM, GMGM, NMGM - ACNA, TCNA, CCNA, GCNA, NCNA ACNT, TCNT, CCNT, GCNT, NCNT ACNC, TCNC, CCNC, GCNC, NCNC ACNG, TCNG, CCNG, GCNG, NCNG ACNM, TCNM, CCNM, GCNM, NCNM - AAMN, TAMN, CAMN, GAMN, NAMN AAMG, TAMG, CAMG, GAMG, NAMG - ATMN, TTMN, CTMN, GTMN, NTMN ATMG, TTMG, CTMG, GTMG, NTMG - ACMN, TCMN, CCMN, GCMN, NCMN ACMG, TCMG, CCMG, GCMG, NCMG - MGMN, AGMN, TGMN, CGMN, GGMN, NGMN MGMG, AGMG, TGMG, CGMG, GGMG, NGMG - MNAM, MNAA, MNAT, MNAC, MNAG CNAM, CNAA, CNAT, CNAC, CNAG - MNTM, MNTA, MNTT, MNTC, MNTG CNTM, CNTA, CNTT, CNTC, CNTG - MNCN, MNCM, MNCA, MNCT, MNCC, MNCG CNCN, CNCM, CNCA, CNCT, CNCC, CNCG - MNGM, MNGA, MNGT, MNGC, MNGG CNGM, CNGA, CNGT, CNGC, CNGG	163

Table B.4 Total number of monomers, dimers, monomers in trimer contexts, and dimers in tetramer contexts containing at least one modified base in monomers and dimers. Palindromes are highlighted in bold. Trimers and tetramers with the same central monomer and dimer, respectively, are separated by a dashed line.

Appendix C

Mathematical detail

C.1 Rotations in three-dimensions, $SO(3)$ group

The special orthogonal matrix group, $SO(3)$ represents all proper rotations in three-dimensional Euclidean space, i.e., $\in \mathbb{R}^3$ and is defined as:

$$SO(3) = \{R \in \mathbb{R}^{3\times3} \mid R^T R = RR^T = I \in \mathbb{R}^{3\times3},\ \det R = +1\} \tag{C.1}$$

where I is an identity matrix. R is a proper right-handed rotation matrix with $\{1, e^{\iota\theta}, e^{-\iota\theta}\}$ as eigenvalues, where θ is the angle of rotation, and the real eigenvector of R is the axis of rotation u.

Furthermore, for a given rotation matrix, R, the Euler-Rodrigues formula gives the following relation between the rotation matrix R and a unit axis of rotation u and angle of rotation θ:

$$SO(3) \ni R = \cos\theta I + (1 - \cos\theta)u \otimes u + \sin\theta u^{\times} \tag{C.2}$$

where $I \in \mathbb{R}^{3\times3}$ is an identity matrix, $u \otimes u = uu^T$ is outer-product and $u^{\times}$ is a skew-symmetric matrix satisfying $(u^{\times})v = u \times v\ \forall\ v \in \mathbb{R}^3$ and of the form:

$$u^{\times} = \begin{bmatrix} 0 & -u_3 & u_2 \\ u_3 & 0 & -u_1 \\ -u_2 & u_1 & 0 \end{bmatrix} \tag{C.3}$$

A skew-symmetric matrix $u^{\times}$ can be transformed to its corresponding as $u = \text{vec}(u^{\times})$.

Using Euler-Rodrigues formula in Equation (C.2), a direct relation between R and u, θ is given as:

$$[0, \pi) \ni \theta = \arccos\left(\frac{tr(R) - 1}{2}\right) \text{ and } \mathbb{R}^3 \ni u = \frac{2}{1 + tr(R)}\text{vec}(R - R^T) \tag{C.4}$$

Note that, for $\theta = 0$ and π, the rotation matrix become symmetric which means $R - R^T$ will be a zero-matrix and thus, can't be used for the computation of rotation axis, u. In the case of $\theta = 0$, Q becomes an identity matrix, and any unit vector can be the rotation axis. However, when $\theta = \pi$, the axis of rotation will be the eigenvector of matrix $Q + I$.

C.2 Parameterisation of rotations in cgDNA+ model

Now, to parameterise rotations in cgDNA+ model, we have used Cayley parameters (details in [99, 149, 158]). We have defined the function $cay : \mathbb{R}^3 \rightarrow SO(3)$ as:

$$cay_\alpha(\eta) = I + \frac{1}{4\alpha^2 + |\eta|^2}\left[4\alpha\eta^\times + 2(\eta^\times)^2\right] = R(u,\theta)\ \forall\ \alpha \in \mathbb{R}, \eta \in \mathbb{R}^3 \tag{C.5}$$

where $\mathbb{R}^3 \ni u = \frac{\eta}{|\eta|}$ and $\theta = 2\arctan\left(\frac{|\eta|}{2\alpha}\right)$. The inverse of cay transformation can be defined as cay^{-1}: $SO(3) \rightarrow \mathbb{R}^3$ and is given in Equation (C.6).

$$cay_\alpha^{-1}(R) = \frac{2\alpha}{1 + tr(R)}\text{vec}(R - R^T) \tag{C.6}$$

C.3 Rigid body transformation, $SE(3)$ group

To describe the position and orientation of rigid body, we have used special euclidean group, $SE(3)$ which is defined as:

$$SE(3) = \left\{\mathbb{R}^{4\times4} \ni G = \begin{bmatrix} R & r \\ 0 & 1 \end{bmatrix}\right\} \tag{C.7}$$

where $R \in SO(3)$ is the rotational component and $r \in \mathbb{R}^3$ is the translational component of the rigid body transformation.

The product of $G_1 \in SE(3)$ and $G_2 \in SE(3)$ is given as

$$SE(3) \ni G_1G_2 = \begin{bmatrix} R_1R_2 & R_1r_2 + r_1 \\ 0 & 1 \end{bmatrix} \tag{C.8}$$

and the inverse of an element in $SE(3)$ is,

$$SE(3) \ni G^{-1} = \begin{bmatrix} R^T & -R^Tr \\ 0 & 1 \end{bmatrix} \tag{C.9}$$

C.4 Kullback-Leibler divergence

The Kullback-Leibler (KL) divergence [94], also known as relative entropy, between two continuous pdfs $\rho_1(x)$ and $\rho_2(x)$ defined on $\Omega \subset \mathbb{R}^N$ is given as:

$$D_{KL}(\rho_1(x), \rho_2(x)) = \int_\Omega \rho_1(x)\log\frac{\rho_1(x)}{\rho_2(x)}dx \geq 0 \tag{C.10}$$

where the equality holds if $\rho_1 = \rho_2$.

Some of the key properties of KL divergence are:

- KL divergence is non-symmetric, i.e., $D_{KL}(\rho_1, \rho_2) \neq D_{KL}(\rho_2, \rho_1)$, in general.
- KL divergence doesn't satisfy triangle inequality and doesn't qualify as a metric. It just defines a premetric on the set of pdfs.

- Invariant under re-scaling i.e say X_1, X_2 are random variables associated to pdfs ρ_1, ρ_2 and X_1', X_2' are random variables associated to pdfs ρ_1', ρ_2' where $\{X_i = aX_i'\}_{i=1,2}$, then $D_{KL}(\rho_1, \rho_2) = D_{KL}(\rho_1', \rho_2')$. This invariance of KL divergence under re-scaling allowed an easier re-scaling of rotational coordinates in cgDNA family of models and change of reading strand.
- In case, when the pdfs ρ_1, ρ_2 are normal multivariate distributions, equation (C.10) simplifies to an algebraic form,

$$\begin{aligned} D_{KL}(\rho_1, \rho_2) &= \mathcal{S}(\rho_1, \rho_2) + \mathcal{M}(\rho_1, \rho_2) \\ \mathcal{S}(\rho_1, \rho_2) &= \frac{1}{2}\left[K_1^{-1} : K_2 - \ln\left(\frac{|K_2|}{|K_1|}\right) - I : I\right] \\ \mathcal{M}(\rho_1, \rho_2) &= \frac{1}{2}(\mu_1 - \mu_2)^T K_2 (\mu_1 - \mu_2) \end{aligned} \tag{C.11}$$

  where μ_1 and μ_2 are mean vectors, K_1 and K_2 are inverse covariance matrices, and : represents the standard Euclidean inner product for square-matrices and I is the identity matrix of the size same as K_1 and K_2. $\sqrt{\mathcal{M}}$ is also known as Mahalanobis distance [113].

Moreover, KL divergence can be symmetrised as follows:

$$\begin{aligned} D_{KLS}(\rho_1, \rho_2) &= \frac{1}{2}[D_{KL}(\rho_1, \rho_2) + D_{KL}(\rho_2, \rho_1)] \\ &= \mathcal{S}_S(\rho_1, \rho_2) + \mathcal{M}_S(\rho_1, \rho_2), \\ \mathcal{S}_S(\rho_1, \rho_2) &= \frac{1}{2}\left[K_1^{-1} : K_2 + K_2^{-1} : K_1 - 2I : I\right], \\ \mathcal{M}_S(\rho_1, \rho_2) &= \frac{1}{2}\left[(\mu_1 - \mu_2)^T (K_2 + K_1)(\mu_1 - \mu_2)\right], \end{aligned} \tag{C.12}$$

where $\mathcal{S}_S(\rho_1, \rho_2)$ is symmetrised stiffness contribution, and $\mathcal{M}_S(\rho_1, \rho_2)$ is symmetrised shape contribution of the symmetrised KL divergence.

Appendix D

An involution of 3×3 block structure

Let P be an orthogonal matrix

$$P = \begin{bmatrix} 0 & 0 & P_1 \\ 0 & P_2 & 0 \\ P_1 & 0 & 0 \end{bmatrix} \text{ where } P^2 = I, P^T P = I, P \in n \times n, P_i^2 = I, P_i \in n_i \times n_i \text{ and } P_i^T = P_i \; \forall \; i = 1, 2 \tag{D.1}$$

and a symmetric matrix K

$$K = \begin{bmatrix} A & B & E \\ B^T & C & D \\ E^T & D^T & F \end{bmatrix} \text{ such that } K^T = K \tag{D.2}$$

which is also an involution symmetric matrix such that $K = PKP = P^T K P$. Let's define another orthogonal matrix Q,

$$Q = \frac{1}{\sqrt{2}} \begin{bmatrix} I & 0 & I \\ 0 & \sqrt{2}I & 0 \\ -P_1 & 0 & P_1 \end{bmatrix} \text{ such that } Q^T Q = I \tag{D.3}$$

Now, if P_2 is diagonal matrix with s elements -1 and rest of the r elements +1 (as ± 1 are the only possibility as $P_2^2 = I$) then the involution symmetry of K implies that the orthogonal transformation $Q^T K Q$ yields a 2×2 block structure, i.e.,

$$Q^T K Q = \begin{bmatrix} H_1 & O \\ O & H_2 \end{bmatrix} \text{ where } H_1 \in (n_1 + s) \times (n_1 + s) \text{ and } H_2 \in (n_1 + r) \times (n_1 + r) \tag{D.4}$$

$$PKP = \begin{bmatrix} P_1 F P_1 & P_1 D^T P_2 & P_1 E^T P_1 \\ P_2 D P_1 & P_2 C P_2 & P_2 B P_1 \\ P_1 E P_1 & P_1 B^T P_2 & P_1 A P_1 \end{bmatrix} \tag{D.5}$$

so the symmetry and involution symmetry iff 4 conditions (independent) satisfy

- $A = P_1 F P_1$
- $P_1 B = D^T P_2$
- $E P_1 = P_1 E^T = (E P_1)^T$
- $P_2 C P_2 = C$

Now,

$$
\begin{aligned}
Q^T K Q =& \frac{1}{2}\begin{bmatrix} I & 0 & -P_1 \\ 0 & \sqrt{2}I & 0 \\ I & 0 & P_1 \end{bmatrix}\begin{bmatrix} A & B & E \\ B^T & C & D \\ E^T & D^T & F \end{bmatrix}\begin{bmatrix} I & 0 & I \\ 0 & \sqrt{2}I & 0 \\ -P_1 & 0 & P_1 \end{bmatrix} \\
=& \frac{1}{2}\begin{bmatrix} I & 0 & -P_1 \\ 0 & \sqrt{2}I & 0 \\ I & 0 & P_1 \end{bmatrix}\begin{bmatrix} A - EP_1 & \sqrt{2}B & A + EP_1 \\ B^T - DP_1 & \sqrt{2}C & B^T + DP_1 \\ E^T - FP_1 & \sqrt{2}D^T & E^T + FP_1 \end{bmatrix} \\
=& \frac{1}{2}\begin{bmatrix} A + P_1FP_1 - P_1E^T - EP_1 & \sqrt{2}[B - P_1D^T] & A - P_1FP_1 + EP_1 - P_1E^T \\ \sqrt{2}[B^T - DP_1] & 2C & \sqrt{2}[B^T + DP_1] \\ A - P_1FP_1 - P_1E^T - EP_1 & \sqrt{2}[B + P_1D^T] & A + P_1FP_1 + EP_1 + P_1E^T \end{bmatrix} \\
=& \begin{bmatrix} A - EP_1 & \frac{1}{\sqrt{2}}B[I - P_2] & O \\ \frac{1}{\sqrt{2}}[I - P_2]B^T & C & \frac{1}{\sqrt{2}}[I + P_2]B^T) \\ O & \frac{1}{\sqrt{2}}B[I + P_2] & A + EP_1 \end{bmatrix} \\
=& \frac{1}{2}\begin{bmatrix} H_1 & O \\ O & H_2 \end{bmatrix}
\end{aligned}
\tag{D.6}
$$

where $H_1 := \begin{bmatrix} A - EP_1 & \sqrt{2}B_1 \\ \sqrt{2}B_1^T & C_{11} \end{bmatrix}$ and $H_2 := \begin{bmatrix} C_{22} & \sqrt{2}B_2 \\ \sqrt{2}B_2^T & A + EP_1 \end{bmatrix}$.

Lemma

- Orthogonal similarity doesn't change eigenvalues and systems for H_1 and H_2 are decoupled, i.e., $\lambda(K) = \mu(H_1) \cup \gamma(H_2)$
- If $(\mu_i, \begin{bmatrix} \underline{\mathrm{w}}_i \\ \underline{\mathrm{x}}_i \end{bmatrix})$ is eigenpair for H_1 then corresponding eigenpair for K is $(\mu_i, Q\begin{bmatrix} \underline{\mathrm{w}}_i \\ \underline{\mathrm{x}}_i \\ o \end{bmatrix} = \frac{1}{\sqrt{2}}\begin{bmatrix} \underline{\mathrm{w}}_i \\ \sqrt{2}\hat{\underline{\mathrm{x}}}_i \\ -P_1\underline{\mathrm{w}}_i \end{bmatrix})$ where $\hat{\underline{\mathrm{x}}}_i = \begin{bmatrix} \underline{\mathrm{x}}_i \\ 0 \end{bmatrix}$
- Similarly, if $(\gamma_j, \begin{bmatrix} \underline{\mathrm{y}}_j \\ \underline{\mathrm{z}}_j \end{bmatrix})$ is eigenpair for H_2 then corresponding eigenpair for K is $(\gamma_j, Q\begin{bmatrix} o \\ \underline{\mathrm{y}}_j \\ \underline{\mathrm{z}}_j \end{bmatrix} = \frac{1}{\sqrt{2}}\begin{bmatrix} \underline{\mathrm{z}}_j \\ \sqrt{2}\hat{\underline{\mathrm{y}}}_j \\ +P_1\underline{\mathrm{z}}_j \end{bmatrix})$ where $\hat{\underline{\mathrm{y}}}_j = \begin{bmatrix} 0 \\ \underline{\mathrm{y}}_j \end{bmatrix}$

It explains the sparsity pattern in eigenvectors of a palindromically symmetrized matrix.

Appendix E

Supplementary figures for Comparison of non-local sequence-dependent mechanics of DNA in protein-DNA crystal structures ensemble with cgNA+ model

E.1 Additional figures and tables

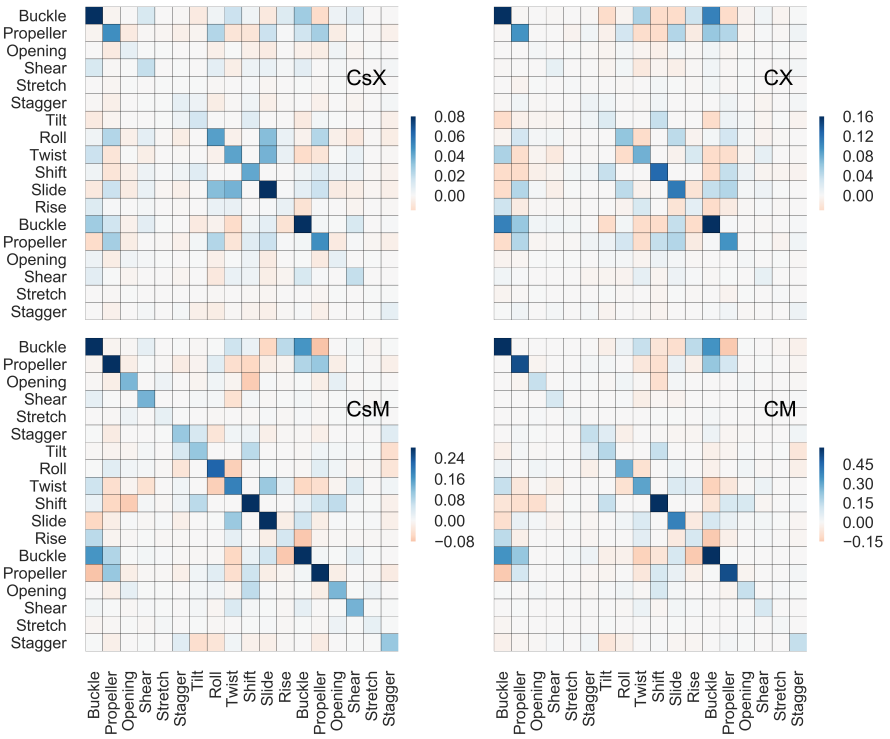


Fig. E.1 Heat map for shape and configuration covariance for X-ray (C_sX and CX) and cgNA+ (C_sM and CM) model data set. The corresponding variances are listed in the table E.1. Note that the scale in all four covariance is different. Scale of configuration covariance is approximately two times that shape covariance in both the data set. Scale in cgNA+ model data set, for both the covariance (shape and configuration), is almost three times than in X-ray data set possibly due less effective temperature in X-ray data set.

Internal Coordinate	Shape variance (X-ray)	Shape variance (CG)	Configurational variance (X-ray)	Configurational variance (CG)
Buckle	0.0799	0.2441	0.4764	0.9488
Propeller	0.0499	0.0956	0.3487	0.5143
Opening	**0.0070**	0.0035	0.1272	0.1335
Shear	0.0191	0.0136	0.1333	0.0954
Stretch	**0.0006**	0.0013	0.0231	0.0136
Stagger	**0.0073**	0.0092	0.1061	0.1361
Tilt	0.0147	0.0223	0.1078	0.1660
Roll	0.0439	0.0629	0.2241	0.2870
Twist	0.0432	0.0767	0.1888	0.3094
Shift	0.0413	0.1238	0.3470	0.6836
Slide	0.1447	0.1127	0.4347	0.3938
Rise	**0.0063**	0.0167	0.0497	0.0964

Table E.1 List of variances (the diagonal elements) for shape and configuration covariances for X-ray and cgNA+ model data set. The unit for the variance can be considered as $\mathrm{\AA}^2$ and $(\mathrm{rad}/5)^2$ for translational and rotational coordinates, respectively. The corresponding full covariance matrix is plotted in figure E.1. Due to palindromic symmetry, the variance for both the base-pairs (in terms of intra coordinates) is identical and, thus, listed once.

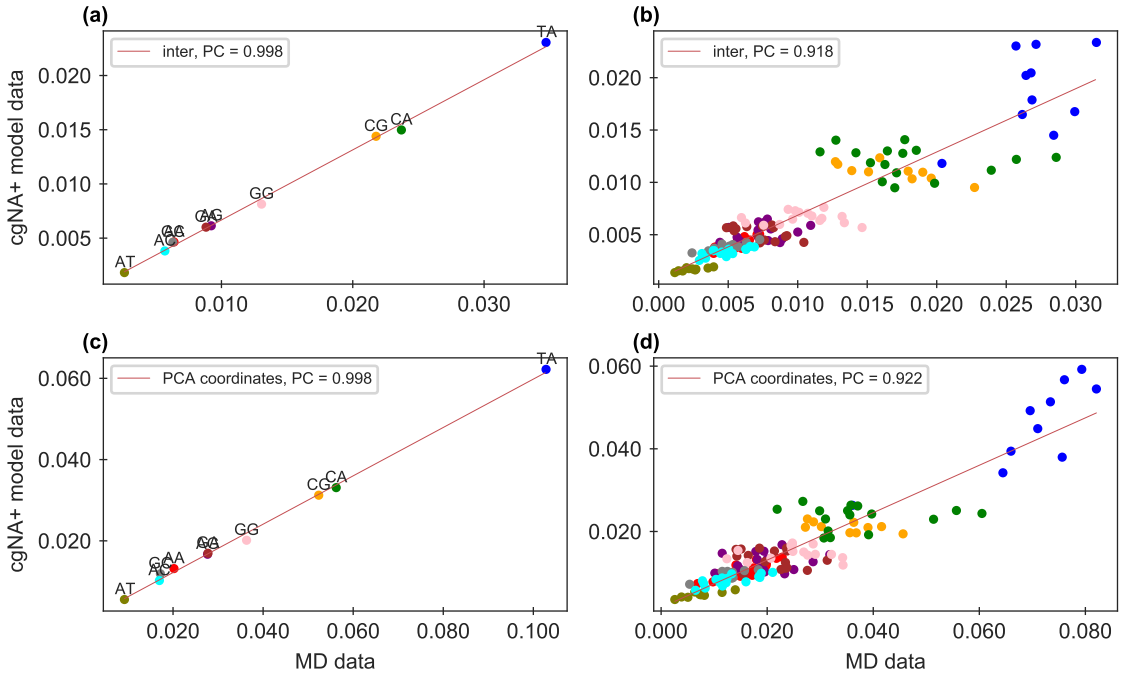


Fig. E.2 Comparison of configurational volume (S) for cgNA+ model covariance vs MD data set covariance a) in inter coordinates for independent dimer steps in average context, b) in inter coordinates for dimers in independent tetramer contexts, c) in PCA coordinates (in eight principal modes of cgNA+ model shape covariance) for independent dimer steps in average context, d) in PCA coordinates (in eight principal modes of cgNA+ model shape covariance) for dimers in independent tetramer contexts. The red line is best-fit line between the two data sets using linear regression.

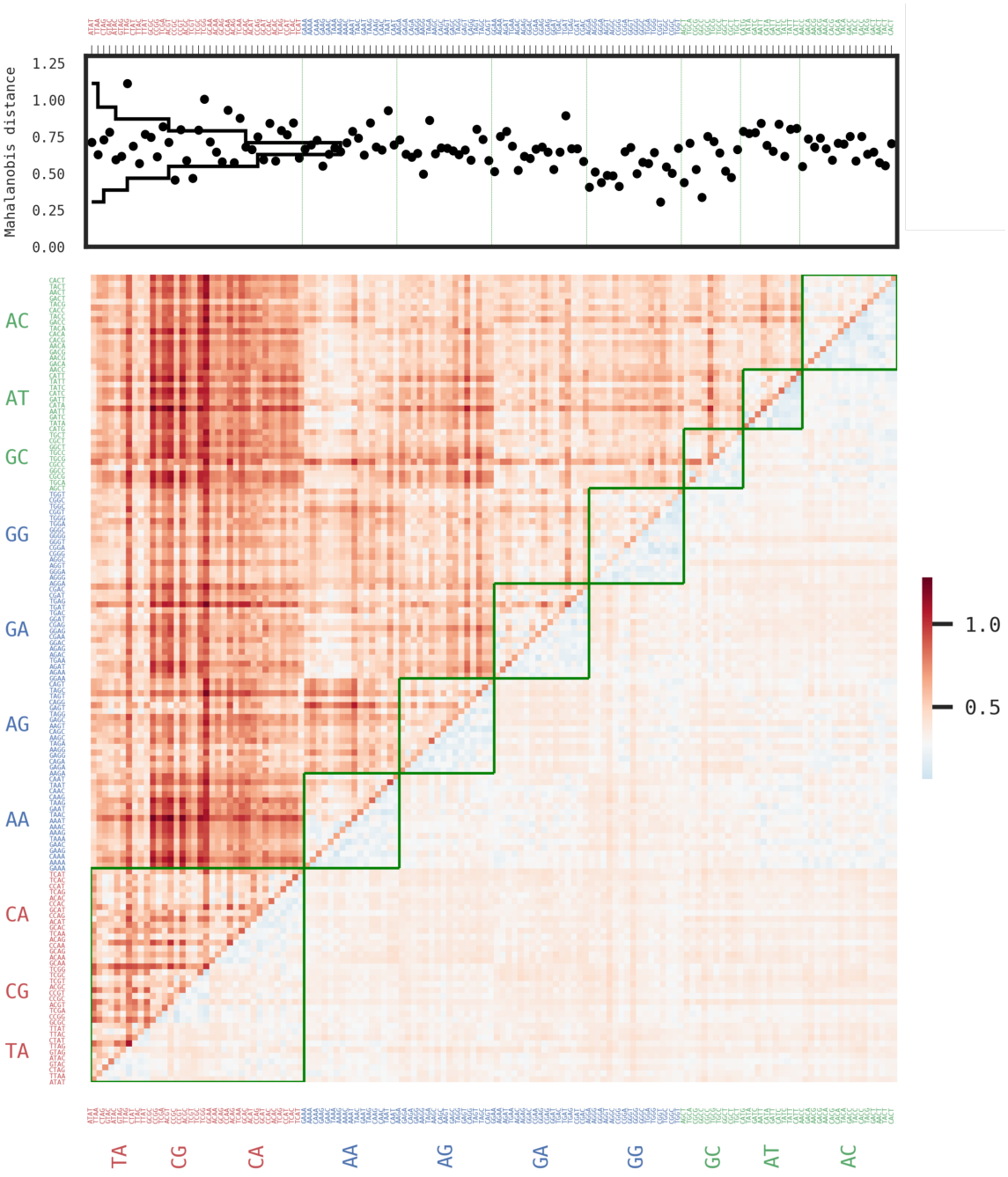


Fig. E.3 The diagonal entries in the heat map (bottom) are Mahalanobis distance between the groundstate of dimers (in 136 independent tetramer contexts) in the X-ray and cgNA+ model data set. Whereas lower and upper off-diagonal entries are Mahalanobis distance between different dimers (in specific tetramer context) within the cgNA+ model and X-ray data set, respectively. The diagonal entries of the heat-map are again plotted in the scatter plot (top) along with the histogram in the same plot. Note that the Mahalanobis distance (as defined in **??**) is computed in the 18 CURVES+ coordinates and using the cgNA+ shape covariance matrix as the weight matrix.

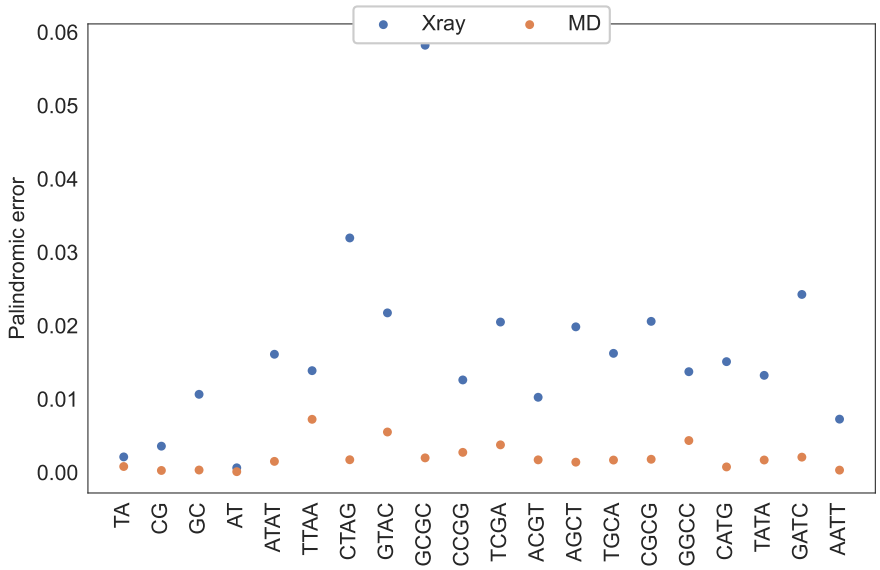


Fig. E.4 Palindromic error (as defined in section 2.5.1) per degree of freedom in the groundstate of palindromic dimer in tetramer flanking context and in average flanking context for X-ray data set and MD simulations (used to train cgNA+ model).

E.2 Comparison of two X-ray data sets with different resolutions and results for case-II

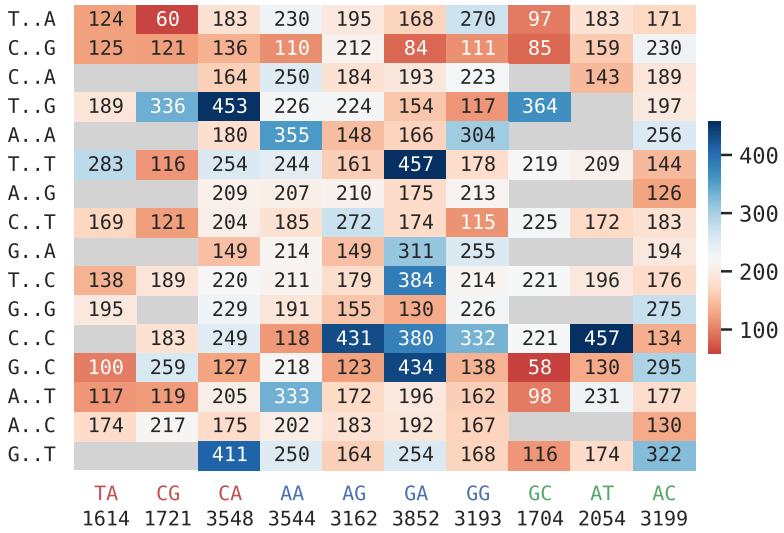


Fig. E.7 Number of appearances of 136 tetrameters in X-ray data set (case-II). Abscissa is middle junction dimer-step and ordinate is tetramer context. Note that the palindromic steps are only read from reading strand.

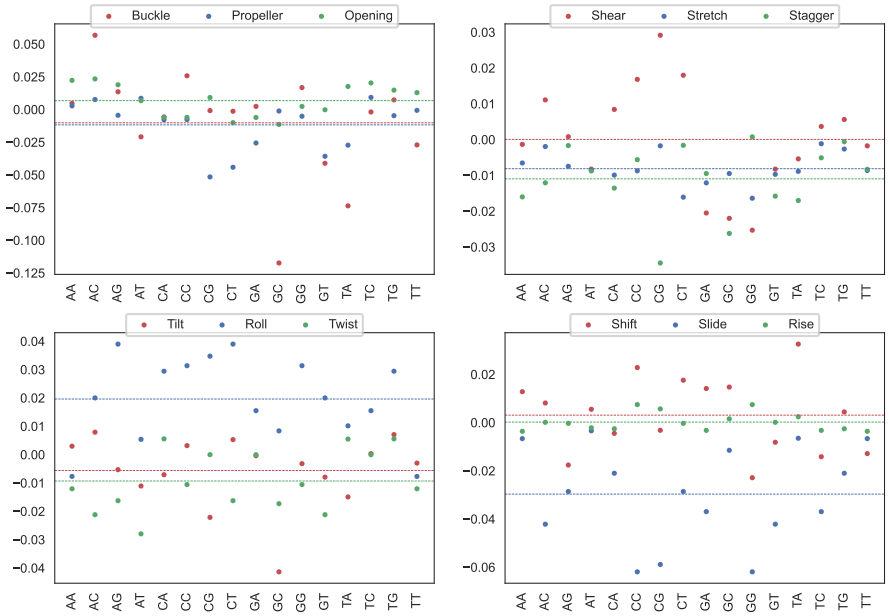


Fig. E.5 Plot comparing the average shape of dimer in two X-ray datasets as defined in case-I and case-II where case-I has no resolution cut-off and case-II has data only resolution better that 3 Å in section 5.2.2. In this figure, we have plotted the difference in average shape of dimers in average context as the scatter plot and **dashed** line is the average difference between two data sets for a given internal coordinate.

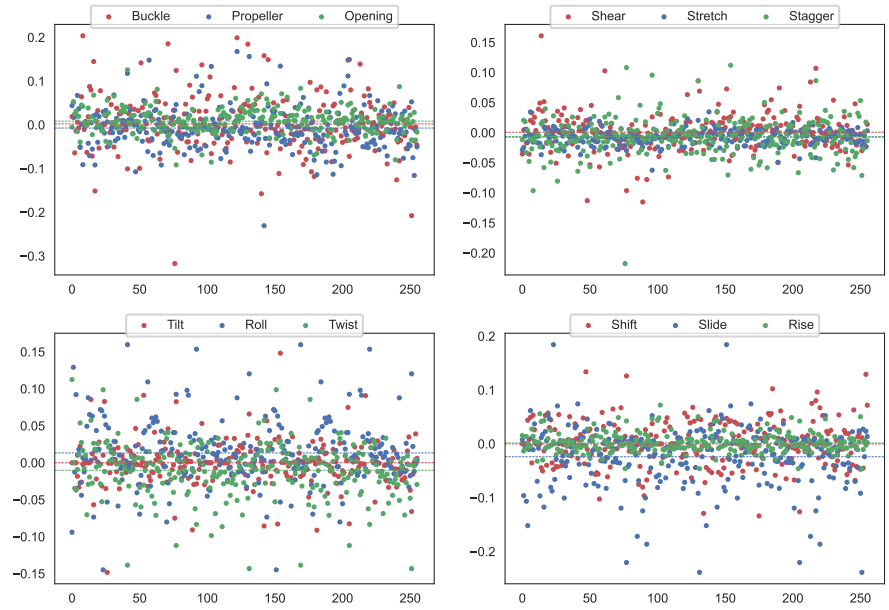


Fig. E.6 Plot comparing the average shape of dimer in two X-ray data sets as defined in case-I and case-II where case-I has no resolution cut-off and case-II has data only resolution better that 3 Å in section 5.2.2. In this figure, we have plotted the difference in average shape of tetramers as the scatter plot and dashed line is the average difference between two data sets for a given internal coordinate.

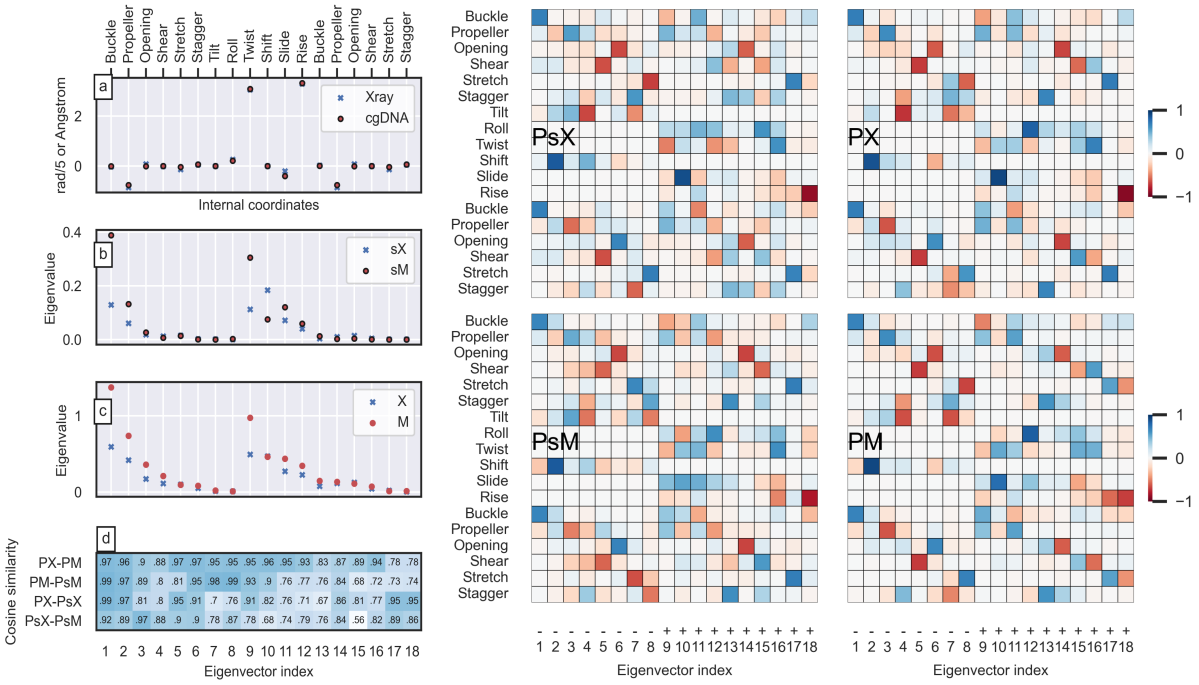


Fig. E.8 a) Plot comparing sequence-independent groundstate (average shape) of dimer coordinates in X-ray (case-II) and cgNA+ model data set. On right, P_sX and P_sM are the associated eigenvector matrices for the shape covariance matrix (denoted by subscript s) describing the directions of variation in groundstate over sequence space for X-ray (denoted by superscript X) and cgNA+ model (denoted by superscript M) data sets, respectively and D_sX and D_sM are corresponding eigenvalues in b). While PX and PM are the eigenvectors of average configuration covariance describing the direction of deformation of DNA in configuration space and DX and DM are corresponding eigenvalues in c). In d), there is cosine similarity index for corresponding eigenvectors in (CX, CM), (C_sM, CM), (C_sX, CX), and (C_sX, C_sM).

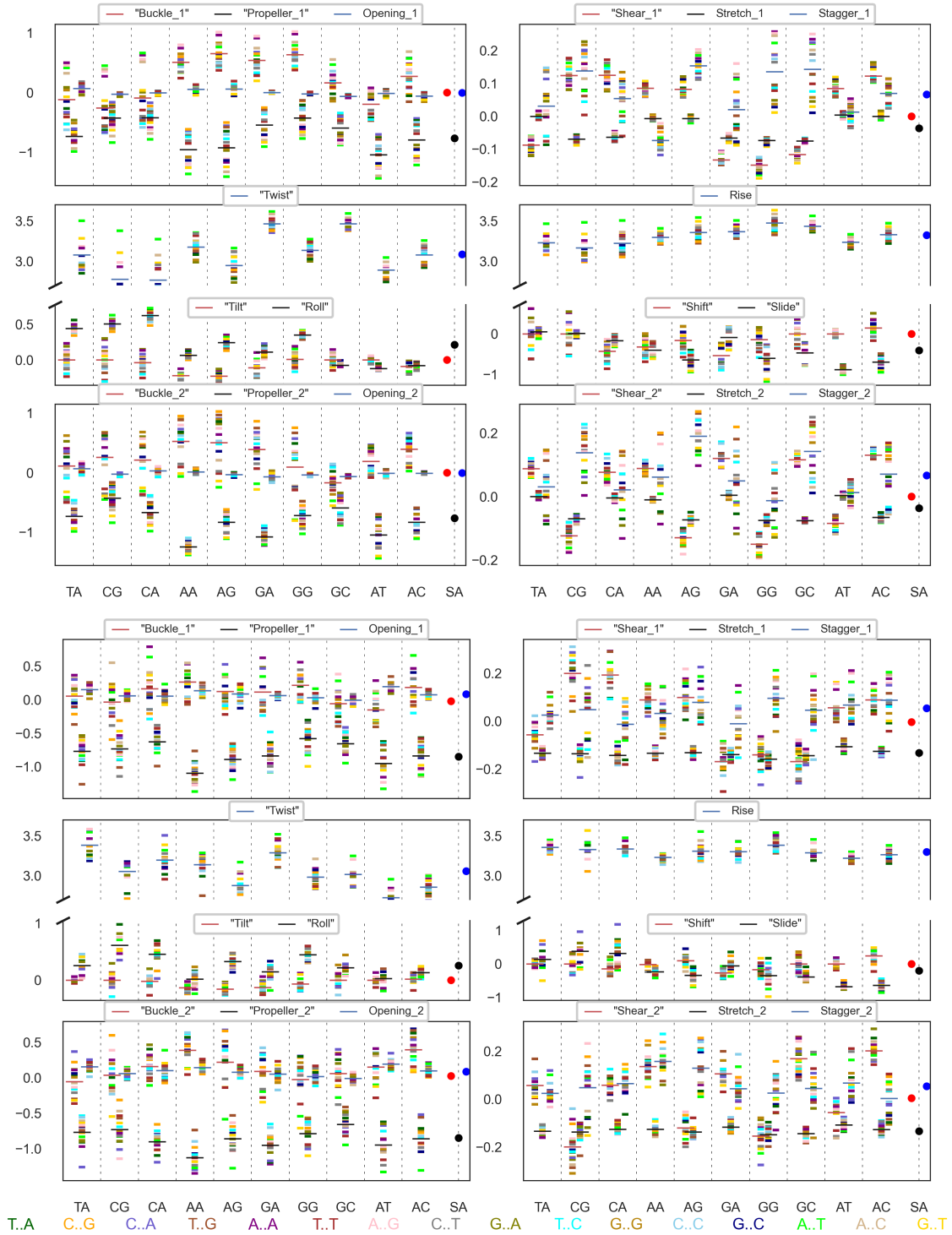


Fig. E.9 Plot of Intras and Inter for X-ray, case-II (bottom) and cgNA+ model (top) data set in which large dash lines depict ICs of a dimer (in average context) while the other smaller dash lines are the ICs for that dimer in a specific tetramer context. For a better and more concise visual representation, the three ICs are slightly shifted on the X-axis in each subplot. Also, various flanking contexts are plotted in different colors, as described at the bottom of the plot. SA is sequence-average groundstate.

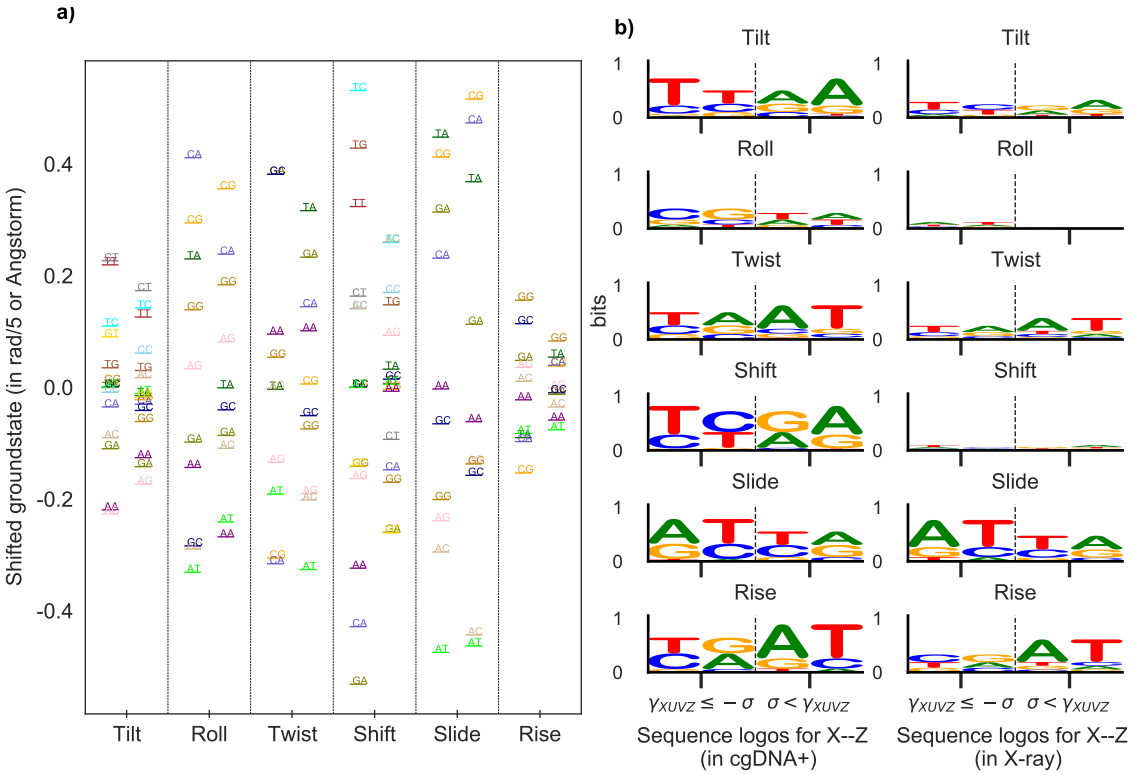


Fig. E.10 a) Inter ICs (shifted with respect to sequence-average groundstate) are plotted for dimers in average flanking context to identify which dimers assume distant values from sequence-average groundstate for a given variable and whether that signal is consistent in the two data sets. The left column is for the cgNA+ model data set for each IC, and the right column is for the X-ray data set. b) Sequence logo plot to statistically quantify the role of tetramer context on the ground-state (in inter variables) of a given dimer. For each internal coordinate (IC), we have defined $\gamma_{\mathrm{XUVZ}} = \mathrm{IC}_{\mathrm{XUVZ}} - \mathrm{IC}_{\mathrm{X_{avg}UVZ_{avg}}}$ as the difference of the internal coordinate of a dimer (UV) in tetramer context (X - - Z) with the same dimer in average context, where X, U, V, Z $\in$ [A, T, C, G]. Then, for each internal coordinate, we have defined positive and negative outliers as, $\gamma_{\mathrm{XUVZ}} < -\sigma$ and $\gamma_{\mathrm{XUVZ}} > +\sigma$, where σ is standard deviation of γ_{XUVZ}. In the sequence-logo plot, we have plotted the information content in the tetramer flanking context (X - - Z) for which γ_{XUVZ} are negative or positive outliers.

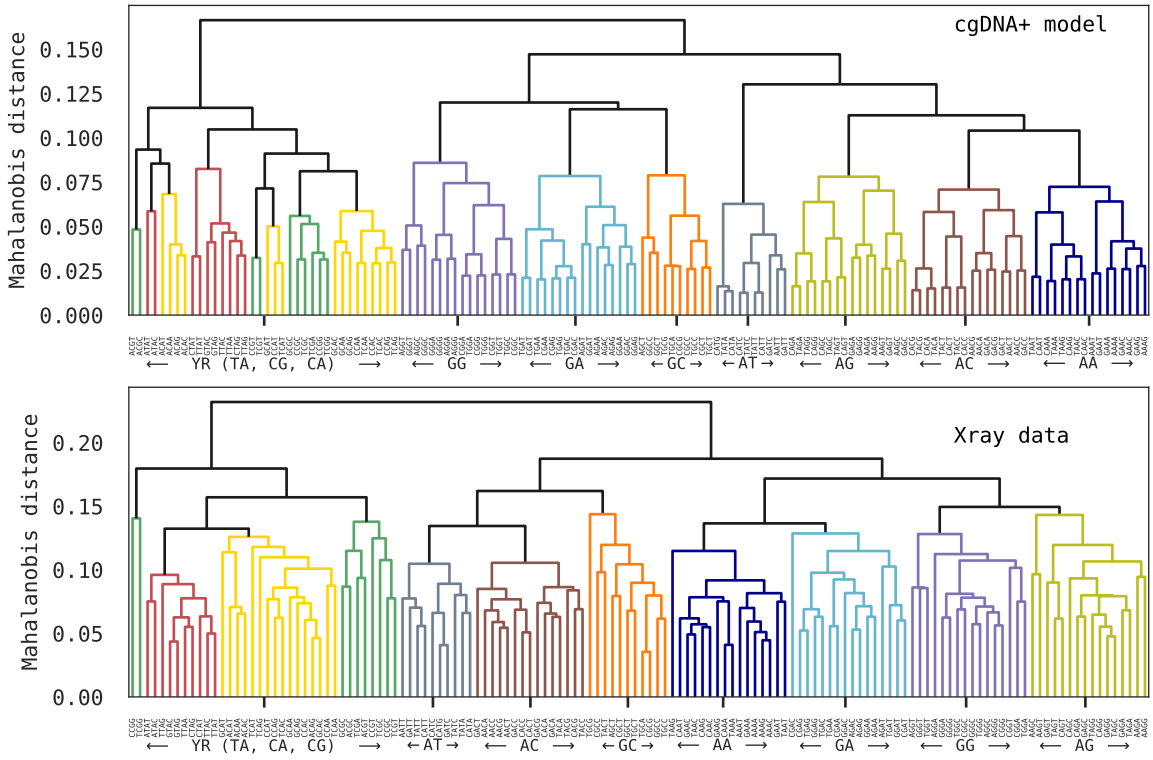


Fig. E.11 Dendrograms using hierarchical clustering on independent tetramers using Mahalanobis distance (taking inverse of sequence-dependent configuration covariance as the weight matrix) and average linkage algorithm section 5.2.3.3.

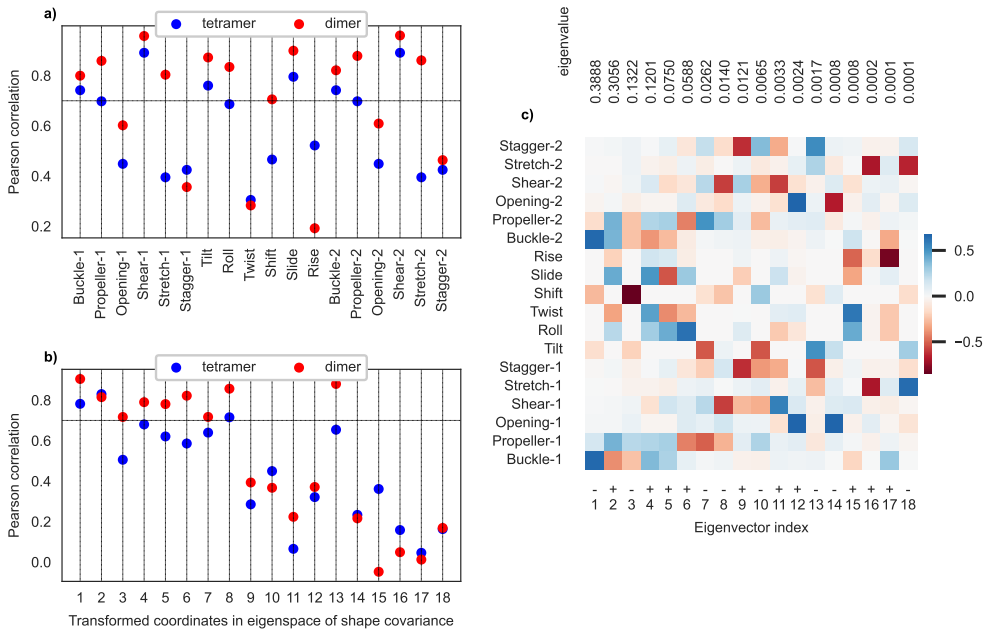


Fig. E.12 Pearson correlation between X-ray (case-II) and cgNA+ data set in a) standard CURVES+ coordinates and b) transformed coordinates in eigenspace of cgNA+ shape covariance and corresponding eigenvectors shown in c) with +/− parity as defined in section 5.2.3.2.

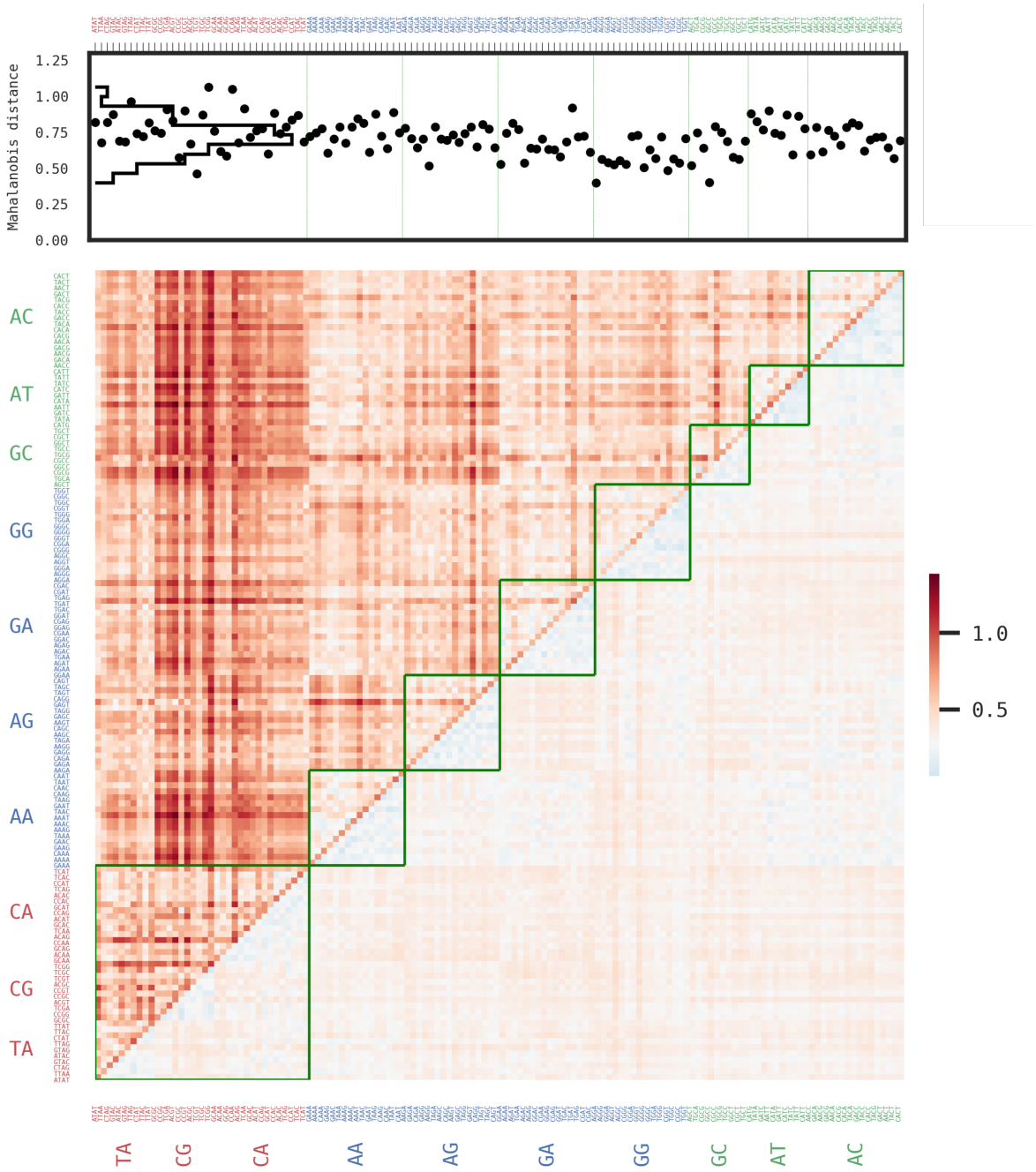


Fig. E.13 In the heat map (bottom), the diagonal entries are Mahalanobis distance between the groundstate of dimers (in 136 independent tetramer contexts) in X-ray and cgNA+ model data set. Whereas lower and upper off-diagonal entries are Mahalanobis distance between different dimers (in specific tetramer context) within cgNA+ model and X-ray data set, respectively. The diagonal entries of heat-map are again plotted in scatter plot (top) along with the histogram in the same plot. Note that the Mahalanobis distance (defined in section 2.5.5) is computed in the 18 CURVES+ coordinates and using cgNA+ shape covariance matrix as weight matrix.

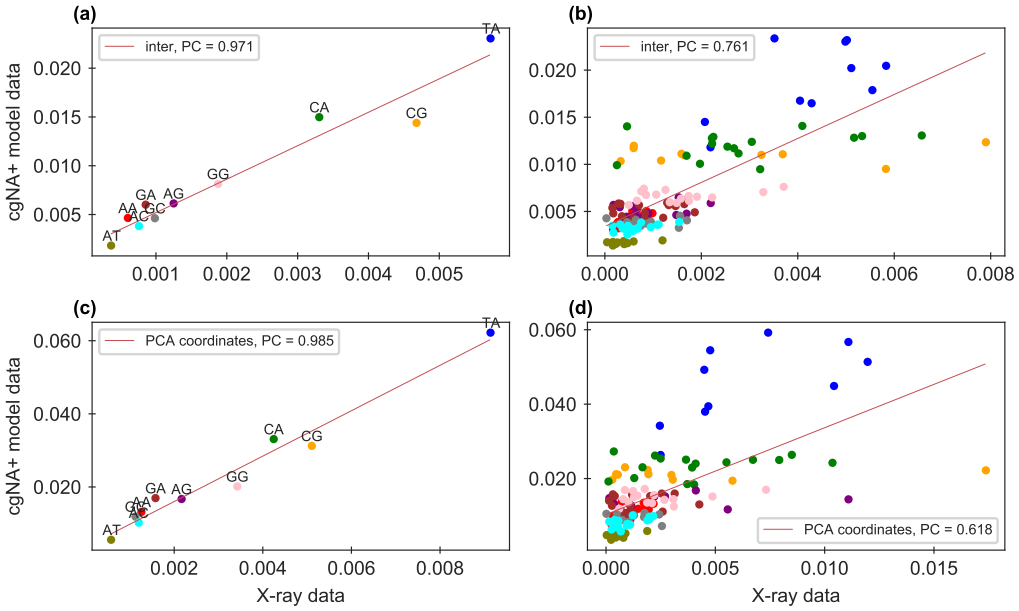


Fig. E.14 Comparison of configurational volume for cgNA+ model covariance vs X-ray data set (case-II) covariance a) in inter coordinates for independent dimer steps in average context, b) in inter coordinates for dimers in independent tetramer contexts, c) in PCA coordinates (in eight principal modes of cgNA+ model shape covariance $\in \mathbb{R}^{18}$) for independent dimer steps in average context, d) in PCA coordinates (in eight principal modes of cgNA+ model shape covariance $\in \mathbb{R}^{18}$) for dimers in independent tetramer contexts.

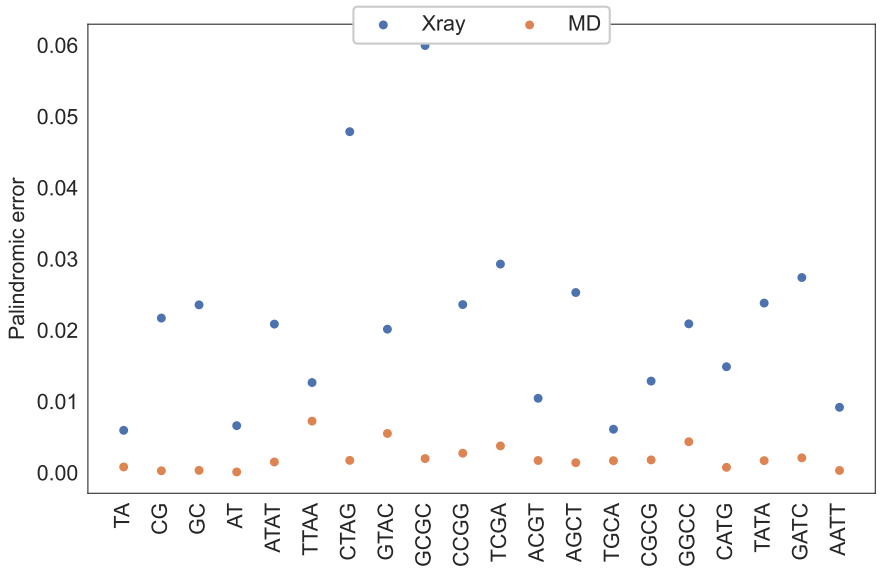


Fig. E.15 Palindromic error (as defined in section 2.5.1) per degree of freedom in the groundstate of palindromic dimer in tetramer flanking context and in average flanking context for X-ray data set (Case-II) and MD simulations (used to train cgNA+ model).

Appendix F

Codes and data availability

The current version of the cgNA+ model is available (in python and MATLAB) at https://github.com/rahul2512/cgNA_plus. This code is straightforward to use and requires a few python libraries to run the python code. The Python version has more functionalities; for instance, the code to obtain the sugar ring from cgNA+ coarse-grained configuration.

The codes used for training the cgNA+ model are available at https://github.com/rahul2512/cgNA_plus_training and https://github.com/rahul2512/cgNAplus_sugar_fitting. Note that these codes require data to train the model, which is not provided on GitHub due to size limits. However, all the data can be obtained on request to rs25.iitr@gmail.com.

Finally, all the codes used for plotting the figures in this thesis can be accessed at https://github.com/rahul2512/rsharma_thesis along with the latex file of this thesis. Note that some of the data must be requested to rs25.iitr@gmail.com due to the size limits on GitHub.

Curriculum vitae

Rahul Sharma hails from Loharu, a beautiful and historic town in the north of India, and was born in 1992. He received his integrated bachelor's and master's degree in Chemistry from the Indian Institute of Technology Roorkee. During his undergraduate studies, he had the opportunity to work on exciting projects studying biomolecules using molecular simulations and developed an interest in computational chemistry. In June 2018, he started his doctoral studies under the supervision of Prof. John H. Maddocks at EPFL.