Atomistic Simulations of Supercoiled Linear DNA Under Tension

Matthew Burman

Doctor of Philosophy

University of York
Physics

June 2023

Abstract

DNA in living beings is constantly subject to torsional stress as a result of processes such as transcription and replication, as well as the action of nucleosomes and nucleiod-associated proteins. This stress is relieved through DNA supercoiling; a process by which turns are added to or taken away from DNA, forming higher-order structures known as plectonemes. These structures are the result of the DNA duplex wrapping around itself in three dimensions, and act to bundle DNA together, compacting it. DNA supercoiling also plays a role in protein recognition and gene regulation, promoting the binding of transcription factors, and forming topological barriers. This thesis presents the results of all-atom molecular dynamics simulations of supercoiled DNA under tension, with the aim of understanding the dynamics of structure formation, as well as the role of sequence in dictating their behaviour.

These simulations re-create the experimental ‘hat-curve’, displaying clear asymmetry between positive and negative supercoiling, and revealing the presence of denaturation bubbles in negatively supercoiled systems. Showing high levels of co-localisation with the tips of plectonemes, these denatured regions confirm the existence of the ‘tip-bubbles’ observed previously in coarse grained simulations. Simulations also unveil the existence of ’tip-bubbles’ in positive supercoiling, which show high levels of curvature.

Also demonstrated is the possible role of sequence in structure formation, with regions of bubble formation clearly identified by predictive methods, but with probability landscapes perturbed by sharply bent plectoneme tips. Results also indicate that plectoneme nucleation locations are influenced by the inherent curvature of the DNA, with highly curved regions of DNA predicting the locations of plectoneme formation.

Finally, a simple model in which both size and ground path curvature is used to predict plectoneme position in positive supercoiling is formulated, with preliminary results from magneto-optical tweezers appearing to be predicted by these two factors.

Contents

Abstract 3

Contents 4

List of Figures 10

1 Introduction 13

1.1 Thesis structure 13

1.2 DNA as a carrier for genetic material 14

1.3 Nucleic acid structure 15

1.4 Base-pair and base-step parameters 16

1.5 DNA supercoiling 17

1.5.1 Formal definitions of DNA supercoiling 18

1.5.2 Structures of supercoiled DNA 19

1.6 Single molecule experiments 20

1.6.1 Visualisation of individual DNA molecules 20

1.6.2 Force transduction techniques on DNA 21

1.7 Modelling DNA under tension and torsion 25

1.7.1 Continuum models 25

1.7.2 Statistical models 29

1.7.3 Coarse-grained modelling 32

1.8 Atomistic molecular dynamics simulations 34

2 Methods 36

2.1 Molecular dynamics 36

2.2 Structure creation and simulation initialisation 36

2.3 Force fields 37

2.3.1 The Lennard-Jones potential 37

2.3.2 The Coulomb potential 37

2.3.3 Bond length and angular potentials 38

2.3.4 Torsional potential 39

2.3.5 The AMBER potential 40
2.4 Solvent models 40
2.4.1 Explicit solvent 41
2.4.2 Implicit solvent 42
2.5 Integrators 45
2.5.1 The leapfrog method 46
2.5.2 Velocity Verlet 46
2.5.3 Choice of time step 46
2.6 Constraints and restraints 47
2.7 Thermostats and barostats 48
2.7.1 Berendsen thermostat and barostat 48
2.7.2 Langevin dynamics 49
2.8 Simulation setup 50
2.8.1 Simulation parameters 50
2.8.2 Application of tension 51
2.8.3 Keeping the strand on-axis 52
2.8.4 Torsional restraint 52
2.8.5 Excluded volume 53
2.8.6 Additional equilibration 55
2.8.7 Summary of simulation methodology 55
2.9 Analysis of simulations 56
2.9.1 Determination of denatured base pairs and bubbles 56
2.9.2 Determination of toroidal and plectonemic loops 57
2.9.3 Determination of twist and writhe 58

3 All-atom MD Simulations Recreate Experimental Hat-curves 59
3.1 Introduction 59
3.2 Simulation setup 61
3.3 Recreation of experimental hat curves 61
3.4 Simulations give insight in to how DNA responds to tension and torsion 63
3.5 Structure and role of bubbles in supercoiled DNA 66
3.5.1 An overview of bubble prevalence 67
3.5.2 Bubble Stability 68
3.5.3 Twist absorption by bubbles 68
3.5.4 Bending in bubbles 70
3.5.5 Structure of bubbles in positive supercoiling 73
3.6 Structure and dynamics of plectonemes 74
3.6.1 Plectoneme size as a function of force 75

3.6.2 Plectoneme formation . . . 75
3.7 Plectoneme-bubble interactions . . . 76
3.7.1 Plectoneme tip-bubbles . . . 77
3.7.2 Toroid and plectoneme migration . . . 78
3.8 Anomalous systems . . . 80
3.9 Sequence dependence in plectoneme and bubble formation . . . 82
3.9.1 Bubble formation . . . 82
3.9.2 Plectoneme formation . . . 83
3.10 Summary . . . 86

4 Sequence Dependence of DNA Structure Formation 87
4.1 Introduction . . . 87
4.2 Failed attempt - AT runs . . . 87
4.2.1 Sequences . . . 87
4.2.2 Problems with these sequences . . . 88
4.3 Semi-randomised sequences . . . 90
4.3.1 Sequence design . . . 90
4.3.2 Behaviour of the ‘centred’ sequence . . . 91
4.3.3 Behaviour of the ‘offset’ sequence . . . 96
4.4 Summary . . . 102

5 Understanding Large Scale Sequence Dependence 103
5.1 Predicting plectoneme location . . . 103
5.1.1 Two-state model of plectoneme formation . . . 104
5.1.2 Algorithmic implementation of the two-State model . . . 105
5.1.3 Testing the new energetics . . . 107
5.1.4 Comparisons to experiment . . . 110
5.2 Summary . . . 113

6 Discussion 114
6.1 Future work . . . 115

Appendix A Sequences and additional analysis 117
A.1 Sequence used in hat-curve simulations . . . 117
A.2 Failed ‘centred’ sequence . . . 117
A.3 Failed ‘offset’ sequence . . . 117
A.4 ‘Centred’ sequence . . . 118
A.5 ‘Offset’ sequence . . . 118
A.6 Kymographs and end-to-end distances for all hat-curve simulations . . . 118

Appendix B Code Listings **127**

B.1 Python code to find bubbles within criteria 127

B.2 Bubble tracking python code 129

B.3 Plectoneme finder python code 134

B.4 Modified igor pro code 135

B.5 Energy penalty code 140

List of Figures

1.1 Schematic structure of DNA 15
1.2 DNA base pair and base step parameters 16
1.3 Schematic of twist and writhe. 18
1.4 Illustration of the stages of Cryo-EM 21
1.5 Illustrations of DNA curtains and TPM. 22
1.6 DNA magnetic tweezers and the experimental hat-curve. 23
1.7 Variants of optical tweezers experimental setups. 24
1.8 Diagram of a freeky jointed chain. 25
1.9 Diagram of the worm-like chain. 26
1.10 Comparisons of DNA stretching models. 28
1.11 Hat-curve predictions from the model of Moroz and Nelson and Marko and colleagues. 28
1.12 Predictions of bubble formation from SIDD, along with the predictions of plectoneme locations from the work of Dekker and colleagues. 31
1.13 Description of the Ox-DNA coarse grained model, as well as hat-curves and observed plectoneme tip-bubbles. 34
1.14 Comparison of AFM and molecular dynamics on DNA minicircles. 34

2.1 Lennard-Jones and coulomb potentials. 38
2.2 Hookean and torsional bonded potentials. 39
2.3 Restraints used prevent supercoiling from being removed from simulated supercoiled DNA. 51
2.4 Restraints used to create an excluded volume in AMBER. 53
2.5 Restraints used to hold a section of DNA straight. 55
2.6 Method used to locate plectoneme crossing points. 57

3.1 Hat-curve recreated in MD simulations, along with plots of twist and writhe. . . 62
3.2 Landscape of representative structures from all simulations. 64
3.3 Plot of base pairs in bubbles for all simulations. 67
3.4 Plot of bubble lifetimes and sizes for all simulations. 68
3.5 Twist absorption of bubbles in all simulations. 69

3.6 Bending profiles in bubbles for all simulations. 70
3.7 Three distinct bubble states observed in simulations. 72
3.8 Bubble structure and groove widths in positive supercoiling. 73
3.9 Base pairs in plectonemes for all simulations. 74
3.10 Plectoneme sizes as a function of force for $\sigma = 0.06$ and 0.08 75
3.11 Dynamics of plectoneme formation in a single simulation. 76
3.12 Plectoneme-bubble separation in all simulations. 77
3.13 Plectoneme mobility in positive supercoiling. 78
3.14 Dynamics of a single migrating toroid in a single simulation. 79
3.15 Plectoneme mobility in negatively supercoiled simulations. 80
3.16 Dynamics of a single anomalous simulation. 81
3.17 Bubble formation locations and predictions at $\sigma = -0.1$ 83
3.18 Predicted ground path curvature for the sequence simulated in this chapter, with curvature measured over a series of base pair steps. 84
3.19 Comparisons of predicted curvature and plectoneme formation. 85

4.1 AT percentages in the failed sequences. 88
4.2 Representative images for the failed sequences. 89
4.3 Repesentation of both block-averaged sequences. 90
4.4 Representative structures for simulations of the centred sequence. 91
4.5 Plectoneme formation in the centred sequence illustrated by strong bending. . . 92
4.6 Bubble locations and predictions in the centred sequence. 94
4.7 Plectoneme positions and predicted curvature in the centred sequence. 95
4.8 Illustration of problems with the projection method in locating plectonemes. . . 96
4.9 Representative structures of all simulations of the offset sequence. 97
4.10 Dyamics of the $\sigma = +0.1$ simulation of the offset sequence. 98
4.11 Dyamics of the $\sigma = -0.1$ simulation of the offset sequence. 99
4.12 Bending in all simulations of the offset sequence, along with predicted ground path curvature. 100
4.13 Bubble locations and predictions for all simulations of the offset sequence. . . . 101

5.1 Values of writhe per base pair calculated from experimental results. 106
5.2 Position-dependent energy penalties in plectoneme formation. 106
5.3 Energy distributions between twist and writhe for a 10kbp duplex. 107
5.4 Demonstration of changes in predicted plectoneme location according to force and supercoiling density. 109
5.5 Fluorescence images of supercoiled DNA obtained from magneto-optical tweezers. 110
5.6 Observed and predicted plectoneme positions from experiments of a 15.6kbp sequence in positive supercoiling. 111

5.7 Observed plectoneme positions in a negatively supercoiled experiment compared to predictions of both bubble and plectoneme position. 112

A.1 Kymographs and end-end distances for $\sigma = -0.1$ 119
A.2 Kymographs and end-end distances for $\sigma = -0.08$ 119
A.3 Kymographs and end-end distances for $\sigma = -0.06$ 120
A.4 Kymographs and end-end distances for $\sigma = -0.04$ 120
A.5 Kymographs and end-end distances for $\sigma = -0.02$ 121
A.6 Kymographs and end-end distances for $\sigma = 0$ 121
A.7 Kymographs and end-end distances for $\sigma = 0.02$ 122
A.8 Kymographs and end-end distances for $\sigma = 0.04$ 122
A.9 Kymographs and end-end distances for $\sigma = 0.06$ 123
A.10 Kymographs and end-end distances for $\sigma = 0.08$ 123
A.11 Kymographs and end-end distances for $\sigma = 0.1$ 124
A.12 Kymographs and end-end distances for the ‘centred’ sequence. 125
A.13 Kymographs and end-end distances for the ‘Offset’ sequence. 126

Acknowledgements

First and foremost I would like to thank my supervisor, Agnes Noy, for her constant support and encouragement over the entire course of this project. Her ability to give helpful and concise guidance in a patient and understanding manner, without being too hands-on, created the ideal circumstances for me to develop as a scientist without getting completely lost in what ended up being a complex and tricky project.

Further, I would like to thank the wider Physics of Life group at York. Every member of the group has been extremely welcoming, kind and willing to offer advice whenever it was needed. I also would like to thank my fellow PhD students, not only for their help and guidance scientifically, but also for their friendship. In particular, I would like to thank Seb, Matteo and Elliot for their constant companionship throughout the course of this PhD, both in and out of the office.

Finally, I would like to thank my family, whose support and encouragement, not only during this PhD, but also in all of the years that led up to it, allowed me to pursue a career for which I have a true passion.

This work has been supported by grants from the Engineering and Physical Sciences Research Council. Calculations were performed on JADE2, the Cambridge Tier-2 computing cluster, Bede, and the local facilities at the University of York, specifically Viking.

Declaration

I declare that this thesis is a presentation of original work and that I am the sole author. This work has not been presented for an award at this, or any other, University. All sources are acknowledged as references.

All simulations in this text, as well as their accompanying analysis (except when presented as examples of prior or ongoing work in the field, and acknowledged with a reference), were performed independently by me, with support and guidance from my supervisor Agnes Noy.

All experimental results presented in this text were gathered using optical-magnetic tweezers, with the technique developed, and results gathered by, Drs Jack Shepherd, Jamie Howard, Sébastien Guilbaud and Jack Zhu, under the supervision of Prof. Mark Leake.

Publications

1. Victor Velasco-Berrelleza, Matthew Burman, Jack W Shepherd, Mark Leake, Ramin Golestanian and Agnes Noy 2020 "SerraNA: a program to determine nucleic acids elasticity from simulation data" *Physical Chemistry Chemical Physics* 22, 19254–1926

2. Emil L Kristoffersen, Matthew Burman, Agnes Noy and Philipp Holliger 2022 "Rolling circle RNA synthesis catalyzed by RNA" *elife* 11:e75186

3. Jack W Shepherd, Sebastien Guilbaud, Zhaokun Zhou, Jamieson Howard, Matthew Burman, Agnes Noy, and Mark C Leake 2023 "DNA plectoneme size, position, and mobility revealed by fluorescence microscopy and combined optical and magnetic tweezers" (Under review)

4. Matthew Burman and Agnes Noy "Atomic description of the reciprocal action between supercoils and melting bubbles on linear DNA" (In preparation)

Chapter 1

Introduction

1.1 Thesis structure

This thesis describes a series of simulations modelling the behaviour of DNA under increasing levels of torque and tensile force, as well as a simple theoretical model for predicting locations of plectoneme formation in long sequences. This chapter provides biological background, including the basics of DNA, the theoretical background of supercoiling, and descriptions of the experimental techniques used to study it. In chapter 2, simulation methodology is described in detail, beginning from a description of the AMBER force field, and moving to a detailed guide to setting up simulations of supercoiled linear DNA. In addition to this are descriptions of the analysis techniques used to understand their behaviour.

Chapter 3 describes a large set of simulations of supercoiled DNA under varying levels of supercoiling density and force. This work displays the ability of atomistic simulations to describe the so-called 'hat-curve', a result that has been seen experimentally several times, and gives a detailed description of the structure formation that gives rise to its shape. In addition, the existence of plectoneme tip-bubbles is confirmed, with plectonemes and bubbles regularly seen to co-localise, with examples even seen in positive supercoiling.

Chapter 4 attempts to use simulations to describe the sequence-dependent behaviour of supercoiled DNA. This work reveals complex interactions between bubbles and plectonemes, and the wide variety of sequence-dependent effects seen in their formation, as predicted by a pair of theoretical models.

Chapter 5 attempts to marry the observations from simulations, experiments and statistical models through the use of a simple two-phase model for supercoiled DNA. This model is able to re-create the movement of plectonemes to the centre of their respective sequence seen in simulations, further reinforced by early experimental results.

Chapter 6 concludes by summarising the results described in the preceding chapters, evaluating the success of the project and the simulations therein, and proposing areas for future work.

1.2 DNA as a carrier for genetic material

The discovery of DNA and the study of its role as the carrier of genetic material has been the source of some of the most important scientific discoveries of the last two centuries. Although the idea of transferring characteristics from generation to generation was first suggested by the “Father of Genetics" Gregor Mendel in 1866 [1] it was not until the isolation of the substance contained within the cell nucleus, at the time called “nuclein", by Friedrich Miescher in 1868 [2] that research on DNA became a source of interest. Following this discovery, Albrecht Kossel [3] was able to identify nuclein as a nucleic acid, and subsequently isolate the four nucleic acid bases adenine (A), cytosine (C), guanine (G) and thymine (T), an achievement for which he would later earn a Nobel Prize. Building on the work of Kossel, Phoebus Levene discovered the repeating“phosphate-sugar-base" order of nucleic acid components, naming these repeating units “nucleotides" [4].

Despite the discoveries of Kossel and Levene, the idea of DNA as a carrier of genetic information was largely dismissed, with its form thought to be too simple and repetitive to contain any significant information. Its was not until the work of Avery in 1944 [5] that DNA was identified as the “transforming principle", that is the substance that can be transferred from dead bacterial cells to living ones, allowing the living cells to express features of the deceased. Avery’s work went on to inspire research by Erwin Chargaff, leading to his formulation of “Chargaff’s rules", which showed that, in DNA, the number of guanine units equals the number of cytosine units, and the number of adenine units equals the number of thymine units [6]. These rules hinted strongly towards the base pair structure of DNA, but it was not until the X-ray diffraction work of Rosalind Franklin [7], and the subsequent proposal for the structure of DNA by Watson and Crick [8] that these ideas were confirmed. The structure proposed in these works was one of two complementary strands arranged in a double helix, with each strand composed of consecutive “phosphate-sugar-base" units bonded covalently, and with nucleotides on opposite strands linked by hydrogen bonds. Inspired by Chargaff’s rules, the bonding of these nucleotides was shown to be specific, with A exclusively binding to T, and G exclusively binding to C. Later, more exotic forms of base pair bonding were proposed by Hoogsteen [9], although Watson-Crick base pairs are by far the most common form.

A key observation that Watson and Crick took from the specificity of base pairing was that any single strand of DNA encodes its entire structure, forming a simple basis for DNA replication. One can easily imagine the replication of a DNA sequence without the need for proteins: the double helix can unwind, exposing hydrogen bonding sites, in time these sites would attract complementary nucleotides which then bind to one another to create a new strand. While in reality the decoding of genomes is much more complex than this, it starts in much the same way, with the binding of nucleotides instead replaced by ribonuceotides, creating a messenger RNA copy of the strand. DNA and RNA are fundamentally similar molecules, only differing in the structure of their backbone (ribose for RNA, deoxyribose for DNA) and in the base complementary to A (Uracil for RNA, Thymine for DNA). This messenger RNA then codes for proteins in three base-pair blocks known as “codons", with each of these blocks coding for a single amino acid. This code is translated by binding with carrier RNA, a carrier of amino acids, which then synthesises the required polypeptide chain, forming a protein [10].

While information storage is often recognised as the sole function of DNA, this is not in fact the case. A large portion of any functional DNA sequence is made up of non-coding regions [11], the function of which is a subject of much debate. Studies on the function of these regions have suggested a series of answers, ranging from the structural maintenance of the genome [12] to gene regulation [13]. Some non-coding regions are also transcribed in to non-coding RNA, which can take many forms. One such form is micro-RNA, which acts as an RNA silencer, regulating gene expression in the post-transcriptional phase [14].

1.3 Nucleic acid structure

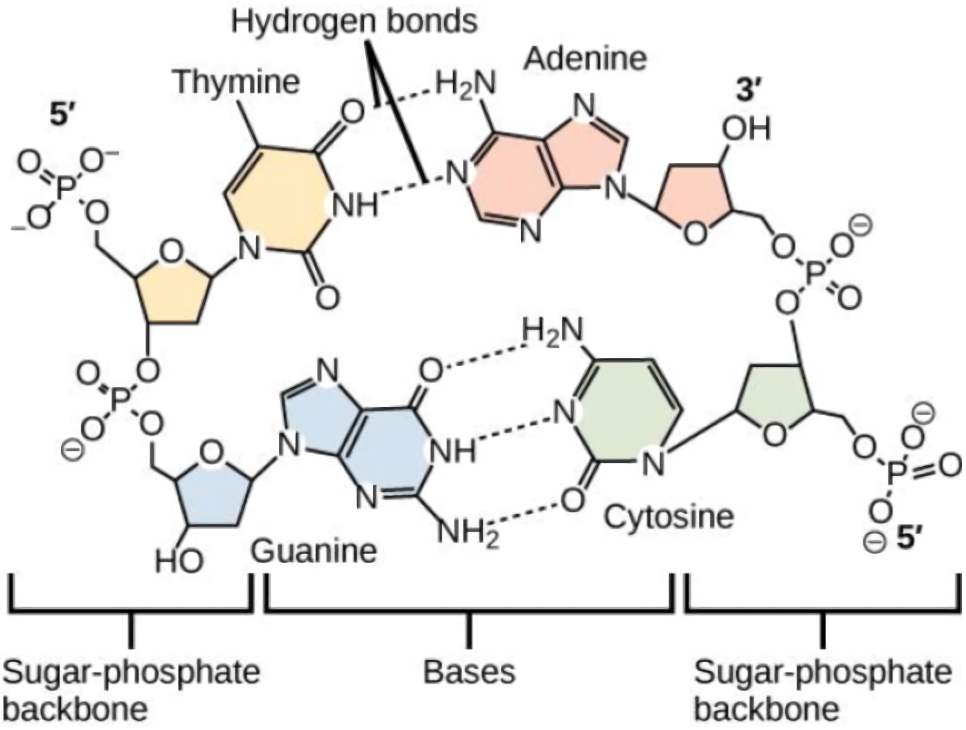


Figure 1.1: Schematic structure of DNA including all four bases as well as the sugar phosphate backbone. Image from [15].

The structure proposed by Watson and Crick, consisting of a double helix with 10 base pairs per turn and a distance between neighboring base pairs of 3.4Å, is named B-DNA. This is by far the most common and biologically relevant DNA form, being the most stable form that the strand can adopt at the temperatures and ion environments seen in living cells. The stabilisation of B-DNA structure relies on two sets of non-covalent molecular interactions: the stacking of pi orbitals around the aromatic rings of the nucleotides, often called “stacking interactions", and the lateral bonding of paired bases, whose strengths vary depending on the bases in question. The variability of the strength of the lateral bonding arises from the difference in number of hydrogen bonds between the two types of base pair, with A-T pairing containing two hydrogen bonds, and G-C base pairing containing three (see figure 1.1).

While B-DNA is the most common form of DNA, it is not the only one, with the duplex able to adopt a variety of conformations dependent on environment, temperature and mechanical stress. The most biologically compatible forms of ‘exotic’ DNA are A-DNA and Z-DNA, The structural parameters of which can be seen in table 1.1. A-DNA is similar in structure to B-

DNA, but with a larger number of base pairs per turn and a greater helical pitch, and occurs in low hydration environments, such as in dehydrated bacteria [16] and in the binding of specific proteins [17]. Z-DNA is distinctly different to both other forms, showing left-handed helicity, major and minor grooves of similar widths, and a repeating unit of 2bp [18]. Z-DNA is mostly present in negatively supercoiled DNA [19], and is thought to play a role in transcription [20] as well as certain genetic diseases [21].

Form	A-DNA	B-DNA	Z-DNA
Chirality	Right-handed	Right-handed	Left-handed
Repeating unit	1bp	1bp	2bp
Base pairs per turn	11	10.5	12
Base pair inclination	+19°	-1.2°	-9°
Base pair rise	2.3Å	3.32Å	3.8Å
Diameter	23Å	20Å	18Å

Table 1.1: Structural parameters of A-, B- and Z-DNA.

1.4 Base-pair and base-step parameters

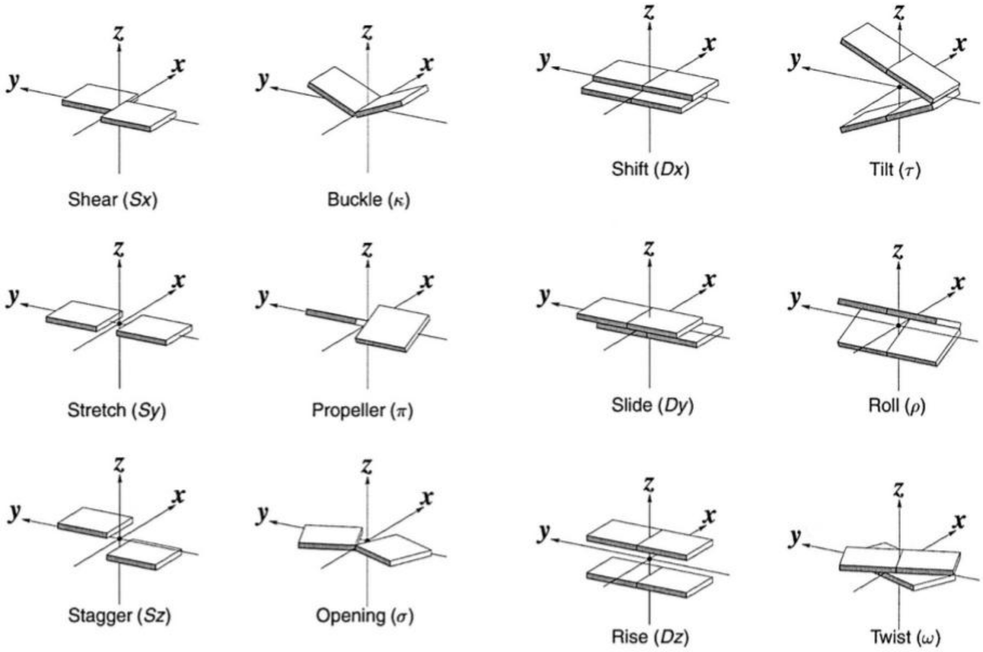


Figure 1.2: (**Left**) DNA translational (S_x, S_y, S_z) and rotational (κ, π, σ) base pair parameters. (**Right**) Base pair displacements (D_x, D_y, D_z) and rotations (τ, ρ, ω). Image from [22].

In order to fully classify the structure of DNA, a set of parameters that describe the relative orientations of paired bases and neighboring base pairs is needed. The most widely used method of calculating base pair and base step parameters is the CEHS (Cambridge university Engineering department Helix computation Scheme) [23], being both mathematically rigorous and reversible [24]. This scheme takes advantage of the planar nature of DNA bases, representing each as a long, thin cuboid, using the relative orientations of these cuboids to calculate the various parameters.

Base-pair parameters describe the details of the relative orientations of paired bases. These are are split into six separate parameters, first are the three translations: shear (S_x), stretch (S_y) and stagger (S_z), which describe the movement of the paired bases relative to one another in the x, y and z directions respectively. Second are the rotations: buckle (κ), propeller twist (π) and opening (σ), which describe the relative rotations of the base pairs in the x, y and z directions respectively.

Base-step parameters describe the relative orientations of neighboring base pairs. These are once again described by six parameters, three displacements: shift (D_x), slide (D_y) and rise (D_z), and three rotations: tilt (τ), roll (ρ) and twist (ω). These parameters, along with the base pair parameters, are shown in figure 1.2.

The ability to accurately measure base-pair and base-step parameters is vital in the close examination of DNA, with their values underpinning the role of sequence in DNA structural properties [25] and correlations between them controlling response to external forces [26].

1.5 DNA supercoiling

The extraction of genetic information from DNA requires the unwinding of its double helical structure, an action which, by its very nature, requires the application of torsion. The application of torsion to DNA causes it to twist, with the direction of the applied torsion dictating the direction of twist. The twisting of DNA is known as DNA supercoiling, with twisting in the same direction as the helix called positive supercoiling, and twisting in the opposite direction called negative supercoiling.

DNA *in vivo* is rarely relaxed, and in both eukaryotes and prokaryotes it is maintained in a negatively supercoiled state [27, 28]. This negative supercoiling reduces the energy required to separate the strands, thereby reducing the energy barrier for the elongation stage of RNA polymerase [29]. The interaction between transcription and supercoiling also works in the opposite direction, with the strand separation generated by RNA polymerase inducing localised supercoiling: positive supercoiling is generated ahead of the transcription site while negative supercoiling is generated behind it [30]. This localised supercoiling, as well as global levels, must be carefully regulated by topoisomerase proteins [31] with failure to remove supercoiled loops ahead of transcription sites resulting in stalling or backtracking before transcription is complete [32].

It is clear from these simple interactions that DNA supercoiling plays a key role in the regulation of metabolic processes [33–36]. Perhaps the most famous example of this regulation is modulation of the *lac* operon in *Escherichia coli*. Here the *lac* repressor binds to three sites (*O1*, *O2* and *O3*) in the *lac* operon to create a 401 base-pair (bp) loop, and this loop acts as a topological barrier, preventing transcription [37].

A deeper understanding of the effects of DNA supercoiling, along with the factors that control it, will have a significant impact on our understanding of the function of DNA and the regulation of genes *in vivo*, and these impacts will be significant in the fields of gene therapy [38] and cellular engineering [39]. The ability to predict and conrol genome topology could open the way for the production of bespoke genetic switches, allowing for the direct control of gene expression. Further to this, DNA supercoiling may allow for the long range transmission of regulatory signals, opening the possibility of novel methods for the control of gene expression without the need to alter nucleotide sequence.

1.5.1 Formal definitions of DNA supercoiling

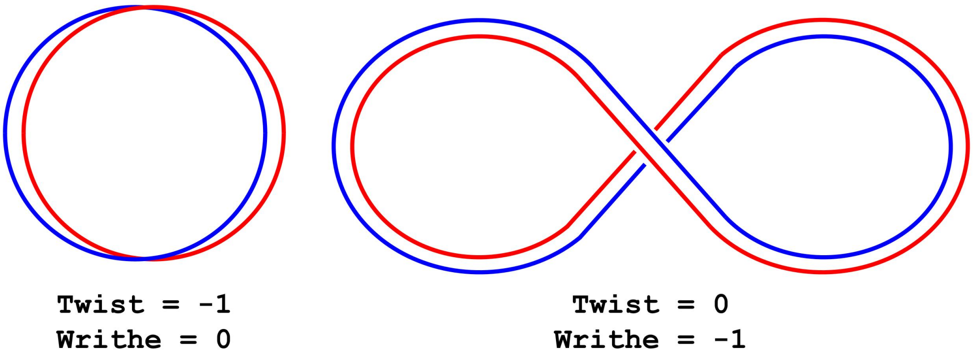


Figure 1.3: Schematic diagram of twist and writhe on a pair of linked loops with linking number -1.

Within a topologically bounded DNA duplex - that is a DNA duplex with torsionally restrained ends, out of which torsion cannot be transferred - the overall degree of supercoiling is constant. The simplest example of such a duplex is a DNA minicircle, in which torsion cannot be relaxed without the breaking of covalent bonds. The fixed number of turns in such a system can be defined mathematically using a quantity called linking number, *Lk*. Linking number is defined as the number of times two closed curves - such as the two strands of DNA in a minicircle - twist around one another. Formally, the linking number of any two closed curves whose coordinates are $\mathbf{r}_1(s)$ and $\mathbf{r}_2(s')$, with curivelinear coordinate s can be written as [40]

$$Lk = \frac{1}{4\pi}\oint_0^L ds \oint_0^L ds' \frac{\mathbf{r}_1 - \mathbf{r}_2}{|\mathbf{r}_1 - \mathbf{r}_2|^3}\left(\frac{d\mathbf{r}_1}{ds} \times \frac{d\mathbf{r}_2}{ds'}\right). \tag{1.1}$$

Fortunately this expression can be simplified for double helical DNA, being written as

$$Lk_0 = \frac{N}{p_0} \tag{1.2}$$

where N is the length of the duplex in bp, and p_0 is the average equilibrium pitch or the number of base pairs per turn (10.4 bp/turn in B-DNA). The addition or subtraction of turns perturbs the linking number, and the amount by which it changes can be defined by the excess linking number ΔLk, written as

$$\Delta Lk = Lk - Lk_0 \tag{1.3}$$

where *Lk* is the current value of linking number.

In a DNA duplex, *Lk* can be accommodated either by the twisting of the strands around one another, or by the writhing of the molecule in 3D (see figure 1.3). Each of these quantities is assigned a number, these being twist (Tw) and writhe (Wr), the sum of which is conserved according to the Fuller-White-Călugăreanu theorem [41–43]

$$Lk = Tw + Wr. \tag{1.4}$$

In the same manner as linking number, expressions for both twist and writhe can be formally defined, with twist written as [40]

$$Tw = \frac{1}{2\pi}\oint_0^L ds \frac{d\mathbf{r}}{ds}\left(\hat{\mathbf{u}} \times \frac{d\hat{\mathbf{u}}}{ds}\right), \tag{1.5}$$

with $\hat{\mathbf{u}}$ the vector connecting the single strands $\hat{\mathbf{u}} = (\mathbf{r}_1(s) - \mathbf{r}_2(s'))/|\mathbf{r}_1(s)\mathbf{r}_2(s')|$, and writhe written as

$$Wr = \frac{1}{4\pi}\oint_0^L ds_1 \oint_0^L ds_2 \frac{\mathbf{r}_1 - \mathbf{r}_2}{|\mathbf{r}_1 - \mathbf{r}_2|^3}\left(\frac{d\mathbf{r}_1}{ds_2} \times \frac{d\mathbf{r}_2}{ds_1}\right), \tag{1.6}$$

which is almost identical to equation 1.1, only now s_1 and s_2 are coordinates along a single curve, as opposed to s and s' which are coordinates along two separate curves.

The formal definitions of linking number, twist, and writhe are not all-together useful for this thesis, however they have been employed formally in the program WrLINE [44], the uses of which will be discussed in a later chapter.

1.5.2 Structures of supercoiled DNA

Linking number represents a global property of a torsionally constrained DNA duplex, but the way in which it is distributed does not need to be homogeneous. The complex structural properties of DNA along with sequence specific effects can lead to the concentration of excess torsion in specific locations, and a useful quantity for exploring these effects is the superhelical density σ, defined globally as

$$\sigma = \frac{\Delta Lk}{Lk_0}. \tag{1.7}$$

This value represents the change in linking number relative to its relaxed value and its usefulness lies in its length-independence, allowing it to be used as a measure of supercoiling at all length scales. In supercoiled DNA, twist represents the simplest way of absorbing excess linking number, with changes in the number of base pairs per turn able to account for for twisting in either direction. These over- or under-twisted duplexes are otherwise identical to their relaxed, canonical counterparts, with the only visual difference being in the tightness of the double helix itself.

As $|\Delta Lk|$ increases DNA duplexes begin to form writhed structures, in which the DNA wraps around itself in 3D space. The simplest of these structures is represented in figure 1.3, in which a linking number of -1 is accommodated through the formation of toroids (or more accurately teardrops). Decreasing the linking number further results in the formation of more tightly wrapped structures, known as plectonemes, in which multiple crossing points are seen. As more turns are added or subtracted from the duplex the structure of plectonemes becomes more and more compact, eventually resulting in the formation of highly compact rods.

The ability of twist and writhe to absorb torsion while maintaining the canonical form of the duplex is limited, and at high enough levels of supercoiling defects begin to form. The two most common forms of defects are kinks - in which the stacking interactions between neighboring base pairs break down, resulting in the formation of a sharp bend - and bubbles - in which the hydrogen bonding between base pairs breaks down, allowing for the localisation of large amounts of twist. The formation of these bubbles and kinks is thought to be reversible [45], allowing structures to effectively recover from disruption once torsion has been relaxed, for example by nicking the backbone.

1.6 Single molecule experiments

The study of supercoiled DNA *in vivo* can give insight on the impact of supercoiling on molecular processes. However, the study of DNA in cells is extremely difficult, owing to the plethora of actors in the cellular environment, coupled with high levels of crowding. As such, the supplementation of *in vivo* experiments with *in vitro* experiments is vital to developing our understanding of the behaviour of individual bio-molecules. DNA in one such bio-molecule - its high levels of stability and complex structural features make it the perfect target for these kinds of *in vitro* experiments.

1.6.1 Visualisation of individual DNA molecules

The ability to view single bio-molecules in great detail is vital to developing an understanding of their function, and as such techniques with a resolution that allows for detailed examination of their structure are of great importance. It is often the case that to obtain the highest amount of detail one must sacrifice the ability to view dynamic behaviour. Often techniques that are able to view bio-molecules with the greatest detail require them to be stuck in their ground state, whether it be by crystallisation, freezing or some other technique.

1.6.1.1 X-ray crystallography

X-ray crystallography requires a molecule to be arranged in a periodic crystal, a form in to which biomolecules can be prepared through a complex series of purification and crystallisation. X-rays with a wavelength similar to interatomic spacings are scattered through the layers of the crystal, diffracting and interfering with one another as they go. The diffraction patterns obtained can then be analysed to reconstruct an atomic representation of the molecule.

Crystallographic methods are extremely widespread, having been used by Rosalind Franklin to obtain the first image of the structure of DNA [7]. The use of X-ray crystallography is still widespread today, having been used to obtain most of the ≈100,000 structures found in the Protein Data Bank [46].

1.6.1.2 Cryogenic electron microscopy (Cryo-EM)

Cryo-EM uses the diffraction of electrons to map the charge density of a sample, this data can then be used to construct high resolution structures of a variety of molecules [48]. In contrast to X-ray crystallography, the samples used in cryo-EM do not need to be crystallised, rather the molecules are rapidly cooled, with low temperatures trapping the molecules in their physiological conformation. 3D structures are then obtained by combining images of identical

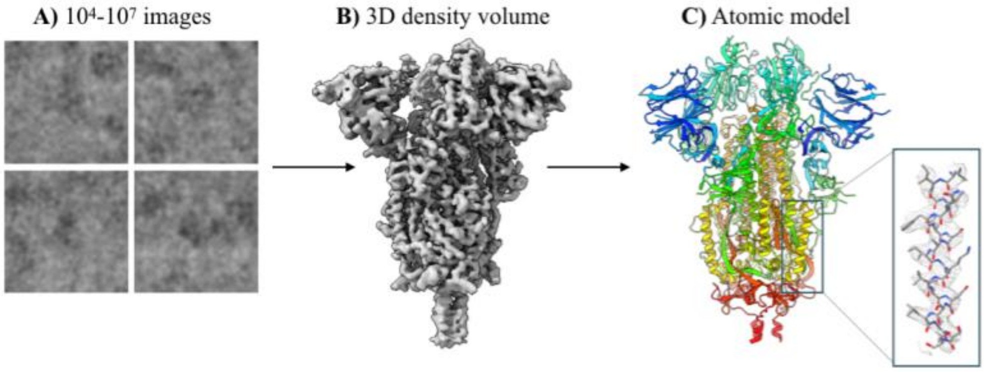


Figure 1.4: Construction of the structure of SARS-CoV2 spike glycoprotein using cryo-EM, from initial imaging (**A**) to atomic structure (**C**). Image from [47]

molecules in different orientations, with molecular structures then found by combining these 3D structures with atomic models [49].

1.6.1.3 Nuclear magnetic resonance (NMR)

NMR can obtain structures with a similar resolution to crystallography, but without the need for the molecules to be arranged in a crystalline lattice. This is achieved by exciting atomic nuclei with a strong magnetic field and measuring the frequency of the radiation they emit, with the exact frequency depending on the properties of the functional groups. These frequencies can be directly mapped to specific groups, combining to give the overall structure. In contrast to cryo-EM and X-ray crystallography, NMR can be carried out in liquid, allowing it to be used in the study of dynamics and kinetics in physiologically relevant environments [50]. However, the range of molecular sizes that can be explored with NMR is limited, as the radiation frequencies emitted by the molecules are vulnerable to interference, with larger molecules causing more interference.

1.6.2 Force transduction techniques on DNA

While the techniques above allow for the study of DNA and proteins in high detail, they are extremely limited in their ability to physically manipulate the molecules they are studying. The manipulation of DNA is vital in the study of supercoiling, and as such techniques that allow for the application of torque and/or tension, whether directly or indirectly, are of great interest.

1.6.2.1 DNA curtains

DNA curtains are systems in which hundreds or thousands of DNA molecules are anchored to a lipid bilayer, over which flow can be applied, resulting in extension of the DNA in the direction of the flow [51], an effect which can be seen on the left of figure 1.5. This technique is generally used to image protein-DNA interactions [52], with the large numbers of molecules allowing for the study of an entire landscape of interactions, and the ability to apply flow allowing for the study of these systems under a series of different conditions. While the setup used in DNA curtains experiments could in principle be used to study supercoiled DNA, it

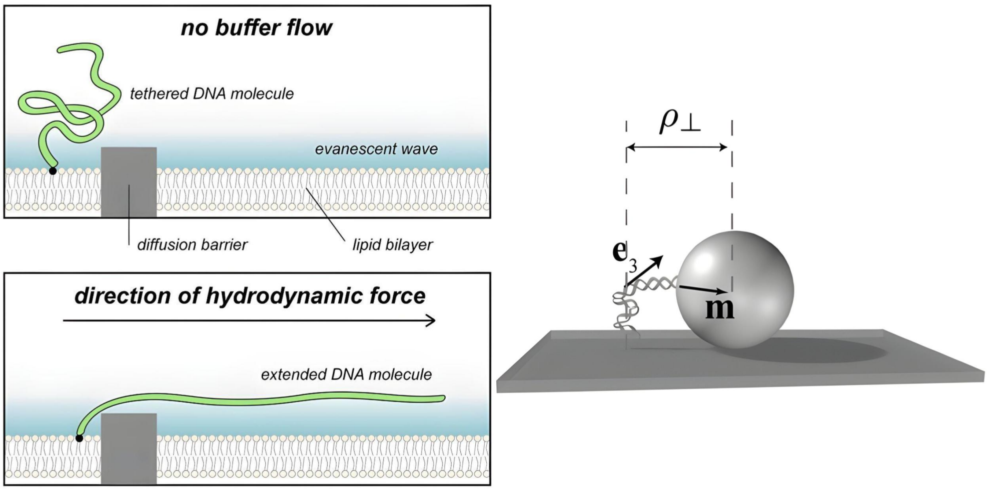


Figure 1.5: (**Left**) Side view of a DNA curtains experiment with applied flow. Image from [53]. (**Right**) Simplistic example of tethered particle motion, the the value $\rho_\perp$ representing the displacement of the bead from the view of a top-down observer. Image from [54].

hasn't been applied in any real experiments. The issue lies in the inability to accurately control the levels of torque and stretching force applied to the DNA, in addition to the challenges in torsionally isolating the duplex.

1.6.2.2 Atomic force microscopy (AFM)

Atomic force microscopy is an imaging technique in which a physical probe moves over a surface and generates a height map by measuring the probes vertical displacement [55]. Specifically, the instrument consists of a sharp tip mounted at the end of a flexible cantilever, as the tip moves over the surface it experiences a force, which directly translates to the movement of the cantilever. The movement of the cantilever is then measured, for example, by slight changes in the amount of laser light it reflects into a photodiode.

Atomic force microscopy has a wide range of uses, both inside and outside of the study of biological molecules, and has for example been used to study the growth of crystals [56] and in the imaging of entire bacteria [57]. Most relevant to this text are AFM studies of DNA-protein interactions [58] and studies of supercoiled DNA minicircles [59], both of which show close agreement between atomistic simulation and experiment, and with the latter giving direct insight into the structure of supercoiled DNA.

1.6.2.3 Tethered particle motion (TPM)

One of the simplest ways to study the mechanical properties of DNA is through tethered particle motion (TPM), a technique in which a molecule of DNA is fixed at one end to a surface and at the other end to a freely moving bead (see figure 1.5). The movement of the freely moving bead can be tracked, and from the distribution of bead positions mechanical properties such as spring constants [60] can be discerned. TPM has also been used to study complex processes such as transcription [61]. In this case RNA polymerase is attached to a surface, and it moves along the tethered DNA, with the movement of the bead allowing for the readout of the movement of the polymerase.

1.6.2.4 Magnetic tweezers

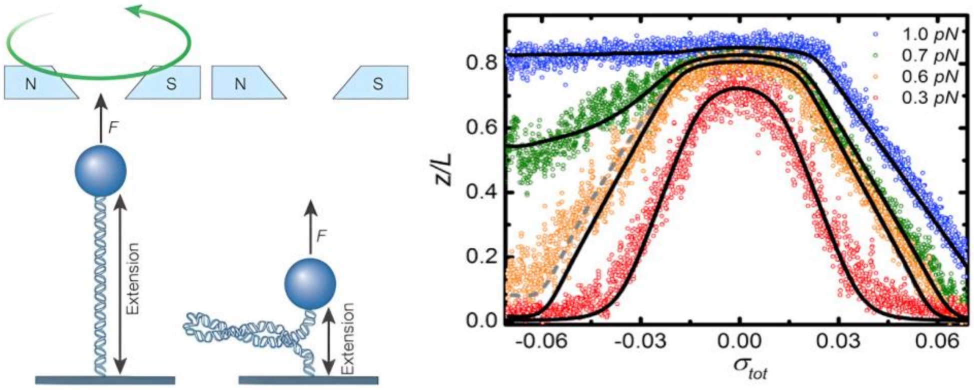


Figure 1.6: (**Left**) Diagrammatic example of a magnetic tweezers experiment. Image from [62]. (**Right**) DNA hat curve including a series of forces and supercoiling densities, with all curves collected at an ion concentration of 0.2M. Image from [63].

While the simple system of TPM can be used to discern some structural properties of DNA, it cannot be used to study behaviour under tension and torsion. In order to do this the bead needs to be fixed in place, with its movement able to be carefully controlled.

One method that is able to achieve this careful control is magnetic tweezers. These replace the inert bead used in TPM with one embedded with a magnetic metal, allowing it to be trapped by a magnetic field. Application of a nonuniform magnetic field to the bead allows for the application of a stretching force, while the application of a rotating field of uniform strength can be used to generate torque.

Since their initial application in the study of the supercoiling of λ-DNA [64] magnetic tweezers have been used to probe the stretching and twisting of a large range of DNA molecules at a variety of salt concentrations and pH levels [63, 65, 66]. All experiments on supercoiled DNA using magnetic tweezers are able to recreate the famous ‘hat-curve’ seen in figure 1.6, in which a clear asymmetry can be seen between positive and negative supercoiling at high forces. Unfortunately, magnetic tweezers do not allow the user to directly view the DNA that is being twisted and stretched, with only the bead being visible. This lack of observation means that, while one can hypothesise the structural features that give the hat-curve its shape, they cannot be experimentally confirmed by this technique.

Magnetic tweezers also have uses outside of the study of bare DNA. Much like TPM, they can be used to study the mechanics of proteins that change the shape of DNA duplexes. Among these proteins are topoisomerases, which cleave DNA [67]; condensins, which bridge and condense DNA [68]; and RNA polymerase, which causes upstream supercoiling during transcription [69].

1.6.2.5 Optical tweezers

The principles of optical tweezing are very similar to those of magnetic tweezing. Here the molecule is attached to one (or more) optically transprent beads, commonly made of

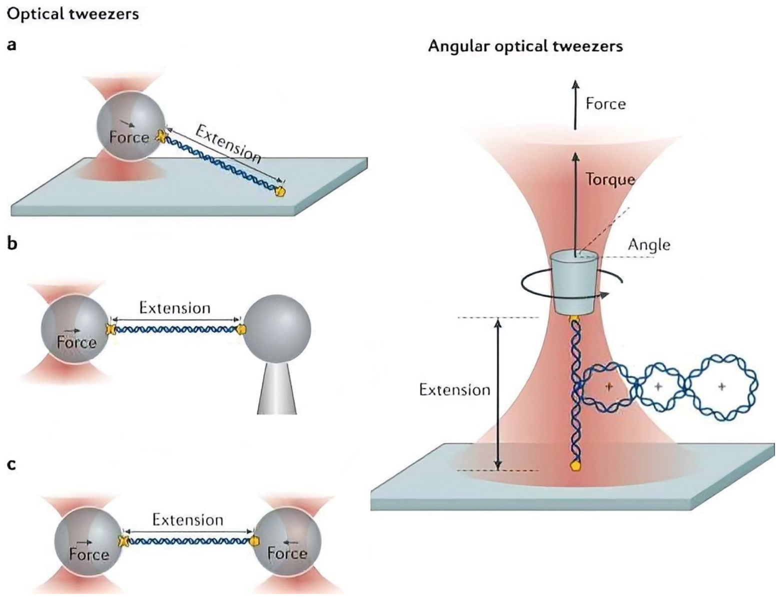


Figure 1.7: (**Left**) Variations of standard optical tweezers experiments. (**Right**) Example of an angular optical tweezers experiment. Yellow shapes indicate functionalised ends. Both images from [70].

polystyrene or latex. Due to their higher optical density relative to their surroundings, these beads are attracted to the centre of a focused laser beam, allowing them to be trapped. In the centre of this beam trap is a region in which the force applied to the bead scales linearly with its displacement in the manner of a simple spring. Force can be applied to the bead either through the use of flow or an electric field, the displacement of the bead can then be measured and forces extrapolated from its distance from the centre of the beam [71].

There are several variants of optical tweezer experiments, with examples being a duplex attached at one end to a bead and at the other to a surface; a duplex being attached at one end to a bead and at the other to a second bead held in place by suction on the end of a micropipette; and finally a duplex attached to optically trapped beads at both ends (see figure 1.7) [72, 73]. The advantage that these techniques hold over their magnetic counterpart is that the DNA can be actively observed when the technique is combined with fluorescence microscopy [74, 75], in principle allowing for the observation of sequence-dependent effects, although in practise the combination of these techniques is highly technical and very challenging [76].

All of the optical trapping techniques shown above lack the ability to apply torque to DNA, confining their usage to DNA stretching experiments. It is possible to achieve torque using optical tweezers however, through the use of angular optical tweezers (AOT), pictured in figure 1.7. Rather than a spherical bead, these tweezers use bifringent quartz cylinders which need to be nanofabricated from quartz wafer [77]. The cylindrical axis of these crystals aligns with the direction of light propagation, this allows the crystal to be rotated about its axis via rotation of the laser polarisation [78].

Unlike their more conventional counterparts, angular optical traps cannot be paired with fluorescence - the requirement for the cylinders to be parallel to the beam of the laser means that the only possible view of the system is top down, parallel to the helical axis of the DNA. This makes them closer in nature to magnetic tweezers, with one key advantage: that of torsional precision. The range of torques that can be applied by AOT is much smaller than their magnetic counterparts, but also much more precise, allowing for much more closely controlled experiments.

1.7 Modelling DNA under tension and torsion

1.7.1 Continuum models

One way to model DNA, which is particularly effective for long sequences and the prediction of macroscopic behaviour, is through the use of continuum models. The relatively large (≈ 50nm) persistence length of DNA allows for assumptions to be made about bulk properties, namely that DNA sequence is irrelevant and the molecule can therefore be modelled as a homogeneous chain. This assumption allows relatively simple models to be applied to the behaviour of DNA under tension and torsion, giving predictions of properties such as force-extension.

1.7.1.1 Kratky-Porod model

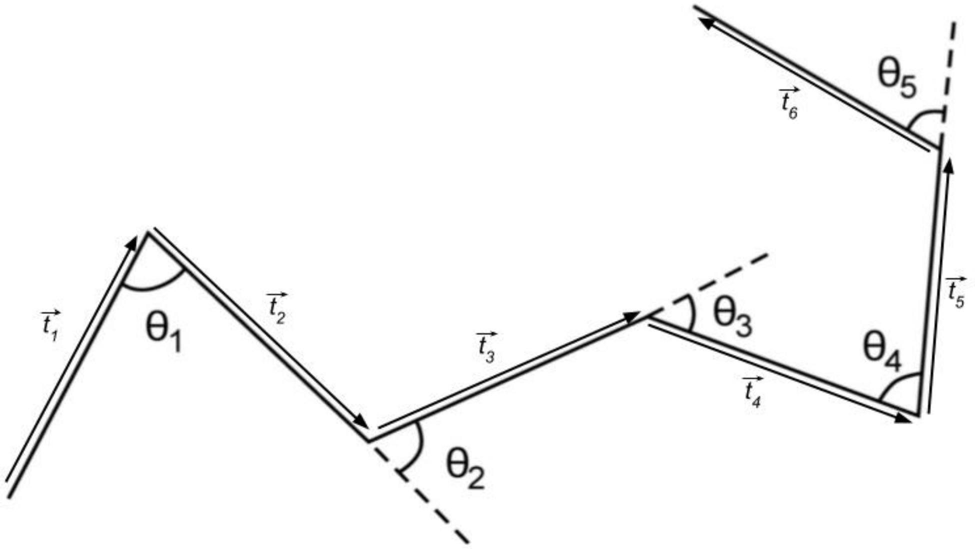


Figure 1.8: A freely jointed chain with all segments of length b. The vector t_i represents the orientation of segment i, and angle θ_i represents the angle between segments i and $i+1$. Image adapted from [79].

The Kratky-Porod model [80], also called the freely-jointed chain, first proposed in 1949, modelled a polymer as a succession of straight segments (see figure 1.8), which are able to freely rotate around one another in two dimensions.

Several useful statistics can be extracted from this model. The first and simplest is the contour length L_c, which is the length of the chain when it is fully extended and is given by

$$L_c = Nb \tag{1.8}$$

where N is the number of segments and b is the length of any individual segment.

Filaments such as DNA are rarely fully extended, as such it is helpful to consider an expression for its end-to-end vector **R**. Due to the freely jointed nature of the chain all of its possible conformations are equally likely, this makes the average value for **R** zero. A more useful measure is the root mean squared (RMS) end-to-end distance

$$\sqrt{\langle \mathbf{R}^2 \rangle} = \sqrt{\sum_{i,j}^{N} \langle \vec{t_i} \cdot \vec{t_j} \rangle} = \sqrt{b^2 \sum_{i,j}^{N} \langle \cos\theta_{i,j} \rangle} = \sqrt{N}b = \sqrt{bL_c}. \tag{1.9}$$

It is worth noting that the end-to-end distance of the filament grows as the square root of the contour length, this is a result of the tendency of such chains to adopt loosely bundled conformations due to their greater entropy.

Another standard measure of the size of a filament is the radius of gyration R_G. This defines the displacement of the vectors from their centre of mass and, due to it considering every segment rather than just the ends, is a more complete measure of structure. In the case of the freely jointed chain, the radius of gyration can be calculated to be [81]

$$\langle R_G^2 \rangle = \langle \mathbf{R}^2 \rangle / 6 \implies R_G = \sqrt{\frac{b^2 N}{6}}. \tag{1.10}$$

A result of the tendency of these chains to adopt loose bundles is that, even though segments are able to freely rotate around one another, the filament displays a mechanical resistance to stretching. This resistance is known as *entropic elasticity*, and at small extensions is given by [82]

$$F = \frac{3k_B T}{bL_c} x \tag{1.11}$$

where x is the end-to-end distance, k_B is Boltzmann's constant and T is the temperature. This is precisely analogous to the filament behaving as an ideal spring with spring constant $3k_B T/bL_c$.

The exact relation between force and extension, valid for both small and large extensions is [83]

$$x = L_c\left(\coth\frac{Fb}{k_B T} - \frac{k_b T}{Fb}\right). \tag{1.12}$$

Note that, according to this expression, as $x \to L_c$ the force F goes to infinity, making this model inappropriate for large forces.

1.7.1.2 Worm-like chain (WLC)

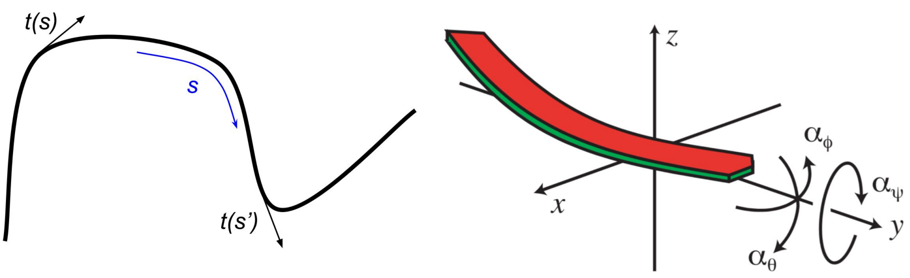


Figure 1.9: (**Left**) Diagrammatic representation of the worm-like chain. (**Right**) Yaw (a_ϕ), pitch (a_θ) and roll (a_ψ) along the contour of a worm-like chain, image from [84].

Although it is clear that large amounts of structural information can be extracted from the freely-jointed chain model, its main issues lie in its nature as a discrete system. Filaments such as DNA do not bend at a series of well-defined kinks, rather their bending is continuous, able to occur at any point along the contour. The 'freely-jointed' nature of the model is also an issue as the bending of most real polymers, including DNA, requires the input of energy associated with various rigidities.

These issues can be overcome by taking the freely-jointed chain to the continuous limit $b \to 0$ and adding a series of energetic penalties associated with flexural and torsional rigidities. Accounting for these rigidities, the bending energy density for the worm-like chain can be written as [85]

$$\epsilon(s) = \frac{\kappa_\phi}{2}[\alpha_\phi - \alpha_{\phi_0}]^2 + \frac{\kappa_\theta}{2}[\alpha_\theta - \alpha_{\theta_0}]^2 + \frac{\kappa_\psi}{2}[\alpha_\psi - \alpha_{\psi_0}]^2 \tag{1.13}$$

where $\kappa_\phi, \kappa_\theta, \kappa_\psi$ represent bending and torsional rigidities, $\alpha_\phi, \alpha_\theta, \alpha_\psi$ represent values for the three Euler angles (see figure 1.9) and $\alpha_{\phi_0}, \alpha_{\theta_0}, \alpha_{\psi_0}$ represent the intrinsic values of these angles.

If values for the intrinsic curvatures are zero then the worm-like chain orientation autocorrelation function, which is an expression for the relation between curvature at one point in the chain s and another point s', is given by [86]

$$\langle t(s) \cdot t(s') \rangle = \langle \cos\theta \rangle = e^{|s-s'|/l_p} \tag{1.14}$$

where lp is the persistence length, a length-independent constant that describes the stiffness of the chain. Roughly speaking, this relation implies that any two points that are within the persistence length of one another are likely to have the same orientation, with those separated by distances much larger than the persistence length are likely to have different orientations.

The persistence length is central in the definitions of many parameters of the worm-like chain model, for example in the RMS of the end-to-end vector [81]

$$\langle \mathbf{R}^2 \rangle^{1/2} = \sqrt{2L_c l_p - 2l_p^2(1 - e^{1-L_c/l_p})}. \tag{1.15}$$

For systems where $L_c >> l_p$ this value approaches $(2L_c l_p)^{1/2}$ in direct agreement with the value from the Kratky-Porod model seen in equation 1.9 (specifically if the segment length is equal to $2l_p$).

For small displacements the entropic elasticity of the WLC is directly dependent on the persistence length

$$F = \frac{3k_bT}{2l_pL_c}x \tag{1.16}$$

once again matching closely with the result from the Kratky-Porodd model seen in equation 1.11. At large extensions the elsticity is strongly approximated by [87]

$$F = \frac{k_BT}{l_p}\left[\frac{1}{4}\left(1 - \frac{x}{L_c}^{-2}\right) - \frac{1}{4} + \frac{x}{L_c}\right] \tag{1.17}$$

with an exact solution for all extensions able to be calculated from the work of Marko and Siggia [87].

Direct comparison of these solutions of the worm-like chain model to experimental results (see figure 1.10) can be used to extract persistence length.

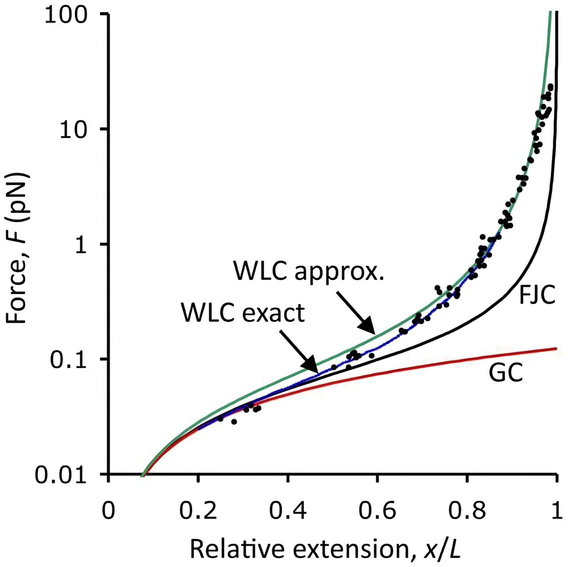


Figure 1.10: Comparison of DNA stretching models, with black dots representing experimental results from [87]. Note that GC here represents a Gaussian chain, equivalent to the predictions of Hooke's law, and FJC represents the freely jointed chain. Image from [85].

1.7.1.3 Twistable WLC

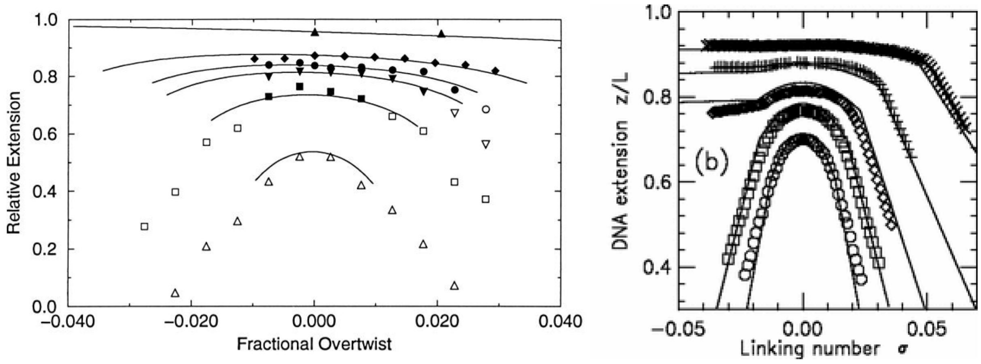


Figure 1.11: (**Left**) Predictions of relative extension vs supercoiling density from Moroz and Nelson. From top to bottom the curves are at fixed forces of 8.0, 1.3, 0.8, 0.3 and 0.1pN. Dots are experimental data, with open dots representing data outside the range of validity of the model. Image from [88]. (**Right**) Predictions from Marko 2007. From top to bottom the curves are at fixed forces 2.95, 1.31, 0.74, 0.44 and 0.25pN. Symbols represent experimental data. Image from [89]. Experimental results in both cases were collected at an ion concentration of 10mM.

While the work of Marko and Siggia [87] is able to quantify the force response of stretched DNA, its predictions do not extend to the behaviour of DNA under torsion. The first successful prediction of DNA behaviour under a combination of tension and torsion was made by Moroz and Nelson [88], who presented a model very similar to that in equation 1.13, with the inclusion of terms for both stretching force and torsion. This model was able to predict the behaviour of twisted DNA in a small range of supercoiling densities ($-0.04 \leq \sigma \leq 0.04$) as seen in figure 1.11.

The predictions of Moroz-Nelson model very rapidly diverge from experimental results, especially in negatively supercoiled systems. This failure comes from a series of issues [1]: the first is the lack of self-interaction - in addition to having a physical size, DNA is also negatively charged, causing short-ranged repulsion that is directly influenced by salt concentrations [90]. Second is the failure to account for disruptions of base pairing that occur in negatively supercoiled systems, resulting the in absorption of twist without a change in extension.

One successful treatment of DNA at a range of forces and supercoiling densities was performed by Marko in 2007 [89], here the DNA is treated in two separate states: plectonemic and stretched. The free energy of the system is then found by the mixing of the energies of these two states, with the amount of mixing depending on the supercoiling density and tension. The predictions of this model can be seen in figure 1.11, with good agreement seen at a range of supercoiling densities and forces. This work has since been built upon further, with the addition of self-interaction via the treatment of DNA as a charged rod, and the modelling of defects such as kinks and bubbles [91, 92].

1.7.2 Statistical models

While the theoretical treatment of DNA using continuum models allows for the prediction of large-scale behaviour, it fails to give any insight into the structural details of the DNA duplex, and in addition lacks any accommodation for sequence. The modelling of detailed structure formation and sequence dependence by their nature must go beyond continuum models - their assumptions of bulk properties such as sequence homogeneity and uniform flexibility make them completely unsuited to predicting behaviour with this level of detail.

One method of treating sequence in an entirely theoretical manner is with statistical models. These models attempt to isolate a subset of possible DNA states, which vary from one another only in the structure of interest (for example in the location of their denatured base pairs), and find their free energies. Once the free energies of all such states are found a partition function can be constructed, and from this the sequence-dependent probabilities of state formation can be calculated.

1.7.2.1 Predictions of bubble formation

One such example of this statistical modelling is Stress-Induced DNA Duplex Destabilisation (SIDD) [93] and later Stress-Induced Structural Transitions (SIST) [94] codes from Benham and colleagues. These aim to predict structural transitions in negatively supercoiled DNA, identifying regions of bubble, Z-DNA and cruciform formation. For the sake of simplicity, and due to its relevance to this text, the focus here will be solely on the prediction of denatured regions.

Consider the distribution of ΔL_k in a supercoiled state at a high force, i.e. without plectoneme formation. Changes in twist can be distributed in three ways: perturbations in twisting the B-DNA duplex, the formation of bubbles and the perturbation of twist at the bubble sites.

---

[1] Note that these issues also exist in the Marko-Siggia model, however they are not as prevalent due to the much simpler nature of force-extension curves.

Formally, these changes can be written as

$$\Delta L_k = -\frac{N}{A} + T_{ss} + \Delta L_{kr} \tag{1.18}$$

where N is the number of denatured base pairs, A is the twist of B-DNA (10.4 bp/turn), T_{ss} is the twist of the single stranded (bubbled) regions and ΔL_{kr} is the residual twist, or the change in twist of the B-DNA. T_{ss} can be written as the sum of the twists of all denatured base pairs

$$T_{ss} = \sum_{i}^{N} \frac{n_i \tau_i}{2\pi} \tag{1.19}$$

where τ_i is the twist of base pair i in the region of N denatured base pairs.

In order to account for sequence the different strengths of base pairing need to be considered. These varying strengths not only arise from the larger number of hydrogen bonds in G-C base pairs compared to A-T [95], but also from nearest neighbor interactions [96]. Assuming that the values of the melting energies associated with these strengths are known, one can calculate the energy associated with the denatured regions

$$E_D = \sum_{i}^{N} b_i + n_b H \tag{1.20}$$

where b_i is the energy required to denature base pair i, n_b is the total number of bubbles and H is the fixed bubble initiation energy, associated with the breaking of stacking at the initiation site [97].

From here a simple expression for the free energy of the system can be constructed using Hookean expressions for the energy of twisting in both the denatured regions and the B-DNA

$$G = \sum_{i}^{N_k} \left(b_i + \frac{C n_i \tau_i^2}{2} \right) + \frac{K \alpha_r^2}{2} + n_b H \tag{1.21}$$

where the spring constants C and K are the torsional stiffnesses of ssDNA and B-DNA respectively.

Using this expression, one can find the energy of any state containing one or more denaturation bubbles. Finding the energy of all such states allows for the construction of a partition function, from which the probability of any given state occurring can be found, with these probabilities then directly associated with the denaturation probability of each base pair (see figure 1.12). In reality, finding the free energy of every individual variation in bubbles size and position is impractical, requiring huge amounts of time and computational power for even medium-length sequences. Fortunately, practical applications of this method can be optimised: this is done by only considering states within 12kcal/mol of the lowest energy state, and by restricting the maximum number of bubbles to 4. These optimisations massively reduce the amount of time required for computation while having little to no impact on the final predictions, this is simply due to the fact that high-energy states have such low probabilities that they can be considered negligible.

SIDD and its successor codes have been used to identify destabilisation regions in a wide range of sequences, ranging from replication initiation regions in yeast [98] to the study of the

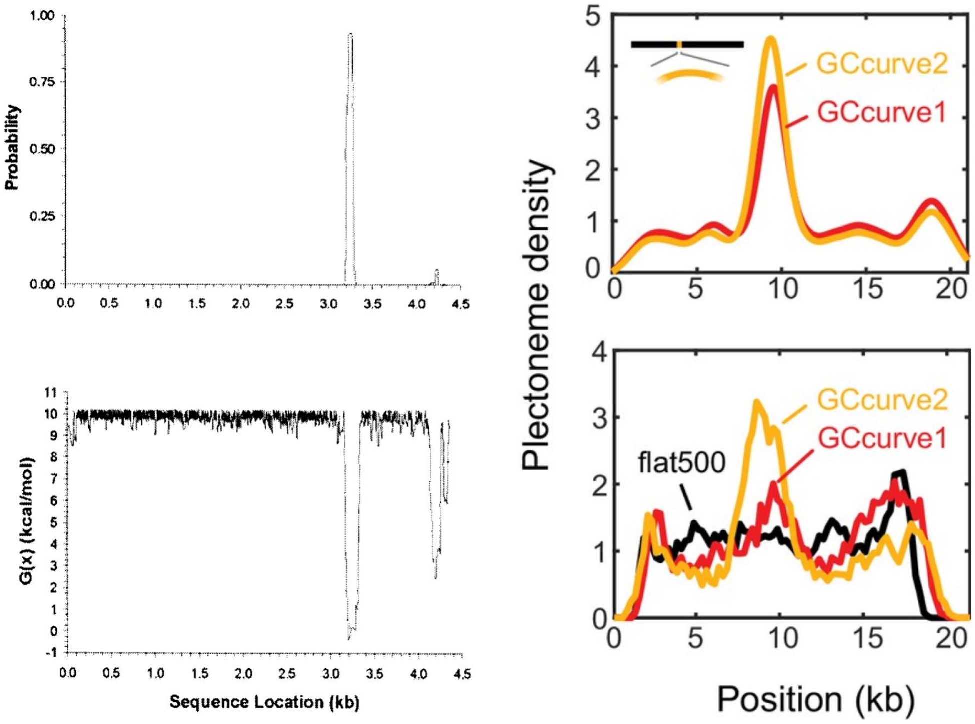


Figure 1.12: (**Left**) Bubble formation profile (**top**) and associated total free energies (**bottom**) from SIDD. Analysis performed on a 4.5kbp sequence with ΔL_k = -26. Image from [93]. (**Right**) Predicted (**top**) and experimental (**bottom**) plectoneme formation probabilities in a pair of sequences with a highly curved GC-rich insert. Black line indicates the sequence without any inserts. Image taken from [100].

role of pentanucleotide repeats in human genetic diseases [99]. Although the utility offered by SIDD is undeniable, it is limited in its functionality, namely in its inability to model DNA with plectoneme formation. In the study of DNA supercoiling, this restricts its use solely to systems at extremely high forces, and gives no insight into the interactions between plectonemes and bubbles.

1.7.2.2 Prediction of plectoneme location

In addition to its direct effects on bubble formation, DNA sequence also determines structural properties. These include local variations in persistence length due to varying stacking strengths [101] and changes in path due to the varying curvature of each base pair step [102]. In principle, these localised variations make certain sequences more suited to the formation of plectonemes than others. For example sequences with higher inherent curvature should be more easily able to form in to the sharply bent tip of a plectoneme.

A statistical model of DNA that aims to determine the effects of sequence-dependence on DNA deformation is that of Dekker and colleagues [100]. This model aims to predict the locations of plectoneme formation by focusing on the strong bending at their tips, associating the energy of tip formation with a combination of innate DNA curvature and local persistence lengths.

Consider the energy required to bend a length of B-DNA in to a circular loop [81]

$$E_{360} = \frac{2\pi^2 k_B T L_p}{rN} \tag{1.22}$$

where r is B-DNA rise (3.4 Å) and N is the number of base pairs in the duplex. This energy can be scaled to find the energy of any specified bend angle θ by

$$E_\theta = E_{360}\left(\frac{\theta}{360^\circ}\right)^2. \tag{1.23}$$

In order to predict the energy of plectoneme tip formation, these expressions need to be extended to three dimensions, and have the ability to account for inherent curvature and localised changes in persistence length. The first step in creating such a function is to consider the local bending angle $\theta(N,i)$ where N is the size of the window over which the bend occurs and i is the base pair in the centre of this window. In three dimensional space, this localised bending also requires a direction of bending ϕ which defines the plane in which the bend takes place. ϕ_B is the direction that aligns with the intrinsic curvature, that is the plane of bending when the DNA is at rest. The localised persistence length is then $L_p(N,i,\phi)$ - a function of the position at which it is taken, the size of the bending window and the direction fo bending. Putting this all together, and using the law of cosines to account for bending in any direction, the energy required to bend by an angle θ in the direction ϕ is [100]

$$E_{bend} = \left(\frac{\theta}{2\pi}\right)^2 \frac{2\pi^2 L_p(N,i,\phi)}{rN}\left[1+\frac{\theta(N,i)}{\theta} - 2\left(\frac{\theta(N,i)}{\theta}\right)cos(\phi-\phi_B)\right]. \tag{1.24}$$

Using this function, the bending energy at every position along the duplex can be calculated for every possible window size, allowing for the construction of a full energy landscape and the subsequent prediction of plectoneme location. The results of such a prediction are seen in figure 1.12, where a highly curved GC-rich sequence is inserted in to an otherwise flat sequence. This insertion results in the promotion of plectoneme formation, with this promotion accurately predicted by the above model.

While the predictions of this model appear to be accurate, it is not without limitation. First and foremost of these limitations is the models inability to predict plectoneme formation in negative supercoiling, owing directly to its inability to account for the formation of bubbles. In order to extend this model to negative supercoiling, one would need to combine its predictions with those of SIDD, accounting not only for the ability of bubbles to absorb twist, but also the increased flexibility at bubble formation sites. Another limitation of this model is its improper accounting for end effects: the model simply assumes that the size of a plectoneme is fixed at 1000 base pairs, it then suppresses plectoneme formation within this many base pairs of each end of the sequence. While practical at small lengths or low levels of supercoiling, this accounting starts to collapse once plectonemes grow much larger than the proposed size, perhaps leading to the over-promotion of strongly bent sequences near the ends of strands.

1.7.3 Coarse-grained modelling

While continuum and statistical models can be successful in using free energy landscapes to understand DNA at equilibrium, they give little to no insight into how the DNA reaches this state. Add to this their inherently restrictive nature, where each model focuses on a particular structural aspect, and it is clear that another technique is needed to fully understand DNA dynamics. Molecular dynamics techniques, which aim to model the evolution of the duplex over time using basic mechanical concepts, can offer invaluable insight into the time-evolution of DNA in a massive range of conditions.

Coarse-grained models attempt to compromise between the accuracy of modelling every individual atom and the practicality of continuum models by grouping atoms in to single objects which interact with one another. The manner in which this grouping or 'coarse-graining' occurs depends heavily on the characteristics of the system being modelled and the details which need to be captured.

Perhaps the simplest forms of coarse-grained DNA modelling are those that model DNA as a semiflexible chain; models of this form often behave much like the Kratky-Porod model [80], with beads attached to one another with harmonic potentials. These models are mainly used to simulate large, complex systems, and have, for example, been used in the prediction of whole-genome structure [103] and in modelling the effects of crowding on supercoiled plasmids [104]. The simplicity of these models makes them perfect for large systems, but also neglects fine detail, and any effects that are specific to DNA base pairing or double helical structure are completely ignored.

The majority of complex coarse-grained models of DNA are so-called "bottom-up". These are models that start with high-resolution models and remove unimportant degrees of freedom, leaving a much simpler system with only the necessary degrees of freedom remaining. When modelling DNA, these kinds of models often reduce individual bases to one or more beads, with these beads able to interact with one another and their environment electrostatically, and with the strength of these electrostatic interactions based on the atoms which make up the bead. There are a wide range of these bead-like models, designed to accurately capture different kinds of DNA behaviour. These range from those designed specifically to reproduce the thermodynamics of melting and bubble formation [105, 106], to those designed to accurately model DNA-ion interactions [107], and those designed to model the interactions between DNA and specific proteins such as chromatin [108, 109].

1.7.3.1 OxDNA

OxDNA [112, 113] is a nucleotide-level coarse-grained DNA model. Each nucleotide is represented as a rigid body with three interaction sites: the backbone interaction site, the stacking interaction site and the hydrogen bond interaction site, shown in figure 1.13. OxDNA parameters are set using a top-down approach: first a model is created with a set of phenomenological interactions (hydrogen bonding, stacking etc.), then the parameters representing each of these interactions is changed to match experimental values for the various thermodynamic, structural and elastic parameters of DNA.

The ability of OxDNA to accurately represent DNA while being computationally efficient has led to its use in the study of a multitude of DNA structures including DNA origami [114], DNA hairpins [115] and a variety of other DNA nanotechnological structures and processes [110]. Of particular relevance to this thesis are OxDNA simulations of molecular tweezers experiments [111, 116], in which OxDNA is able to provide semi-quantitative agreement with experimental data. In particular, the model was able to reproduce the 'hat-curve' seen in magnetic tweezers experiments (see figure 1.13), albeit at much higher forces than those observed experimentally. These simulations displayed a unique insight into the interaction of bubbles

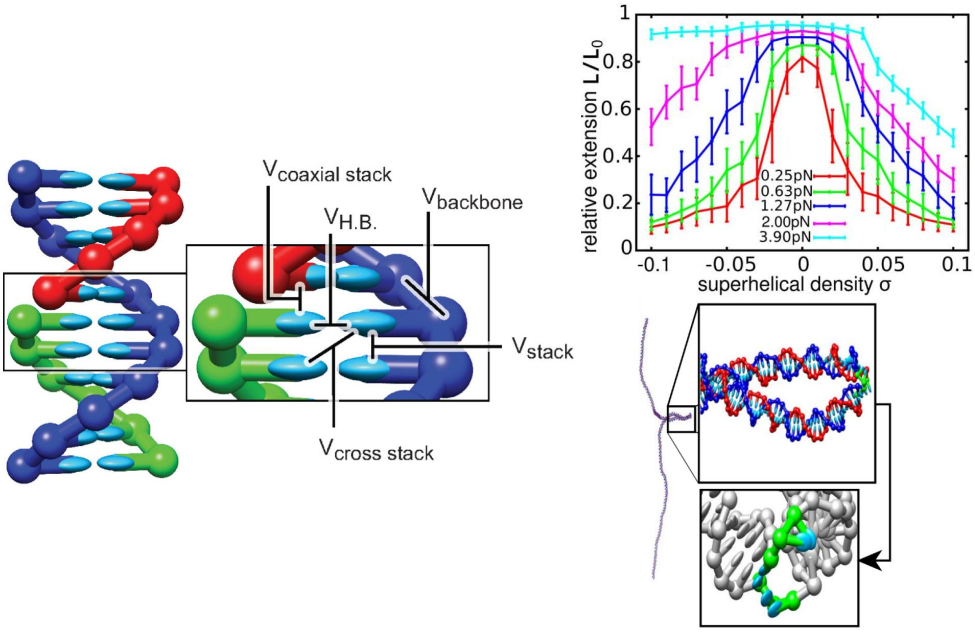


Figure 1.13: (**Left**) A nicked duplex in oxDNA, with a close-up showing the details of interactions for both the backbone and basepair interaction sites. Image from [110]. (**Right**) 'Hat-curve' from oxDNA simulations of a 600bp duplex at a range of forces and supercoiling densities with a salt concentration of 0.5M (**top**), and the formation of a 'tip-bubble' for the same duplex at $\sigma = -0.05$, $F = 1.27pN$ with zoomed-in details of the 3bp bubble (**bottom**). Image from [111].

and plectonemes, unveiling a previously unseen phase at the boundary of the L/P-DNA phases and the supercoiled B-DNA phase, named the tip-bubble phase. In this phase a bubble colocalises on the tip of the plectonemic loop, acting to both lower the energy cost of plectoneme formation and pin the plectoneme in place, significantly impeding its ability to diffuse. These simulations also showed that plectoneme nucleation sites are highly sequence dependent, in agreement with experimental [100] predictions.

1.8 Atomistic molecular dynamics simulations

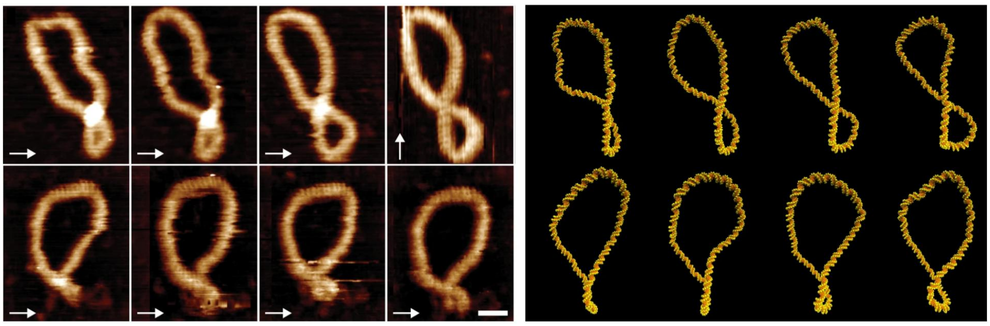


Figure 1.14: (**Left**) Time-lapse AFM images of natively supercoiled ($0.03 \leq \sigma \leq 0.06$) 339bp DNA minicircles. (**Right**) Chronological snapshots from 500ps simulations of the same 339bp minicircle, with a supercoiling density $\Delta L_k = -1$. Both images from [59].

The focus of this thesis will be in the recreation of molecular tweezers experiments using atomistic molecular dynamics. While this atomistic study of supercoiled linear DNA will be the first of its kind, there have been previous atomistic studies of supercoiled DNA, specifically of DNA minicircles [59, 117, 118]. These simulations reveal the relationship between structural properties and supercoiling density, with the appearance of dynamic bubbles, kinks and wrinkles. These simulations also show very close agreement with experimental data, attesting to the accuracy and utility of molecular dynamics simulations in the study of small supercoiled constructs (see figure 1.14).

The aim of this thesis is to combine the molecular tweezers method employed in oxDNA simulations with the detail seen in atomistic simulations. This will give unprecedented insight into the behaviour of supercoiled DNA under both tension and torsion, allowing for the discovery of as-yet unseen details of the interaction between plectonemes and bubbles.

Chapter 2

Methods

2.1 Molecular dynamics

The principles of molecular dynamics (MD) simulation are straight forward: take a set of distinct components with well-defined positions and evolve them forward in time using a set of simple physical rules. In the case of the simulations performed in this thesis these distinct components are atoms, although, as seen in coarse-grained simulations, they do not need to be.

One can imagine the motion of a simple system with a single particle moving with a velocity v_i, being acted on by a force $F(x_i)$, a function of the particles current position. Simple application of Newton's second law, $F = ma$, over a small time-step, ϵ, allows for the calculation of velocity at the next time step v_{i+1}

$$v_{i+1} = v_i + \epsilon F(x_i)/m \tag{2.1}$$

where m is the mass of the particle. The position of the particle at time $i+1$ can be computed in a similar way

$$x_{i+1} = x_i + \epsilon v_i. \tag{2.2}$$

These simple equations of motion lay out the theoretical groundwork for molecular dynamics, with the force F adjusted according to the those acting in the system of interest.

Of course, the reality of designing and implementing an MD system that is able to simulate real systems is orders of magnitude more complex than this simple description makes it appear. Accurately designing potentials to match physical interactions, defining proper time-steps and correcting for quantum interactions are among the many challenges that real MD programs have to address.

2.2 Structure creation and simulation initialisation

The first step in any MD simulation is the creation of the structure(s) that you want to simulate, with the correctness of the starting structure being vital to the stability and accuracy of your simulation [119]. Fortunately the creation of B-DNA is simple, due to its predictable double-helical structure and known geometric attributes. Ideal B-DNA structures can be generated with programs such as the Nucleic Acid Builder (NAB) [120] from the AMBER software suite [121]. Structures generated using nab use known values for the DNA base-step parameters to generate an idealised strand of straight B-DNA, this can then be used as a starting structure for an MD simulation.

Of particular importance to this thesis is the generation of supercoiled DNA structures, that is those with a value for twist that differs from the known value of B-DNA. Fortunately, the generation of such structures using NAB is trivial, and structures with values of twist that differ from the known value is as simple as changing a single option.

2.3 Force fields

Once an initial structure has been defined, the forces which act upon it can be implemented. The forces experienced by any individual molecule range widely depending on the system being simulated.

Relevant to this thesis are the force fields of DNA, defined by covalent interactions between bonded atoms, and non-bonded interactions as a result of electrostatic and Van-Der-Waals forces.

2.3.1 The Lennard-Jones potential

The most well-studied and widely used interaction in MD is the Lennard-Jones potential [122]

$$V_{LJ} = 4\epsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right], \tag{2.3}$$

where ϵ is the depth of the potential well which sets the strength of the interaction, r is the distance between interacting particles and σ is the value at which the potential is zero (effectively the “diameter" of the atom in question).

The term proportional to r^{-6} arises from Van Der Waals forces, caused by fluctuations in partial charges, with the power of 6 found to best represent force dispersion. The term proportional to r^{-12} approximates the Pauli exclusion principle, effectively giving each atom a radius in to which others cannot pass. The choice of the power 12 is a purely heuristic one, speeding up calculation by allowing the Lennard-Jones potential to be computed as $V = 4\epsilon[\alpha - \alpha^2]$ with $\alpha = (\sigma/r)^6$, and ϵ is once again the depth of the potential well.

The Lennard-Jones potential alone is sufficient for the modelling of inert noble gases such as argon [122, 123], and, with some modification, the modelling of swarm dynamics [124]. However, in systems of molecules with more complex interactions, the Lennard-Jones potential acts only as a small part of the landscape of interactions.

2.3.2 The Coulomb potential

While the Lennard-Jones potential is useful for the modelling of small partial charges, it is far from sufficient to model the interactions between charged particles such as ions. The modelling

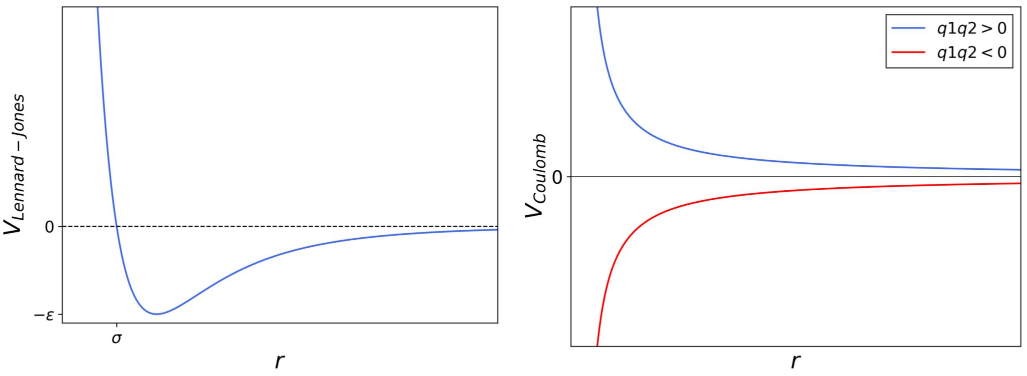


Figure 2.1: The non-bonded potentials Lennard-Jones (**left**) and Coulomb (**right**).

of such particles is well represented by the Coulomb potential

$$V_C = \frac{q_1 q_2}{4\pi\epsilon_0 r}. \tag{2.4}$$

Here q_1 and q_2 are the charges of the two interacting particles, r is the distance between them and ϵ_0 is the permittivity of free space. The form of this potential is seen in figure 2.1.

In principle, a combination of the Coulomb potential and Lennard-Jones potential would be sufficient to model a system of charged, non-bonded atoms. In real molecular dynamics simulations, fine tuning of both the Lennard-Jones and Coulomb parameters gives rise to physically accurate hydrogen bonding [125].

2.3.3 Bond length and angular potentials

In order to expand to physically relevant systems, interactions between covalently bonded atoms need to be accurately modelled. Unlike the non-bonded interactions described above, these potentials apply specifically between pairs or groups of bonded atoms, and cannot act over a distance. Another notable feature of these bonded interactions is their binary nature- atoms are either bonded or not, and therefore the breaking of these bonds is not possible in most MD simulations. The construction of bonded interactions can be approached in a dimensional manner, with one interaction per spacial dimension (three total). The simplest of the bonded interactions is the one-dimensional bond length interaction, which represents the vibrations of pairs of covalently bonded atoms. The potential used to represent these vibrations is a simple Hookean, with the bond length l oscillating about its equilibrium value l_0

$$V_l = \frac{k_L}{2}(l - l_0)^2 \tag{2.5}$$

where k_L is the spring constant, the value of which varies depending on the atoms in question.

In order to define a two-dimensional potential a bonded triplet needs to be considered, with the variable in this case being the angle between the three atoms (see figure 2.2) . The relative orientations of electron orbitals means that moving the angle (θ) away from its equilibrium value (θ_0) incurs an energy cost, with the potential given in the same manner as the bond length interaction

$$V_\theta = \frac{k_\theta}{2}(\theta - \theta_0)^2 \tag{2.6}$$

where k_θ is, once again, a spring constant.

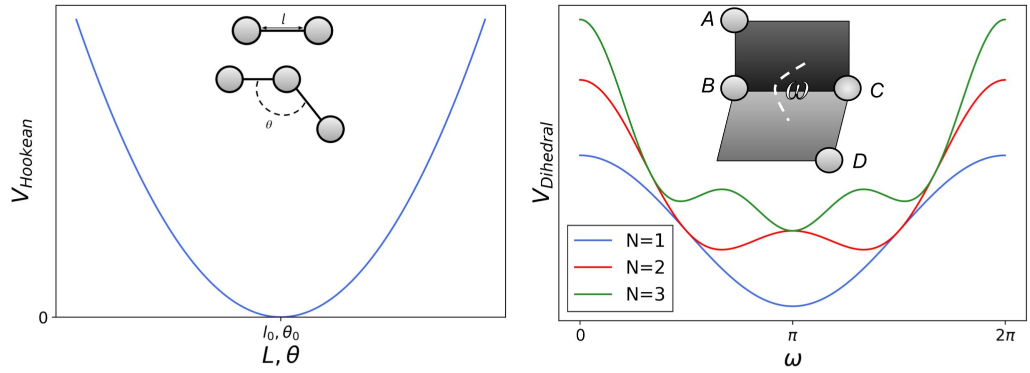


Figure 2.2: The bonded potentials Hookean (**left**) and torsional (**right**) along with diagrammatic representations of the relevant variables.

2.3.4 Torsional potential

The definition of a three-dimensional potential requires four atoms, with the variable being the dihedral angle. Take 4 atoms A, B, C and D such that A is bonded to B, B is bonded to C and C is bonded to D, then take the two planes ABC and BCD, the dihedral angle ω is then defined as the angle between these two planes (see figure 2.2).

In contrast to the two simpler bonded potentials, the dihedral potential is not Hookean, rather it is expressed in the form of a Fourier series

$$V_T = \sum_n \frac{V_n}{2}[1 + \cos(n\omega - \gamma)] \tag{2.7}$$

where n is the multiplicity and V_n is the amplitude of term n, which represents the magnitude of the torsional energy barrier and γ is the phase angle of the atomic triplet in question.

The multiplicity n varies according to the periodicity of the bond in question, and varies greatly depending not only on the atoms in question, but also the nature of their bonding to one another and others in the chain. Take for example bonds involving a pair of tetravalent sp^3 carbon atoms, these have multiplicity $n = 3$ and phase $\gamma = 0$, whereas bonds between a pair of trivalent sp^2 carbon atoms has multiplicity $n = 2$ and phase $\gamma = 180°$ [126].

2.3.5 The AMBER potential

Combining the terms listed above and summing over all atoms in the system gives the generalised AMBER potential for N atoms [125]

$$V_{AMBER} = \sum_{\substack{\text{Bonded}\\ \text{pairs}}} V_l + \sum_{\substack{\text{Bonded}\\ \text{triplets}}} V_\theta + \sum_{\substack{\text{Bonded}\\ \text{quadr.}}} V_T + \sum_{\substack{\text{Non-bonded}\\ \text{pairs}}} (V_{LJ} + V_C) \tag{2.8}$$

$$= \sum_{i=1}^{N_B} \frac{k_{l,i}}{2}(l_i - l_{0,i})^2 + \sum_{i=1}^{N_A} \frac{k_{\theta,i}}{2}(\theta_i - \theta_{0,i}) + \sum_{i=1}^{N_D} \sum_{n} \frac{V_{i,n}}{2}[1 + \cos(n\omega_i - \gamma_{i,n})] \tag{2.9}$$

$$+ \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left(4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \right)$$

where N_B,N_A and N_D are the number of bonded pairs, triplets and quadruplets respectively.

The calculation of every non-bonded interaction in a system is often impractical, with the number of calculations scaling as $O(N^2)$. As such, in some simulations, specifically those with large numbers of water molecules, a cutoff is used. This cutoff means that all Lennard-Jones interactions outside of a distance r_C are not calculated, cutting computation time significantly without affecting accuracy.

The final challenge in the construction of this potential is in establishing values for the various constants. The values of these constants will vary depending on the atom or atoms involved in the specific interaction. Creating a force field in which every possible atom and combination of atoms is fully parameterised would be completely impractical, fortunately the elements and molecules seen in biological systems is fairly restricted.

One method for extracting the relevant force field constants is using first principles quantum-mechanical calculations. This method has been successfully employed in both the creation [127] and refinement [128, 129] of force fields.

A very recent addition to the field of force field creation and refinement is the application of machine learning (ML) techniques. These techniques apply machine learning methods to density functional theory (DFT) calculations in order to rapidly calculate accurate force field parameters for small molecules. ML methods have been successfully applied in the creation of force fields for drug discovery [130] and in the construction of new coarse grained techniques [131].

2.4 Solvent models

While creating precise structures of biomolecules and ensuring the physical accuracy of their dynamics is important, they are not the only molecules present *in vivo* or *in vitro*, rather they exist in an aqueous environment. The makeup of this environment is vital to the structure and stability of biomolecules, DNA for instance is unstable without the presence of counterions. In order for simulated biomolecules to be physically accurate it is vitally important that their solvent environment is properly represented, and modern techniques represent this enviroment in one of two ways: explicit solvation, in which solvent molecules are explicitly simulated and represented with their own unique force field parameters; and implicit solvent, in which a bulk representation of the solvent with specific dielectric properties is used.

2.4.1 Explicit solvent

The presence of water in biological systems is universal and, in explicity solvated simulations, accounts for the majority of system volume. Take for example a cubic system with side length L, any biomolecule in the system will likely be of length L, with the number of atoms it contains scaling roughly as $N_{solute} \propto L$. All empty space in this system must then be filled with water (also any molecules of salt, but the number of these is negligible when compared to the number of water molecules), meaning that the number of molecules scales significantly with the size of the system, such that $N_{water} \propto L^3$. Despite occurring in massive numbers, the behaviour of the water is largely uninteresting, and so it is pertinent to model the water molecules in the most efficient manner possible.

The key simplification used when modelling water molecules is the assumption that the vibrations of the O-H covalent bonds have no effect on the behaviour of the system, allowing for both bond length and angle terms to be discarded by the water molecules' potential. This simplification leaves only two contributions to the potential of water atoms: van der Waals interactions and Coulomb interactions. A further simplification can be derived from the fact that the strength of the Lennard-Jones interactions in water comes almost entirely from the oxygen, meaning that the L-J term of the potential only needs to include the contribution from oxygen. The most widely used water model in modern molecular dynamics simulations is the TIP3P model [132], which employs a three-point model, with one interaction site per atom. Models with additional interaction sizes, namely four-point, five-point and six-point do exist [133], with these additional sites accounting for additional features of the water molecules, such as their polarisation.

2.4.1.1 Boundary conditions

The single most important optimisation of explicitly solvated simulations, and the key to making them computationally viable, is the implementation of periodic boundary conditions.

The study of biomolecules both *in vivo* and *in vitro* almost always involves the study of systems containing many thousands of the molecule of interest, with the use of flourescence and spectroscopy allowing for the examination of structural features. In contrast to this, all-atom and (some) coarse grained simulations involve the study of a single molecule, and, due to constraints on computational power, the size of the systems must be limited. It neither makes sense nor is computationally viable to simulate solvent that is sufficiently separated from the molecule of study, but the system also cannot be allowed to exist in a vacuum - if it were, the solvent would slowly diffuse in to this vacuum, leading to a density that goes to zero with time.

In order to both minimise computational complexity and create a system that makes physical sense, periodic boundary conditions are employed. Using these conditions the single simulated solvent box can be replicated an arbitrary number of times without any increase in computational complexity. In this scheme, any molecule that passes outside of the boundary of the pre-defined box simply re-enters the system on the other side. The new coordinates of any molecule that passes outside of the box are simply the old coordinates *modulo* the side length of the box.

2.4.2 Implicit solvent

As an alternative to explicitly simulating each water and solvent molecule, MD simulations can employ implicit solvents, where the solvent is represented as a continuous medium with specific dielectric properties. Consider the solvent free energy ΔG_s, which can be divided into two separate terms, one accounting for electrostatics ΔG_{el}, and another non-electrostatic term ΔG_n, such that [134]

$$\Delta G_s = \Delta G_{el} + \Delta G_n. \tag{2.10}$$

The non-electrostatic term ΔG_n is a function of the solvent-accessible area, solute-solvent Van der Waals interactions and solvent-solvent cavity interactions, arising from the need to create a cavity in to which the solute is placed. The electrostatic term ΔG_{el} arises from the charge of the solute and the effects of the implicit dielectric medium. A general form of the electrostatic free energy is given by

$$\Delta G_{el} = \frac{1}{2}\sum_{i=1}^{N} q_i \psi(\mathbf{r}_i) \tag{2.11}$$

where q_i is the charge of ion i with position $\mathbf{r}_i$, and ψ is the electrostatic potential. In the bulk this is a solution to the Poisson-Boltzmann equation [135]

$$\Delta^2 \psi = -\frac{\rho_e}{\epsilon_0 \epsilon} \tag{2.12}$$

where ρ_e is the local electric charge density, ϵ_0 is the permittivity of free space and ϵ is the dielectric constant of the solvent. Combining this with the Boltzmann equation [136], the local ion density c can be determined

$$c = c_0 \exp\left(\frac{-W}{k_B T}\right) \tag{2.13}$$

with c_0 being the bulk concentration and W the work required to move an ion from an infinite distance. This is given by $W = \pm e\psi$, where e is the charge of an electron and the sign depends on the charge of the ion in question. Plugging this expression for work in to equation 2.13 gives

$$c_{\pm} = c_0 \exp\left(\frac{\mp e\psi}{k_B T}\right). \tag{2.14}$$

Local charge density is therefore

$$\begin{aligned}\rho_e &= e(c_+ - c_-)\\ &= c_0 e\left[\exp\left(\frac{-e\psi}{k_B T}\right) - \exp\left(\frac{e\psi}{k_B T}\right)\right]\\ &= -2c_0 e \sinh\left(\frac{e\psi}{k_B T}\right).\end{aligned}$$

Subbing this in to equation 2.12 gives

$$\nabla^2 \psi = \frac{2c_0 e}{\epsilon\epsilon_0}\sinh\left(\frac{e\psi(\mathbf{r}_i)}{k_B T}\right). \tag{2.15}$$

For most geometries, the solution to this equation is a numerical one, but in some specific cases it can be solved analytically. Of greatest interest among these is a planar surface which is infinite in both the y and z coordinates, meaning that the potential can only change in x. In this geometry $\psi(\mathbf{r}_i)$ becomes $\psi(x)$ and, under the assumption that $e|\psi| << k_BT$, equation 2.15 is solved by

$$\psi(x) = \psi_0 \exp\Big(-\sqrt{\frac{2c_0e^2}{\epsilon\epsilon_0 k_BT}}x\Big) \tag{2.16}$$

in some MD simulations this low-potential approximation is valid and therefore used.

Although specialised numerical Poisson-Boltzmann solvers exist [137], they are far too computationally inefficient for application in large-scale molecular dynamics. In order to find a solution to this equation that can be rapidly computed, further approximations are needed. The most popular of these approximations is the implicit Born model, which begins with the Born equation

$$\Delta G_{solv}(R_i) = -\Big(1 - \frac{1}{\epsilon}\Big)\frac{q^2}{2A} \tag{2.17}$$

where A is the ion radius and q is its charge.

Alone, this is a so-called ‘spherical cow’ model, and is one of many exact solutions to the Poisson-Boltzmann equation for spherical geometries with a single charge centre [138, 139]. Unfortunately, biomolecules have neither spherical geometries nor single charge centres, making the above solution insufficient to represent them.

A more general model based on the Born equation, the generalised Born model, aims to solve the problems presented by the simplicity of equation 2.17. Imagine a molecule consisting of N charges $q_1...q_N$ embedded in spheres of radius $a_1...a_N$. If the separation r_{ij} between any two spheres is sufficiently large compared to their radii, then the free energy can be approximated by a set of Born terms and pairwise Coulomb terms [140]

$$\Delta G_{el} \approx \sum_i^N -\frac{q_i^2}{2a_i}\Big(1 - \frac{1}{\epsilon}\Big) + \frac{1}{2}\sum_i^N\sum_{j\neq i}^N \frac{q_iq_j}{r_{ij}}\Big(\frac{1}{\epsilon} - 1\Big). \tag{2.18}$$

In real molecules, atomic spheres are not necessarily far from one another, so this equation needs to be expanded upon further. First, the linearity of the Poisson-Boltzmann equation ensures that ΔG_{solv} will be quadratic in the source charges, meaning that equation 2.18 can be generalised to

$$\Delta G_{el} \approx -\Big(1 - \frac{1}{\epsilon}\Big)\frac{1}{2}\sum_{ij}\frac{q_iq_j}{f_{ij}^{GB}} \tag{2.19}$$

where f^{GB} is some simple function. The diagonal $i = j$ terms of f^{GB} can be thought of as the effective Born radii, while the off-diagonal terms become an effective interaction distance. One effective and popular choice for f^{GB} is that of Still *et al* [141]

$$f_{ij}^{GB} = \Big[r_{ij}^2 + R_iR_j\exp\Big(\frac{-r_{ij}^2}{4R_iR_j}\Big)\Big]^{1/2} \tag{2.20}$$

where R_i terms are effective Born radii, dependent on both a_i and the relative positions of all other atoms. Being conformation dependent, the values of R_i change at every time step, therefore it is vital that an efficient method of computing them is used. The method used in the AMBER software suite is the Coulomb field approximation [142], which replaces the true electric displacement around atoms by the coulomb field. In this regime, an approximate expression for R_i is

$$R_i^{-1} = \rho_i^{-1} - \frac{1}{4\pi} \int \theta(|\mathbf{r}| - \rho_i) r^{-4} d^3\mathbf{r} \tag{2.21}$$

$$= \rho_i^{-1} - I_i \tag{2.22}$$

where the integral is over the solute volume surrounding the atom i. Numerical computation of the integral in equation 2.21 is highly inefficient, owing to the fact that, for any real molecule, definition of the molecular surface is far from trivial. Here once again approximations are needed, with the most prominent among these being the pairwise screening approximation described by Hawkins, Cramer and Truhlar [143] in the GB-HCT model. This method returns to the use of spheres to describe solute surfaces, and as a result has a tendency to underestimate the effective radii for buried atoms. This issue was largely mitigated by Onufriev, Bashford and Case [142] in the GB-OBC by scaling the Born radii of buried atoms using a set of empirically derived parameters α,β and γ as follows

$$R_i^{-1} = \tilde{\rho}_i^{-1} - \rho^{-1} \tanh(\alpha\upsilon - \beta\upsilon^2 + \gamma\upsilon^3), \tag{2.23}$$

where $\tilde{\rho}_i$ is a scaled version of ρ_i, and $\upsilon = \tilde{\rho}_i I_i$. The approximations of these models are further corrected in the GB-neck model of Mongan *et al* [144] which, in addition to integrating over the spheres, also integrates over the 'neck' regions formed by the molecular surface between pairs of nearby atoms.

The final addition required for a complete implicit force field is an ability to account for the effects of electrostatic screening by monovalent ions. This can be accounted for by the addition of a Debye-Hückel screening parameter κ [129], given by

$$\kappa = \sqrt{\frac{8\pi S_{ionic}}{\epsilon k_B T}} \tag{2.24}$$

where S_{ionic} is the ionic strength of the solution. Modifying equation 2.19 then yields the new expression for free energy

$$\Delta G_{el} \approx -\frac{1}{2} \sum_{ij} \frac{q_i q_j}{f_{GB}} \left(1 - \frac{\exp[-\kappa f_{GB}]}{\epsilon}\right). \tag{2.25}$$

The implicit Born model, with the above approximations, provides the most efficient approximate solution to the Poisson-Boltzmann equation, it is therefore the standard choice for implicit solvation in MD simulations. Provided that the Born radii can be accurately estimated, a subject of much refinement [145], the Gibbs free energy is simple to integrate in to a force field, adding to the efficiency of the solvent model.

2.5 Integrators

Having created the system that we wish to study, and defined the rules by which both it and its solvent evolve, we now need to define the method by which the system will progress in time. Doing so in a continuous manner is not possible, as it would require exact solutions to all the complex differential equations which define the system's behaviour. Instead, time must be modelled by a series of discrete steps Δt, and by extension the system must move in a discrete manner.

The simplest method of numerical integration is the Euler method, this is precisely the method described by equations 2.1 and 2.2, with the constant ϵ replaced by a discrete time step Δt

$$x_{i+1} = x_i + \Delta t v_i \tag{2.26}$$

$$v_{i+1} = v_i + \Delta t a_i. \tag{2.27}$$

This method, while simple, is not suitable for the modelling of physical systems. Consider the small angle approximation of a mathematical pendulum, with motion described by

$$\frac{d\theta^2}{d^2t} + \frac{g}{l}\theta = 0 \tag{2.28}$$

taking a starting angle θ_0 and assuming zero angular velocity upon release, the solution is

$$\theta(t) = \theta_0 \cos\left(\sqrt{\frac{g}{l}}t\right). \tag{2.29}$$

The energy of this system can be checked analytically, taking $\omega = \frac{d\theta}{dt}$:

$$E = E_{kin} + E_{pot} = \frac{ml}{2}\omega^2(t) + \frac{mgl}{2}\theta^2(t) = \frac{mgl}{2}\theta_0^2 = \text{constant}. \tag{2.30}$$

Now, instead of solving this analytically, apply the Euler method

$$\theta_{i+1} = \theta_i + \omega_i \Delta t \tag{2.31}$$

$$\omega_{i+1} = \omega_i - \frac{g}{l}\theta_i \Delta t \tag{2.32}$$

plugging this into equation 2.30 gives

$$E_{i+1} = \frac{ml^2}{2}\left[\omega_{i+1}^2 + \frac{g}{l}\theta_{i+1}^2\right] \tag{2.33}$$

$$= E_i \frac{mgl}{2}\left(\frac{g}{l}\theta_i^2 + \omega_i^2\right)(\Delta t)^2 \tag{2.34}$$

meaning that application of the Euler method breaks energy conservation, regardless of step size. The method applied here can also be applied to the Hookean springs used in MD simulations, meaning that application of the Euler method to MD would result in systems which do not conserve energy, and therefore are unrepresentative of real physical systems. Fortunately, some relatively simple changes can be made to this method in order to fix the issue of energy conservation without significantly increasing the complexity of calculation (and by extension the time taken to compute each time step).

2.5.1 The leapfrog method

One such solution is to simply update position and velocity at staggered intervals - called the “leapfrog method". The new versions of equations 2.26 and 2.27 are then

$$v_{i+\frac{1}{2}} = v_{i-\frac{1}{2}} + a_i \Delta t \tag{2.35}$$

$$x_{i+1} = x_i + v_{i+\frac{1}{2}} \Delta t. \tag{2.36}$$

Provided that Δt is constant, and that $\Delta t \leq \frac{2}{\omega}$, the leapfrog method is stable for harmonic motion, and requires the same number of calculations per time step as the Euler method.

While the leapfrog method efficiently solves the problem of energy conservation it is not without its issues. Inspection of equation 2.35 shows that, upon initialisation of a system, one needs to determine $v_{\frac{1}{2}}$ which in turn requires the calculation of $v_{-\frac{1}{2}}$, an ill-defined quantity. In MD simulations that rely on the leapfrog method, this is solved by the equilibration phase, which follows the minimisation phase and involves gradually heating the system form 0K to the desired temperature. In addition to the issues with calculating $v_{\frac{1}{2}}$, the staggering of position and velocity calculations mean that any calculation of system energy will be inherently incorrect, although with a small enough time step the difference between real and calculated energies is minimal.

2.5.2 Velocity Verlet

Another solution that is similar to the leapfrog method is the velocity Verlet method, which uses higher-order versions of equations 2.26 and 2.27 that are as follows:

$$x_{i+1} = x_i + v_i \Delta t + \frac{1}{2} a_i \Delta t^2 \tag{2.37}$$

$$v_{i+1} = v_i + \frac{1}{2}(a_i + a_{i+1}) \Delta t. \tag{2.38}$$

The interleaved nature of this method arises from the presence of multiple acceleration terms in equation 2.38, making this a half-step method much like the leapfrog method.

In contrast to the leapfrog method, here position and velocity are aligned, making energy calculations accurate. Where the disadvantages arise is in the higher-order nature of the method, making it less computationally efficient than both the Euler and leapfrog methods.

The application of integrators in the AMBER software suite is a non-trivial mixture of the leapfrog and velocity Verlet methods.

2.5.3 Choice of time step

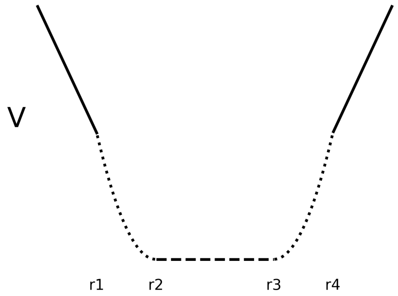
For any order-N integrator, system error propagates as the Nth power of Δt, the timestep. It therefore stands to reason that the most optimal choice for the timestep is simply the smallest one possible, and, while this is true in principle as it will take the system to the continuum limit, it is practically impossible. Extremely small timesteps are, in reality, completely unreasonable due to the massive computational cost of reducing Δt, with compute time scaling directly with the quotient of the time step.

Rather than proposing a minimum value for Δt a more practical approach is to find its maximum value, that being the largest value at which the system remains stable. Practical calculation of this has been performed [146] and, in general, the maximum Δt is found to be proportional to the mass of the simulated atoms. This relationship arises from the fluctuations of the covalent bonds that these atoms form - the lighter the atom the higher the frequency of its bond fluctuations, meaning that the minimum time step is defined by the fluctuation of bonds involving hydrogen (hydrogen being the lightest of the atoms).

2.6 Constraints and restraints

Methods of restraint in AMBER come in multiple forms, the first and most universal of these are SHAKE constraints, as defined by the SHAKE algorithm [147]. These are constraints applied to bonds with hydrogren atoms, acting to fix their length rather than allowing them to oscillate. As stated above, the vibrations of these hydrogen atom bonds typically account for the highest frequency oscillations in all-atom simulations, and therefore present the first source of energetic instability, resulting in a significant reduction in the maximum time step. Fixing the lengths of these bonds allows for the time step to be increased and, due to their soft nature, has little to no effect on system dynamics.

Restraints, rather than fixing a value completely as SHAKE constraints do, allow a value to oscillate around a pre-defined value, with the strength of these oscillations varied according to the strength of the applied restraint. Restraints are often called NMR restraints, a result of their original use in the refinement of structures obtained using NMR spectroscopy, which applies upper and lower bounds on certain lengths. In contrast to SHAKE, which simply fixed bond lengths, restraints allow for the restriction of lengths, bend angles and torsional angles - three of the potentials that contribute to the overall AMBER potential. The form of these restraints, regardless of the value they are restraining, is generic, and can be defined as follows:



$$\begin{cases} k_1(r1-R) & R<r1 \\ k_1(r2-R)^2 & r1\leq R<r2 \\ 0 & r2\leq R<r3 \\ k_2(R-r3)^2 & r3\leq R<r4 \\ k_2(R-r4) & R>r4 \end{cases}$$

where k_1, k_2 are user-defined interaction strengths, and in the case of the plot above $k_1 = k_2$. Any value of R between $r2$ and $r3$ is the 'target value', that is the minimum of the potential. Outside of these target values the form of the potential is dependent on the distance between R and its target.

In the range $r1 \rightarrow r2$ and $r3 \rightarrow r4$, the potential takes the form of a Hookean, with its value varying quadratically with the distance between R and the value $r2$ or $r3$. The use of a Hookean has two relevant effects: first, it makes the potential continuous and differentiable, both of these are vital for numerical calculation; second, it creates a ‘soft’ transition between zero and non-zero potentials. Recalling that $F = dV/dR$, the use of a hookean gives a force that goes to zero as R goes to $r2$ or $r3$, meaning that there is no sudden transition between zero force and high force.

In the regions $R < r1$ and $R > r4$ the potential is linear, dependent only on the distance between R and the value $r1$ or $r4$. This creates a constant force, the value of which is simply $\pm k$, the user-defined interaction strength, with sign dependent on the direction of force.

2.7 Thermostats and barostats

While the positions of, and interactions between, atoms define the important dynamics of MD simulations, it is vital to also assess their macroscopic properties, aligning these as closely as possible with experiments. A basic MD simulation might be performed in an NVE microcanonical ensemble, this is an ensemble in which number of particles (N), volume (V) and energy (E) are fixed. In contrast, experiments are performed in an NVT ensemble in which, instead of energy, temperature (T) is fixed. From a statistical mechanical perspective these two ensembles are distinct, and it is therefore desirable to create a microcanonical MD simulation in which temperature is fixed, this can be achieved with a thermostat.

2.7.1 Berendsen thermostat and barostat

For any system of particles, the equilibrium temperature (T_0) can be related to the time-averaged kinetic energies according to the equipartition theorem

$$\langle E_k \rangle = \frac{3}{2} N k_B T_0. \tag{2.39}$$

While the total energy of the system is conserved, kinetic energy can vary at each time step, as it is converted in to potential energy and back. Taking this kinetic energy at each time step the effective temperature (T) can be calculated. In order to maintain as close to a constant temperature as possible the system can be weakly coupled to a heat bath at a constant temperature T_b, which is the target temperature of the system. This coupling can then be used to suppress fluctuations in kinetic energy by forcing it to exponentially decay to T_b with a time constant τ

$$\frac{\Delta T}{\Delta t} = \frac{T_b - T}{\tau}. \tag{2.40}$$

Temperature is not a parameter of MD simulations, instead the above equation can be used to determine a scaling factor λ by which the velocities are adjusted

$$v \rightarrow \lambda v, \tag{2.41}$$

where

$$\lambda = \sqrt{1 + \frac{\Delta t}{\tau}\left(\frac{T_b}{T} - 1\right)}. \tag{2.42}$$

This method of velocity adjustment is called the Berendsen thermostat [148]. This method of temperature maintenance, by nature of its velocity suppression, can only produce results that converge on the canonical ensemble in systems with numerous collisions between particles, making appropriate only for simulations in explicit solvent in which it is the preferred thermostat due to its computational efficiency.

In explicitly solvated systems with continuous boundary conditions, maintaining the pressure of systems is non-trivial, and a method is needed to control it. One such method is the Berendsen barostat. The method of this barostat is similar to the Berendsen thermostat, scaling lengths in place of velocities. The position coordinate of particles is adjusted using a scaling factor μ such that

$$\mathbf{r} \rightarrow \mu \mathbf{r} \tag{2.43}$$

this time the factor μ scaled according to the dimensions of the system. For a cubic system it is

$$\mu = \left[1 + \frac{\Delta t}{\tau_p}(P - P_b)\right]^{1/3} \tag{2.44}$$

with the pressure P for a box of volume V approximated by

$$P = \frac{1}{V}\left(Nk_B T + \frac{1}{3}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} \mathbf{F}_{ij} \cdot \mathbf{r}_{ij}\right) \tag{2.45}$$

where $\mathbf{F}_{ij}$ is the force between particles i and j and $\mathbf{r}_{ij}$ is the vector connecting the two.

The result of these pressure adjustments is a system that agrees with the properties of an NPT system, rather than the assumed NVT.

2.7.2 Langevin dynamics

Due to the aforementioned issues with the Berendsen thermostat in small systems, simulations using implicit solvent require an alternative solution for maintaining temperature, the simplest and most widely used of these is provided by Langevin dynamics [149].

In real systems, the transfer of kinetic energy occurs almost entirely through collisions with the solvent, and these collisions are not present if the solvent is modelled implicitly. The generalised Born model only accounts for electrostatic screening, and therefore offers no solution to this lack of collisions.

Consider the friction that the movement solvent molecules applies to the solute, the force of this friction is defined by

$$\mathbf{F}_{Fric.} = -\gamma \mathbf{v} \tag{2.46}$$

where $\mathbf{v}$ is the velocity of a given particle and γ is the solvent viscosity, which, in most MD simulations, is defined in terms of a collision frequency. A subset of these solute-solvent collisions will cause a significant perturbation in the motion of the particles, and these perturbations are modelled as an additional random force

$$\mathbf{F}_{Rand.}(t) = \sqrt{2\gamma k_b T}\mathbf{R}(t) \tag{2.47}$$

with $\mathbf{R}(t)$ a stationary Gaussian process with

$$\langle \mathbf{R}(t) \rangle = \mathbf{0} \tag{2.48}$$

$$\langle \mathbf{R}(t) \cdot \mathbf{R}(t') \rangle = \delta(t - t') \tag{2.49}$$

where $\delta(t - t')$ is the Dirac delta function. The result of this force is a series of random kicks, the magnitude of which is dependent on the system temperature and the collision frequency, which provide no net acceleration. In general, the solute viscosity γ should be kept small, as the result of an excessively viscous solvent is an overdamped, Brownian system with no net acceleration.

In some implicitly solvated simulations, including those shown in this text, the solvent viscosity γ is set to a value ($\gamma = 0.01 ps^{-1}$) [150, 151] that is multiple orders of magnitude lower than that of water ($\gamma_{water} = 50 ps^{-1}$) in order to maximise the speed of system dynamics.

2.8 Simulation setup

2.8.1 Simulation parameters

In this work, all simualtions were performed using versions 16 to 18 of the AMBER software suite [120].

Due to the size of the DNA constructs simulated, this work uses exclusively implicit solvent models. Specifically, constructs were solvated using the implicit generalised Born model [152] at a salt concentration of 0.2M of monovalent salt with GBneck2 corrections [144], mbondi3 Born radii [153] and with no cutoff. Langevin dynamics was employed for temperature regulation with $T = 300K$ and a collision frequency of $0.01 ps^{-1}$ in order to reduce solvent viscosity, thereby increasing the speed at which the conformation space is explored. The use of implicit generalised Born models, along with low solvent viscosity, has been seen to provide an up to 100-fold speedup in the sampling of conformational space of large DNA-based systems when compared to their explicit counterparts [150].

Creation of DNA with varying supercoiling densities was performed using the Nucleic Acid Builder (NAB) [121] from the AMBER software suite [120]. Using an equilibrium twist value (i.e the twist of relaxed B-DNA) of 33.214 degrees, a value set using a $20ns$ unconstrained simulation of the linear strand studied in the following chapter. Strands were created using the `wc_helix` function, with the value of `twist` varied according to the required supercoiling density, and all other variables maintained at their default B-DNA values. This process creates perfectly straight strands of B-DNA that already have the required supercoiling density, with the supercoiling presenting itself simply as a value of twist that is perturbed from its rest value.

In order to simplify the simulation setup, as well as more closely mimic experimental conditions, one end of the duplex is fixed in place. More specifically, the coordinates of the $O3'$ and $O5'$ atoms in the backbone of the final base pair of the strand are kept constant over the entire course of each simulation. This is achieved by Cartesian restraints with the maximum possible restraint weight of 100 $kcal/mol\AA^2$. One important note is that the fixing of the backbone means that the bases at this fixed end are still free to move, denature etc.

It is well known in MD simulations of DNA that nucleotides at the ends of the strand are significantly less stable, giving rise to end effects. In order to prevent end effects at this fixed end of the duplex a pair of extra GC bases are added, these base pairs are then simply ignored during analysis. For the remainder of this chapter this end of the duplex will simply be referred to as the 'fixed end'.

In order to allow for the shortening of DNA end-to-end distance, and by extension the formation of plectonemes, the other end of the duplex must be allowed to move in space. This end is also the one to which tension is applied in order to study the effects of force. This movement in space presents a series of complex issues when it comes to maintaining supercoiling, and as such a series of restraints must be applied. The remainder of this section will lay out this series of restraints, and detail how they can be implemented in a real AMBER simulation. Henceforth, this end of the duplex will be referred to as the 'mobile end'.

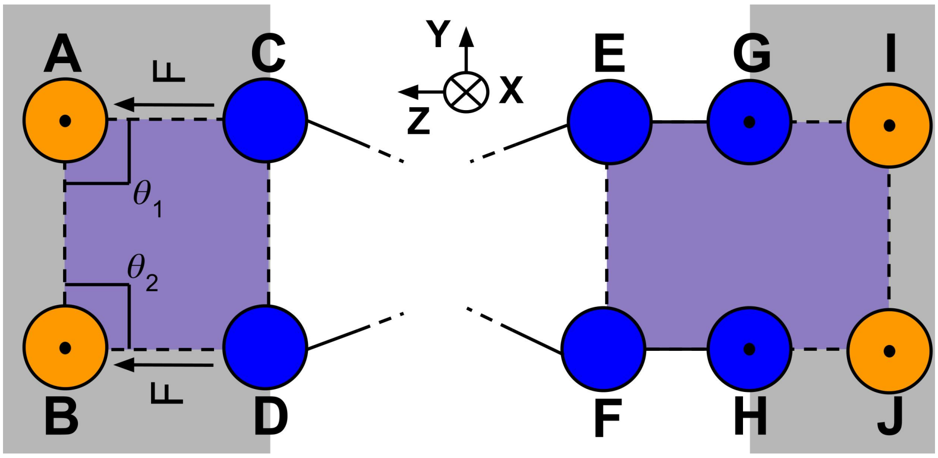


Figure 2.3: Restraints applied to both the mobile (left-hand side) and fixed (right-hand side) ends of the duplex. Atoms **A B**, **I**, **J** (in orange) are reference points, the coordinates of which are fixed. Atoms **C** and **D** represent the $O5'$ and $O3'$ atoms at the mobile end of the strand, the motion of which is limited to the Z axis. Atoms **G** and **H** are the $O5'$ and $O3'$ atoms at the fixed end of the DNA, the coordinates of which are fixed. Atoms **E** and **F** are the second-last $O5'$ and $O3'$ atoms at the fixed end of the DNA. The shaded purple areas represent the planes in which the triplets **ABD**-**BAC** and **JIE**-**IJF** are fixed such that both ends of the strand are torsionally constrained. The forces **F** represent the tensile force, applied along the Z axis. The angles θ_1 and θ_2 are fixed in order to prevent the movement of atoms **B** and **C** in the Y direction. The grey areas represent the excluded volume, in to which the bulk of the strand cannot move.

2.8.2 Application of tension

All of the restraints that will follow in this section will need points, fixed in space throughout the entire course of the simulation, through which they will be applied. In order to achieve this, dummy atoms parameterised in Henriksen *et al* [154] will be used. These atoms have zero charge, meaning that they do not interact with the DNA strand in any meaningful way. In order to act as reference points, these atoms are fixed in place using Cartesian restraints with the maximum possible restraint weight of 100 $kcal/mol Å^2$. For the remainder of this section these atoms will be referred to as 'reference points'.

In order to study the effects of tension on supercoiled linear DNA, a method by which tensile force can be applied to the duplex is needed. Fortunately, using a pair of reference points, applying tension to the strand is simple. The reference points **A** and **B** (figure 2.3) are placed 10Å from the $O5'$ and $O3'$ atoms of the mobile end (**C** and **D** of figure 2.3) at full extension (i.e when end-end distance = contour length). A pair of distance restraints are then applied to these atoms, pulling them towards their respective reference points. Provided that these distance restraints are linear, the force applied to the strand is constant, giving the desired constant tension. This tension is displayed as F in figure 2.3, note that force is applied to each strand separately, meaning that if a total force $F_{tot.}$ is needed, then $F = F_{tot.}/2$.

2.8.3 Keeping the strand on-axis

In experiments such as optical or magnetic tweezers the movement of the ends of the DNA duplex is largely confined to the axis along which the DNA strand runs at full extension. This axis will, from this point onwards, be referred to as the 'setup axis'. In simulation, the linear strands are floating freely in aqueous solution, with no surface or bead to confine their movement. In order to best match experimental setups, the ends of the linear strands simulated in this work also need to be confined to moving only along the setup axis.

The method by which this is achieved is pictured in figure 2.3. For the sake of simplicity, and indeed as a mirror to the real simulation setup, the 'setup axis' here is simply the Z-axis. In order to prevent movement in the Y-direction the angles θ_1 and θ_2 are fixed using strong angular restraints. Another pair of reference points and restraints (not pictured) are then implemented in precisely the same manner to prevent movement in the X-direction. The combination of these restraints confines the movement of atoms **C** and **D** to the Z-axis.

2.8.4 Torsional restraint

In experiments, the DNA being studied is bound at one end to a bead and at the other to a surface (or another bead), these bindings mean that the duplex is inherently torsionally constrained - the ends are not free to twist around their setup axis. In MD simulations this is not the case, the structures studied are floating freely in an aqueous environment, with no surface or bead to torsionally constrain them[1]. The free nature of the strand means that any supercoiling density within the strand upon its creation will simply dissipate in simulation, as the strand is free to simply twist in the direction opposite to the direction of supercoiling. In order to prevent supercoiling being removed, the strand needs to be restrained.

Torsional restraint of the mobile end is done by ensuring that the last $O3'$ and $O5'$ atoms and their respective reference points are co-planar. The restraints used to achieve this are represented by the purple area in figure 2.3. By fixing the angle between the planes **ABD** and **BAC**, and assuming that the atoms **C** and **D** are only able to move in the Z-direction, this restraint prevents the mobile end of the strand from twisting around the setup axis, while maintaining its ability to move parallel to it.

---

[1] Although it is technically possible to add surfaces that mimic beads or surfaces to MD simulations, this is completely impractical due to the massive number of additional atoms that would need to be simulated

At the fixed end of the strand the two added base pairs need to be prevented from twisting, thereby minimising their contribution to overall system dynamics. First, the freezing of the backbone position of the final base pair means that it cannot twist and therefore cannot contribute to supercoiling. The second-last base pair then needs to be torsionally restrained, an effect which is achieved in precisely the same way as the mobile end. The restraints used to achieve this are shown in figure 2.3, this time the angle between the planes **JIE** and **IJF** is fixed.

2.8.5 Excluded volume

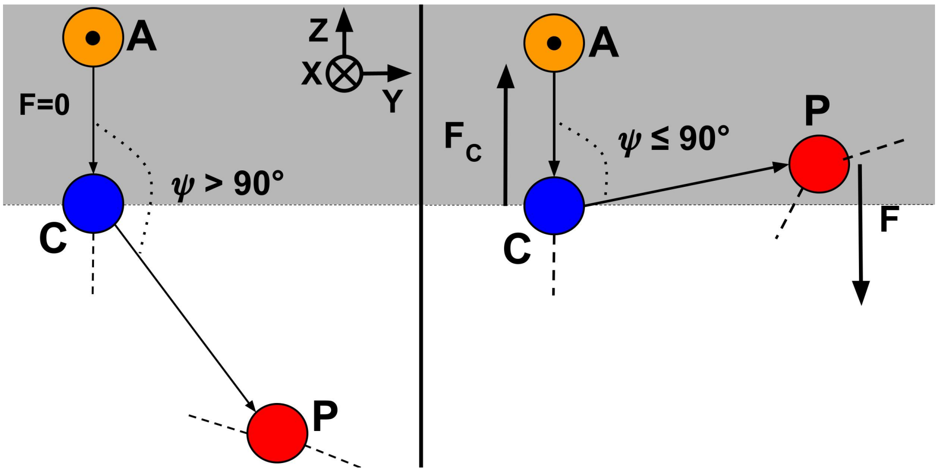


Figure 2.4: An example of an angular restraint used to create the excluded volume interaction at the mobile end of the duplex. Shown are the states in which the phosphorus atom **P** is outside (**left**) and inside (**right**) of the excluded volume. Atom **A** is a reference point, atom **C** represents an O5' (or O3') atom at end of the duplex, the movement of which is confined to the Z axis, and atom **P** is any phosphorus atom in the bulk of the duplex. The passing of any atom **P** in to the grey area results in the application of a force **F** to this atom. This force acts to push the atom **P** out of the grey area. Due to the double-sided nature of restraints in AMBER, the application of the force **F** on **P** also results in the application of a force $\mathbf{F}_c$ on the atom **C**, which acts along the Z axis in precisely the same way as the applied tension.

While the restraints described in subsections 2.8.4 and 2.8.3 prevent the relief of supercoiling by rotation around the setup axis, they are not sufficient to completely prevent the strand from relaxing its supercoiling. Consider the untying of a knotted string - the passing of one end of the string through one of the knotted loops, if done correctly, can lead to the knot coming undone. The same idea applies to supercoiled DNA - the passing of a loop (for example a plectoneme) formed in the duplex over one of its ends will cause the relaxation of supercoiling. From this point onwards this mechanism of supercoil relaxation will be referred to as 'untying'.

In order to prevent this untying, the bulk of the duplex cannot be allowed to pass over the ends, a condition that is most efficiently achieved via the implementation of an excluded volume. This excluded volume needs to be implemented such that any atom in the strand that passes one of the ends in the direction of the Z-axis is immediately pushed back towards the bulk of the strand. This excluded volume is represented by the grey areas in figures 2.3 and 2.4.

Implementation of this excluded volume in AMBER is complex due to its lack of any explicit excluded volume restraints[2], but, thanks to the restraints defined in section 2.8.3, it is possible. The excluded volume relies on a series of angular restraints using the reference points, $O3'/O5'$ atoms and all phosphorus atoms in the bulk of the chain, with this angle represented as ψ in figure 2.4. The restraint placed upon ψ is asymmetrical, and only applies a force when $\psi < 90°$. As long as **A** and **C** share both X and Y coordinates, restraint placed on the angle will only act when **P** passes **C** in the Z-axis. Applying this restraint to every phosphorus atom in the chain (keeping **A** and **C** the same) creates an effective excluded volume along the Z-axis.

Each phosphorus atom in the duplex has a restraint of this exact nature applied to it twice - once for each end of the DNA. Figure 2.4 is an example of the mobile end of the strand. The case of the fixed end of the strand is exactly the same, but the force $\mathbf{F}_c$ (see below) is not a concern as the $O5'/O3'$ atom is fixed in place using Cartesian restraints.

2.8.5.1 Problem with the excluded volume

While the restraints described above do create an excluded volume that prevents the passing of loops over the ends of the DNA, they also create their own problems, specifically at the mobile end of the duplex.

Restraints in AMBER are inherently two-sided and, in the context of the angular restraints pictured in figure 2.4, this means that any force (**F**) pushing **P** out of the excluded volume will also push **C** further in to it with an equal amount of force ($\mathbf{F}_c$). As **C** is fixed in both the X and Y axes, the force $\mathbf{F}_c$ acts only along the Z axis, in precisely the same direction as the applied tension.

The result of this force $\mathbf{F}_c$ is a system in which a constant tensile force cannot be reliably applied - this is a huge problem for simulations whose entire purpose is to measure the dynamics of supercoiled DNA under fixed tensile forces. The simplest way to solve this issue is to ensure that these excluded volume restraints are triggered as infrequently as possible, a solution that can be implemented by adding some kind of buffer between the bulk of the duplex and the mobile end.

2.8.5.2 Dummy DNA

The simplest way to add a buffer between the bulk of the strand and the mobile end is to add a stretch of additional DNA to this end. This DNA needs to be torsionally isolated from the rest of the duplex and held straight (i.e. prevented from moving in the X and Y-directions) in order to prevent it from contributing to plectoneme formation. In the simulations performed in this text, a series of 60 GC base pairs was added to the mobile end of the strand. For the remainder of this section these 60 base pairs will be referred to as the ‘dummy base pairs’, while the original 300bp will be referred to as the ‘main strand’.

Torsional isolation of these dummy base pairs is simple - take the planar restraints shown in figure 2.3 and, instead of applying them to the first base pair of the mobile end, apply them to the first base pair in the main strand (which is the 61st base pair in the entire construct). This

---

[2] It would likely be easier with restraints that allow for larger numbers of atoms, however in the version of AMBER used in this text these restraints are only available on CPUs, there are no CUDA versions.

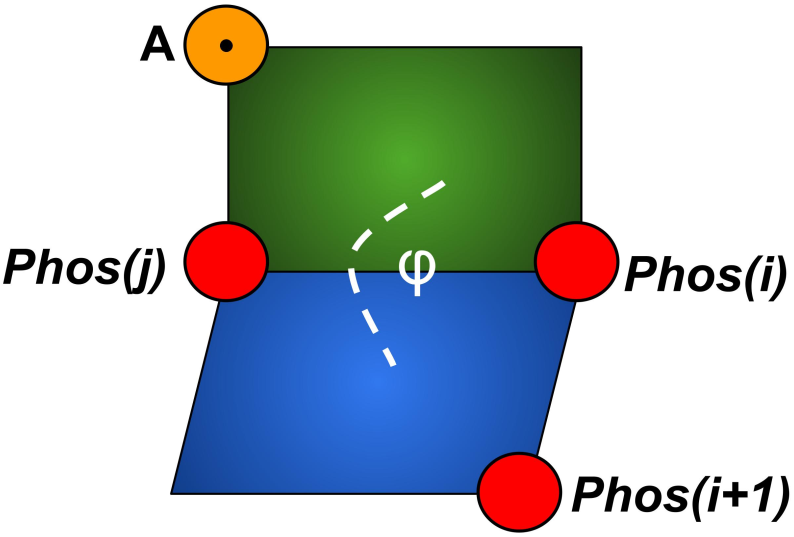


Figure 2.5: Dihedral restraint applied to all phosphorus atoms in the string of 60 dummy base pairs. Atom **A** is a reference point, red atoms represent phosphorus atoms within the dummy base pairs. The angle ϕ is fixed in order to prevent bending.

means that the dummy atoms are free to rotate around their setup axis, while the main strand remains torsionally constrained.

Keeping these dummy base pairs straight requires the implementation of a series of dihedral restraints, pictured in 2.5. Using the pre-existing reference points, these restraints are applied to each complementary pair of phosphorus atoms in the dummy base pairs and a single phosphorus atom in their neighbouring base pair. These restraints act to completely suppress bending in both the major and minor grooves, and by extension in the dummy base pairs as a whole. Applying a pair of these restraints (one for each strand) to each base pair in the dummy base pairs ensures that they do not contribute to plectoneme formation.

2.8.6 Additional equilibration

In order to ensure the formation of plectonemes, all hydrogen bonds between complementary bases are constrained for the first $40ns$ of each simulation. During these first 40ns, all of the above restraints are applied with the exception of the tensile force. This lack of force, combined with the forced maintenance of base pairing, allows for the formation of plectonemes in all systems. In test simulations in which hydrogen bonds were not held all negatively supercoiled systems formed bubbles almost instantaneously, thereby negating the need for plectoneme formation at all forces, a process that is clearly unphysical.

2.8.7 Summary of simulation methodology

Combining all of the above restraints and modifications creates a system in which a single linear DNA duplex can be completely torsionally constrained. This bespoke simulation setup allows the duplex to form complex supercoiled structures and dynamically adjust its distributions of twist and writhe over long time scales, without the fear of gradual torsional relaxation.

Briefly, the full setup procedure is as follows:

- Create the 300 + 60 + 2 bp sequence with the desired supercoiling density using NAB.
- Edit the .pdb file to add 6 dummy atoms: 4 at the mobile end of the strand, two at the fixed end of the strand.
- Create the Cartesian restraints, using the maximum allowed strength to hold the dummy atoms and $O5'$ and $O3'$ atoms of the fixed end in place.
- Create the nmr restraints with the following strengths:
  - Maximum strength for the torsional and angular restraints, applied to the $O5'$ and $O3'$ atoms of the last base pair of the main strand and the last base pair of the dummy base pairs respectively. Torsional restraints applied to the $O5'$ and $O3'$ atoms of the second-last base pair of the fixed end are also set to maximum strength.
  - A weight of 1 $kcal/mol\AA^2$ for the torsional restraint applied to all dummy base pairs (figure 2.5) and the excluded volume restraints for both ends of the duplex. A lower restraint weight was chosen here due to the large number of restraints being applied, in an effort to lower the overall system energy.
- Generate hydrogen bond restraints for use in the first 40ns of simulation.
- Run the additional 40ns of equilibration.
- Remove the hyrdrogen bond restraints and add the tensile force according to requirements, with a conversion rate of 1 $kcal/mol\AA = 69.4789pN$.

2.9 Analysis of simulations

2.9.1 Determination of denatured base pairs and bubbles

DNA denaturation is the disruption of the local structure of the double helix, determined by the breaking of hydrogen bonds between complementary bases, the separation of the helical strands and changes in groove widths. A region of denatured DNA, with multiple consecutive denatured base pairs, is often referred to as a ‘bubble’. The ability to accurately determine the locations of bubbles while discounting stochastic DNA ‘breathing’ [155, 156] - in which local conformational changes within the double-stranded structure lead to the breaking of base pairs at a temperature well below the melting temperature - is of vital importance to this text. As such a robust definition of bubbles which considers not only hydrogen bonding, but also the DNA base parameters, is of vital importance.

Bubble locations are easily identifiable upon visual inspection, however translating this into a consistent, qualitative definition is not easy. Fortunately there are a series of programs which can be used not only to determine hydrogen bonding in MD simulations, but also the base parameters, among these is the cpptraj program from the AMBER software suite [157]. The `nastruct` routine in cpptraj outputs both the number of hydrogen bonds between complementary bases along with base-pair and base-step parameters, which are calculated using the 3DNA procedure [22]. Using these, a strict, concrete definition of denaturing can be built.

In this thesis a base pair is defined to be denatured if: 1) There are no hydrogen bonds between the known donor-acceptor sites. 2) The values of propeller twist, opening and buckle are, on average, at least 2σ from their reference values, where σ is the standard deviation of their values over the course of the entire simulation. Reference values were determined using a simulation of the same DNA construct at zero supercoiling density and force. Values for these three base pair parameters are normally distributed about their means in the absence of disruption, meaning that one can expect to find 95% of their values within two standard deviations. Values outside of this range are a strong implication that significant structural disruption has occurred.

Once a proper definition of denaturation is found, expansion to the definition of a bubble is trivial: a bubble must contain at least three consecutive denatured base pairs. Often the definition of a bubble will be further restricted to include only those with a lifetime greater than 1ns. Implementation of this lifetime limitation is complex due to the tendency for bubble location to fluctuate over time. Take for example a bubble containing three base pairs at positions a, $a+1$ and $a+2$, over time this bubble might move to positions $a+1$, $a+2$ and $a+3$. As long as this transition was continuous, meaning that it didn't involve repair and re-nucleation, this is the same bubble and therefore still counts towards the lifetime requirement. As such, any analysis that requires a minimum lifetime must be able to account for these migration events, along with effects such as the merging of multiple bubbles.

The python scripts used to find both the locations of denatured base pairs and bubbles are included in the appendix (sections B.1 B.2).

2.9.2 Determination of toroidal and plectonemic loops

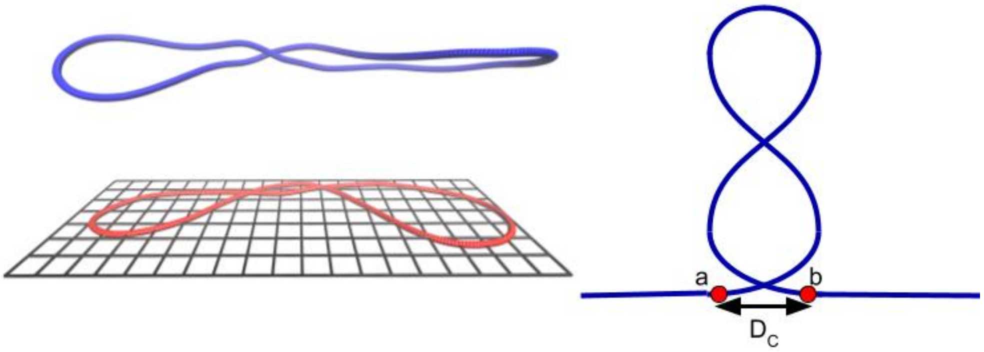


Figure 2.6: (**left**) DNA minicircle (blue) projected on to a 2D plane (red). Image from [158]. (**right**) Schematic representation of plectoneme start and end points.

Formally, a toroid is defined as a surface of revolution with a hole in the middle through which an axis can pass without intersecting the surface. In supercoiled DNA this is extended to loops in which a single DNA duplex crosses itself in 3D space, with a single crossing point defining a toroid and multiple crossing points defining a plectoneme.

In the same manner as bubbles, finding plectonemes upon visual inspection is simple, but creating a quantitative definition is challenging. The method used to find the start and end points in this text is heavily inspired by previous work in OxDNA [113], where crossing points are found by measuring the distance between base pairs that are separated by at least 40 steps, in our case however this methodology is complicated slightly. Plectonemes in OxDNA have clearly defined crossing points in which distal base pair co-localise in all three spacial dimensions, in our simulations this is not the case. Instead we often see distal base pairs co-localise in two of the three spacial dimensions, this means that, while they give the appearance of a plectoneme or toroid, the distance between them remains large. In order to overcome this issue simulation trajectories are projected on to a best-fit plane using SerraLINE (see figure 2.6), this 2D fitting makes crossing points share coordinates in all three spacial dimensions, making it possible to systematically locate them.

Once the DNA has been projected on to a plane the procedure is as follows:

- Starting at the end of the strand, loop over all possible base pair doublets (a,b) (i.e. (bp 1,bp N)...(bp N/2, bp N/2+1)), finding the cartesian distance $D_c = |\mathbf{a} - \mathbf{b}|$ and the contour distance $N_c = index(\text{a}) - index(\text{b})$.
- Set values for the maximum separation of the base pairs at the crossing points (D_{max}) and the minimum number of base pairs in the loop (N_{min}). These numbers are entirely set by the user, in this work they will be $D_c = 3\text{Å}$ and $N_{min} = 40bp$.
- If $D_c \leq D_{max}$ and $N_c \geq N_{min}$ then the base pairs a and b are marked as the start and end of the plectoneme respectively (see figure 2.6).
- The index of centre of the plectoneme is then $p_C = \frac{index(\text{a})+index(\text{b})}{2}$ rounded down.

The python code used to find plectoneme positions using the method above is listed in the appendix (section B.3).

2.9.3 Determination of twist and writhe

While the measurement of twist using the 3DNA definition is sufficient for finding regions of denaturation, the values its produces within these regions do not match real twist. This mismatch is due to the nature of the 3DNA definition, which assumes that base pairing is intact in order to measure twist.

In order to measure the twist absorption of bubbles a different definition is required, with the best option being topological twist. This definition is simpler than the 3DNA definition, it simply measures the number of times the backbones of the two strands cross one another. This measurement has the advantage of being independent of the behaviour of the bases themselves, focusing solely on the backbone. The WrLINE molecular contour [44] gives a measure of per-bp topological twist, meaning that, once the locations of bubbles have been identified, measuring their twist is simple.

The WrLINE molecular contour also measures DNA writhe, however many of the assumptions made in its methodology hinge on the DNA in question being circular. While some of these assumptions have been altered to allow for the measurement of writhe in short sequences [159], these modifications do not extend to the long seqeuences studied in this text. Instead, working on the assumption that supercoiling density is constant in all simulations, writhe will simply be calculated directly from topological twist using the Fuller-White-Călugăreanu theorem [41–43] (equation 1.4).

Chapter 3

All-atom MD Simulations Recreate Experimental Hat-curves

3.1 Introduction

As discussed in chapter 1, supercoiling is known to have an effect on a number of important regulatory processes, and to be the product of a large number of complex interactions [160, 161]. While these are often the consequence of interactions with proteins, it is still important to understand what role pure DNA plays within them, and therefore how it behaves under different levels of stress.

Naked DNA is able to absorb excess linking number through the unwinding or overwinding of the double helix (twist), or the formation of three-dimensional supercoiled structures (writhe). Continuous regions of denatured base pairs, called bubbles, form locally in undertwisted DNA, and the energy required to nucleate them is associated with the breaking of base pairing, stacking, and the unwinding of the double helix. This energy is roughly constant with respect to both tension and supercoiling density [162], only affected by factors such as salt concentration. In contrast, the formation of large scale supercoiled structures, called plectonemes, is inherently dependent on tension. Plectoneme formation, by its very nature, requires a shortening of end-to-end distance, thereby doing work against applied tension. It is the balance of energies between plectonemes and bubbles, as well as the direction in which excess twist is applied, which controls the behaviour of supercoiled linear DNA.

Provided that, at zero force, the energy of plectoneme formation is lower than that of bubble formation, there should be a range of forces in which plectonemes are more stable than bubbles, along with a point of transition in which the two are energetically similar. The nature of the transition between the two states is dependent on the manner in which they interact with one another: if the two act completely independently then a drastic transition between the two might be expected, giving rise to a sharp increase in the extension of the strand. Conversely, if the two are able to interact in a meaningful way, one might expect more of a smooth transition in extension, with the presence of intermediate states.

In magnetic tweezers experiments, the nature of the transition has been shown to be dependent on the direction of supercoiling, with bubbles thought to play little or no role in positively supercoiled systems. In negatively supercoiled systems, experiments show a force-dependent transition with what appear to be intermediate states, in which the strand is not fully extended, but also not as compressed as its positively supercoiled counterpart [163, 164]. Importantly, these states exist at $0.5 \leq F \leq 1pN$ and $\sigma \leq 0.1$, known to be in the range of forces and densities seen in the cell [165–167]. The existence of these intermediate states implies a form of cooperative behaviour unique to negative supercoiling, likely influenced by the presence of bubbles.

The resolution of tweezers experiments is limited, and while in principle one can observe plectonemes [168], the existence of bubbles cannot be confirmed, nor can any interaction between the two states be observed. The reality of force-extension experiments is that one often has to choose between torsionally restraining the DNA (magnetic tweezers) or its direct observation (optical tweezers), making the creation of a system in which one can twist the strand and observe the position of plectoneme nucleation exceptionally challenging.

As a result of the limitations of experimental techniques, simulations have been used to probe the behaviour of supercoiled DNA on a base pair level. One such study is that of supercoiled linear DNA using OxDNA by Matek *et al* [111], in which a wide range of supercoiling densities and tensions were explored. Here, observations of so-called ‘tip bubbles’ were made, forming sharply bent plectoneme tips. The presence of these structures was found to be highly force-dependent, appearing in systems with intermediate (in this case around $2pN$) levels of tension. While these simulations shed light on the interactions between plectonemes and bubbles, and were able to re-create the experimental hat-curve, they are still limited in their resolution and accuracy. First, transitional forces in these systems were much higher than the experimentally observed values, with the flattening of the curve only beginning at $F = 2pN$, significantly higher than the experimentally observed force [63]. In addition to this, the coarse grained nature of OxDNA means that some potentially influential features of DNA are omitted, for example innate curvature and non-canonical interactions. The coarse graining also results in a much flatter energy landscape, with minima in free energy often very clearly defined. While this is helpful in reaching a steady state in a short amounts of time, it also omits any potential stochastic effects, such as the spontaneous breaking of base pairs.

Another *in silico* study of DNA supercoiling is that of Pyne *et al* [59], in which the topology of supercoiled DNA minicircles was studied with atomistic simulations in conjunction with AFM experiments. This study observed the formation of sharply bent regions in the tips of plectonemes, synonymous with stack breaking, in both simulations and experiments, for all systems with $\sigma \leq -0.06$. However, due to the circular nature of the DNA, no force could be applied and therefore the effect of tension could not be studied.

The lack of atomistically detailed data regarding supercoiled linear DNA leaves space for exploration using the simulation methodology described earlier. This allows for the simulation of linear DNA at a wide range of supercoiling densities and forces using atomistic techniques, and will be employed here in an attempt to recreate the experimental hat-curve, as well as furthering observations made in previous simulations, with the aim to shed light on the behaviour of tip-bubbles, as well as give insight in to the influence of sequence.

3.2 Simulation setup

As a first step towards exploring this space a sequence must be chosen, the nature of which must allow for the formation of both plectonemes and bubbles, without excessively favoring one or the other. To this end a random sequence with a length of 300 bp and a 50/50 distribution of AT and GC was generated. The choice of length was made to strike a balance between simulation time and plectoneme growth potential; the length chosen is roughly two times the persistence length of B-DNA, allowing strong bending to be facilitated while keeping the sequence as short as possible. Due to the requirements of the simulation protocol the overall length of the strand used is 362 base pairs, with 300 of these being the aforementioned sequence and the remaining 62 being the dummy base pairs, all of which are GC in order to minimise disruption in their structure. The full sequence can be found in appendix 1 of this text (specifically section A.1).

The choice of supercoiling densities and forces was inspired directly by the work of Meng *et al* [100] with densities ranging from $\sigma = -0.1 \rightarrow 0.1$ and forces of $0, 0.3, 0.6, 0.7$ and $1pN$. The choice of supercoiling densities is simple, with the range chosen to represent a spectrum of biologically relevant values while avoiding extreme levels of supercoiling which may lead to the formation of exotic DNA structures such as P-DNA. The forces were chosen to maximise the number of systems at interesting, biologically relevant forces. The forces 0.6 and $0.7pN$ represent a range in which the DNA is in a transitionary state, being neither bubble nor plectoneme dominated. Here the aim was to capture the cooperation between plectonemes and bubbles, which should be the most prevalent at these intermediate forces, where both states can exist simultaneously. Forces at either end of the range were chosen to capture pure states, with plectonemes at low forces ($0, 0.3pN$) and extruded bubbles at high forces ($1pN$). Forces exceeding $1pN$ were avoided in order to minimise the effect of DNA stretching, which has been shown to cause the formation of denaturation bubbles in systems without supercoiling [169].

3.3 Recreation of experimental hat curves

The first step in examining the viability of atomistic simulations, and specifically the simulation setup described in the previous chapter, is to assess its ability to re-create well-known experimental results, specifically the 'hat-curve'. Recreation of this result will strongly indicate that the behaviour seen in simulation is an accurate representation of the behaviour of real supercoiled DNA, and give an unprecendented insight into the dynamics of these systems.

Positive supercoiling is the simpler of the two regimes, with twisting in the same direction as the helix restricting the range of possible conformations. This simplicity is highlighted by the much less chaotic nature of the twist and writhe of positively supercoiled systems when compared to their negatively supercoiled counterparts (figure 3.1 **B,C**). Predictably, as supercoiling density increases, so do both the writhe and twist of the systems. At lower levels of positive supercoiling ($\sigma < 0.06$) there is a clear divide between twist-dominated systems at low tension and writhe-dominated systems at high tension. The systems also appear to display a soft maximum for the amount of supercoiling that can be absorbed by twist, with twist leveling off at $\sigma = 0.08$, while writhe continues to increase.

Negatively supercoiled systems are far more complex, with twisting in the opposite direc-

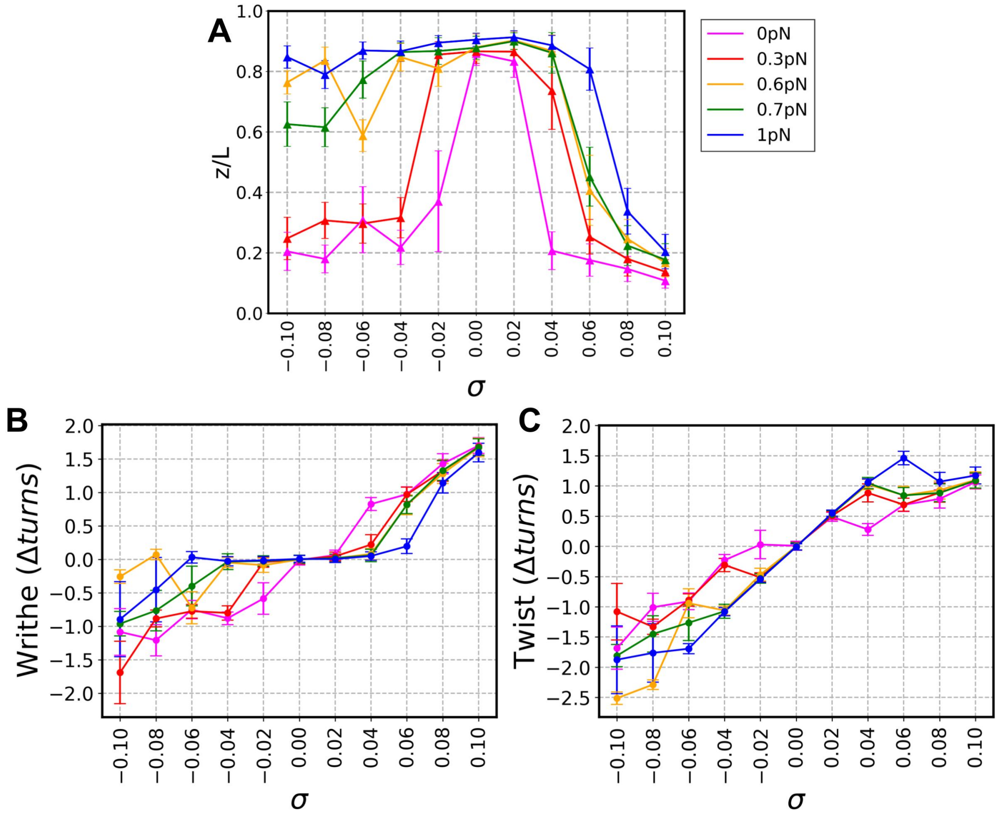


Figure 3.1: **A** Recreation of the experimental hat-curve using MD simulations. With the exception of anomalous results at $0.6pN$, the results quantitatively agree with those seen in experiment. **B,C** Writhe and Twist respectively, calculated using WrLINE. Behaviour in positive supercoiling is largely predictable, whereas behaviour in negative supercoiling, especially at higher densities, is much more complex.

tion to the helix causing a decrease in helical stability, enabling the formation of non-canonical structures [170, 171]. Broadly speaking, lower force systems are dominated by writhe, while higher force systems are dominated by twist, with critical tensions at 0.6 and $0.7pN$ marking σ-dependent transition points, after which the systems are extruded. In agreement with experiment, negatively supercoiled systems mirror their positively supercoiled counterparts at low supercoiling densities ($|\sigma| \leq 0.04$). As force and σ increase this symmetry begins to vanish, with all higher force, higher density systems showing completely different extensions to their positive counterparts. This difference is also reflected in values of twist and writhe, where, in general, twist becomes the much more dominant state at higher forces. However, unlike positive supercoiling, the separation between twist and writhe is much less clear and at higher densities ($|\sigma| \leq 0.06$) the partition of supercoiling between the two becomes much less clear, with most systems showing a strong mixture of both twist and writhe. Additionally, negatively supercoiled systems do not display the 'soft maximum' of twist seen in positive supercoiling, with the value of twist continuing to become more negative as σ does the same.

Although there are clearly some anomalous systems, particularly at $F = 0.6pN$, the system shows good overall agreement with experimental hat curves (figure 3.1 **A**), with asymmetry between positive and negative supercoiling as well as flattening of extension at high forces for $\sigma < 0$. The features of this force-extension curve strongly imply that simulations are recreating experimental behaviour, and therefore the high level of detail they offer should explain the behavior that gives rise to the shape of this curve.

3.4 Simulations give insight in to how DNA responds to tension and torsion

Combining the statistics in figure 3.1 with structural images seen in figure 3.2 gives a clear insight into the structural features controlling the behaviour of supercoiled DNA at a variety of densities and tensions. Given that this insight is unique to simulations, it is important to take this opportunity to explore it in detail.

Starting with weakly negatively supercoiled systems ($0 \geq \sigma \geq -0.04$), there is a clear separation into two regimes - buckled and unbuckled. This separation of states is demonstrated in figure 3.1 **B** and **C**, where a clear split between twist-dominated and writhe-dominated states is displayed. At $\sigma = -0.02$, there is a clear split between $0pN$, where all superhelical stress is released by writhe, and all higher forces, where this stress is relieved by twist. All of these states can be seen in figure 3.2, where a clear toroid is seen in the $0pN$ case, while the remainder of the states are fully extended. Separation is also clearly evident at $\sigma = -0.04$, this time occurring between 0.3 and $0.6pN$. Throughout the regime of low force, low σ systems, toroids are seen instead of plectonemes, this is a simple consequence supercoiling density only requiring a single crossing point to absorb all of the necessary strain. The size of these toroids is on the same scale as the entire strand, this implies that the ideas suggested by classical polymer physics [81] stand true - the size of the toroid is maximised in order to minimise the energy required to form it. At high forces, bubbles begin to emerge at $\sigma = -0.04$, implying that the ability of B-DNA to absorb changes in twist is being pushed, with bubbles becoming the more energetically favorable state.

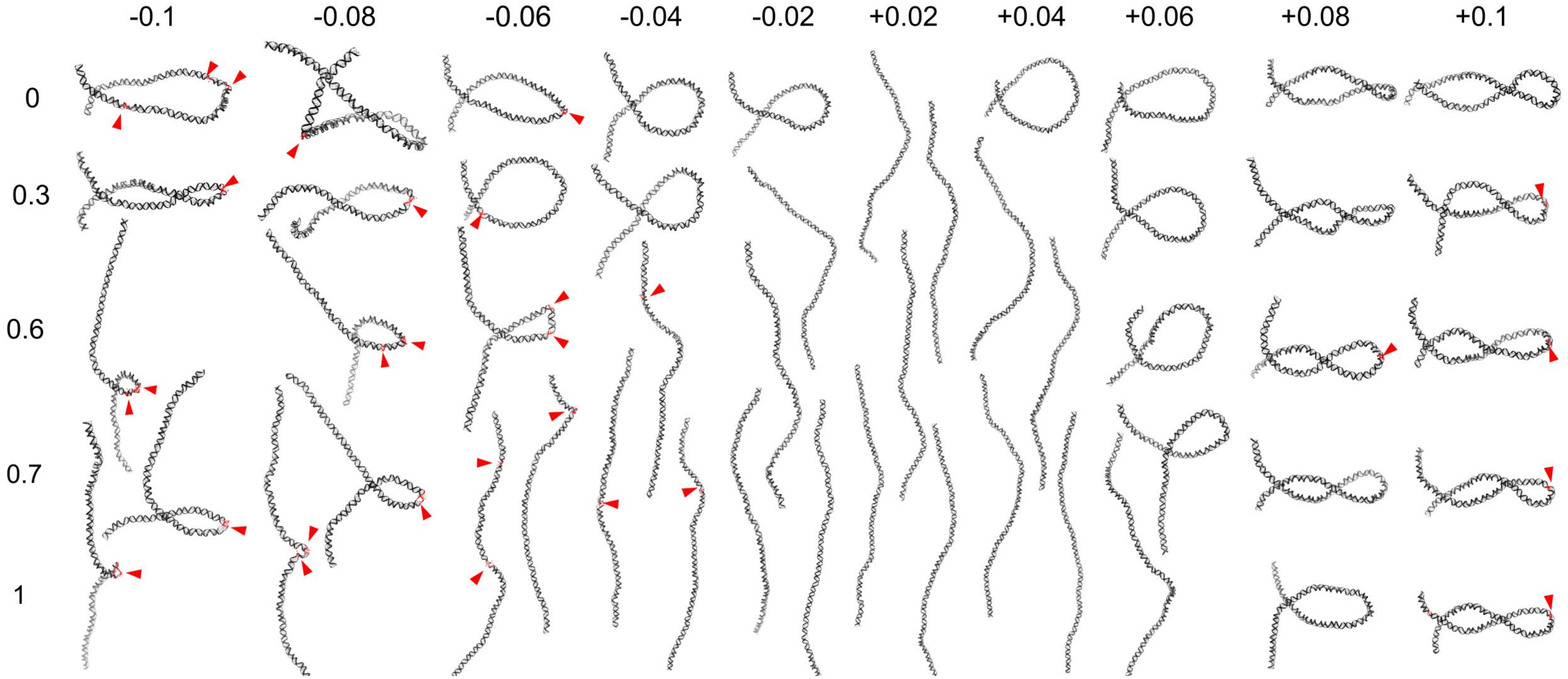


Figure 3.2: Full landscape of representative structures for all simulations. Bubbles are highlighted red and marked with a red arrow. In the same manner as the hat-curve, the clear asymmetry between negative and positive supercoiling is displayed here, with structures at high force and high $|\sigma|$ contrasting one another. Here we also see the first clear presence of plectoneme tip-bubbles, with bubbles and plectonemes co-localising, allowing the bubbles to absorb a significant portion of the strong bending at the tips of plectonemes.

Moving to the weakly positively supercoiled systems, a clear separation in states is once again displayed, this time emerging only at $\sigma = 0.04$. Behaviour at this density reflects that of $\sigma = -0.02$, with toroid formation arising only at $0pN$, and is reflected in plots of twist and writhe. The asymmetry between negatively and positively supercoiled systems at these low densities is a surprising result. Intuitively one might expect it to be reversed, with negatively supercoiled systems buckling at higher densities. This would be easily explained by the presence of bubbles, with small amounts of disruption able to absorb excess twist, nullifying the need for plectoneme formation. However, what is seen in simulations is the opposite of this - it seems that plectoneme formation in positively supercoiled systems is somehow more difficult. This is not the first time this asymmetry has been observed, having been seen in simulations of supercoiled DNA minicircles [45]. One possible explanation for this asymmetry could be self-interaction between the DNA at the crossing point. There is a distinct difference between the crossing points depending on the direction of supercoiling, and this difference could lead to stronger repulsion in one regime vs another. This idea can be dismissed by [172] however, here it is shown that, electrostatically, the left-handed crossing seen in negative supercoiling is actually more energetically favorable than the right-handed crossing seen in its opposite. Another possible explanation for the unexpected asymmetry is an entirely energetic one - it is possible that the ability of the helix to absorb twist is inherently asymmetric, with twisting in the positive direction requiring less energy than its opposite, thereby making the state of purely twisted B-DNA more energetically favorable in positively supercoiled systems.

At high levels of positive supercoiling ($\sigma \geq 0.06$), the plectonemic and extended states are once again seen, with the former dominating all systems with $\sigma \geq 0.08$. The behaviour of these higher σ values is exactly as expected; plectonemes are the only structures able to practically absorb large amounts of excess superhelical stress, and as such it is expected that extension will decrease as σ increases, even in the presence of high tension. The main influence of tension in positive supercoiling seems to be the σ values at which the DNA buckles, with $1pN$ preventing bucking for all σ values up to and including 0.06. At values of σ where buckling is seen universally ($\sigma > 0.06$), tension has only a small and predictable effect on extension, with high tension systems showing a slightly lower level of extension when compared to low tension. At high positive supercoiling and force the extension is increased without any significant change in twist or writhe compared to lower force systems at the same density. This implies that this tension simply tightens the plectonemes, decreasing their size without changing the number of crossing points. At the highest levels of positive supercoiling ($\sigma \geq 0.08$) regions of significant base pair disruption begin to appear, and these regions of disruption manifest themselves in the form of bubbles. The representative structures of figure 3.2 imply that these bubbles appear exclusively at the tips of plectonemes, this suggests that the strong curvature of the plectoneme tip may be directly influencing their formation. Interestingly, positively supercoiled systems in which bubbles are present show little increase in twist compared to those in which bubbles are not present, suggesting that the ability of bubbles to absorb excess twist is hindered in this regime.

Behaviour at high levels of negative supercoiling ($\sigma \leq -0.06$) is far more complex, with interplay between plectonemes and bubbles appearing to control dynamics. Here once again the plectonemic and extended states are seen, but there also appears to be a transitionary state which represents a hybrid of the two. In plectonemic states, bubbles are universally present, implying cooperation between the two states, and are almost always present at the tips of plectonemes. Take for example $\sigma = -0.08$ and -0.1 at $0.3pN$: these are states that largely reflect their positively supercoiled counterparts, but with bubbles at the tips of their plectonemes. These tip-bubbles have two possible effects: First, they may allow for sharper bending at the tip of the plectoneme, with their increased flexibility lowering the energy required to nucleate the plectoneme tips. Secondly, they may act to absorb twist, with the lower torsional stiffness of single stranded DNA making them ideal focal points for torsional stress. The role of these tip bubbles in negative supercoiling is more wide-ranging than their positively supercoiled counterparts - they appear at lower supercoiling densities, and are more numerous at higher densities. This is owing to the energetic favorability of bubbles in negative supercoiling in general: the formation of a denaturation bubble requires the straightening of the DNA double helix - a process which is favorable in systems where torsion is applied in the opposite direction to the helical twist.

Transitionary states are more difficult to concretely define than their extended or plectonemic counterparts, they act more as a continuum between the two, representing all states that don't fall under the banner of the others. These transitionary states are dominant in highly negatively supercoiled systems at medium to high forces, with some good representations of them seen at $\sigma = -0.06, F = 0.6pN$ and $\sigma = -0.1, F = 0.7pN$. These states are usually defined by the presence of a toroid with one or more bubbles within it, although many also have bubbles outside of the toroid as well. The role of these states, as well as the plectonemic and extended states need to be explored in greater detail, a process that will require much more in-depth statistical analysis of the systems.

3.5 Structure and role of bubbles in supercoiled DNA

A Note on Analysis

As discussed in section 2.9.1, the definition of bubbles in supercoiled systems is far from trivial, with significant noise arising from DNA breathing as well as low levels of stochastic bubble formation in all simulations. It is vital when discussing bubbles that the defnition used to classify them is clear, and the definition used will therefore be re-stated here.

First, denatured base pairs are found using the criteria described in section 2.9.1, these are then further refined by requiring that any bubble must contain *at least* three consecutive denatured base pairs. Finally, and most importantly to the following analysis, a requirement that the bubble have a lifetime of *at least* $1ns$ is implemented, as this aids dramatically in the reduction of noise due to stochastic bubble formation, while maintaining bubbles with lifetime long enough to be relevant to system dynamics.

Unless otherwise stated, the above definition will be used for all analysis of bubbles for the remainder of this chapter.

In addition to the above definition of bubbles, in order for the data presented here to be as representative of equilibrium states as possible, data used will be taken only from the last 400ns of each simulation. If this is not the case for a specific plot the amount of simulation used will be explicitly stated.

3.5.1 An overview of bubble prevalence

The ability of bubbles to adopt highly bent and twisted conformations is key to their role in supercoiled DNA. The lack of both hydrogen bonding and stacking interactions in denatured regions of DNA makes them both less resistant to bending ($l_p \approx 2nm$ for ssDNA [173] vs $l_p \approx 50nm$ for dsDNA) and more torsionally flexible than B-DNA (the torsional stiffness associated with interstrand twisting of ssDNA has been measured to be around $1/80$ that of B-DNA [97, 174]). This allows bubbles to absorb significant amounts of both bending and superhelical stress with little energetic cost. Balancing the energy savings associated with increased flexibility is the initial energy investment required to nucleate a bubble, with this balance driving force behind their formation. The initial cost of bubble formation - a result of the breaking of hydrogen bonding and stacking, requires a significant energy investment, presenting a barrier to bubble formation (10-12kcal/Mol [97]). In systems where increased flexibility is more useful - namely those with higher levels of supercoiling - the probability of bubble formation is significantly higher.

Figure 3.3 shows the average number of base pairs in bubbles for all simulations, and demonstrates the clear contrast between negative and positive supercoiling. The formation of bubbles in positively supercoiled systems is inhibited by the need to open up the denaturation site, a process that requires the straightening of the backbone and in turn induces localised negative supercoiling. As a result of this the balance of energies, positively supercoiled systems are swayed significantly against bubble formation, instead preferring to from writhed structures, even at higher forces. Contrast this with negative supercoiling, where the formation of the localised negative supercoiling at the bubble site acts to absorb some of the global supercoiling, therefore skewing the energy landscape in favor of bubble formation.

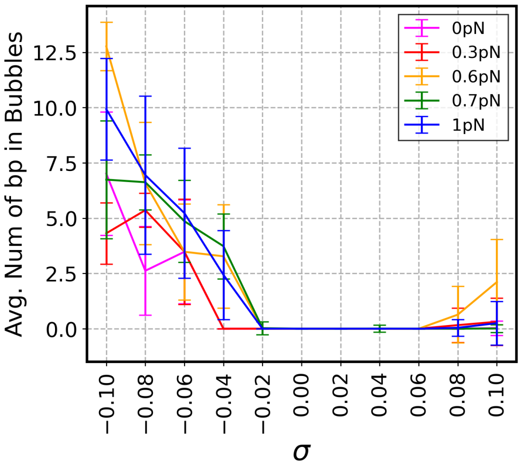


Figure 3.3: Number of base pairs in bubbles for all simulations, with standard deviations indicated. The contrast between positive and negative supercoiling is clear - bubbles are far less prolific in positively supercoiled systems when compared to their negatively supercoiled counterparts.

Comparison of figure 3.3 with figure 3.1 begins to reveal the role of bubbles as loci for twist, especially in negative supercoiling. In general, the number of base pairs in bubbles correlates with both extension and twist, implying that larger or more numerous bubbles can act to absorb large amounts of torque, reducing the need for plectoneme formation. In positive supercoiling this role seems much less significant. Take for example $\sigma = 0.1, F = 0.6pN$, this system has the largest number of base pairs in bubbles of any positively supercoiled system, and yet its values of extension and twist are no larger than any other system at the same supercoiling density. In order to fully understand the role of bubbles in these systems, a more in-depth analysis of their twist is required.

3.5.2 Bubble Stability

Figure 3.3 does not give a full view of the role of bubbles in these systems - a larger average number of base pairs in bubbles could be the result of smaller bubbles, less stable bubbles, or a mixture of the two. It is therefore vital that the stability of bubbles be analysed such that any conclusions made regarding their role are as accurate as possible.

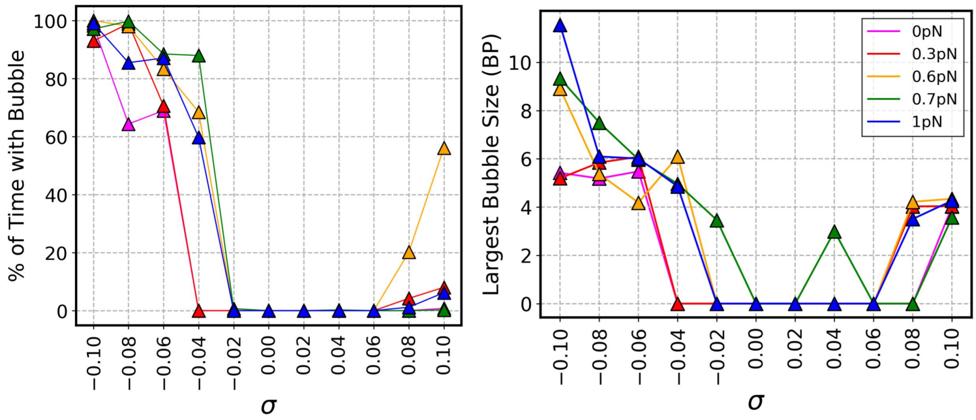


Figure 3.4: Stability and sizes of largest bubbles across all simulations. (**left**) Percentage of simulation time spent with a bubble present. (**right**) Average size of the largest bubble. Both figures show clear asymmetry between positively and negatively supercoiled systems, with bubbles in positive supercoiling being both smaller and shorter lived.

Figure 3.4 shows both the amount of each simulation spent with a bubble and the average size of the largest bubble in each simulation. This figure makes it clear that, in the positively supercoiled regime, the lower averages seen in figure 3.3 are the result of bubbles being both smaller and less stable. The smaller size of bubbles suggests that, even when they are present, their effects on the system will be minor, an assertion backed up by the lower extensions and twist values seen in figure 3.1. The lower lifetimes of bubbles in positive supercoiling is perhaps more significant than their smaller sizes, it suggests that, while bubble formation is commonplace in strongly supercoiled systems, the time for which these bubbles last is much lower than their negatively supercoiled counterparts. The cause of this reduced stability may be the structure of the bubbles themselves, with the need to increase interstrand twist meaning that the denatured regions maintain much of the same structure as B-DNA, making it much easier for the regions to re-form back in to B-DNA.

3.5.3 Twist absorption by bubbles

The ability of bubbles to absorb twist varies massively depending on the direction of supercoiling, as evidenced in figure 3.5. In negatively supercoiled systems, the ability to absorb twist is key to the function of bubbles, and is their sole role in fully extended systems. Take for example systems at $1pN$, here we see significant twist absorption in denatured regions without significant increase the twist of the regions of B-DNA. In positively supercoiled DNA, the ability of bubbles to absorb twist is stunted, with only a slightly higher average twist value in denatured regions when compared to B-DNA. This stunting is likely due to self-interaction in the DNA backbone, with the values of twist seen in positively supercoiled bubbles representing

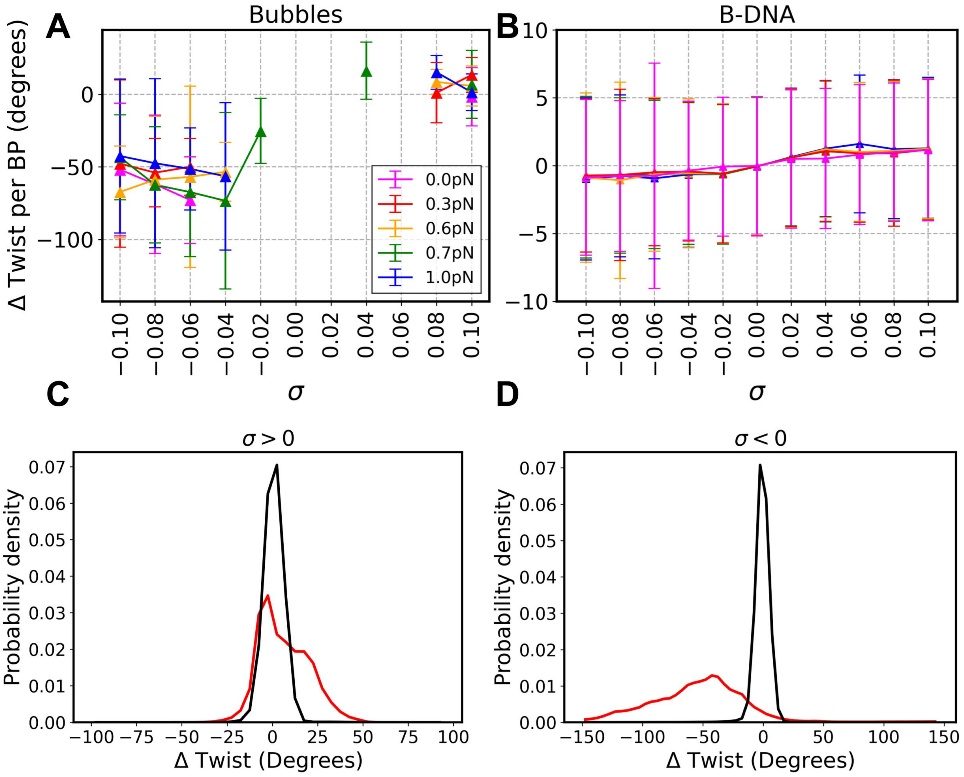


Figure 3.5: Twist absorption by bubbles in all systems. **A, B** Average change in twist for bubbles and B-DNA respectively. Values for bubbles show a clear divide between direction of supercoiling, with the massively decreased torsional stiffness of bubbles allowing for much higher values of twist. B-DNA values show little variation from values at $\sigma = 0$. **C, D** Distribution of twist values across all simulations, split by direction of supercoiling, with black curves representing the distribution of values in B-DNA, and red curves representing the distribution of values in bubbles. Here the contrast between the two directions is seen, with bubbles in negative supercoiling able to absorb significant amounts of twist, while those in positive supercoiling differ little from the values for B-DNA.

the maximum amount that the backbones of the two strands can wrap around one another before excluded volume and electrostatic interactions become too strong. This extremely limited twist absorption strongly implies that the role of bubbles in positively supercoiled systems is almost entirely to their role as loci for strong bending at the tips of plectonemes.

Values for the twist of bubbles seem to be independent of the value of $|\sigma|$, changing only when the direction of supercoiling is changed. This suggests that existence of a soft maximum for the amount of twist that can be absorbed per base pair, with higher levels of supercoiling resulting in the growth of bubbles or the formation of additional bubbles (as seen in figures 3.3 and 3.4) rather than an increase in the amount of twist absorbed per denatured base pair.

Values of twist in negative supercoiling are also extremely noisy, with a huge range of values displayed in 3.5 **D**. This is the result of the extremely high torsional flexibility of the denatured regions, meaning that the amount of twist can vary massively without too large of an energy cost being incurred. This massive variation is not present in positive supercoiling, as evidenced by 3.5 **C**, due to the aforementioned maximum twisting of the two backbones.

3.5.4 Bending in bubbles

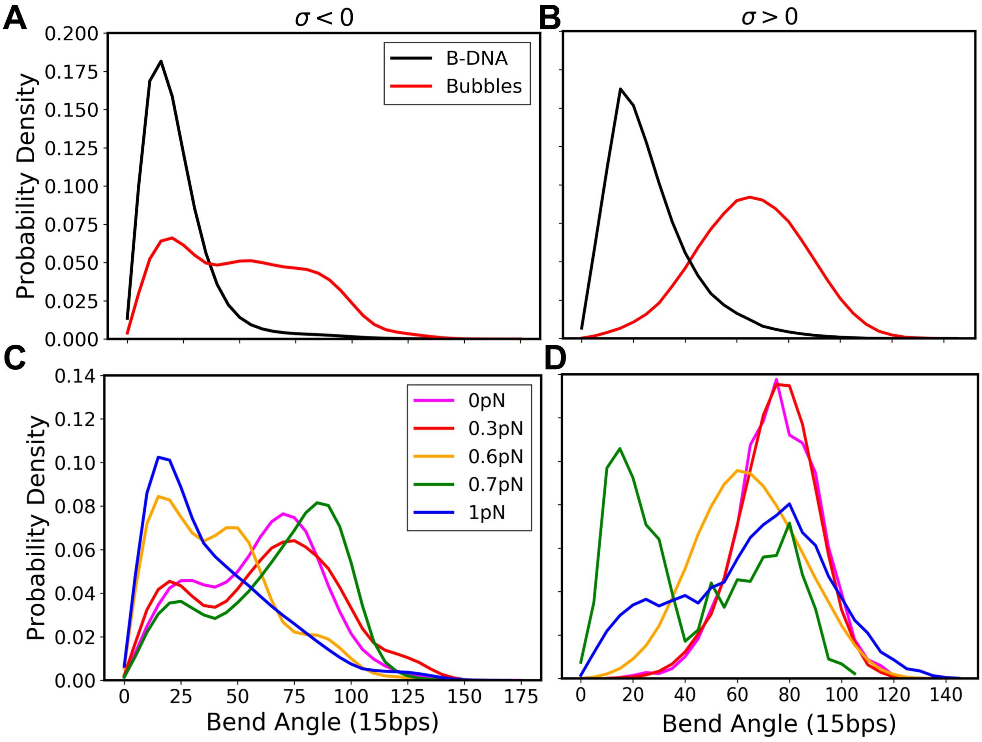


Figure 3.6: **A,B** Strong bending in bubbles, split by direction supercoiling. **C,D** Bending in bubbles further sub-divided by force. Bending in negative supercoiling is complex due to the multiple states in which bubbles are seen, bubbles in positive supercoiling almost always occur at the tips of plectonemes, acting as focus points for the strong bending.

The absorption of strong bending is the second key role of bubbles in supercoiled systems, with the lower stiffness of denatured regions allowing them to bend strongly with little to no energy penalty. This ability is particularly relevant in the tips of plectonemes, where the strong bending of the plectoneme tip acts as the main energy barrier for plectoneme formation.

As implied by the lack of strong twisting, the role of bubbles in positive supercoiling is the absorption of the strong bending in the tips of plectonemes, this is confirmed by the distribution of bend angles seen in figure 3.6, in which a single peak at around 75 degrees is seen, aligning with angles observed in previous studies [59]. The only exception to this rule is the $F = 0.7pN$ case, where a secondary peak that aligns with the values for B-DNA is seen. This peak is largely irrelevant however, and arises from some stochastic bubble formation in the extruded system of $\sigma = 0.04, F = 0.7pN$ (similar stochastic formation is also the result of the increased width of the $F = 1pN$ peak). As demonstrated by the representative structures of figure 3.2, these tip-bubbles appear independently of tension, meaning that supercoiling density is the driving force behind their formation. Distributions of twist and writhe strongly suggested that, in positively supercoiled systems, the amount of twist had a soft upper limit. The formation of these bubbles is likely the result of the system becoming saturated - the amount of twist the B-DNA can absorb is at a maximum, and the amount of writhe that can be formed is at its upper limit due to the size of the plectoneme being equal to the size of the system. This leaves only bubbles as a source of extra torsional relaxation, with their formation allowing the size of the plectoneme to shrink slightly, thereby furthering its ability to absorb writhe.

The distribution of bending in negative supercoiling is much more complex, with multiple peaks, such as those at 20,50 and 85 degrees. The shape of this distribution becomes clearer once it is broken down by force. Starting at $0pN$, a bimodal distribution of bend angles is seen with peaks at around 25 and 75 degrees. These peaks represent two bubble states - extruded and tip-bubbles, with bend angles aligning with those seen in B-DNA and positively supercoiled bubbles respectively. The relative size of the peaks in this case suggests that, while tip-bubbles are dominant, the presence of bubbles outside plectonemes is significant. The story at $0.3pN$ is largely the same, with a slight increase in the size of the 25 degree peak, once again showing that bubbles outside plectoneme tips play a significant role. Moving to $0.6pN$, the systems display three peaks, two aligning with the previously seen states, and a third between the two. This 'intermediate' state is exclusive to this force, and may be the origin of the anomalous extensions seen at this force. $0.7pN$ once again returns to the bimodal behaviour, with tip bubbles more prominent here, while the secondary peak at 25 degrees is the result of bubbles in extruded systems. Finally, at $1pN$ the 25 degree state is completely dominant, owing to the fact that all the systems at this force are extruded. The long tail of the peak at this force is the result of strong curvature in the bubbles of the $\sigma = -0.08, 0.1$ systems.

The bubbles seen in these systems can be broadly broken down in to three states: tip-bubble, extruded, and intermediate (figure 3.7). The tip-bubble and extruded states represent the significant peaks in figure 3.6, while the intermediate state represents all others that do not fall under these two umbrellas, mainly those seen in $F = 0.6pN$.

The extruded state, defined by bending between 20 and 30 degrees, is seen with varying amounts of significance, at all forces. This state is most prevalent in systems with high force, owing to the extension of these systems not allowing the formation of plectonemes. Outside these high force systems, extruded bubbles are also seen in systems where bubbles form outside the tips of plectonemes, including in those that have multiple bubbles, for example $\sigma = 0.1, F = 0pN$ (see figure 3.2).

The tip-bubble state, in which bubbles co-localise with the tips of plectonemes, accounts for the strongest observed bending. These states are seen in systems with higher supercoiling densities, where the bending at the plectoneme tips is the strongest. Interestingly, there are also tip-bubbles present in lower σ systems (for example $\sigma = 0.06, F = 0pN$) despite the weak bending of toroids compared to plectonemes. The formation of this bubble is likely the result of stochasticity, the breaking of a single base pair overcomes a large amount of the initial bubble formation energy, and from here a full sized bubble, which in turn absorbs twist, is formed. These tip-bubbles are also the only state that appears consistently in positive supercoiling, although the structure of the bubbles in positively supercoiled systems is likely different to that seen in figure 3.7.

The intermediate states are a consequence of bending localised to bubbles in otherwise extruded systems. The system seen in figure 3.7 has a small toroidal region focussed around the denatured region. This structure allows for the absorption of writhe without the need for significant changes to extension. Capturing this third transitionary state within a single figure is impossible as, in general, the state is hard to define. The only consistent definition of this state is that it is neither fully extended nor a tip-bubble, making it a definition by omission.

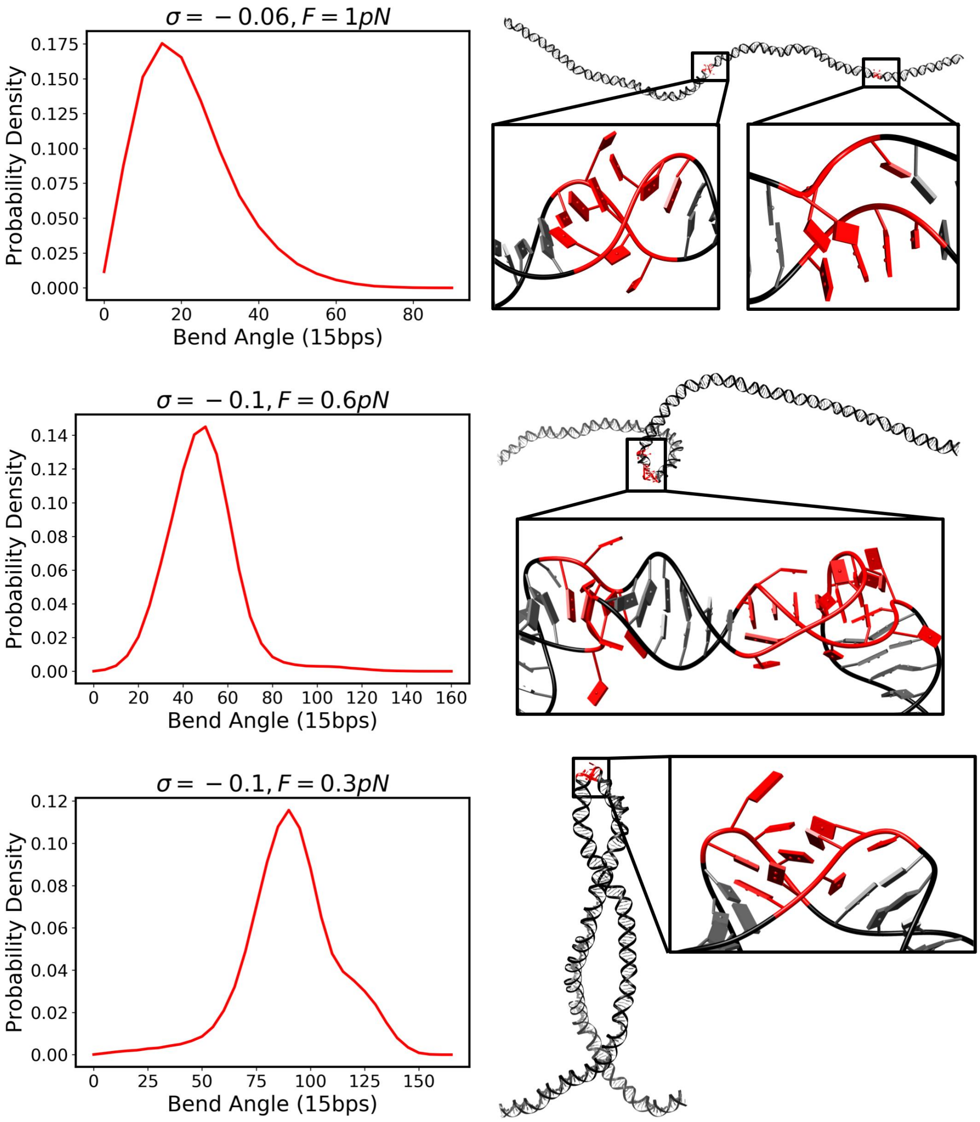


Figure 3.7: Representative examples of three observed bubble states along with their corresponding bend profiles. (**Top**) Extruded bubbles are seen in systems with high forces, as well as those with bubbles outside of plectonemes, these bubbles have bend profiles similar to that of B-DNA. (**Middle**) Intermediate bubbles are seen in $F = 0.6pN$ systems, these have bend angles higher than those of extruded bubbles, but lower than tip-bubbles. (**Bottom**) Plectoneme tip-bubbles are seen in low force, mid/high supercoiling density systems and exhibit the highest bend angles. These tip-bubbles are also the only denatured states observed consistently in positively supercoiled systems, where they rise independently of force at high densities.

3.5.5 Structure of bubbles in positive supercoiling

As stated above, analysis from figures 3.5 and 3.6 strongly suggests that the role of bubbles in positively supercoiled systems is to act as loci for bending at the tips of plectonemes. It is also clear from figure 3.4 that bubbles in positively supercoiled systems are unstable, this instability suggests a structure that is unlike those seen in figure 3.7, where the double helix is greatly perturbed, and more akin to canonical B-DNA.

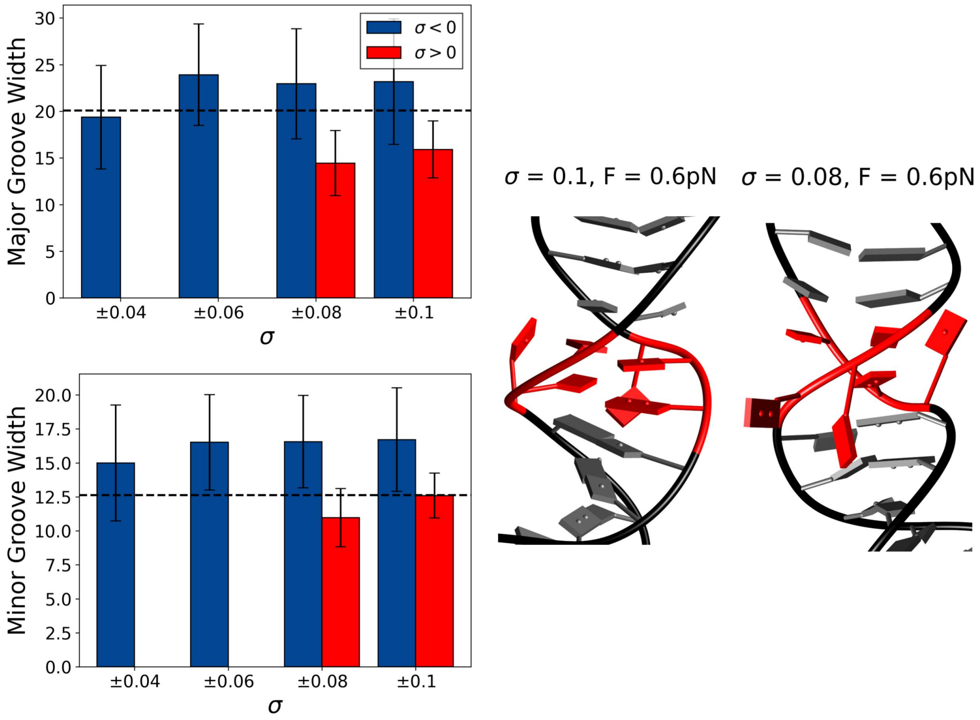


Figure 3.8: The structure of bubbles in positive supercoiling is much closer to that of B-DNA than their negatively supercoiled counterparts. Major and Minor groove widths of bubbles in positive and negative supercoiling (**left**) next to structures of two of the longest-lived bubbles seen in positive supercoiling (**right**).

The structure of bubbles in positive supercoiling is shown in detail in figure 3.8. As seen in section 3.5.3, positively supercoiled bubbles have very little excess twist. Still, they present a reduced width in the major groove, likely a consequence of the lack of stacking interactions, giving rise to a slightly more compressed structure. With the exception of this compression, the structure of these bubbles is very similar to that of B-DNA. This similarity is likely the underlying cause of the decreased stability of the bubbles - with the backbone structure largely in-tact, all that needs to happen in order for these bubbles to repair themselves is for the complementary bases to come close enough to each other to re-bind. It is likely that the most stable positively supercoiled bubbles are those in which the compression of the major groove is the most significant, here the base pairs will be unable to re-insert between the backbones, with the other base pairs acting as a physical barrier.

3.6 Structure and dynamics of plectonemes

Toroids and plectonemes are the structures by which supercoiled DNA absorbs σ through writhe. The atomistic simulations performed in this text offer a unique insight into these structures, allowing for observation of features such as size (and how this changes as a function of force and supercoiling density), and dynamics such as short-ranged migration.

Figure 3.9 shows average plectoneme sizes as a function of both force and supercoiling density. It is clear from this figure that plectoneme size is strongly dependent on the applied force, while dependence on supercoiling density varies accoding to the direction of supercoiling.

Plectonemes are the only structure through which positively supercoiled systems can absorb excess superhelical stress, and as such the sizes of plectonemes are predictable. In all but the $0pN$ systems, sizes of plectonemes grow in a roughly linear fashion as σ increases, with a maximum size at around 250 base pairs. This behaviour can be understood with a simple energetic argument - the energy required to form a plectoneme is a function of the work done against tension, meaning that larger plectonemes are only energetically favorable in systems with large amounts of B-DNA twist. In the absence of tension there is no energetic penalty associated with plectoneme growth, as such plectonemes will always act to maximise their size. It's worth noting here that this is a feature exclusive to short strands such as the one studied here. In longer DNA molecules entropy becomes a driving factor, and entropic elasticity would result in the probability of a plectoneme that is the size of the entire sequence being extremely low.

Plectoneme behaviour in negative supercoiling is significantly more complex due to the presence of bubbles. In negative supercoiling, plectoneme growth is stunted by the ability of the systems to form bubbles in their place. The work done against tension can be completely negated if there are no plectonemes formed, and they are instead replaced by bubbles. The balance of energetics between this work and the initial cost of bubble formation results in plectonemes vanishing at higher ($F \geq 0.6pN$) forces, where bubble formation becomes almost universally more favorable. This balance is still present in systems with $F = 0.3pN$, as evidenced by the lack of plectoneme growth at higher values of σ. This suggests that, at this force, the balance of energies allows plectonemes to grow to around 200bp in size before bubble formation becomes more energetically favorable.

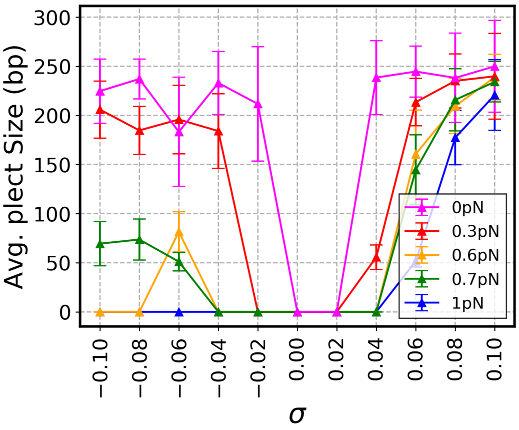


Figure 3.9: Average number of base pairs in plectonemes for all simulations, with standard deviations indicated. Contrast between positive and negative supercoiling is once again clear, however both seem to share a maximum plectoneme size at around 250bp. Data from the last 400ns of each simulation.

3.6.1 Plectoneme size as a function of force

Understanding the effect of tension on plectoneme size is important for predicting their behaviour [175], and as such systems with plectonemes only, namely those in positive supercoiling, should be studied in detail.

Plectoneme sizes as a function of force are shown in detail in figure 3.10. The choice of $\sigma = +0.08, +0.06$ was made to capture the behaviour of systems that are largely free of the influence of bubbles. These systems show a decrease in plectoneme size as a function of force, implying that, while writhe is largely unaffected, the shrinking of plectonemes does in fact occur. This is attributed solely to the higher tensile force, as work done against this force is directly proportional to the size of the plectoneme. The effect is much more significant in $\sigma = 0.06$, likely due to the toroidal nature of the systems - the writhed structures do not require a second crossing point, and are therefore much less tightly wrapped and more free to shrink.

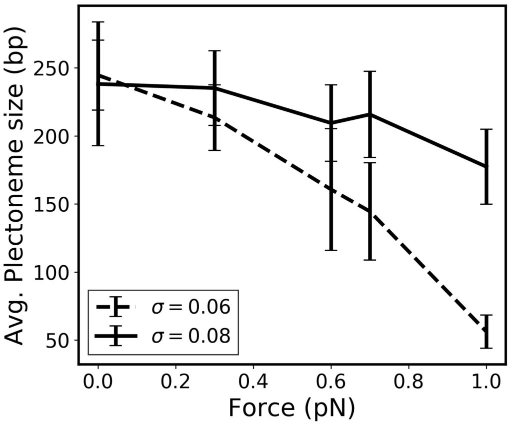


Figure 3.10: Number of base pairs in plectonemes for simulations at $\sigma = 0.06, 0.08$, with standard deviations indicated. Variation in plectoneme size is clearly more significant in the lower density system, however both systems show significant reductions in size as a function of force.

While the difference in size between $0pN$ and $1pN$ at $\sigma = 0.08$ is certainly statistically significant (p=0), the actual difference in size is fairly small, at around 50 base pairs. This suggests that the plectonemes are already close to their limit in terms of tightness, likely due to the small size of the system. This limit also explains the formation of bubbles at $\sigma = 0.1$ - the plectonemes formed are unable to decrease in radius, and the bubbles offer a vector by which they can shrink, with the sharp bending of the bubbles decreasing the size of the plectoneme tip without affecting the rest of the structure.

Another notable feature is the difference between 0.6 and $0.7pN$ at $\sigma = 0.08$. While this difference is small, it is certainly statistically significant (p=2×10^{-156}). The difference in size can be attributed to the presence of a bubble in the $0.6pN$ system, which acts to reduce the size of the plectoneme tip, thereby reducing the overall size of the plectoneme.

3.6.2 Plectoneme formation

The formation of plectonemes can occur via two different mechanisms: extrusion and rebuckling. The first is a process by which the size of the plectoneme increases at the same rate as writhe - each turn is sequentially extruded from the strand. In the second the formation of an initial large toroid occurs, with this toroid then re-buckling to form a second crossing point.

The process of plectoneme formation varies from system to system, however most appear to display similar behaviours, with the initial formation of a toroid (often accompained by short-ranged migration) and subsequent extrusion of further crossing points. In many systems,

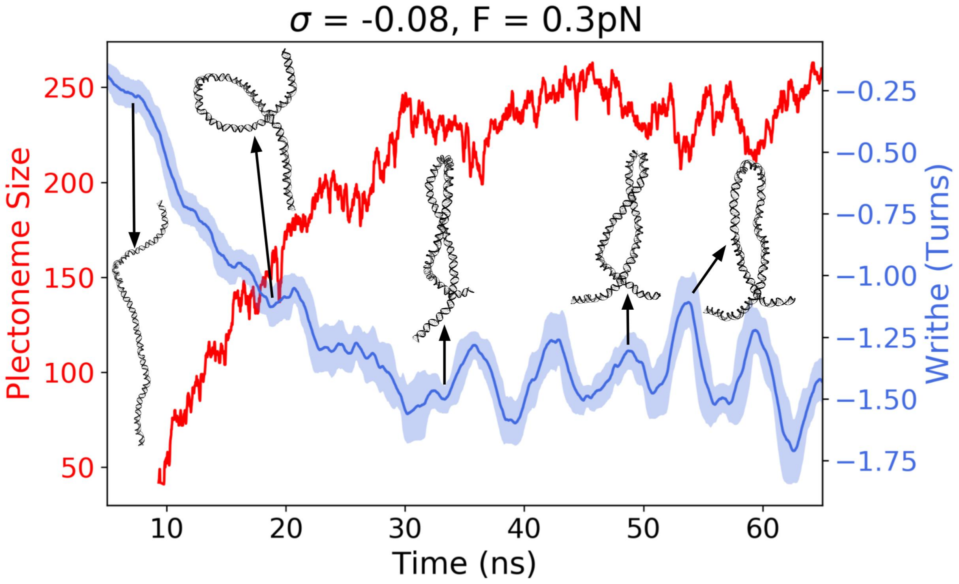


Figure 3.11: Size of plectoneme and writhe over time for the first 65ns of a single simulation, accompanied by representative structures. This system displays initial extrusion followed by a series of unbuckling and rebuckling events. Writhe is calculated as a moving average over a window of $1ns$, with the standard deviation over this window displayed by the shaded area. Note that, for this particular figure, the first $40ns$ is included, during this time the hydrogen bonds between base pairs are held together.

such as the one shown in figure 3.11, this initial behaviour is followed by a series of small rebuckling events. Specifically, this system experiences an initial extrusion between 5 and 30 ns, in which the size of the plectoneme grows close to its maximum, along with writhe. Following this initial extrusion is a series of unbuckling and rebuckling events, for example at around $52ns$. These events cause a significant change in writhe, while only at small change (<25bp) is seen in the size of the plectoneme. This corresponds to the disappearance and reappearance of the second crossing point (displayed in the representative structures), which is effectively a transition between the toroidal and plectonemic states. These events imply that, once the initial extrusion is complete, and provided that the loop is already near its maximum size, loop size does not strictly correlate with writhe, indicating only that the system is in a buckled state.

3.7 Plectoneme-bubble interactions

While the examination of bubbles and plectonemes as separate entities gives clear insight into their behaviour, it is in the interactions between the two that the most interesting behaviour is seen. It is clear from figure 3.2 that the two structures are deeply related to one another, and that these simulations recreate the tip-bubbles first seen in OxDNA simulations [111]. However, it is also apparent that bubbles, specifically in negative supercoiling, are not restricted to the tips of plectonemes, and that they appear at much lower forces than previously observed.

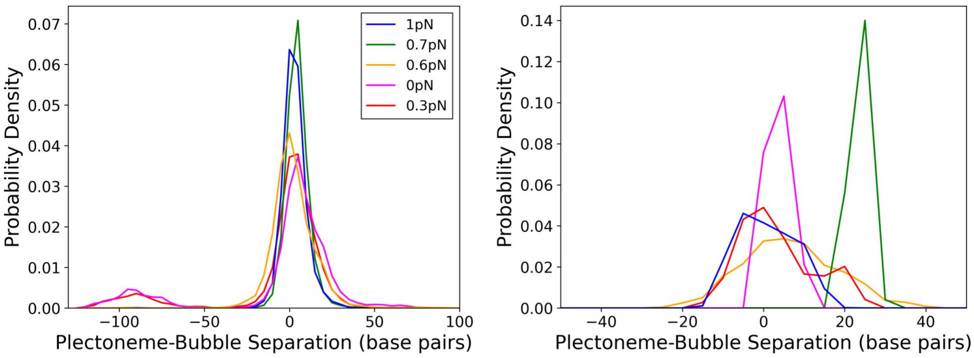


Figure 3.12: Distance between the centres of plectonemes and bubbles for negatively (**left**) and positively (**right**) supercoiled systems. Both systems show clear co-localisation, however the lack of data in the positive case creates significantly more noise.

3.7.1 Plectoneme tip-bubbles

Figures 3.2, 3.6 and 3.7 gave a clear indication of the existence of tip-bubbles in multiple systems across a series of forces and supercoiling densities. It is clear from these figures that these states are fairly common at our chosen levels of supercoiling and force, and that they play an important role in the dynamics of the systems. However, it is also clear from figure 3.2 that the interactions between these structures are not universal i.e the simultaneous existence of a plectoneme and a bubble in a system does not automatically mean the presence of a tip bubble.

The role of tip-bubbles in positive supercoiling is clear, with their formation only seen in systems at the highest level of positive supercoiling, and independently of force. The presence of bubbles in these systems is likely a consequence of the plectoneme size reaching a soft maximum (figure 3.10), after which decreasing the radius of plectonemic wrapping becomes less energetically favorable than the formation of a bubble, this in turn reduces plectoneme size without the need to decrease the radius. This co-localistation is demonstrated in figure 3.12, where co-localistation is seen across all forces. The exception to this co-localisation is the case of $F = 0.7pN$, where a single simulation gives rise to the off-centre peak.

In negative supercoiling, the role of bubbles is more complex, with nucleation seen at lower levels of supercoiling and force than previously observed, as displayed in figure 3.2. This increased complexity owes itself to the ease of bubble formation in negative supercoiling. As mentioned previously, bubbles play a dual role, both reducing plectoneme size and absorbing twist. This makes the balance of energetics between bubble and plectoneme formation far more subtle than in their positively supercoiled counterparts. In systems in which plectonemes and bubbles exist simultaneously, the most common form of bubbles is the tip bubble (figure 3.12), where the flexibility of bubbles allows them to bend sharply with little to no energetic penalty. However, while they are the dominant form, tip-bubbles are not the form of bubbles observed, with the formation of a clear secondary peak at around 100bp from the plectoneme apex. This is seen specifically in low-tension $(0, 0.3pN)$ systems. The lower tension of these systems means that the work done in nucleating the plectoneme is lower, and therefore the need to minimise the plectoneme size is smaller. As a result, the need for bubbles to absorb strong bending at the plectoneme tip (therefore reducing its overall size) is smaller, allowing them to nucleate up or down stream of the tip. The consistency in the location of this secondary bubbles implies a strong link with sequence [1], suggesting that the energy required to form a bubble at this specific location is lower.

3.7.2 Toroid and plectoneme migration

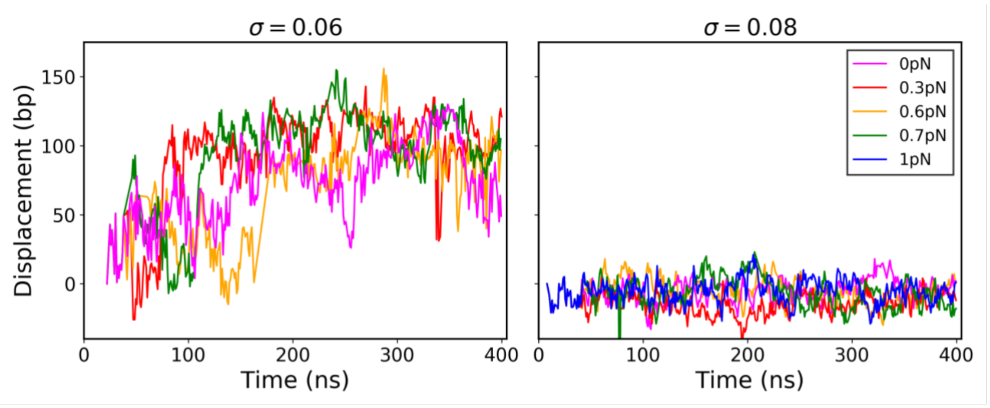


Figure 3.13: Displacement of plectoneme location for the first 400ns of simulations. (**Left**) Simulations at $\sigma = 0.06$ show significant mobility, owing to the toroidal nature of their writhed structures. (**Right**) Simulations as $\sigma = 0.08$ show little to no mobility, owing to the size of the plectoneme forcing it to localise in the centre of the strand.

Plectoneme hopping events (in which a plectoneme nucleates in one location, then slowly vanishes and is replaced by a plectoneme in a new location) as well as migration events (in which the location of the plectoneme shifts over short distances) have been observed both experimentally [176], and predicted using two-phase models [177]. It stands to reason that, with the short sequences examined in this text, instances of plectoneme hopping will be non-existent, but small-scale diffusion events may very well be present.

Figure 3.13 shows the movement of plectonemes and toroids at two different positive supercoiling densities, free from the influence of bubbles. It is clear from this figure that writhed structures are able to move a significant distance through the system, however this mobility vanishes once the system moves from a toroidal to a plectonemic state, with the size of the plectoneme resulting in its location being fixed near the centre of the sequence.

Figure 3.14 shows the migration of a single toroid in a system with $\sigma = 0.06$. Here the movement is clear - the toroid forms close to the mobile end of the strand, then it moves to the centre, by way of toroid migration, over a time period of around $100ns$. This observation matches with those seen in polymer models [177], where plectonemes are able to propagate over short distances without significant disruption of their structure.

3.7.2.1 Plectoneme pinning by bubbles

While observations of plectoneme hopping and migration are well understood at large length scales, what is less well understood is the influence of bubbles, and whether these bubbles influence both the mobility and size of plectonemes. Previous observations from OxDNA

---

[1] This assumes that the locations of the plectonemes is constant throughout all of the analysed simulations, which, in this case, it is.

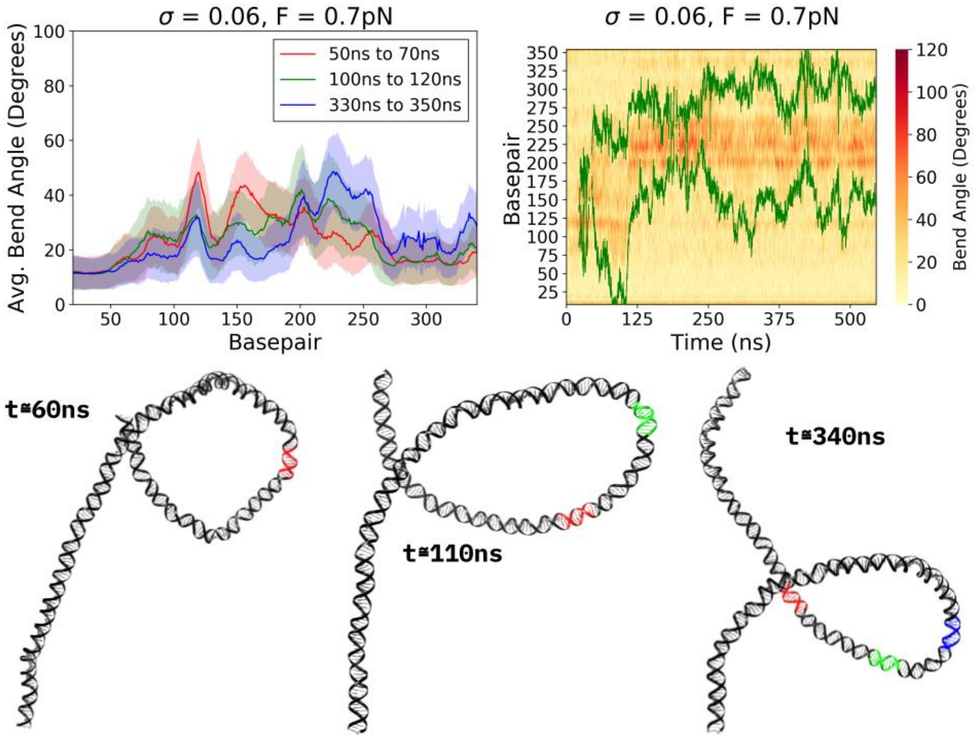


Figure 3.14: Bending and migration in a single simulation. (**Top left**) Average bending angles over three $20ns$ windows, with standard deviations represented by shaded regions. (**Top right**) Kymograph of plectoneme start and end points (green lines), with the background heatmap representing local bending angles. (**Bottom**) Representative structures taken from the same windows as the bending profile, with regions of strongest bending highlighted in their respective colours.

[111] indicate that the co-localisation of bubbles and plectonemes can suppress plectoneme movement almost entirely, with the tip-bubble causing mean squared distances to drop close to zero.

Examining the detailed effects of bubbles on the locations of plectonemes in our simulations is more challenging than one might imagine - there are only one or two examples of systems with both tip bubbles and mobile plectonemes. Instead, it is more enlightening to once again return to systems with $\sigma = -0.06$, where the presence of a toroid allows for significant mobility as well as direct comparison to the mobile toroids seen in positive supercoiling (figure 3.13).

The presence of bubbles in systems at $\sigma = -0.06$ has a profound effect on the mobility of the toroids, as evidenced by in figure 3.15. Here, the two pinned systems with $F = 0pN, 0.6pN$ have significantly more stable positions when compared with their positively supercoiled counterparts, with the standard deviations of plectoneme position dropping from 23.47bp and 18.27bp to 14.4bp and 6.7bp for 0 and $0.6pN$ respectively. For these two systems, the effect of bubble formation is made clear by the examination of kymographs in figure 3.15. For the $F = 0pN$ system, there are clear fluctuations in the start and end positions of the toroid, likely a result of the large flexible bubble allowing for massive variations in toroid size without any significant change in bending energy. Despite these massive variations, the mid-point of the toroid is fixed at the bubble, with the bubble consistently acting as the focus point for strong bending. The $F = 0.6pN$ system has even clearer pinning within the bubble,

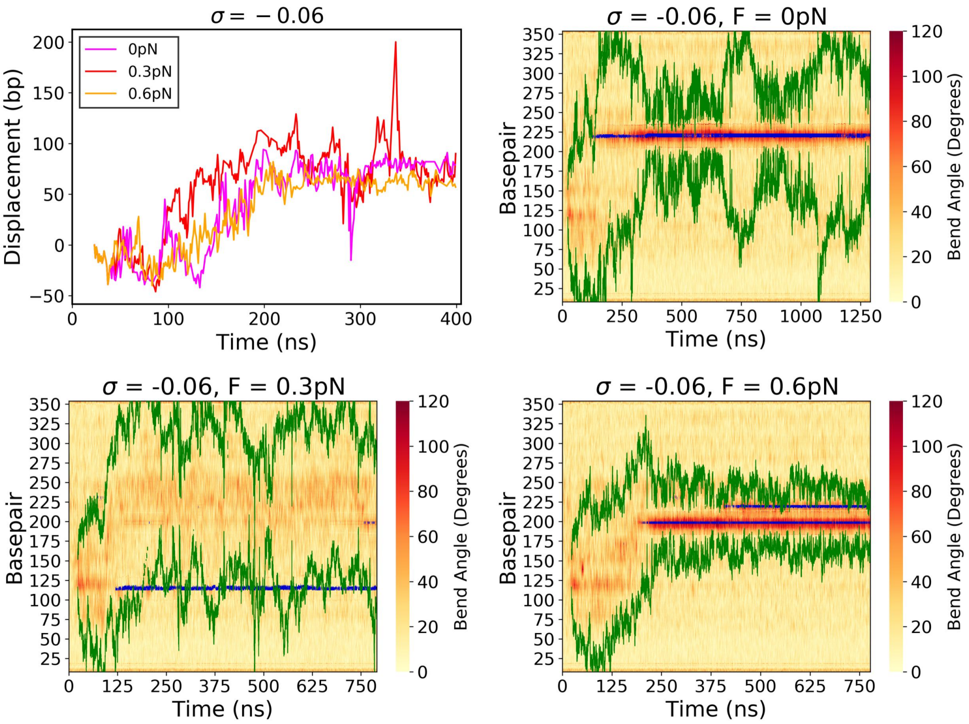


Figure 3.15: Toroid migration and pinning in a series of simulations, all at $\sigma = -0.06$. Once bubbles form, toroid location remains stable despite significant changes in size, the exception to this is the $F = 0.3pN$ system, where the bubble forms outside of the toroid. Kymographs here include bubble location, signified by blue points.

with the fluctuations seen at $0pN$ no longer present due to the significant tensile force. Here the initial formation of the bubble coincides perfectly with the migration of the toroid - almost immediately after this bubble forms the toroid re-positions to its location. Once this migration is complete, the start and end positions of the now significantly smaller toroid remain remarkable stable, even after the formation of a new bubble a few bp upstream of the original.

The most interesting of these systems is $F = 0.3pN$, in which a bubble forms at around base pair 120, and, rather than the apex of the plectoneme moving and stabilising at this position, the plectoneme moves to the centre of the sequence, with the bubble aligning with its crossing point. The consequence of this misalignment is a significantly more dynamic plectoneme, with the standard deviation of position in this case being comparable to its positively supercoiled counterpart (16.1bp for positive vs 20.5bp for negative). The alignment of the bubble with the crossing point presents a potentially interesting state - the crossing point of the plectoneme is likely the area with the second-most amount of bending in the system (behind the plectoneme tip), and may present an area in which the energy required to nucleate a bubble is reduced. It is however clear from the kymograph that the existence of this bubble does not stabilise the position of the crossing point, implying that this alignment may be purely coincidental.

3.8 Anomalous systems

While the simulations above are able to broadly recreate the experimental hat-curve, as well as the structural features seen in OxDNA studies, they are not without issue. A prime example

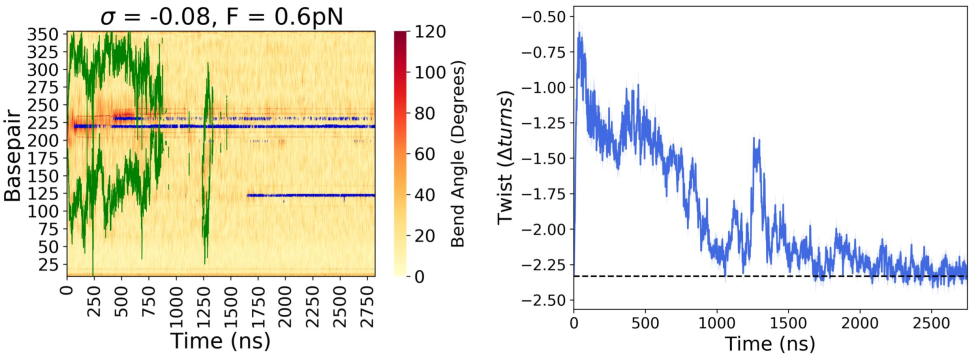


Figure 3.16: Extension and twist statistics in an anomalous system. The system has an unusually large number of base pairs in bubbles (blue lines), coupled with the disappearance of plectonemic loops. The dashed line represents the value of twist at full extension.

of these issues is the simulation of $\sigma = 0.08, F = 0.6pN$ which, as evidenced in figure 3.1, has significantly higher extension than expected. This higher extension aligns, as expected, with uncharacteristically high values for twist, with writhe moving to zero. The unusual behaviour of this system can be attributed to the unusually high number of base pairs in bubbles, as seen in figure 3.3, with these bubbles acting as focal points for large amounts of twist. This excess twist absorption is pictured in figure 3.16. The kymograph displays the nucleation of multiple bubbles, and the disappearance of any plectonemic structure. Noting that a sigma value of -0.08 translates to the removal of approximately 2.3 turns (see figure 3.16), it is clear that by $t \approx 1750ns$ all torsion had been partitioned in to twist, with its value returning to, and remaining at, its initial value. Precisely why this system behaves in this manner is difficult to quantify. It is likely that this full extended state with large numbers of denatured base pairs corresponds to a local minima in energy, with the formation of two bubbles early on in the simulation (at $t \approx 450ns$) moving the system in to this state. It is clear from earlier figures, as well as experimental results [63], that the forces of 0.6 and $0.7pN$ represent a critical regime at these supercoiling densities, with the systems balanced between bubble and plectoneme dominance. This suggests an energy landscape that is far rougher than their fully extended or fully plectonemic counterparts, with the single conformation representing the lowest possible free energy state difficult to distinguish from local minima.

The case of $\sigma = -0.1, F = 0.6pN$ is almost indistinguishable from the one above, with an unusually high number of base pairs in bubbles resulting in the partition of torque entirely in to twist. It is tempting here to read too much in to the fact that both anomalous systems are at $F = 0.6pN$, perhaps trying to relate this force to some kind of strange critical behaviour. The likely case is that these anomalous systems are merely a consequence of our extremely limited exploration of the phase space; having only a single replica for each combination of force and supercoiling density. We believe that if more replicas were performed they would likely display the expected extension behaviour, with these particular systems falling in to states with both bubbles and toroids.

3.9 Sequence dependence in plectoneme and bubble formation

Exploration of sequence dependence in supercoiled linear DNA has thus far been successful in understanding the formation of a single type of structure, for example bubbles [98] or plectonemes [100]. Where these models fail is in the interfaces between regimes, where, for instance, bubbles and plectonemes are able to interact in meaningful ways, perhaps influencing one another's formation. Work in OxDNA [111] has been able to explore sequence dependence, but any conclusions made are tempered by the lack of vital sequence-dependent structural features, such as DNA curvature. To this end, the simulations presented here can offer a one-of-a-kind insight in to sequence dependence, albeit in very short sequences.

3.9.1 Bubble formation

The simplest sequence dependence in these simulations is seen in the formation of bubbles - the locations at which they form should be largely unaffected by the short nature of the sequence, as they form over the space of a few base-pairs, instead of tens or hundreds. It is worth noting however that the landscape of bubble formation probabilities will be significantly noisier than those of longer sequences, such as those seen in [98], due to the effects of small fluctuations in sequence. For example, a run of four consecutive AT base pairs would be insignificant in long (<1000bp) sequences, but in a sequence of the length examined here, this region would represent a significant probability of bubble formation. These fluctuations will likely make any real prediction of the values for formation probability near impossible, but regions of interest should still be identifiable.

In the work of Benham *et al*, the probabilities of bubbles formation are significantly damped in small systems due to the implementation of a significant nucleation energy ($G_{bub} = 10.4 kcal/Mol$, H in equations 1.20 and 1.21) [97]. For short sequences such as the one examined here (which is much shorter than the minimum recommended length of this program), this nucleation energy suppresses bubble formation probabilities at all but the highest supercoiling densities. Fortunately, the matter of interest here is the position and relative values of bubble formation, rather than the absolute values of nucleation probability, and as such, while probability values should be ignored, the positions of peaks and their relative values should still be predictors of bubble location.

The ability of the SIDD algorithm to predict the formation of bubbles is evident even in this short sequence, with this ability evidenced in figure 3.17. Here, the locations of bubble formation align with the locations of predicted peaks, but with massive variation in real probabilities. These shifts in location could be a simple result of limited sampling - the sampling performed in simulations is likely only a small part of the wider landscape, a result of only having a single replica for each combination of force and supercoiling density. A more likely, and more interesting, explanation is the influence of plectonemes. The SIDD algorithm, by its very nature, assumes that the DNA strand is completely extended, with bubble formation unaffected by curvature, and with bubbles the only vector outside the twisting of B-DNA by which torsion can be absorbed. In simulations however we see clear interaction between plectonemes

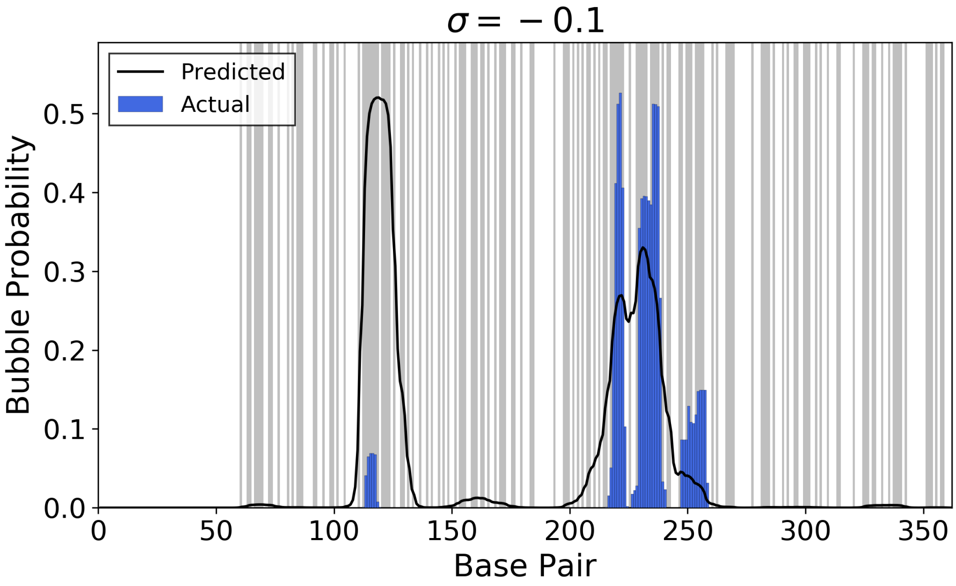


Figure 3.17: Predicted and observed bubble formation probabilities for $\sigma = -0.1$. Observations from simulations align with predicted locations, but with massively varying values of probability. Background of the plot is a visualisation of sequence, with grey lines representing AT base pairs.

and bubbles, and it is likely this interaction that causes the shift in the bubble landscape. The high curvature at the tips of plectonemes may act to lower the nucleation energy of the bubbles, resulting in the promotion of bubble forming regions that are in and around the tip of the plectoneme. This explains the shift in probabilities seen in figure 3.17; we have established previously that, especially at high supercoiling densities, plectonemes shift to the centre of the sequence - the exact shift seen going from the predicted to real bubble probabilities. The interplay between bubbles and plectonemes is further highlighted by the differences in the predicted and real sizes of bubbles - the SIDD algorithm predicts the formation of bubbles much larger than those seen - this is the result of plectonemes acting to absorb a large amount of the excess torque, especially in systems with lower force values.

Close examination of figure 3.17 also gives insight in to the exact sequences in which bubbles form. The locations at which bubbles form align precisely with runs of consecutive AT base pairs, rather than AT-rich regions, suggesting that small, localised variations in sequence can give rise to large shifts in formation probabilities. Additionally, the presence of multiple small bubbles rather than a few large ones seen in figure 3.17 (and in specific systems such as $\sigma = -0.06, F = 0.6pN$ seen in figure 3.15) suggests that, at least in the implicit solvent used in these simulations, the high nucleation energy used in SIDD is not representative. Perhaps a lower nucleation energy paired with a greater difference in the denaturation energies of AT vs GC may give more accurate predictions in this particular case.

3.9.2 Plectoneme formation

Predictions of plectoneme location from the work of Dekker *et al* [100] will be difficult to directly compare to the simulations seen here. The sequence simulated is simply too short to exist within the regime predicted by this code. However, this does not mean that sequence-

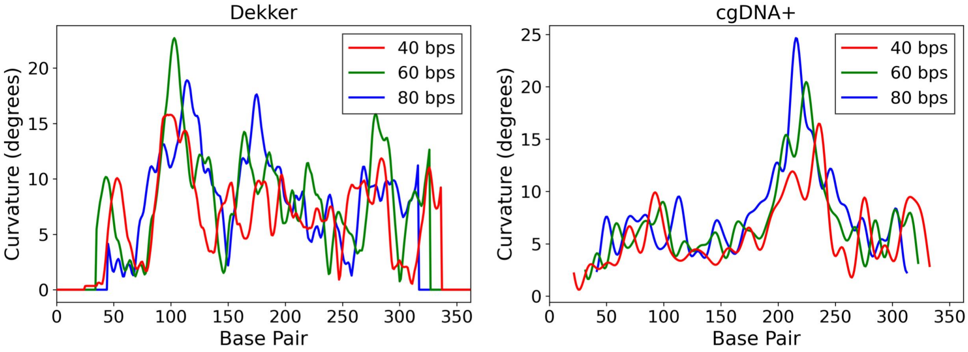


Figure 3.18: Predicted ground path curvature for the sequence simulated in this chapter, with curvature measured over a series of base pair steps. Note the distinct movement of peaks in the Dekker case (left), as well as the drastic changes in curvature measured in cgDNA+ (right). Fourier smoothing was applied to measurements from cgDNA in order to eliminate high-frequency noise arising from the double-helical structure.

dependent behaviour is completely out of reach - the ground path curvature of the sequence can still be calculated, with this being the primary predictor of plectoneme location according to [100].

For the remainder of this text, two primary methods will be used to predict ground path curvature. First is predictions taken directly from the code listed in [100], from which values of curvature can be directly extracted. The second is cgDNA+[178, 179], which uses a rigid body representation of base pairs to trace the DNA ground path, this ground path can then be analysed with SerraLINE [59] to find values for bending, in the same manner as the kymographs seen in multiple figures listed in this text (for example figure 3.16). It should be noted here that there is no reason why measurements made by these two methods should agree; not only do they use different methods of tracing the ground path, they also use different dinucleotide parameters to trace this path.

The key challenge in comparing DNA ground path with simulation data is choosing the window over which to measure ground path curvature. The distance between the base pairs over which curvature is measured not only impacts the magnitude of the measured curvature, but also the overall bending profile. The effect of changing this measurement window is shown in figure 3.18. Here, particularly in the 'Dekker' predictions, the locations of predicted peaks shift significantly as the window over which curvature is measured changes. In order to address this issue, the windows over which plectoneme tips form need to be measured, however this is much easier said than done. We could, for example, use the projection method described in section 2.9.2, finding the crossing point representing the tip of plectonemes/toroids, then use this to infer the average size of a plectoneme tip. In reality this method is simply not practical. First, any simulations in which a tip-bubble is present would have to be eliminated, leaving only a small number of viable systems. Second, the assumption that bending is evenly distributed over an entire plectoneme tip is false, take for example the toroid seen in figure 3.14, measurements of plectoneme size have this structure at over 100 base pairs in the $330-350ns$ case, but strong bending is clearly taking place over a much smaller window (between 60 and 70 base pairs) than this. Unfortunately, there is no concrete solution to this issue, and any choice made for curvature measurement is arbitrary. For the remainder of this text a window size of $60bps$ will be used as it seems to best represent the bending profiles (as seen in figure 3.14) seen in simulations, and is close to the $70bps$ suggested as the optimum size by [100].

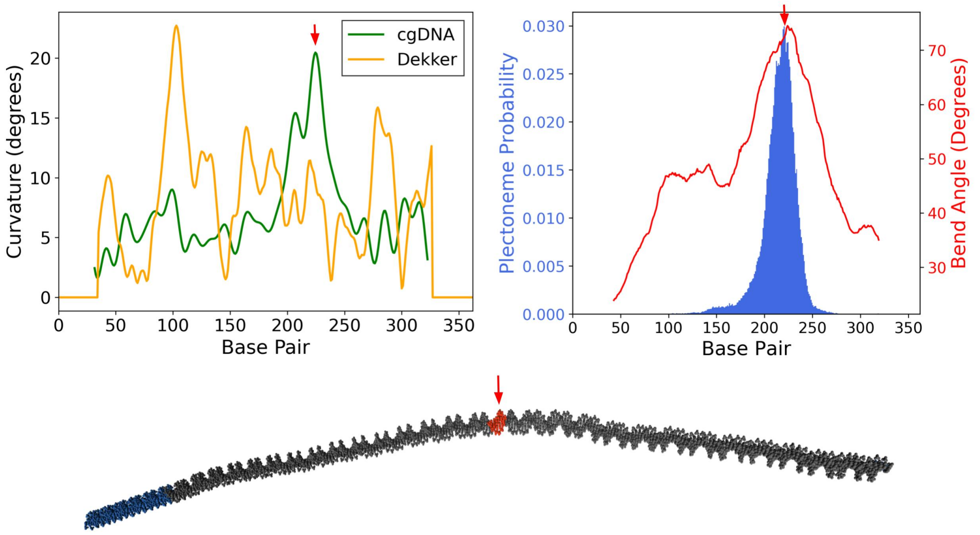


Figure 3.19: Predicted and real DNA curvature for all bubble-free simulations. In all cases the location of strongest bending observed in simulation, that is the region of base pairs 220-225, is indicated by a red arrow. Predictions of curvature (**left**) indicate multiple possible regions of high curvature, with the most significant region predicted by cgDNA also being the location of plectoneme formation shown in simulation (**right**). Also included is a visualisation of the cgDNA ground path (**bottom**) with the region of highest curvature highlighted in red.

As seen earlier in this chapter, bubbles can act as focal points for strong bending. As such any analysis of the effects of DNA curvature on plectoneme formation should only use simulations in which bubbles are not present. Such analysis is seen in figure 3.19. Here the dominance of centralised plectonemes is complemented by the regions of localised curvature. Specifically, the vast majority of plectoneme formation is centred around the region between base pairs 215 and 225, with the modal value of plectoneme position being base pair 220. This aligns well with the region of high curvature predicted by cgDNA, with the modal curvature seen in simulations being within 5bp of the peak of highest curvature. The 'Dekker' predictions are much more poorly aligned with observations from simulation, with the region of high predicted curvature at bp 100 failing to align with any significant increases in curvature seen in simulation.

It should be noted here that, despite appearances, the predictions of the 'Dekker' method are far from a failure. Take for example the dynamics of the $\sigma = \pm 0.06$ systems (see figures A.3, A.9 in the appendix), in both of these systems toroids are seen to nucleate in the region between base pairs 100 and 150 before migrating to the centre of the sequence. This region of nucleation aligns with the predicted region of high curvature from the 'Dekker' method. This could be an indication of the tendency for plectonemes and toroids to migrate to the centre of the sequence in order to maximise their size, a behaviour which is over-exaggerated by the low length of the simulated sequences.

The contrasting predictions of the two models means that making any concrete conclusions regarding sequence-dependence in this case is impossible. On the right of figure 3.19 we see that, even in the absence of bubbles, plectonemes are moving to the centre of the sequence.

Whether this is due to the high localised curvature in this region, as predicted by cgDNA, or simply a consequence of the nucleated structures maximising their size, remains an open question, the answers to which can only be found by examining additional sequences.

3.10 Summary

With a few notable exceptions, and despite the relatively short length of the simulated DNA duplex, we are able to recreate the famous DNA ‘hat-curve’ under a range of supercoiling densities and tensile forces. Simulations show the clear asymmetry between positive and negative supercoiling at higher tensions, and insights in to the detailed dynamics of the DNA allow these phenomena to be explained through the contrasting numbers and behaviours of bubbles. Additionally, these simulations confirm the previously observed co-localisation of plectonemes and bubbles, in the form of so called ‘tip-bubbles’, with the flexibility of the denatured structures allowing the formation sharp bends with little to no energetic penalty. The existence of tip-bubbles in positively supercoiled systems is also observed, with these acting solely as points over which strong bending is seen, while failing to absorb a significant amount of twist.

These simulations also allow for speculation regarding the role of sequence in the formation of these structures. In systems of this size, it is clear that the interactions between plectonemes and bubbles are the driving force behind dynamics. While the effects of sequence of each of them individually are clearly displayed, the tendency of plectonemic structures to localise at the centre of the strand is the dominant force. This localisation not only fixes the locations of plectonemes (as their sizes near that of the entire duplex), but also acts to promote the formation of bubbles, with the strong bending of the plectoneme tip reducing their nucleation energy.

The true importance of sequence is difficult, if not impossible, to explore with a single sequence, and any conclusions made must be tempered by an extreme lack of diverse sequence information. Fortunately, the method described in chapter 2 of this text can be trivially applied to any sequence, and as such the information in this chapter can be used to design and build duplexes with the specific intention of understanding sequence dependence.

Chapter 4

Sequence Dependence of DNA Structure Formation

4.1 Introduction

It is clear from chapter 3 that sequence plays a significant role in the formation of both bubbles and plectonemes in supercoiled systems. The relative ease by which AT base pairs are denatured in comparison to their GC counterparts is vital to the formation of bubbles and, as a result, runs of consecutive AT base pairs create ideal regions for bubble formation. The role of sequence in plectoneme formation is much harder to define with only a single sequence, and any sequence-dependent effects prove hard to decouple from the formation of highly flexible bubbles.

The sequence studied in chapter 3 was, by mere coincidence, excellent at accommodating both large plectonemes and plectoneme tip-bubbles, with a region of high bubble formation probability, defined by several runs of consecutive AT base pairs, close to its centre. This leaves the following questions unanswered: whether these tip-bubbles are a universal property of supercoiled systems, or whether they are merely a consequence of that specific sequence, and whether plectonemes in strongly supercoiled systems will always migrate to its centre, or if their positions are strongly influenced by sequence-dependent factors. Additionally, questions regarding the effectiveness of predictive models such as those proposed by Benham and colleagues [98] and Dekker and colleagues [100] cannot be answered by a single sequence, especially one as short as that studied here. It is therefore a natural next step to create bespoke sequences that allow for further exploration into the role of sequence dependence.

4.2 Failed attempt - AT runs

4.2.1 Sequences

As an initial attempt, a pair of 300bp sequences were created, both randomly generated with an initial AT percentage of 40%. These sequences were then modified, added to the first was a run of 15 AT base pairs, placed in the centre, with the aim of creating a close to idealised system in which bubbles and plectonemes naturally colocalise. This will be known as the 'centred'

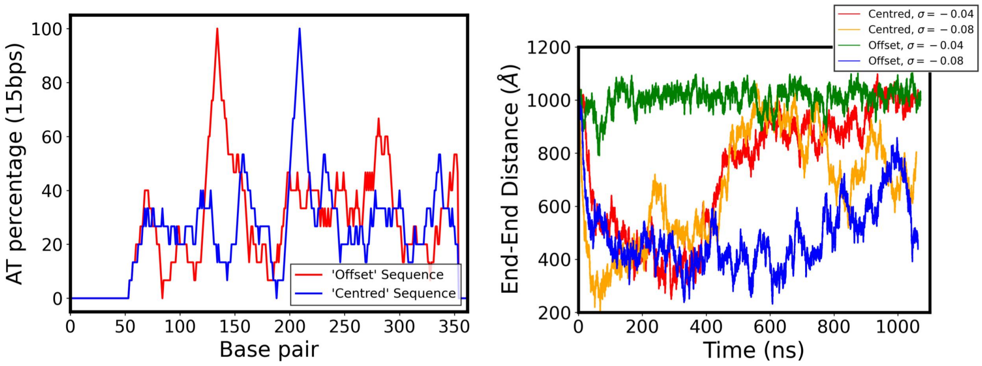


Figure 4.1: (**left**) Rolling AT percentage for a window size of 15 base pairs for both the 'offset' and 'centred' sequences. (**right**). End-to-end distances for all centred and offset simulations. Centred systems show clear increases in end-to-end distance as time goes on, indicating the disappearance of the toroids/plectoneme formed in both cases. Offset sequences show clear contrast between supercoiling densities, indicating differences in twist/writhe distribution.

sequence. The second also had a run of 15 AT base pairs added to it, this time placed a third of the way along the sequence. The aim of this sequence was to create a frustrated system in which plectoneme formation at high levels of supercoiling will occur in a different location to bubble formation. This will be referred to as the 'offset' sequence. Both of these sequences are pictured on the left of figure 4.1, where the locations of both 15bp AT runs are seen clearly.

The intuition behind both of these sequences is clear - add a segment into the sequence that acts as a clear promoter for bubble formation, suppressing small fluctuations in sequence. A clearly defined bubble nucleation location should allow for the separation of sequence-dependent and generic effects. For example, if the nucleation region is placed away from the centre of the sequence, is it still effective at promoting bubble formation, and if so, does this bubble still act to pin plectonemes?

The range of supecoiling densities and forces studied with these sequences must be much more limited than in the previous chapter due to the extremely large amount of time that any individual simulation takes. A pair of supercoiling densities were chosen, those being $\sigma = -0.04, -0.08$. A constant a force of $0.3pN$ was used in all cases. $\sigma = -0.04$ was chosen to represent systems in which toroid formation should occur consistently, with these toroids being small enough that they do not need to move to the centre of the duplex. $\sigma = -0.08$ was chosen to represent systems in which plectoneme formation is clear but tip-bubbles are not absolutely necessary to absorb all supercoiling. The force of $0.3pN$ was chosen to ensure the formation of toroids and plectonemes, but also facilitate possible bubble formation.

4.2.2 Problems with these sequences

While in principle these sequences offer a simple method for studying sequence dependence, the reality of their dynamics are far from ideal. The issue with these sequences lies in the ease with which bubbles form in the inserted 15 bp AT runs, acting, in some cases, to completely suppress the formation of plectonemes. The effects of this plectonemic suppression are strongly hinted at by the right of figure 4.1, where both centred systems show notable increases in end-to-end distance towards the end of their simulations. These increases suggest a collapse of the plectonemic/toroidal structures. The behaviour of the offset sequences is slightly more predictable, with the $\sigma = -0.08$ system showing clear signs of plectoneme formation, while the $\sigma = -0.04$ appears to completely lack the formation of any writhed structure, a clear hint towards the dominance of bubble formation.

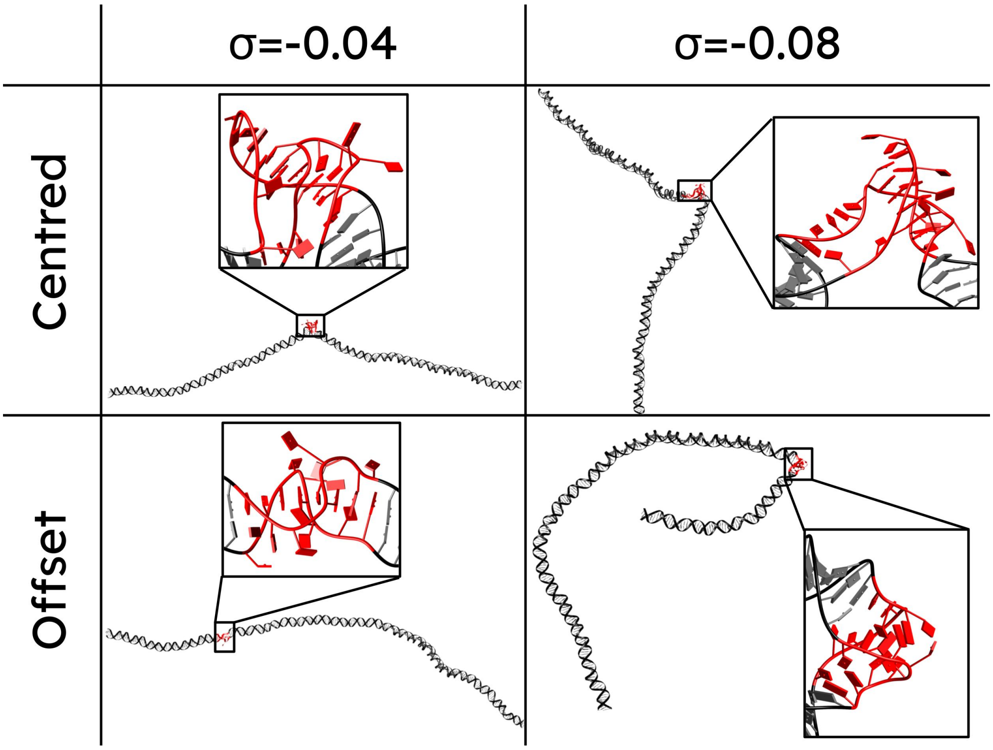


Figure 4.2: Representative structures for all Centered and Offset sequences. Bubbles (in red) are large and show significant disruption of backbone structure. Larger than expected end-to-end distances align with a lack of distinct plectoneme formation in any of the systems, with the $\sigma = -0.08$ offset simulation being the only one to show any indication of writhed structure formation.

The presence of the long run of AT base pairs in both sequences results in the formation of large bubbles in all simulations. Both simulations at $\sigma = -0.04$ display almost total extrusion of writhed structures, with only small amounts of bending seen in the centred case, and none seen in the offset case. This indicates that, with the presence of this long run of AT base pairs, bubbles are universally favored over plectonemes, even at this low level of tensile force. The case of $\sigma = -0.08$, while showing some writhed structures, shows much more extrusion than was seen in the randomised sequence studied in the previous section. The near-toroid seen in the offset simulation forms, as expected, with the bubble at its apex, suggesting that the presence of such a large bubble completely nullifies the need to form a plectoneme at the centre of the duplex, with the majority of supercoiling absorbed by the large denatured region.

Unfortunately, due to the universal bubble formation and lack of any strongly writhed structures, these simulations do not function as good measures of the sequence dependence of plectoneme formation. The fact that such strong bubble formation is seen at such a low force means that any meaningful conclusions regarding the influence of curvature is completely lost, with all bending focussed in the massive regions of base pair disruption.

4.3 Semi-randomised sequences

It is clear from the above simulations that the manipulation of sequence is far from trivial, and that, in order to properly understand sequence-dependence, more consideration must be given to sequence design. The design of sequences at the length scales simulated here, without the kind of brute forcing seen in the ones above, proves to be highly challenging, with the relatively low length of the duplex leaving it extremely vulnerable to localised fluctuations - any small run (4 or 5 consecutive base pairs) of AT base pairs can result in the formation of a bubble in a region with an otherwise low AT percentage. While this can be somewhat negated through manual modification of the sequence (i.e. the removal of any long AT runs), excess modification leads to a system that is highly artificial, resulting in sequence dependence that is uninformative.

4.3.1 Sequence design

As a compromise between randomised and completely artificial sequences, a method of 'block averaging' was employed, in which the sequence is generated in blocks, with each of these blocks having a different AT percentage. However, this presents issues regarding the low sequence length - if a block is too small then any 'randomly generated' sequence with a fixed AT percentage will be far from random. It is important therefore to maximise the block sizes. This however leaves open the possibility of long AT runs in otherwise low AT percentage blocks.

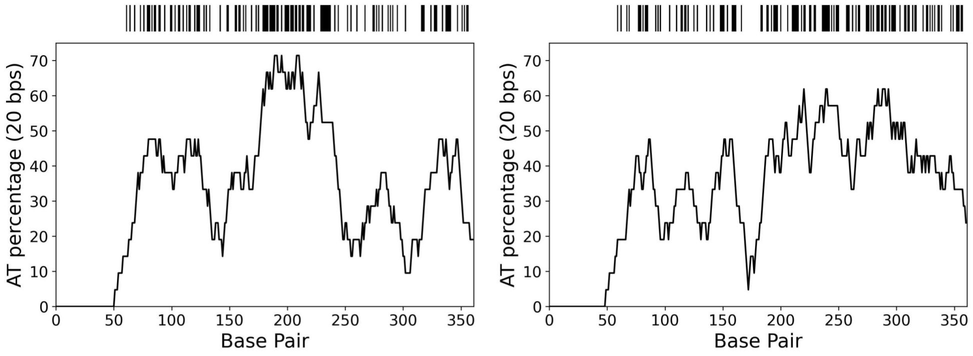


Figure 4.3: The two block averaged sequences. The centred sequence (**left**) has a clear peak in AT percentage between base pairs 150 and 250 with the aim of promoting plectoneme and bubble formation in this region. The offset sequence (**right**) has an increased AT percentage between base pairs 250 and 360, with the aim of promoting plectoneme and bubble formation in this region. Barcodes above plots give an indication of sequence specifics, with black lines representing AT base pairs.

In the same manner as the failed simulations above, two sequences were simulated - one 'centred' and another 'offset'. The 'centred' sequence is designed to promote bubble and plectoneme formation at its centre, with a region of higher AT percentage at the centre of the duplex. Specifically, this sequence was generated in three 100bp blocks: the first and last having an AT percentage of $\sim 40\%$, and the central block having an AT percentage of $\sim 60\%$. The second 'Offset' sequence is designed to promote plectoneme and bubble formation near the fixed end of the duplex and was generated in two 150bp blocks: the first with an AT percentage of $\sim 40\%$ and the second having an AT percentage of $\sim 60\%$. Both of these sequences are visualised in figure 4.3.

In both cases a total of four simulations will be performed, all at a force of $0.3pN$ in order to promote the formation of both bubbles and plectonemes. The supercoiling densities will be $\sigma = +0.1, +0.06, -0.06, -0.1$. These densities were chosen for a few key reasons: first, they are a response to the lessons learned from the failed sequences above, specifically $\sigma = -0.1$ was chosen to in the hope of seeing plectoneme formation regardless of sequence; second, positive supercoiling was chosen to offer a clear contrast between the two systems, the key idea is that positively supercoiled systems should be largely unaffected by bubble formation, making the sequence dependence of their plectoneme formation clearer. As in the previous chapter, all simulations were run for between $500ns$ and $3\mu s$, depending on the convergence of the end-to-end distances.

4.3.2 Behaviour of the ‘centred’ sequence

In principle, the centred sequence should be the simpler of the two, with plectonemes and bubbles forming in the AT-rich region regardless of supercoiling density.

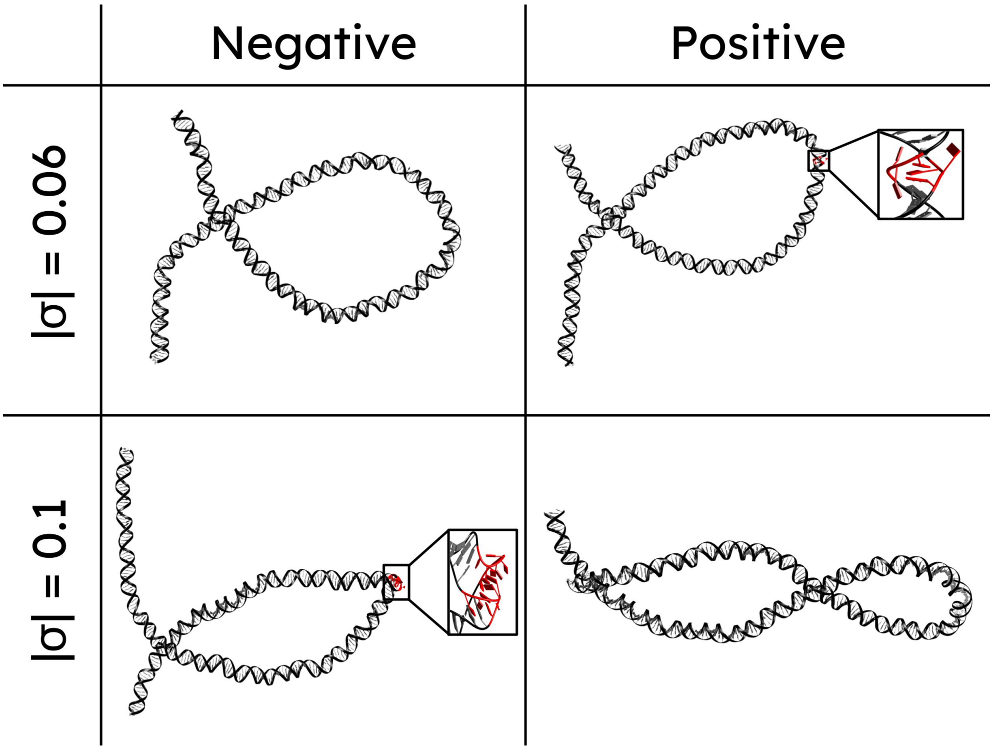


Figure 4.4: Representative Structures for all centred simulations. Negatively supercoiled systems show similar amounts of writhe, with the presence of a large bubble in the $\sigma = -0.1$ system accounting for the difference between the two. Positively supercoiled systems are visually similar to those seen in chapter 3.

The representative structures of figure 4.4 indicate that, fortunately, simulations of this sequence have significantly higher plectoneme stability than those seen in section 4.2, with consistent toroid and plectoneme formation seen at all supercoiling densities.

Similar to the structures seen in the previous chapter, simulations at the lower supercoiling density ($|\sigma| = 0.06$) show the same behaviour, with the formation of a large toroid and a lack of significant bubble formation (there is a bubble in the positively supercoiled system however its impact on system dynamics is minimal). Simulations at the higher density ($|\sigma| = 0.1$) show divergence in behaviour, with the formation of a large bubble in the negatively supercoiled case; the presence of this bubble results in visibly less writhe when compared to the positively supercoiled system.

4.3.2.1 Plectoneme/toroid formation in the centred sequence

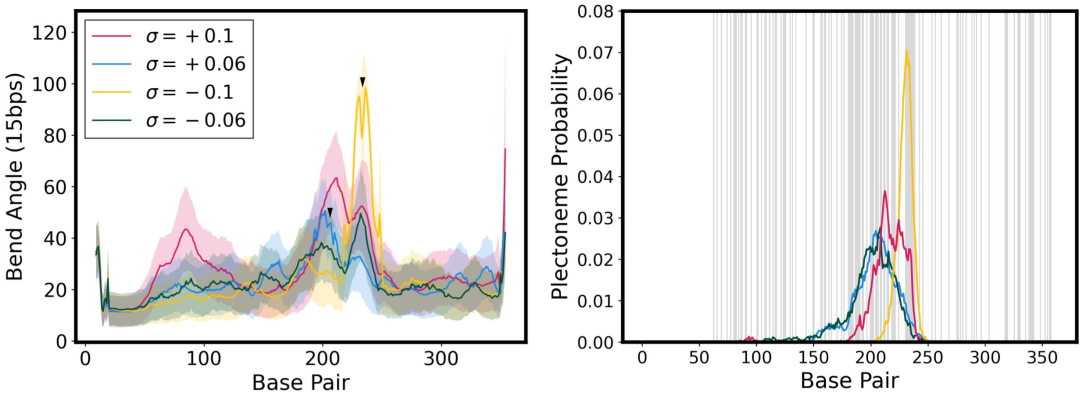


Figure 4.5: Plectoneme formation in the centred sequence indicated by both curvature and projected plectoneme position. (**left**) Bending angles for the last 400ns of all centred simulations, with angles measured over a window of 15 base pair steps. Locations of the bubbles highlighted in figure 4.4 are marked by black triangles. (**right**) Plectoneme positions indicated by the projection method described in section 2.9.2 with the shaded background indicating sequence, where grey lines are AT base pairs.

The combination AT-rich centre of this sequence and the low force ($0.3pN$) of all simulations results, as seen in figure 4.4, in the formation of plectonemes/toroids in all cases. This plectoneme formation should allow for a closer examination of sequence dependence, helping to further our understanding of the interplay between the formation of large loops and the nucleation of plectonemes in highly curved regions.

In all cases plectoneme formation occurs close to the centre of the sequence, as evidenced in figure 4.5. The strong bending of the bubble in the $\sigma = -0.1$ case shows the bubble acting as a clear focus point for the plectoneme tip, with bending in the base pairs surrounding the bubble significantly lower than in the bubble-less simulations. This figure also further reinforces the idea that the bubble seen in the $\sigma = +0.06$ system is inconsequential, with no significant increase in localised bending at its site.

Plectoneme formation is clearly occuring in the AT rich central region in all cases, with locations influenced by the formation of bubbles in the $\sigma = -0.1$ case. Within this region there seems to be some minor conflict when it comes to precise location - the region which should represent the highest bubble formation probability is the run of AT base pairs at around bp235, the location of the sharp bending in the $\sigma = -0.1$ case. This region is in conflict with the region between base pairs 200 and 230, which is directly in the centre of the duplex, and in which there are a large number of AT base pairs, but no significant runs. This creates minor conflict between the formation of bubbles and the maximising of plectoneme size.

The key takeaway from figure 4.4 is that the sequence appears to be behaving precisely as expected, with plectonmes/toroids forming in the AT-rich centre, and with the nucleation of bubbles also taking place in this region. The only simulation that may contradict this is the $\sigma = +0.1$ system, where a region of high bending is seen at the mobile end of the strand, centred around base pair 80. This bending however is unlikely to be the result of sequence-dependent effects, rather being the result of the dynamics of the system during plectoneme formation. In this particular case the plectoneme begins nucleating at the mobile end of the DNA, undergoing a migration similar to those seen in the previous sections. This formation at the mobile end of the DNA is not uncommon, and is more than likely the result of simple inertia - the closer to the mobile end of the DNA the plectoneme forms, the less DNA it has to drag along with it, resulting in lower inertia and therefore faster dynamics.

4.3.2.2 Predictions of bubble and plectoneme location

As in the previous section, direct comparison of theoretical predictions from the work of Benham and colleagues [98] and Dekker and colleagues [100] is difficult for a variety of reasons, all owing to the small sequence lengths examined in these simulations. However, we can once again expect to see agreement in the general pattern of structure formation - there is no reason that the bubble sites identified by SIDD should not also be site of bubble formation in these simulations, and there is also no reason why regions of the strand which are highly curved should not align with plectoneme formation.

In comparing prediction and simulations the simpler of the two structures are bubbles, with the areas identified with SIDD directly comparable to areas of bubble formation in simulation. The comparison seen in figure 4.6 shows some disagreement, with the single run of AT base pairs at around base pair 230 (see barcode on the left of figure 4.3) being the most prominent location for bubble formation in simulation, while the AT rich region between base pairs 175 and 210 is the most prominent region in prediction. This mismatch can be remedied somewhat by lowering the energy G_{bub} (H in equation 1.21) in the SIDD software - this is the sequence-independent part of the bubble nucleation energy. Lowering this number from the known value of $10.84 kcal/mol$ to a value of $5 kcal/mol$ [1] (see section 3.9.1) results in the promotion of the AT run at base pair 240. The mismatch is better explained by examining the free energies, shown at the bottom of figure 4.7. This figure shows the change in free energies as a function of σ, and shows the region around base pair 230 has the lowest free energy at lower supercoiling densities ($\sigma \leq -0.07$) and that the region between 175 and 210 only becomes more favorable as density increases.

It is worth noting that, due to the presence of writhe in all cases, the amount of twist absorbed by bubbles is much lower than the predictions of SIDD. If, for example, we have a system in which a total of two turns have been removed, but one of these is absorbed by writhe, then only a single extra turn is absorbed by either twisting of B-DNA or bubble formation. As a result, the predictions made at $\sigma \geq -0.07$, in which the region of bubble formation in simulation is the most prominent of the predicted regions, are likely more representative of the probabilities of bubble formation at $\sigma = -0.1$ in simulation. This point is further affirmed by the free energy heatmap, with the region in which bubbles are seen to form representing the lower free energy up until $\sigma = -0.08$. It is also worth re-stating here that these simulations sample

---

[1] It should be noted that this value has no physical basis, and is used here only demonstratively.

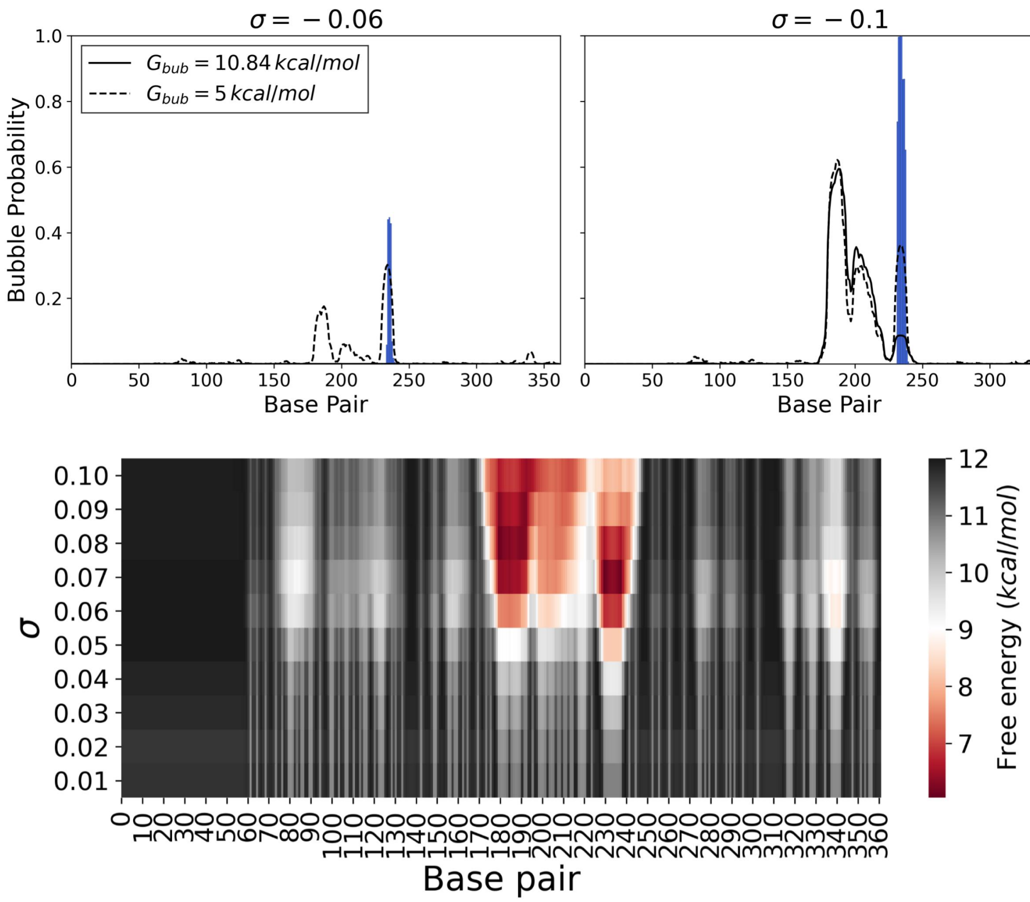


Figure 4.6: Predictions of bubble position from SIDD, with black lines representing predicted results, and blue bars representing results from simulations. (**top**) Bubble probabilities for two supercoiling densities with two different nucleation energies. (**bottom**) Relative free energies of bubble formation at increasingly negative supercoiling densities. Both plots highlight three important regions, between base pairs 175 and 190, 195 and 210, and 230 and 240. Simulations only show bubble formation in the last of these regions, between base pairs 230 and 240, with this region predicted to be more heavily favored at lower supercoiling denisites and in cases where the sequence-independent bubble nucleation energy G_{bub} is reduced.

only an extremely small subset of the total phase space of possible conformations, and so it is likely that, if more replicas of this system were to be simulated, the other regions identified by SIDD would see significant bubble formation.

Comparison of plectoneme locations to prediction proves the more complex and far less clear of the two states. Figure 4.7 shows the locations of plectonemes in all simulations overlaid with the DNA curvature predicted by the method described in [100] and by cgDNA [178, 179]. The key feature to note here is generally high curvature between base pairs 175 and 275 predicted by the 'Dekker' method, while cgDNA shows no significant curvature in any specific region. Plectoneme formation in simulation, with the exception of the $\sigma = -0.1$ case, takes place around the centre of the sequence, in the highly curved region predicted by the 'Dekker' method. Making any more detailed conclusions here is difficult, however the lack of any specific alignment between the predicted curvature and actual plectoneme locations once again hints towards a system which is dominated by the preference of plectonemes and toroids for moving to the centre of the sequence. The sequence studied in the previous chapter seemed

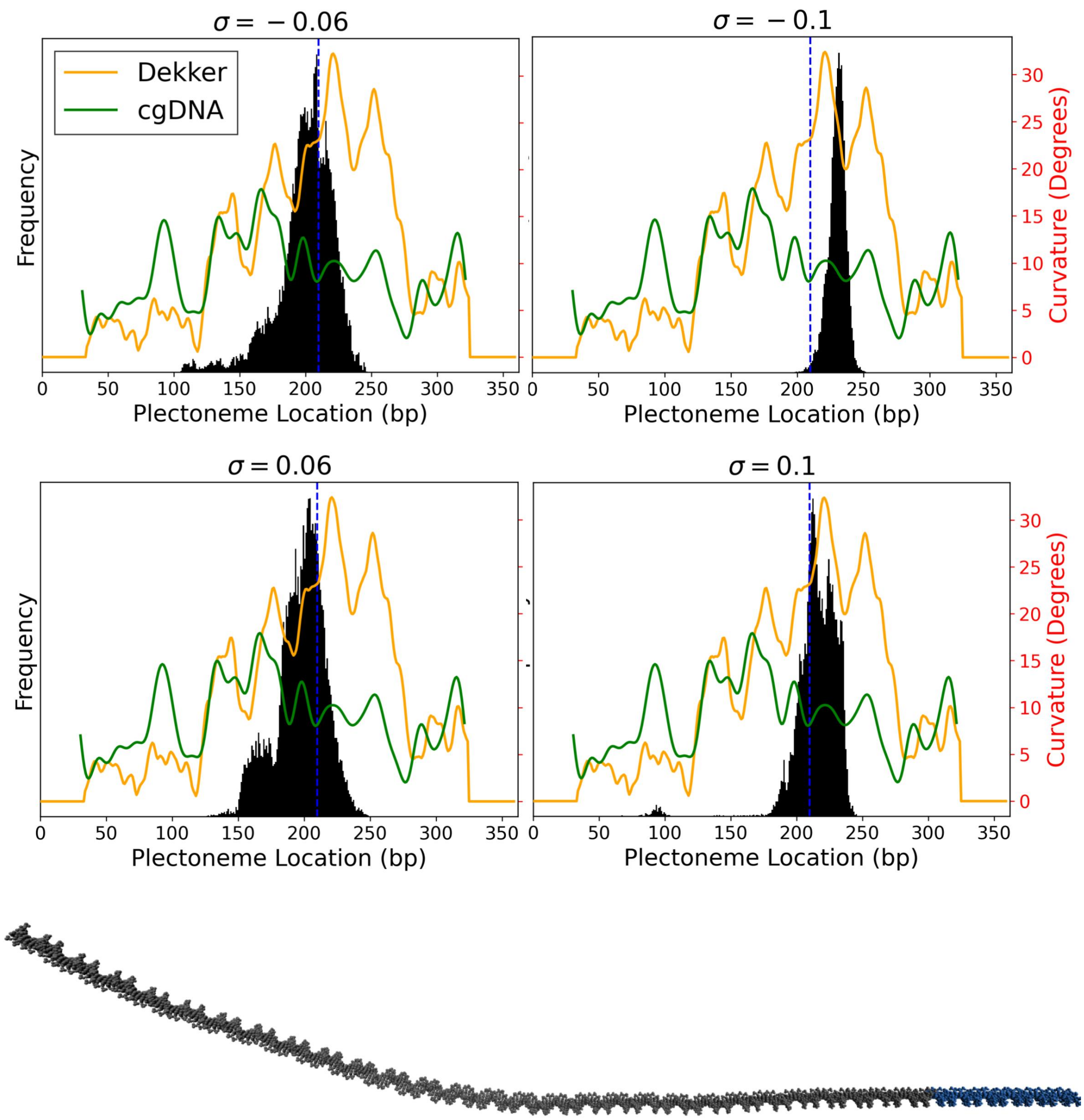


Figure 4.7: Predicted (green and orange lines) and observed (black histogram) results for plectoneme position across all centred simulations. Predictions from the 'Dekker' method show a large highly curved region in the centre of the sequence, between base pairs 175 and 275, while predictions from cgDNA show few signific curved regions, with the representative structure (bottom) showing generalised curvature across the entire sequence. The centre of the sequence is represented by a dashed blue line in each plot, while blue base pairs in the representative structure represent the dummy base pairs.

to be better predicted by cgDNA, while this sequence aligns more closely with the predictions of the 'Dekker' method. What these two systems have in common is movement to the centre suggesting that, in these sequences with no regions of high curvature/bubble formation near their ends, this effect is the main predictor of plectoneme location.

Before moving forward, it is worth highlighting the issues presented by the $\sigma = 0.1$ system in regard to finding actual plectoneme positions. The size of the plectonemes causes issues for the method of projection used when identifying them, as displayed in figure 4.8. Here, upon projections the structure has overlapped the end of the strand, causing the initial crossing point to be lost and resulting in the detected plectoneme start/end points moving to the second

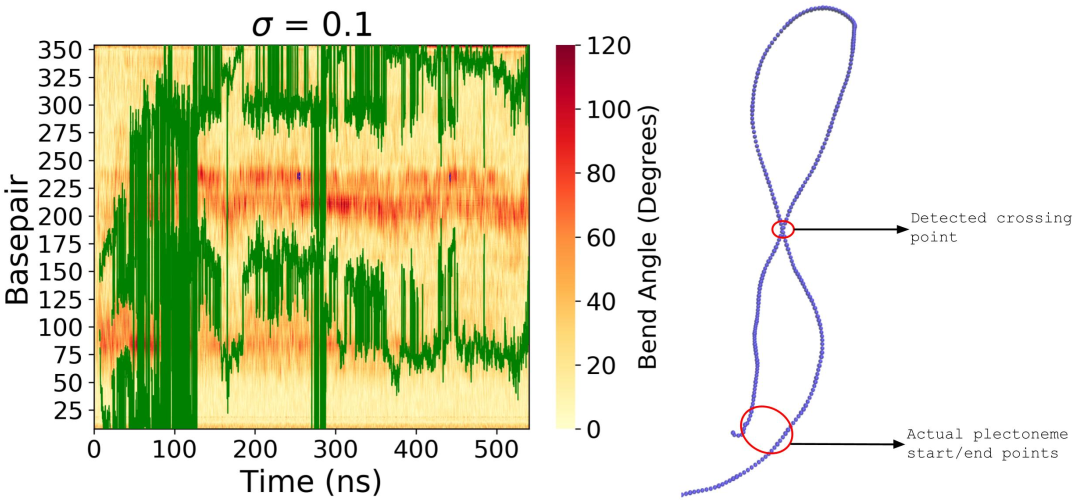


Figure 4.8: Issues with capturing plectoneme position in the $\sigma = +0.1$ case, projection of the structure on to a 2d plane causes the initial crossing point to become separated. As a result the second crossing point is detected as the start and end of the plectoneme.

crossing point. Fortunately, the analysis of curvature seen in figure 4.5 indicates that, in this case, the positions of plectonemes detected by the projection method aligns with the real values. This projection problem, while able to be overcome here, will present a significant issue going forward, with the 'Offset' sequence suffering severely as a result.

4.3.3 Behaviour of the 'offset' sequence

The 'offset' sequence is designed to display complex, sequence dependent behaviour, and as such it is expected to display much more chaotic behaviour. The consistent use of a tensile force of $0.3pN$ is designed to allow all states to form toroids/plectonemes, while also allowing for the nucleation of bubbles. Formation of both of these structures in this frustrated, complex sequence should shed light on the detailed interplay between the two, and the conditions under which they are not necessarily complementary.

The representative structures of figure 4.9 show strongly σ-dependent behaviour, as well as clear asymmetry between positive and negative supercoiling. The structure of the $|\sigma| = 0.06$ systems is largely as expected, and has parallels with those of the previous sequences, with the formation of large toroids and the nucleation of a tip-bubble in the negative case. The structures of the $|\sigma| = 0.1$ systems are more complex than any seen in previous cases, with both showing unexpected conformations. Starting with $\sigma = 0.1$, a plectoneme with two crossing points is seen, mirroring the structures seen in both previous systems at this density, however, in this case the plectoneme appears much closer to the mobile end of the strand, as indicated by the blue base pairs in figure 4.9. Formation of the plectoneme near the mobile end is accompanied by the formation of a second strongly bent region close to the fixed end, near the first crossing point. The structure of the $\sigma = -0.1$ system appears to contrast strongly with its positive counterpart, with writhe expressed in a much more chaotic manner, and the formation of a pair of bubbles at opposite ends of the strand. However, similarly to the $\sigma = 0.1$ system, this system also appears to have a pair of strongly bent regions, the alignment of which will require further analysis.

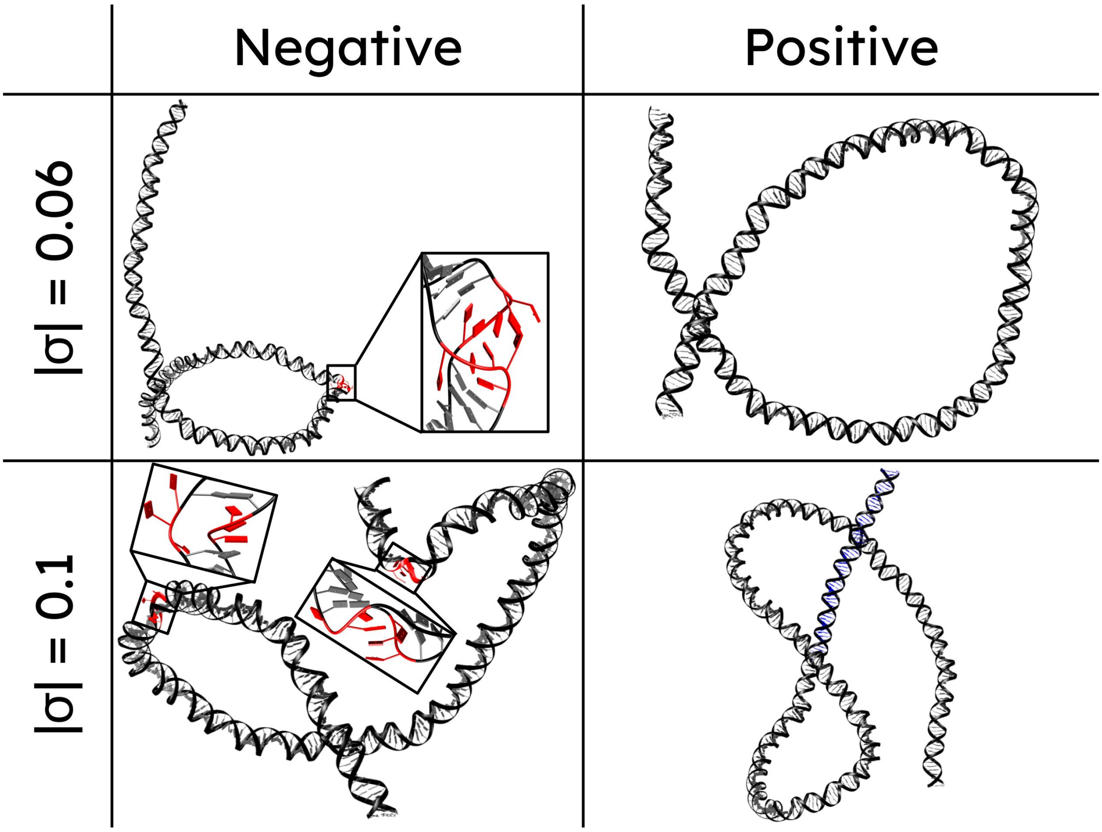


Figure 4.9: Representative structures for all simulations of the ‘offset’ sequence. Simulations at $|\sigma| = 0.06$ display largely predictable structures, with the formation of a toroid in both cases, and the formation of a strongly bent tip-bubble in the negative case. Simulations at $|\sigma| = 0.1$ show highly unpredictable structures unlike any observed in previous simulations. The positive case displays a pair of highly curved regions, one at the plectoneme tip and another near the crossing point. This case also suggests that the dummy base pairs are playing a role, shown here in blue. The negative case shows an extremely chaotic system, with the formation of two distantly separated bubbles and, like the positive case, a pair of highly curved regions.

It is worth noting here that the representative structure of the $\sigma = -0.1$ system is extremely viewpoint-dependent, with the view in figure 4.9 chosen to give a clear view of both the curved sections and bubbles. Unlike almost all other structures, including the others in figure 4.9, the apparent toroids formed in this particular system are not planar, and as a result the appearance of the structure can vary wildly depending on the angle from which it is viewed.

4.3.3.1 Dynamics of the $\sigma = 0.1$ system

Behaviour of the $\sigma = 0.1$ system is unlike any previously observed positively supercoiled system, with the formation of a pair of highly curved regions, and a plectoneme that forms away from the centre of the duplex. As displayed in figure 4.10, the nucleation of the plectoneme in this system begins very close to the mobile end of the strand, at around base pair 80 (base pair 20 if the dummy base pairs are ignored). A secondary curved region then forms between base pairs 150 and 200, while base pair 80 is maintained as the region with the highest curvature. Finally, a tertiary region of curvature forms between base pairs 225 and 275, becoming the apparent tip of the plectoneme. The behaviour of this system is in complete contrast to those

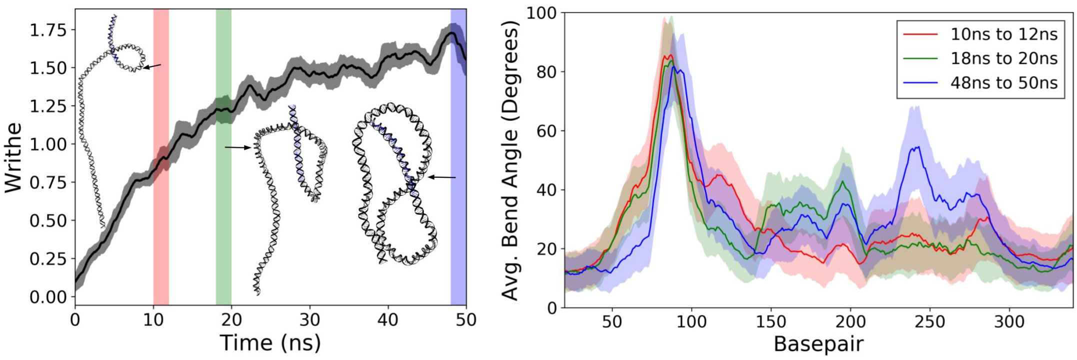


Figure 4.10: Behaviour of the $\sigma = +0.1$ system over the first 50ns of simulation. (**left**) Writhe over time, along with structures representing initial toroid formation, migration, and final structure. (**right**) Average bend angles for same three structures, displaying the migration of the toroid, as well as consistent strong bending near the mobile end of the duplex. Shaded regions in the left-hand plot match the times over which bend is measured on the right.

seen in the previous chapter, in which curvature would migrate through the system in the form of a toroid, usually settling in the formation of a plectoneme in the centre of the sequence.

This system also highlights a possible issue with our simulation setup - the structural images displayed in both figures 4.9 and 4.10 seem to indicate that the dummy base pairs are contributing to plectoneme formation in a non-trivial way. While they are still held straight and torsionally isolated from the bulk of the strand, these base pairs are clearly a part of one of the arms of the plectoneme in this system, possibly allowing it to grow at a location away from the centre of the duplex. The contribution of these base pairs is almost certainly a consequence of the excluded volume restraints - due to the location of the excluded volume restraints at the end of the dummy base pairs, the plectoneme has more space to grow in this direction, allowing it to adopt the folded conformation seen in figures 4.9 and 4.10.

4.3.3.2 Dynamics of the $\sigma = -0.1$ system

In the same vein as the $\sigma = +0.1$ system, behaviour of the $\sigma = -0.1$ system is unlike any previously observed system. The representative structure displayed in figure 4.9 displays the same dual-curvature as it positively supercoiled counterpart, with the added complexity of clear denaturation bubbles in these regions.

This system can be roughly divided into three separate macroscopic states: the single bubble state, between the times of 0 and 200 nanoseconds; the intermediate state, between the times of 225 and 275 nanoseconds and the dual-bubble state, between the times for 300 and 500 nanoseconds. The reasoning behind the choice of these three time periods is clear from the plot of writhe, where the value of Wr moves from an average of approximately -1.5 to an average of approximately -1.0 when transitioning through these three states.

The first state is dominated by its single bubble, which forms at around base pair 80, extremely close to the dummy base pairs at the mobile end of the duplex. As expected, this region acts as a strong focus point for bending, becoming the centre of the initial toroidal region. The structural image in figure 4.11 indicates the presence of a secondary bubble downstream of the strongly bent region, however this bubble has a short lifetime ($< 1ns$) and has no significant

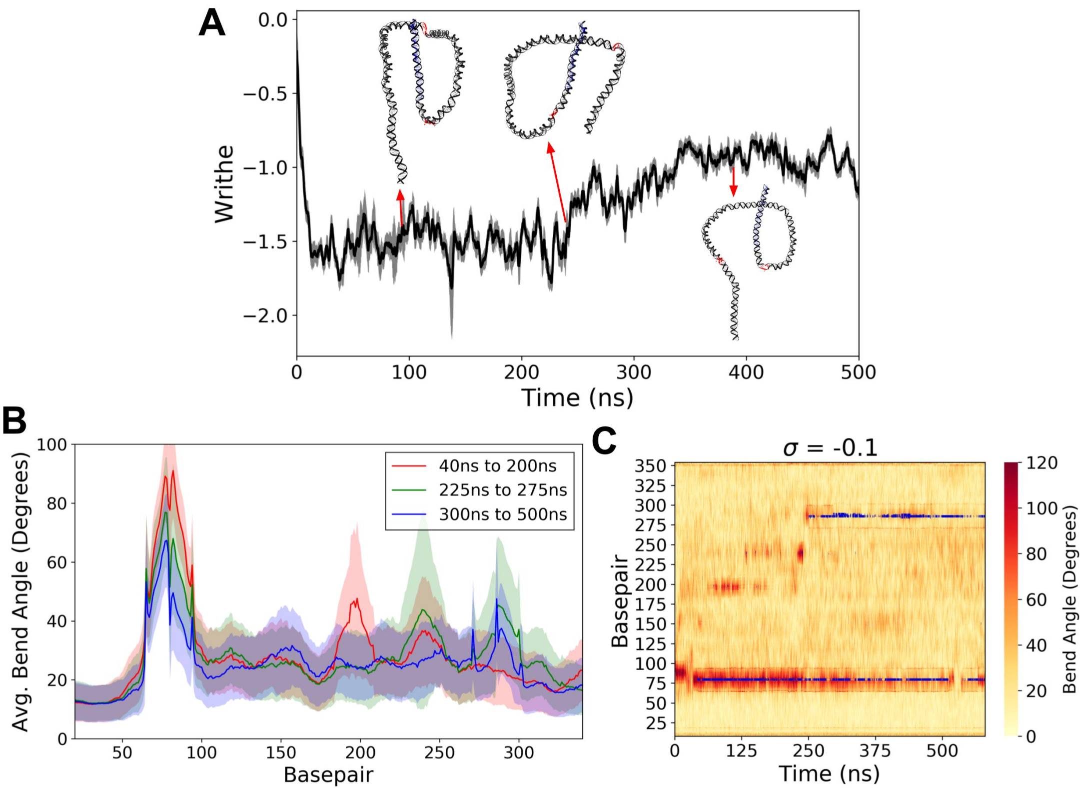


Figure 4.11: Behaviour of the $\sigma = -0.1$ system over the entire course of simulation. (**A**) Writhe over time with corresponding structural images. (**B**) Curvature over three distinct time windows. (**C**) Kymograph of bubble formation accompanied by localise curvature. Here plectoneme locations are excluded due to issues with the projection method described in section 4.3.2.2.

impact on system dynamics.

The transition from the −1.5 to the −1.0 writhe state is clearly defined by the nucleation of a secondary bubble in the region of base pair 290. This bubble however does not become a focus for strong bending, rather acting as the focus for a softly bent inflection point. Once this bubble forms the system likely reaches an energy minimum, after which no further writhe, and by extension no further strong bending, is required.

In this system, as in its positively supercoiled counterpart, the dummy base pairs appear to be contributing to system dynamics in a non-trivial way. This is a consistent weak point of the technique used in this text - in systems in which plectoneme or bubble formation is favored near the mobile end of the DNA the presence of the dummy base pairs allows for the development of higher order structures.

4.3.3.3 Sequence dependence in the offset sequence

Despite the presence of clearly defined denaturation bubbles in the negatively supercoiled case, both simulations of $|\sigma| = 0.1$ show remarkably similar behaviours, with the formation of a pair of strongly bent regions, and with dominance of the region close to the mobile end of the DNA. These similarities hint towards strong sequence-dependence outside of simple bubble formation probabilities.

Due to the specific behaviour of this sequence, it is more informative to begin analysing sequence in terms of curvature and plectoneme formation. However, due to the issues with the projection method defined earlier, it is simpler to consider the behaviour of these systems in terms of curvature, and use this curvature to infer plectoneme positions.

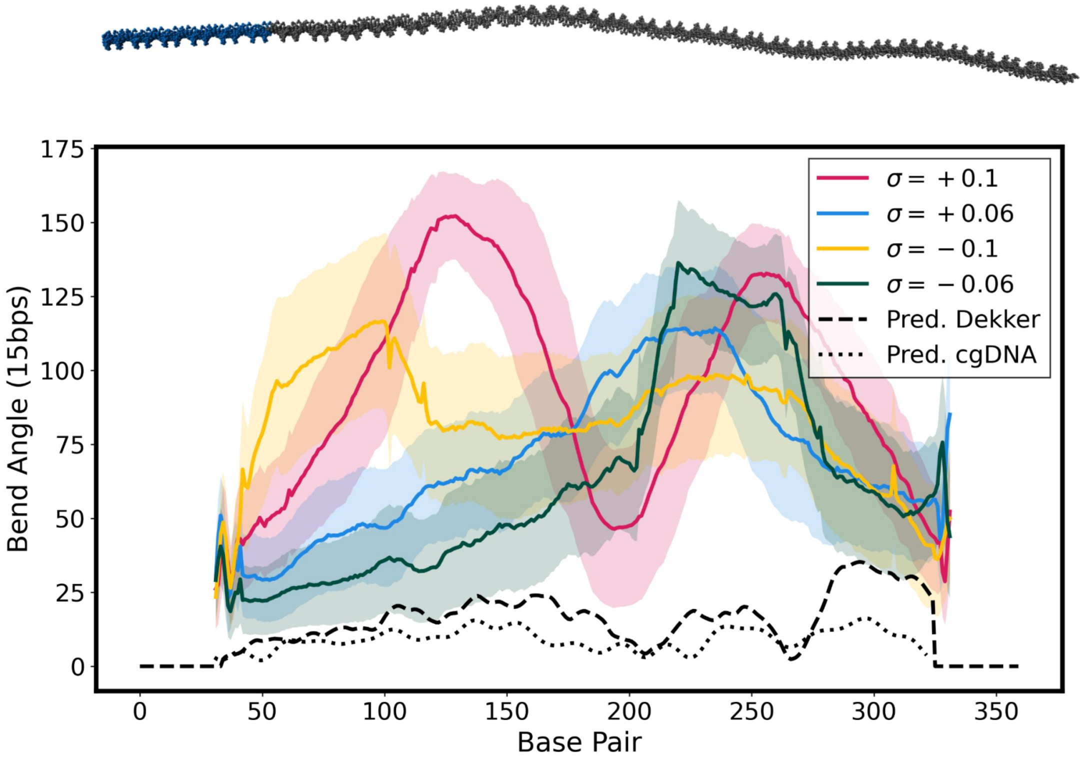


Figure 4.12: Distribution of bending over all simulations of the second offset sequence, paired with predicted curvatures along with the cgDNA ground path. All curvatures from simulation are taken over a 60bp window, in line with the predicted curvatures. Predicted curvatures display the cause of the unusual dynamics of this system - both highly curved regions are away from the sequence centre.

The unusual behaviour of this sequence is elucidated by the examination of its ground path curvature, as displayed in figure 4.12. Here, particularly in the case of $\sigma = +0.1$, the effects of the dual-curved structure can be seen, with a pair of peaks representing the two highly curved regions seen in the representative structures of figure 4.9. Comparing the locations of these two regions with the predicted curvature appears to show alignment. Both the 'Dekker' method and cgDNa show increased curvature away from the centre of the sequence, with curvature in the regions of 100 to 175 and 250 to 320 having higher curvature than the region at the centre of the sequence. These regions appear to align with the two regions of high curvature seen in the $\sigma = +0.1$ simulation. These regions also roughly align with the regions of high curvature in the $\sigma = -0.1$ system, although curvatures in these regions are likely altered by the presence of bubbles.

In the case of the $|\sigma| = 0.06$ simulations the relationship with curvature is less clear, with the regions of high curvature forming closer to the region of low ground path curvature in the centre of the sequence. The behaviour of the $\sigma = -0.06$ system can be explained by the presence of a tip-bubble, as seen in figure 4.9, however there is not tip-bubble present in the

σ = +0.06 system. There are two potential explanations for the behaviour of this system, the first is that the region with slightly higher curvature between base pairs 200 and 250 is acting as a nucleation point, the second is once again the movement of the toroid to the centre of the sequence.

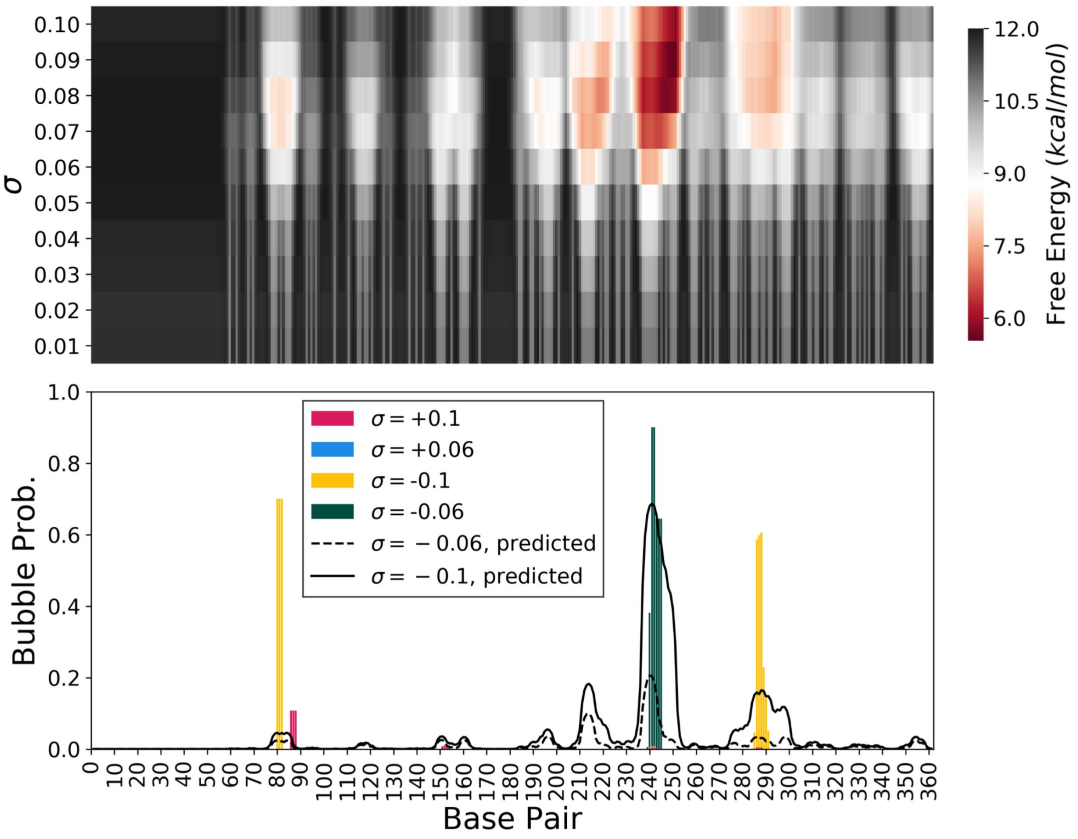


Figure 4.13: Bubble formation in the second offset sequence. Heatmap of free energies (**top**) shows most favorable formation in the region between base pairs 240 and 250, with some secondary regions. Bubble positions in simulations (**bottom**) vary drastically according to supercoiling density, with formation in the $\sigma = -0.06$ system aligning with the highest probability region, while higher densities are show formation in secondary regions.

Locations of bubble formation in this system vary significantly based on supercoiling density, as demonstrated in figure 4.13. In the $\sigma = -0.06$ system bubble formation is largely as expected, with formation in the low-energy, high-probability region between base pairs 240 and 250. This bubble is co-localised with the tip of the toroid, resulting in the bend angles seen in figure 4.12. Bubble formation in the $\sigma = -0.1$ system (and to a much lesser extent the $\sigma = +0.1$ system) seems to have been directly influenced by the formation of the highly curved regions seen in figure 4.12. Bubble formation in this system occurs in regions with probability significantly lower than the central 240-250 region, but with values significantly above zero. These regions align directly with the regions of high curvature seen in the ground path, implying that the formation of writhe in these regions directly resulted in the formation of bubbles, specifically in the area with the highest bubble formation probability within the curved regions.

It is worth noting here that the interplay between bubbles and plectonemes in the $\sigma = -0.1$ system may be the result of our method of equilibration. By employing a 40ns window in which hydrogen bonds between paired bases are restrained we allow for the formation of plectonemes before bubble nucleation is possible. This plectoneme formation results in the formation of strong curvature which may, in turn, reduce the energy required to form a bubble, therefore promoting bubble formation in the region of strong curvature.

4.4 Summary

Despite the relatively low number of simulations, the systems studied in this chapter have proven to be highly informative with regards to studying sequence-dependent behaviour in supercoiled DNA, both in the design of the sequences and in their behaviour.

We were able to demonstrate that sequence design is a non-trivial process, even on this small scale, and that trying to brute-force behaviour by inserting large numbers of consecutive AT base pairs results in anomalous behaviour, specifically the formation of lage denaturation bubbles in all cases.

Learning lessons from these initial failures, we were able to successfully design sequences using a block-averaging technique. These sequences were designed specifically to show contrasting behaviours, with one sequence designed to promote bubble-plectoneme cooperation, and the other designed to cause systematic frustration. Both of these designs proved to be highly successful, with the so-called ‘centred’ sequence displaying plectoneme formation in region of high AT percentage at its centre, and with consistent bubble formation in a single region identified by SIDD. As intended the so-called ‘offset’ sequence displayed behaviour unlike any seen previously, with the formation of a pair of highly curved regions. These regions were well predicted by the DNA ground path curvature, and appeared to promote bubble formation, suggesting that bubble-plectoneme interactions can be dynamic, and are driven by both sequence and torsional stress.

Chapter 5

Understanding Large Scale Sequence Dependence

At its outset, the purpose of this project was to adapt the theoretical frameworks laid out by Benham and colleagues in SIDD[98] and Dekker and colleagues [100] in to a single, unified predictive model of plectoneme and bubble formation in supercoiled DNA. Alone, these frameworks have proven successful in the prediction of structure formation under specific circumstances - in the case of SIDD the framework is highly effective in fully extended, high force systems, while the work of Dekker and colleagues proves highly efficient at predicting plectoneme formation in long, positively supercoiled duplexes at low force. However, both of these frameworks fail to capture behaviour at the critical regions of force and density seen in the simulations in this text, with the interactions between plectonemes and bubbles proving to be the driving force in these regimes. The combination of these two codes, paired with experimental data from optical-magnetic tweezers, in which both force and torque can be reliably applied, aimed to develop a framework able to predict the behaviour of supercoiled DNA at all physiologically relevant forces and densities. In reality, this objective proved to be far too optimistic, with the complexities of combining the two frameworks offering a series of extreme challenges, and the stalling of experimental results due to the covid-19 pandemic forcing a change in direction. This chapter will describe a simple model, designed as a base from which the ideas of combining plectoneme and bubble predictions can be developed, as well as some preliminary experimental results.

Credit for all experimental results discussed in this chapter goes to fellow members of the Physics of Life group at York, specifically Drs Jack Shepherd, Jamie Howard, Sébastien Guilbaud and Jack Zhu, under the supervision of Prof. Mark Leake.

5.1 Predicting plectoneme location

The model proposed by Dekker and colleagues [100] functions on the assumption that DNA curvature is the defining factor in plectoneme formation, with regions of the duplex with higher inherent curvature acting as regions in which the energy of plectoneme tip formation is minimised. The simulations described in the previous chapters appear to support this hypothesis, but also strongly suggest that a failure to consider the formation of bubbles, the sizes of plectonemes and the interactions between them make the model an over-simplification of reality.

The first and most relevant over-simplification is the near-complete ignoring of plectoneme size. In order to fully understand the nature of this issue one must first understand how the model works and the assumptions it makes. First, the model assumes that plectonemes are, on average, 1000bp in size, and adjusts the probabilities near the end of the strands accordingly. Second, the model calculates the free energies of plectoneme tip formation for a range of sizes, from 40bp to 120bp, based on curvature and persistence lengths over this range. Following this, the model assumes a constant tensile force of $3pN$, a force well above the normal physiological range, and adds the work done against this force to the overall energy of tip formation. The larger the plectoneme tip the greater the reduction in extension and, therefore, the more work done against tensile force. All of these calculations are completely independent of supercoiling density, meaning that, for any location more than 1000bp from either end of the sequence, plectoneme size is irrelevant. The reasoning behind this methodology is clear - ignoring considerations of plectoneme size and supercoiling density results in a much simpler model while still capturing key information on the influence of curvature and persistence length. However, as the simulations of previous chapters have demonstrated, ignoring the amount of space available for plectoneme formation, as well as the influence of changing force on overall plectoneme size, fails to capture the behaviour of the vast majority of systems, regardless of superhelical density.

It is also worth noting at this point the difficult nature of the code used in [100]. First and foremost, the code is written in igor pro, a premium software package designed for data analysis, making it both difficult to use and modify. Within this code are multiple smoothing functions, the exact method for which I could not determine, meaning that any attempts at re-implementation of the code in a more sensible language (in this case python) led to small but significant deviations in the predictions produced. As such, all predictions produced in this chapter will be made using the original code, and, due to the difficulty in modifying this code, the analysis included may not be as thorough as desired.

5.1.1 Two-state model of plectoneme formation

The simplest way to account for σ-dependence in plectoneme formation is to examine the partition of free energy between twist and writhe. Consider a supercoiled system with two possible macroscopic states - extended and buckled, with the extended state representing a system in which all stress is partitioned in to twist, and the buckled state representing that in which some or all of the stress is partitioned in to writhe. The balance of twist and writhe is dictated by the relative energies of these states, with this balance influenced by sequence, specifically the sequence at which the plectoneme tip forms, and by applied tensile force.

The transition of DNA between the extended and buckled states in positively supercoiled DNA can be modelled using a simple two-state model as described in the work of Brutzer *et al* [65]. Here, the free energy of the DNA in its pre-buckled, extended state is simply described by the energy of twisting B-DNA

$$E_{pre} = \frac{1}{2}\frac{C_s}{L_0}(2\pi)^2\Delta L_k^2 \tag{5.1}$$

where L_0 is the DNA contour length, ΔL_k is the number of added turns, and C_s is the effective torsional modulus [180]

$$C_s = C\left[1 - \frac{C}{4l_p k_b T}\left(\frac{k_b T}{l_p F}\right)\right]^{1/2} \tag{5.2}$$

which describes the response of DNA to over-stretching. Here C is the torsional modulus of B-DNA and l_p is its persistence length, with values of $100k_B T nm$ and $45nm$ respectively [181].

The post-buckling energy is then the partition of some of this twist in to writhe, accounting for the energies of plectoneme formation (E_b)

$$E_{post} = E_b + \frac{1}{2}\frac{C_s}{L_0}(2\pi)^2(\Delta L_k - \Delta L_{Wr})^2 \tag{5.3}$$

where ΔL_{Wr} is the number of turns partitioned in to writhe.

The energy of plectoneme formation E_b is the sum of two terms, the first representing the energy of nucleating the plectoneme tip, and the second representing the work done against tensile force during nucleation

$$E_b = E_{tip} + W.$$

In order to incorporate sequence dependence, the energy required to nucleate the tip E_{tip} can be computed directly from the method described in [100] (see equation 1.24).

Work done against force is

$$W = L_p F \times z(F,0) \tag{5.4}$$

where L_p is the length of the plectoneme (including the tip), F is the tensile force and $z(F,0)$ is the extension of the DNA at zero torque from the work of Moroz and Nelson [88]

$$z(F,0) = 1 - \frac{1}{2}\left[\frac{l_p F}{k_B T} - \frac{1}{32}\right]^{-1/2}. \tag{5.5}$$

Combining the above gives a full expression for the free energy of a buckled state in this simple two-phase model

$$E = E_{tip} + \frac{1}{2}\frac{C_s}{L_0}(2\pi)^2(\Delta L_k - \Delta L_{Wr})^2 + L_p F \times z(F,0). \tag{5.6}$$

5.1.2 Algorithmic implementation of the two-State model

Now that the theoretical groundwork for accounting for plectoneme size has been established, the next step is to implement it algorithmically, allowing for it to be integrated in to the pre-existing method described in [100]. Fortunately the most difficult part of this predictive framework, accounting for sequence dependence, is already done (see section 1.7.2.2), and the addition of the model described above is simple.

Algorithmically, the model described above can be used to find an energy ‘penalty’ which is a function of force and tip position, added as an additional term to equation 1.24. For a given base pair i at position N_i along the DNA, the maximum size of a plectoneme (in base pairs, including the tip) which can nucleate at i is $2N_i$ (the factor of two is a result of the symmetry of the plectoneme, it is effectively a length of N_i on each side of the position i).

In order to find the amount of writhe that can be absorbed by a plectoneme of size $2N_i$, we need to know the amount of writhe that can be absorbed by each individual base pair. For the sake of simplicity, we will assume that this number, from here on called λ, is constant. It should however be noted here that, in reality, this number is far from constant, with both experimental and theoretical studies suggesting that it is σ-dependent [182–184]. Unless otherwise stated, the number used will be $\lambda = 0.01$, based on experimental data (see figure 5.1).

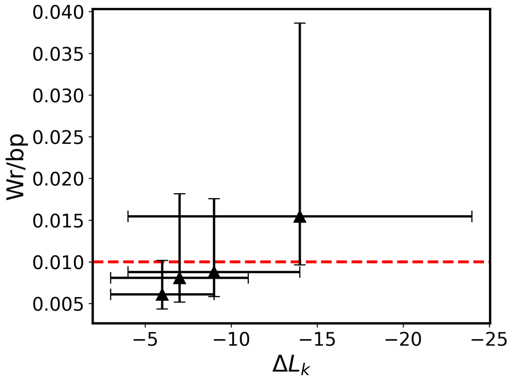


Figure 5.1: Writhe per base pair calculated from [182] (black) along with the chosen value (red). Large uncertainty arises form error in both plectoneme radius r and opening angle α.

In principle, the value of λ could be measured directly from the simulations shown earlier in this text, or from explicitly solvated simulations of DNA minicircles [59]. However, the issue with this lies in the relative size of the plectoneme tips and arms. In these simulations the size of the plectoneme tip varies significantly, either due to the presence of bubbles or simple stochastic effects, and, because the plectoneme tip makes up a significant portion of the total size of the plectoneme (due to the presence of only two or three crossing points in total) any measurements of writhe per base pair in the legs of the plectonemes will be highly unreliable.

Once a value for λ has been established the amount of writhe absorbed by any plectoneme of length l_p is simply $\Delta L_{Wr} = \lambda * l_P$, a value which can be plugged directly in to equation 5.6. Using this calculated energy, the energy for any system with any plectoneme size can be calculated. Examples of this energy are shown in figure 5.3. This figure clearly displays the strong force-dependence of the system energy, owing to the work done against tension, as well as the favorability of mixed plectonemic and twisted states in lower force systems. In all cases the minimum energy is for values of writhe less than the total number of turns (in the case of 0.5pN the minimum energy is at a writhe value of approximately 38), indicating that, at the force values relevant to this text ($F \leq 1pN$), the most energetically favorable systems are those in which there is a balance between the formation of plectonemes and the twisting of the B-DNA.

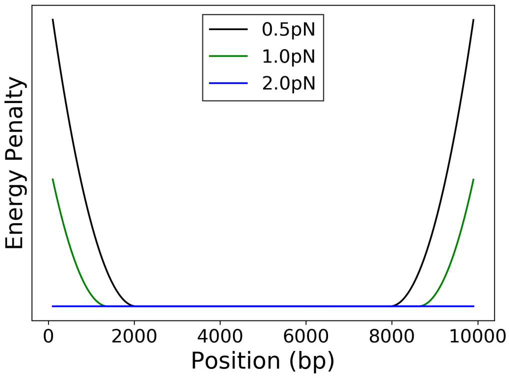


Figure 5.2: Position-dependent energy penalty for a system of 10kb with $\Delta L_k = 50$. Regions in which energy penalties are applied shrink as force is increased due to a lowering of plectoneme size.

Once this energy is calculated, the next step is to insert it into the framework of [100]. This is achieved by calculating the maximum plectoneme size for every position i along the strand, and using this size to calculate an energy penalty. This energy penalty will be set relative to the minimum of equation 5.6, called E_0, which is given by a plectoneme of size L_0. Positions i

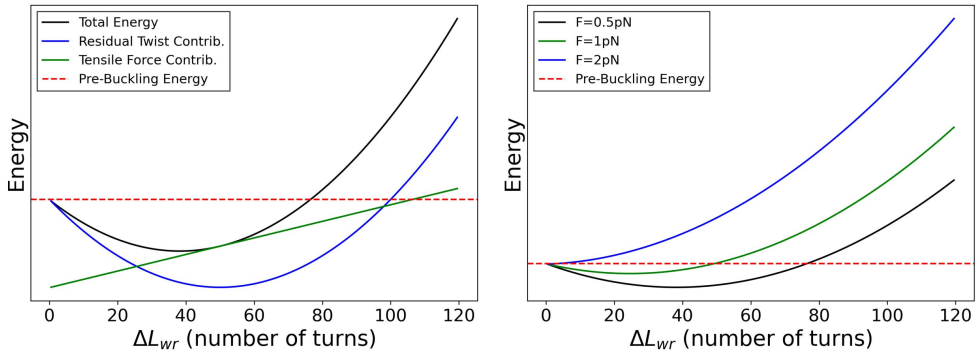


Figure 5.3: Energies for a 10kbp system with $\Delta L_k = 50$ turns. (**Left**) Energies for a system with a force value of $0.5pN$ showing the individual contributions of both residual twist and work done against tension. (**right**) Total energies for three different forces. Both plots show the pre-buckling energy, that is the energy of the system at full extension, when all superhelical stress is partitioned in to the twisting of B-DNA.

such that $2N_i \geq L_0$ are those in which the plectoneme can grow to at least a size of L_0, thereby minimising equation 5.6, meaning that they require no energy penalty. All positions outside this range are forced to have smaller plectonemes, and by extension, more superhelical stress partitioned in to B-DNA twist. As a result, these positions have a higher overall energy, given in the form of an energy penalty. This penalty is simply $E_i - E_{min}$, and is pictured for a range of forces in figure 5.2.

Note here that, by using a single minimum energy for each position along the DNA, we are assuming a fixed plectoneme size for each base pair - this ignores fluctuations in plectoneme size around the minimum, by extension ignoring entropic effects. The choice to use only a single plectoneme size, rather than finding the energies in a range around the minimum is an entirely practical one - the more possible plectoneme sizes, the higher the overall number of states whose energy we need to find. Finding energies for multiple plectoneme sizes would increase the overall computational time by a factor of N, where N is the total number of possible plectoneme sizes. By using only the conformation that represents the minimum energy we are still able to capture the important effect of position on plectoneme formation, without a significant increase in the overall computational time. It is also worth noting here that we will not be exploring the possibility of multi-plectoneme systems, although they have been observed experimentally in long DNA molecules (in this case $21kbp$) [185].

5.1.3 Testing the new energetics

By computing the energy penalties seen in figure 5.2, the method described above can be directly integrated in to the code from [100]. A full listing of the python code used to compute the energy penalties, as well as the modified version of the original plectoneme prediction code can be found in the appendix 2 of this text (specifically B.4 and B.5). Before beginning predictions, it is worth noting that all code for the remainder of this chapter, whether it be the original code or code modified with the new energy penalty, has been altered slightly from its original form. This modification is in the form of the amount of smoothing performed on the data. In the original text the authors use a Gaussian filter to transform raw predictions of plectoneme density to results comparable to experiment. Within the code itself this smoothing is implemented in the form of a series of 64 boxcar smooths, with a window size of 300 base pairs. The consequence of this is that, in smaller systems ($< 10kbp$), predictions are massively over-smoothed, giving results that have no predictive capabilities. All results in this chapter will have an amount of smoothing that is adjusted according to need, with the number of boxcar smooths reduced, but the window size of 300 maintained.

Results for a single sequence, with a pair of highly-curved modifications, are shown in figure 5.4. Systems in which plectoneme formation is not favored, that is those where the pre-buckling energy is lower than all post-buckling energies, are marked with red diamonds. These systems act as a baseline and show that, while the inserts give a noticeable increase in plectoneme density in their regions, the sequence has a region with even greater plectoneme density close to its end (at around base pair 4500). Due to its proximity to the end of the DNA, plectoneme formation in this region is immediately suppressed in regimes in which even the formation of small plectonemes ($\Delta L_k = 5, F = 0.1pN$,$\Delta L_k = 10, F = 0.5pN$) is energetically favorable, with density migrating towards the centre of the DNA in the unaltered sequence, and towards the highly curved inserts in both the SeqA and SeqB systems.

Systems with $F = 0.1pN$ have the largest plectonemes, and therefore show the most exaggerated migration towards the centre of the duplex. These systems give the best demonstration of how the location of the highly curved regions affects behaviour as ΔL_k increases. At $\Delta L_k = 5$ the region of high density (at around bp 4500) in the unaltered sequence is still present, although its density is significantly lower than in the systems where plectoneme formation is not favored. This region then vanishes almost entirely in the $\Delta L_k = 10$ system, and from this point onwards the plectoneme density in the unaltered sequence becomes almost entirely predicted by proximity to the centre of the duplex. The SeqB insert system sees similar behaviour to the unaltered system, only in this case the density in the highly curved region at bp 1000 vanishes only at $\Delta L_k \geq 20$, after which density predictions match those of the unaltered sequence. In contrast, plectoneme density in the SeqA system is consistently predicted by the location of highly-curved insert. The addition of the insert close to the centre of the duplex means that, even in systems with the largest plectonemes, this region persists as the location of the highest plectoneme density.

At higher forces (0.5 and 1 pN) behaviour is similar but shifted. The highly curved region at bp 4500 does not vanish until $\Delta L_k = 20$, and the highly curved region at bp 1000 in seqB does not vanish until $\Delta L_k = 30$ in the $0.5pN$ case, and doesn't vanish at all in the $F = 1.0pN$ case. The highly curved insert near the centre of seqA remains a focal point for plectoneme density in all cases.

The overall results also appear to align well with the predictions of the original algorithm. In the original code, plectoneme density is suppressed in regions 1000bp at either end of the sequence, in the case of the modified predictions this is roughly equivalent to the formation of a 2000bp plectoneme, or a value of $\Delta L_k \approx 20$. The predictions of the original algorithm show the same suppression of the high-density region at bp 4500, and are able to capture the increased density at the highly-curved inserts, but completely neglect the decreased relevance of the SeqB insert at bp 1000 in higher supercoiling density systems.

The method appears to serve its function as hoped - it is able account for the effects of plectoneme size and predict their movement to the centre of the DNA while sustaining the

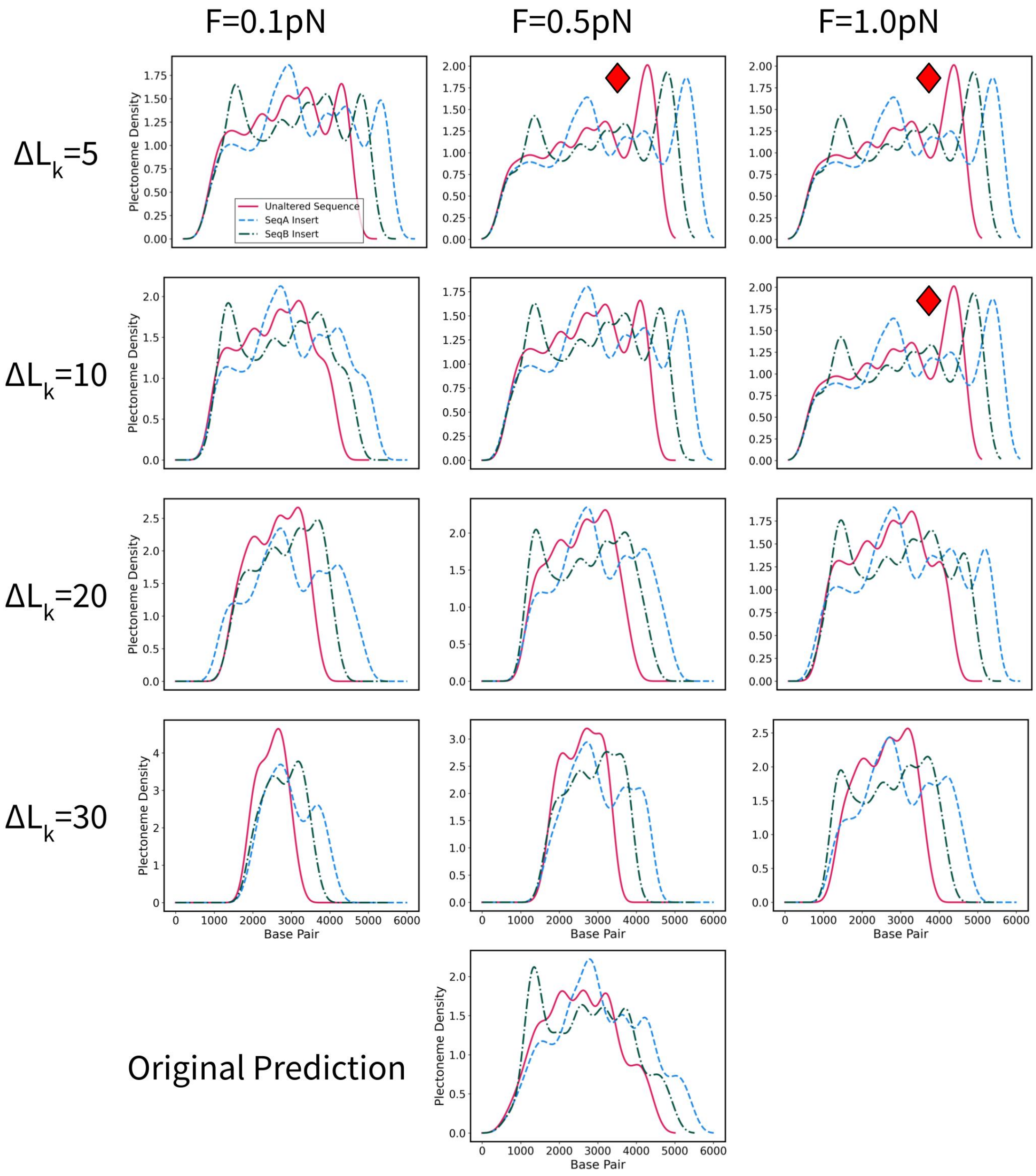


Figure 5.4: Predicted plectoneme positions for a randomly generated 5kbp (50% AT) sequence (solid magenta line) at a range of forces and ΔL_k values. SeqA (1009 bp) and SeqB (514bp) are a pair of highly curved inserts taken directly from [100], inserted at base pair 2000 (dashed blue line) and 1000 (dashed green line) respectively. Predictions of the original program are also listed, and give results similar to those seen at $\Delta L_k = 20$. Plots marked with red diamonds are identical, and are those in regimes where the free energy of the extended state is lower than any plectonemic state. Results smoothed using 5 sequential boxcar smooths.

significant sequence dependence calculated by the original algorithm. The movement of plectoneme density to the centre of the sequence is consistent with observations from simulations, but the lack of data for longer sequences makes this hard to confirm.

5.1.4 Comparisons to experiment

When constructing models such as the ones described above it is vital to test the effectiveness of their predictions in a wide range of systems. So far, we have consistently seen the migration of plectonemes to the center of the sequence in higher-density systems in simulations, and we have been able to implement this movement in to predictions of plectoneme density. What we lack is any data on larger systems, namely those long enough that predictions of plectoneme position can be made. Systems of this size are outside the scope of atomistic MD simulations, and the lack of sequence-dependent features such as inherent curvature mean that current coarse-grained methods are also inappropriate. In order to gather data for direct comparison to the predictive framework described above experimental results are required - specifically results from an experimental setup that allows for fine control of both torque and tension, as well as direct observation of plectoneme position.

5.1.4.1 Magneto-optical tweezers

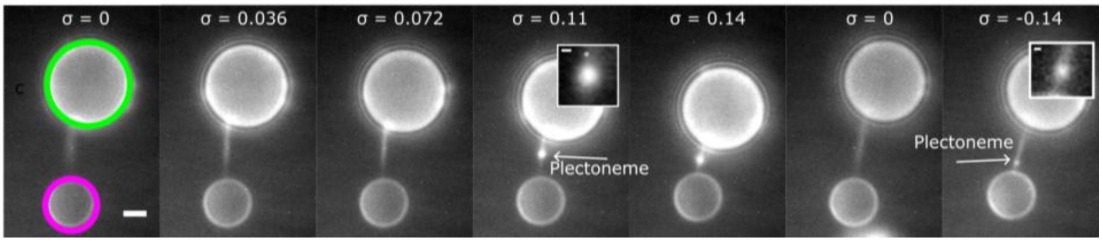


Figure 5.5: Fluorescence images of 15.6kbp DNA tethers in an optical-magnetic trap, with beads indicated with green and magenta circles. In all cases the applied force is $\approx 1pN$. Plectoneme growth is seen as increasing levels of positive supercoiling are applied. The far right panel depicts the same construct under negative supercoiling. Credit J. Shepherd, J Howard and S Guilbaud.

As described in chapter 1, optical tweezers allow for the application of tension, as well as direct observation of the DNA duplex, but lack the ability to apply torsion. In contrast, magnetic tweezers allow for both the application of tension and torsion, but prevent direct observation of the DNA strand. In order to directly observe plectoneme position in torsionally constrained DNA under tension a combination of these techniques is required, and a technique, COMBI-Tweez, which combines both optical and magnetic tweezing, has been developed for this exact purpose. The details of how this technique is achieved are beyond the scope of this work, and won't be explored here. While still in its early stages, this technique has yielded some early results, and has successfully been able to trap DNA, allowing for the application of both controlled tension and torsion.

Figure 5.5 depicts multiple trapped DNA molecules under varying levels of supercoiling, both positive and negative. Evidenced in these images is the expected hat-curve behaviour, with plectonemes in the $\sigma = +0.11, +0.14$ systems appearing significantly larger than the plectoneme present in the $\sigma = -0.14$ system. It is worth noting that, were this system following the pattern of our simulations, we would not expect a plectoneme to be present at $F = 1pN$ in negative supercoiling. There are multiple explanations as to why one is present here though, first and foremost is that the value $\sigma = -0.14$ is beyond the range we explored, and so it is possible that we would observe some significant writhe at this density in simulation. Second is the effect of the dye on the structural properties of the DNA - the dye used here is sybr gold, which is known to increase the persistence length of DNA, making it more resistant to extension [186].

5.1.4.2 Positively supercoiled experiments

At the time of writing, results using the magneto-optical tweezers are limited, with the three successful tethers depicted in figure 5.5 ($\sigma = -0.14, +0.11, +0.14$) being the only successful experiments. Two of the three successful tethers are in strongly positively supercoiled regimes. Plectoneme position in these systems, as well as predictions made by the modified curvature model described previously, are shown in figure 5.6. Note that, in order to predict plectoneme positions in these large, high σ regimes, the value of λ had to be adjusted, with the value increased from 0.01 to 0.02 - meaning that, in these predictions, there is assumed to be 50 base pairs per crossing point.

Results from both predictions and experiments are shown in figure 5.6. Plectonemes appear to display the same movement to the centre of the DNA as observed in simulation, with the +200 turns system nucleating closer to the centre of the DNA than the +150 turns system. Although the lack of sampling prohibits us from making any conclusions regarding sequence dependence, it is still worth noting that the two regions of highest plectoneme density in the base prediction are close to both ends of the DNA, regions in which the observed plectonemes do not form. Outside of this simple observation, there is no clear alignment with predicted peaks, with the regions of normalised intensity being too large to concretely display the location of the plectoneme apex.

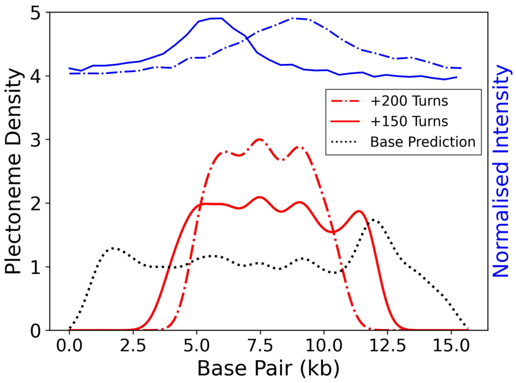


Figure 5.6: Predicted (red) and observed (blue) plectoneme positions for a $15.6kbp$ DNA construct. Experimental results appear to show the same movement to the middle as observed in simulations, with no apparent plectoneme formation at the peak of $12.5kbp$ predicted by the original code. All predictions use 32 boxcar smooths.

5.1.4.3 Negatively supercoiled experiments

The third and final successful tether was performed on a system in a strongly negatively supercoiled ($\Delta L_k = -200$) system, once again with a force of approximately 1pN. As depicted in figure 5.5, the plectoneme nucleated in this system is small, strongly implying the presence of a bubble somewhere in the system, and is far from the centre of the DNA, hinting towards possible sequence-dependence.

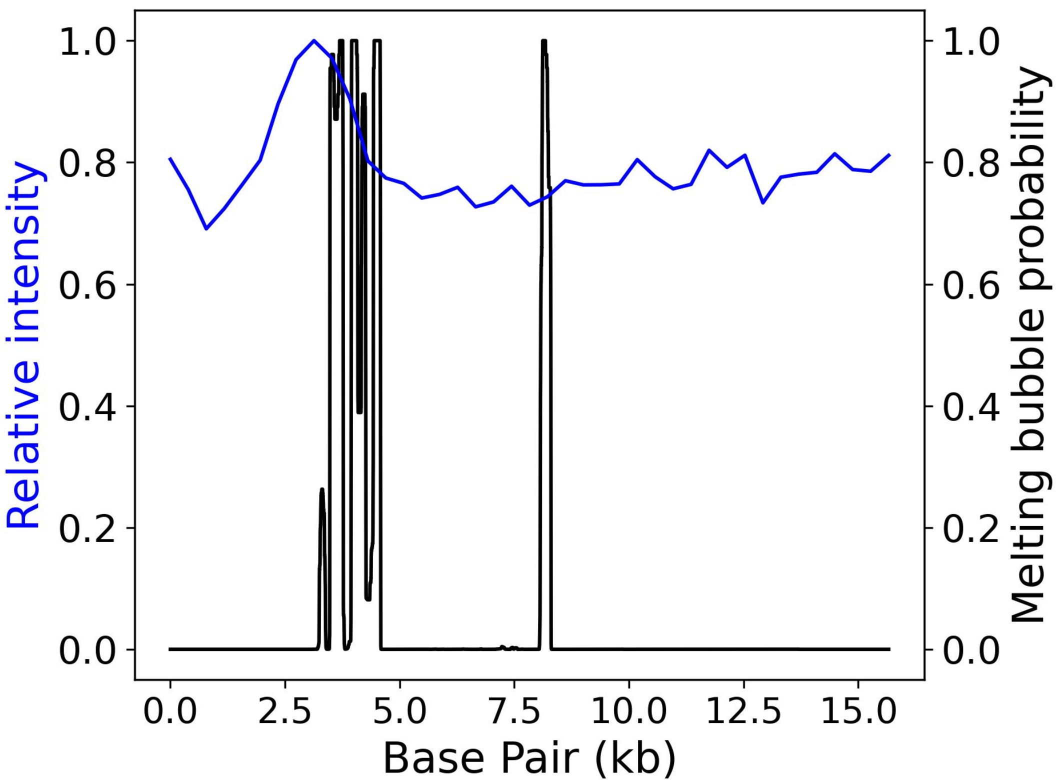


Figure 5.7: Observed plectoneme position in the $\sigma = -0.14$ tether, overlaid with predicted bubble probabilities from SIDD (sold black), as well as predicted plectoneme position (dashed black, from the original model, without the modifications described earlier). Plectoneme position in this system appears to be strongly predicted by the bubble formation.

Figure 5.7 shows the position of the plectoneme nucleated in this system, and compares this position to predictions of both plectoneme and bubble position. The position of the plectoneme seen here aligns strongly with the region of high bubble formation probability identified by SIDD in the region between 3.5 and $5kbp$. The alignment between this region and the plectoneme position is strong, but not perfect, with most of the bubble positions shifted slightly to the right. The likely explanation for this is inaccuracies in the predictions of SIDD, arising from its failure to account for the existence of writhe. Bubble formation at just one of the peaks identified would be sufficient to localise a plectoneme, with that plectoneme absorbing a large portion of the torsional stress and therefore nullifying the need for further bubble formation.

The alignment with bubbles instead of high curvature seen in plectoneme position strongly suggests the existence of a tip-bubble in this case. The existence of this tip-bubble, as predicted by simulation, likely acts to pin the plectoneme in place, as well as reduce its overall size compared to its positively supercoiled counterpart.

This experiment, while only a single example, has wide-reaching implications for any future work attempting to predict the locations of plectoneme formation. Any attempts to do so cannot ignore the formation of denaturation bubbles, and must account for the ability of bubbles to not only absorb twist, but also create strong bending at the tips of plectonemes, pinning them in place.

5.2 Summary

By combining the pre-existing model of plectoneme prediction by curvature with a simple two-phase elastic model of DNA, we have created a theoretical framework for predicting plectoneme location in positively supercoiled DNA. Specifically, this framework allows supercoiling density, and by extension plectoneme size, as well as force, to be accounted for in prediction, and as a result better reflects behaviour seen in simulation. Evidence for the need to account for σ is then further reinforced by initial experimental results, in which a plectoneme is seen to migrate closer to the centre of DNA as its writhe increases, nullifying curvature-based prediction.

The predictions of this model are strongly contrasted by an experimental observation from a negatively supercoiled system, in which a plectoneme is seen to form far from the centre of the sequence, despite the high supercoiling density of the system. Simulations have shown that the presence of bubbles can act both to reduce the overall sizes of plectonemes and pin them in place, and this also appears to be the case in this experiment. The location of the plectoneme observed is also strongly predicted by the bubble-formation predictions of SIDD [98], suggesting that, in this case, any effects of curvature are nullified by the formation of a highly flexible denaturation bubble.

It is worth mentioning here that the theory shown in this chapter is highly simplistic, and was originally intended to be a small part of a large theoretical framework, designed to predict the formations of both bubbles and plectonemes simultaneously. Unfortunately, the design and construction of this framework presented a series of challenges that could not be solved in the limited time remaining in the project. While the full list of practical problems is too long to list here, the key to them lies in the non-trivial way in which plectonemes and bubbles interact, and the order in which they form. We have seen in simulations that the formation of bubbles seems to be promoted by the presence of the high localised curvature seen in the tips of plectonemes. In order to model this, this promotion would have to be quantified analytically, perhaps with the reduction of the sequence-independent bubble nucleation energy. However, we have also seen that the presence of bubbles acts to pin plectonemes, and it stands to reason that, were bubbles to nucleate before plectonemes, these plectonemes would nucleate at the site of bubble formation, in the same manner as defects [92], mismatches [116] or thymine dimers [187]. Making another genuine attempt at building this predictive framework would require properly quantifying these non-trivial interactions, and finding the associated energetics. These challenges alone present what would likely be an entire PhD projects worth of work.

Chapter 6

Discussion

This work constitutes the first simulations of supercoiled linear DNA at the atomistic level, made possible by advances in GPU technology and the increased availability of computational resources. The size of the DNA constructs and the time scales over which they are simulated are at the upper limit of what is currently possible with atomistic methods, with the examination of dynamics in such systems enabled only through the use of state of the art implicit solvation techniques. Due to the great difficulty in designing a system by which a linear strand of DNA can be torsionally constrained and placed under tensile force, the technique used in this text is entirely unique, with the number of required restraints rivaling even the most complex of systems. By demonstrating both the viability of simulations of this size, and the ability of implicit solvent techniques to accurately recreate experimental observations, it is hoped that this work can be used to guide further study of systems of this size using atomistic methods, with the detail afforded by atomistic techniques being used to inform the development of larger-scale models.

Although plectoneme tip-bubbles have previously been observed in OxDNA simulations [111], this work not only acts to confirm their existence, but also allows their dynamics to be examined in greater detail. Analysis of the role of tip-bubbles in atomistic detail unveils their role as foci for both strong bending and twist, as well as their ability to pin plectonemes. These simulations also reveal the existence of bubbles outside plectoneme tips, specifically in low-tension systems, showing that co-localisation is not only dependent on torsional stress, but also tensile force. The existence of tip-bubbles in positive supercoiling is also revealed for the first time, arising only at the highest levels of superhelical stress. The role of these bubbles is shown to be unlike their negative counterparts, with their sole role seeming to be as focus points for strong bending, with their ability to absorb twist severely hindered. Additionally, the structure of these bubbles is unlike conventional ideas of denaturation bubbles; the base pairs remain largely within the confines of the helix, rather than flipping outwards, and the double-helical structure of the backbones remains largely in-tact, with only a low level of major groove compression.

The level of detail afforded by atomistic simulations also allows for in-depth exploration of sequence dependence, a feature that is currently unique to atomistic models. We studied sequence-dependent behaviors across three sequences, with two specifically designed to promote the formation of bubbles and plectonemes in specific regions. Simulations showed strong sequence dependence in all cases, and, using evidence from simulation, we are able to assert that bubble-plectoneme interactions play a vital role in system dynamics. In many cases plectoneme formation was predicted by DNA ground path curvature [100], with plectoneme formation in both the ‘centred’ and ‘offset’ sequences aligning with regions of strong curvature. Favoring of strongly curved regions was, in some cases, overwritten by a need to maximise plectoneme size, resulting in the movement of the plectoneme to the centre of the DNA. Bubble formation was strongly predicted by SIDD [93, 94, 97], with regions of consecutive AT base pairs identified as the ideal candidates for bubble formation. Predictions of bubble formation were however perturbed by the presence of highly curved plectonemes or toroids, with bubble formation promoted in these highly curved regions.

In order to account for both the formation of plectonemes in highly curved regions of the DNA ground path, and the need for plectonemes in highly supercoiled systems to maxmimise their size, a modified version of the model formulated in [100] was proposed. Using a simple two-state model for supercoiled DNA, this model was able to recreate the movement of plectonemes to the centre of sequences, while maintaining dependence on ground path curvature. Results from preliminary magneto-optical tweezers experiments were shown to be complementary to this model, with the apparent movement of plectonemes to the centre of the DNA constructs seen in positively supercoiled systems, although precise analysis of sequence-dependence requires a large amount of additional experimental data. The single negatively supercoiled experiment shows results which contrast previous observations, with plectoneme location far from the centre of the DNA, and away from any highly-curved regions. Plectoneme formation in this system was well predicted by bubble nucleation locations predicted by SIDD, strongly implying the presence of a large tip-bubble.

All told, this work has shown the viability and usefulness of implicitly solvated atomistic simulations in the study of DNA supercoiling. These simulations have shown an ability to recreate experimental results, as well as confirming previous results from coarse-grained systems. Additionally, the combination of observations from simulation and experiment, paired with large-scale statistical models, has proven to be valuable in understanding the sequence-dependent behavior of supercoiled DNA.

6.1 Future work

The aim of this project, and others like it, is to attempt to understand the behavior of supercoiled DNA, in the hope of explaining its role *in vivo*. However, in order to fully understand this behavior, a multi-scale, multi-disciplinary approach is required, with the combination of atomistic, coarse-grained and statistical models, in addition to experimental results, needed to truly address the problem. The simulations described here represent the bottom of this stack of techniques, offering the greatest detail over the shortest time scales. In addition to their inherent value in understanding the detailed behavior of supercoiled DNA, the simulations described here can be used to inform the development of coarse-grained models, as well as aid in experimental design, offering a baseline with which observations can be compared.

As both hardware and software continue to improve, there is no reason why the technique used here could not be used on increasingly large systems. Simulations of longer sequences would allow for the examination of supercoiling in a more entropy-driven regime, giving even greater insight into bubble-plectoneme interactions and sequence-dependence. Additionally, many proteins are known to interact with supercoiled DNA, with topoisomerases known to induce both positive and negative supercoiling [31], and condensin known to extrude plectonemic loops [68, 188]. The technique described in this text could be directly applied to such protein-DNA systems, with fine control of tension allowing for measurement of the forces over which they maintain function. Another possible application of this technique is in simulations of DNA over-stretching [169], allowing not only for the simulation procedure to be simplified, but also allowing for the application of a much wider range of tensile forces.

It is also worth considering possible improvements to the technique used in this work. The approach used here was very much an initial attempt - the difficulty in establishing the specific simulation setup meant that, once it was working, and due to time constraints in the project, further improvement was not considered. This means that the technique described in this text is inefficient, with the need for such a large number of restraints, as well as the addition of non-functioning ‘dummy’ DNA, slowing simulation time considerably. It is my opinion that, should this work be continued, a re-implementation of the technique in an engine such as openMM [189], which allows for the addition of user-defined potentials, would offer a large uplift in simulation efficiency by allowing the dummy DNA to be removed, and the number of restraints to be reduced from hundreds to a mere handful.

Looking past MD simulations, the model described in the previous chapter could form the starting point of a much wider predictive model for structure formation in supercoiled DNA. It is well known that the positions at which bubbles nucleate in supercoiled DNA change depending on the level of torsional stress [94], and as such any model that attempts to simultaneously predict both plectoneme and bubble positions must be able to predict the distribution of torsional stress between the two. The model described in this text is the first step in this process, allowing for the prediction of plectoneme size, and by extension the level of writhe absorbed by plectonemes, while maintaining sequence-dependent behaviors. While combining this with a bubble prediction method such as SIDD would be a very challenging undertaking, it could result in the formulation of a model which is able to predict the distribution of bubbles and plectonemes in entire topologically-associated domains, or, with sufficient optimisation, perhaps entire genomes. A model such as this would offer an unprecedented view in to the role of structural DNA in gene expression and genome stability.

Appendix A

Sequences and additional analysis

A.1 Sequence used in hat-curve simulations

1 **GCGCGCGCGC GCGCGCGCGC GCGCGCGCGC GCGCGCGCGC GCGCGCGCGC GCGCGCGCG**C

61 TGCAAGATTT GCAACCAGGC AGACTTAGCG GTAGGTCCTA GTGCAGCGGG ACTTTTTTTC

121 TATAGTCGTT GAGAGGAGGA GTCGTCAGAC CAGATACCTT TGATGTCCTG ATTGGAAGGA

181 CCGTTGGCCC CCGACCCTTA GACAGTGTAC TCAGTTCTAT AAACGAGCTA TTAGATATGA

241 GATCCGTAGA TTGAAAAGGG TGACGGAATT CGCCCGGACG CAAAAGACGG ACAGCTAGGT

301 ATCCTGAGCA CGGTTGCGCG TCCGAATCAA GCTCCTCTTT ACAGGCCCCG GTTTCTGTTG

361 **CG**

A.2 Failed ‘centred’ sequence

1 **GCGCGCGCGC GCGCGCGCGC GCGCGCGCGC GCGCGCGCGC GCGCGCGCGC GCGCGCGCGC**

61 CGTCCAGCGG ACCGTCTGAC GGCTGGGCCA CATCGGGCGG TCCGGTCTCG GCGTGTCCGG

121 CCTTAGGACT CGGCGCAGCG CGCTGGCCCG GGTCGAGCTG AAAACACCGG CGCCCAAGAC

181 CAGGCGGGCC CGCCGCGTCG GCCAATTTTA ATATATAACC CGGCCGACGT TCTGAGGAAA

241 GCCCGTGGGA GGGGGACGGG TCACCGCCTG CCAGCGGCGA CAACGATCGG CCGCACCTTC

301 CATCGCCGTG GCGGCGCTCG GATCGCGCGG CAAAGGTGCT TGTGTCCCGG CAGGCTAGCG

361 **CG**

A.3 Failed ‘offset’ sequence

1 **GCGCGCGCGC GCGCGCGCGC GCGCGCGCGC GCGCGCGCGC GCGCGCGCGC GCGCGCGCGC**

61 GACGACGTCC CTGCTAACGC CCCGCCCGCC GGACCGCCCT CGCGATGGGG CGGCCGGGCA

121 CGACCCCATA AAATTTAATT ATGGCCTACT CGTTTAGCCA CCTCGCGGCG GAAGCCGGGG

181 GGACGGCGGC CGCTGCAGAC ACTGTACCGC GACTACGCCA AGCTGAGCTA GCCCCGTGGT

241 CGACTACGCA TCCCTCTGGG CCTCACTCAG CCGGATACAG TGACTTTGAC AGGTTTGCGG

301 GCCACGGCAG CCACCCGCAC AGCCGCGTGC GGGGGGAGCA ACCCTTGGGC GTTAGTATGT

361 **CG**

A.4 ‘Centred’ sequence

1 **GCGCGCGCGC GCGCGCGCGC GCGCGCGCGC GCGCGCGCGC GCGCGCGCGC GCGCGCGCGC**
61 GTGCACCCAC CGGACCTGCT TAGTCTAGGT TCGGACCGGT TGGAGGTACG TCCAGATCCC
121 AGATTCCGTC ACCAGGGGGC CCACGCCCTA CCCCCATGAT CCACTGGCCT CCCTGACGCT
181 ATAAGATTTA CAACCAGGTA AACTTAGTAG TAGATCCTAA TGCAGCGGGA TTTTTTTTCC
241 GTGGTCCCCG GGAGGAGGGG ACGCCGGTCC GGGCATCTCT GGTGCCCTGA CTGGGAGGGC
301 CGTCGGCCCC CCGCCCTTAG GCGGTGCATA CCCCTCCATA AACGGGCTGC CAGTCATGGG
361 **CG**

A.5 ‘Offset’ sequence

1 **GCGCGCGCGC GCGCGCGCGC GCGCGCGCGC GCGCGCGCGC GCGCGCGCGC GCGCGCGCGC**
61 TGCTGCCGAC ACCCGGGCAT AGTCTTAGGG GGGTCACTCG GGGGCACCCG CAGCCAACTT
121 GTCGGGTCCT GCCCGGCTGG TCCTCGGGCT AATGCAGCGA ATTGCCGAGG GCCCGGCCCC
181 GCGCAACGGA ACGTCTTTAG CTCCGGCAGG CAATTAAGGA CAACGCAAGC ATGGCGCATA
241 TAAACAGAGA AACGGGCGAA TGGACCTGTT CCGCCATACA GACGAAAGTA CATGGAGATT
301 ACCGAGCGAT CTACCAGATC GGCGACCATC TGTGAGGTAC TGGGGCCGAG AGGTAACTAC
361 **CG**

A.6 Kymographs and end-to-end distances for all hat-curve simulations

End-to-end distances, along with their cumulative averages for the last 400ns of each simulation, accompanied by kymographs for each simulation. Green lines are the start and end points of the plectoneme, blue points are bubble locations, the background heatmap is the bending over a window of 15bps.

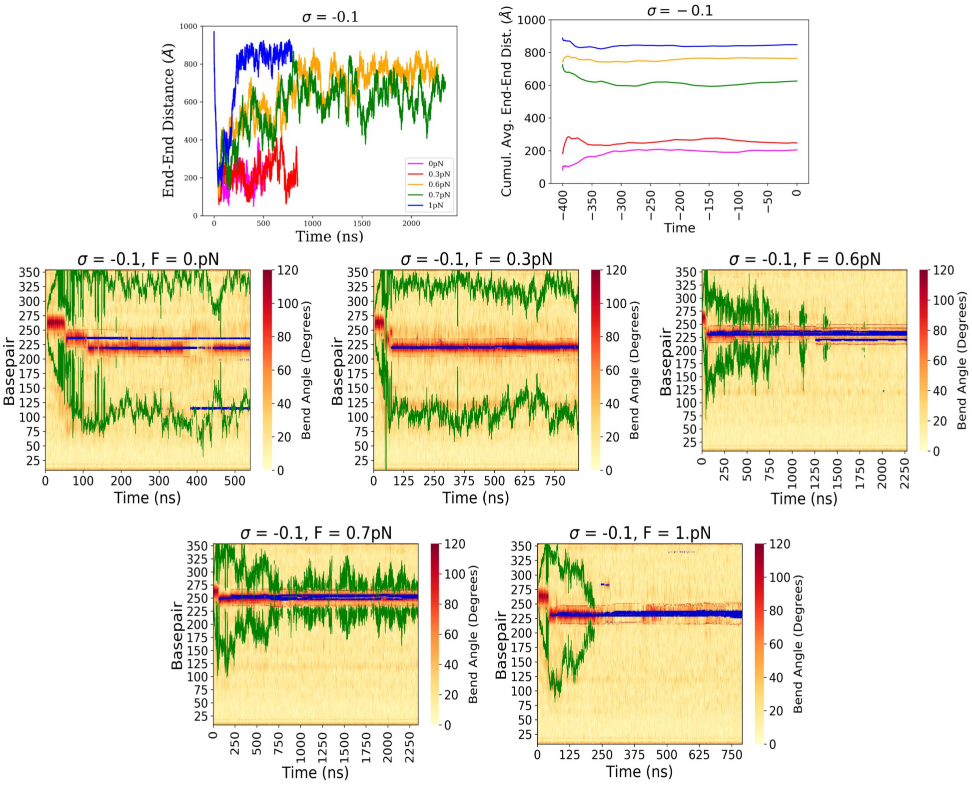


Figure A.1: Details of the dynamics of all $\sigma = -0.1$ systems in the hatcurve sequence.

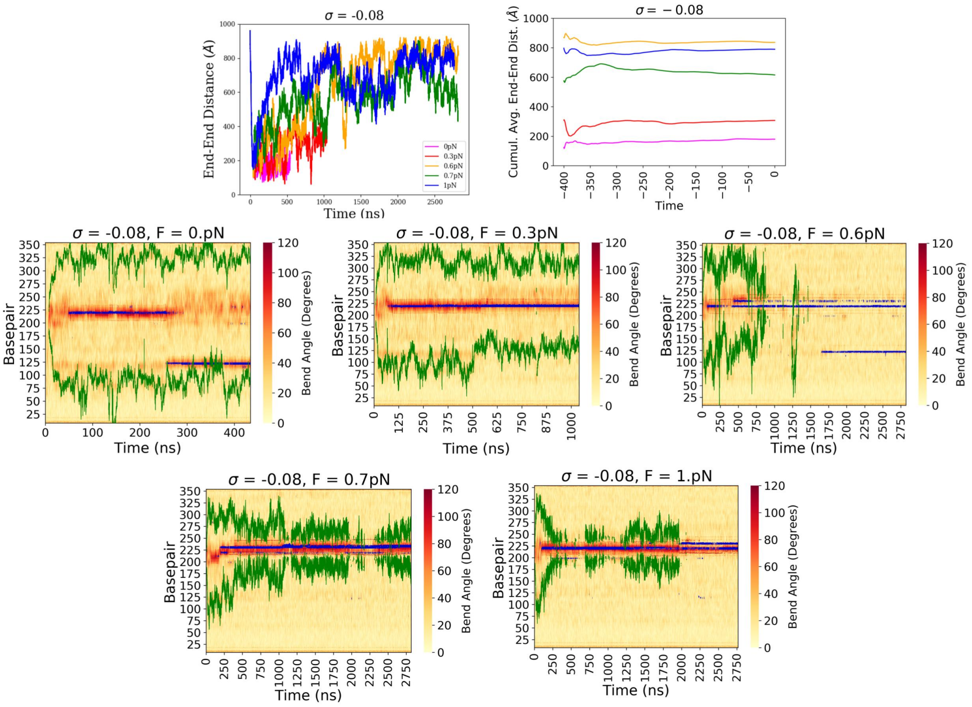


Figure A.2: Details of the dynamics of all $\sigma = -0.08$ systems in the hatcurve sequence.

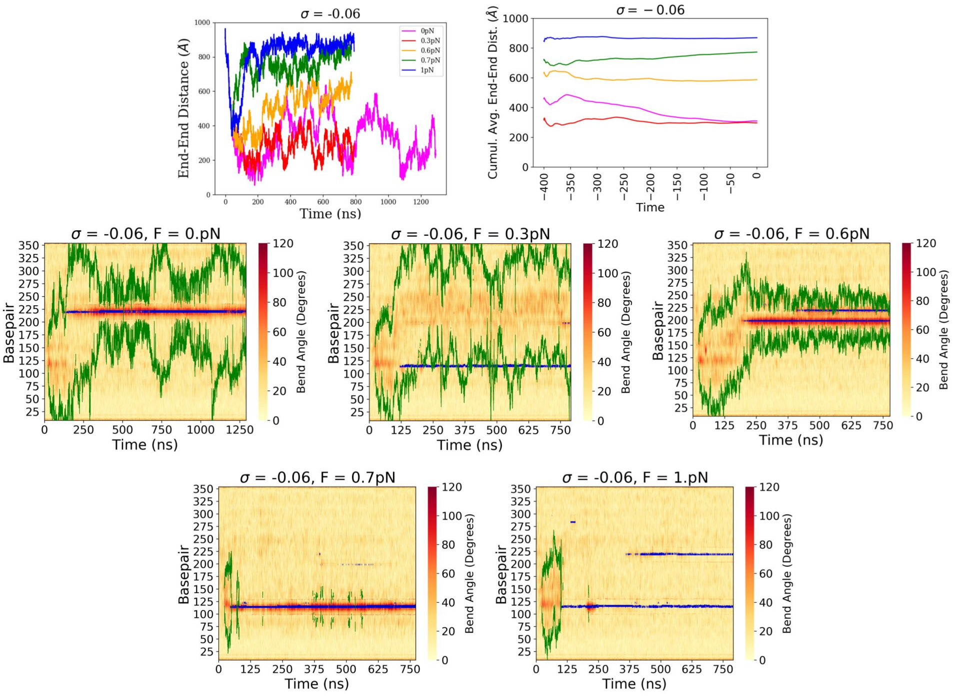


Figure A.3: Details of the dynamics of all $\sigma = -0.06$ systems in the hatcurve sequence.

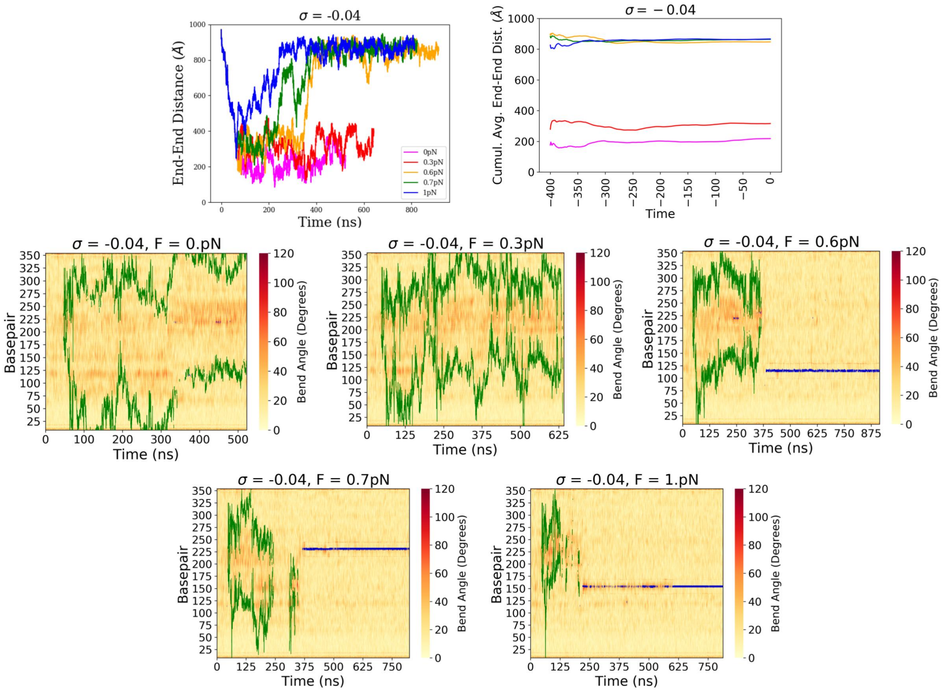


Figure A.4: Details of the dynamics of all $\sigma = -0.04$ systems in the hatcurve sequence.

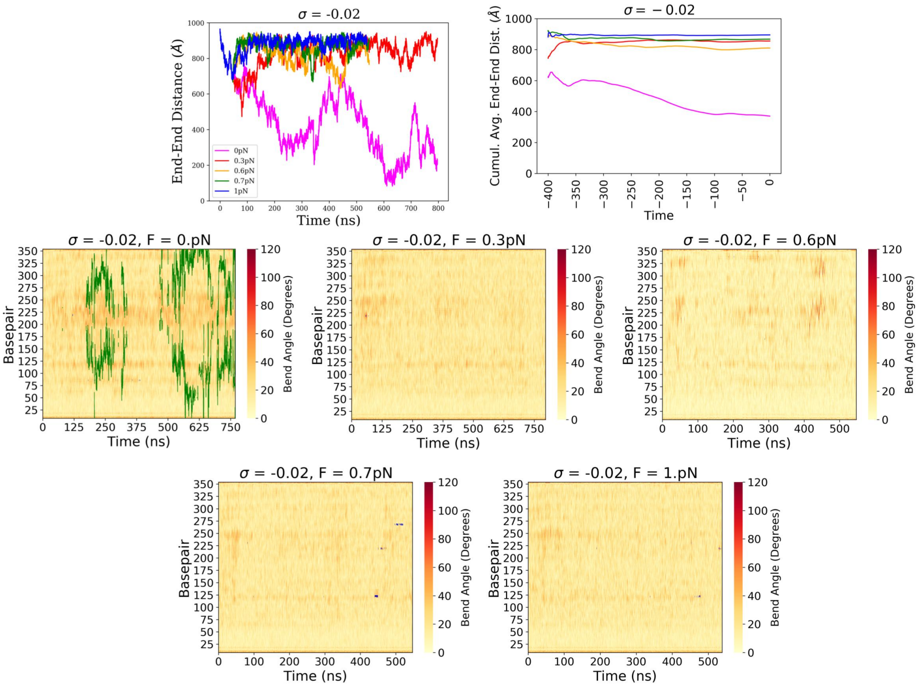


Figure A.5: Details of the dynamics of all $\sigma = -0.02$ systems in the hatcurve sequence.

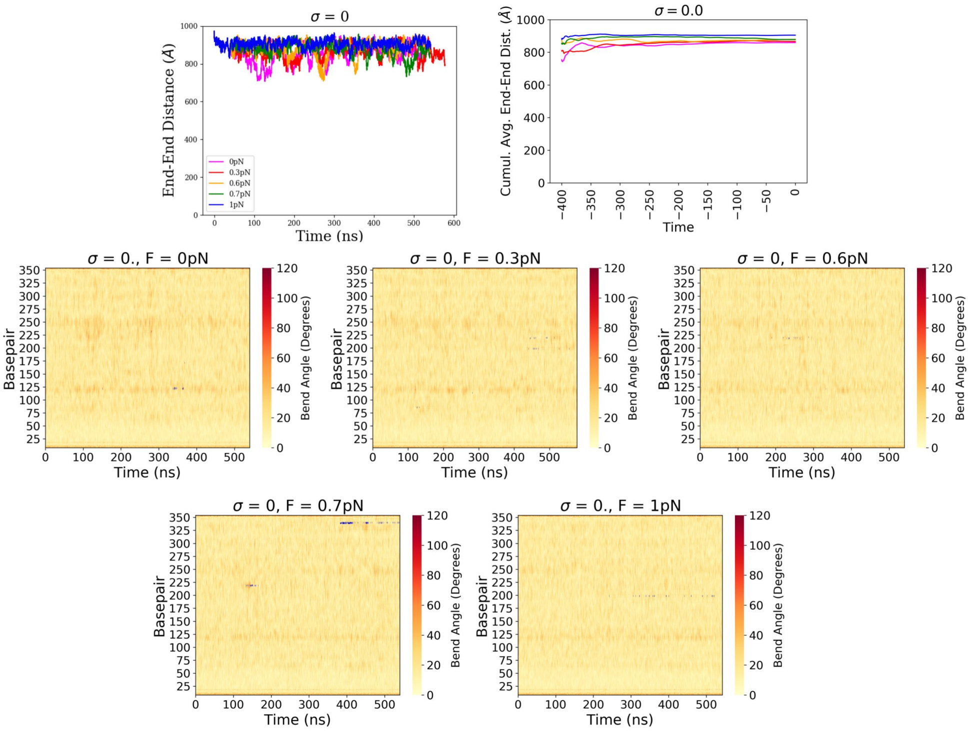


Figure A.6: Details of the dynamics of all $\sigma = 0$ systems in the hatcurve sequence.

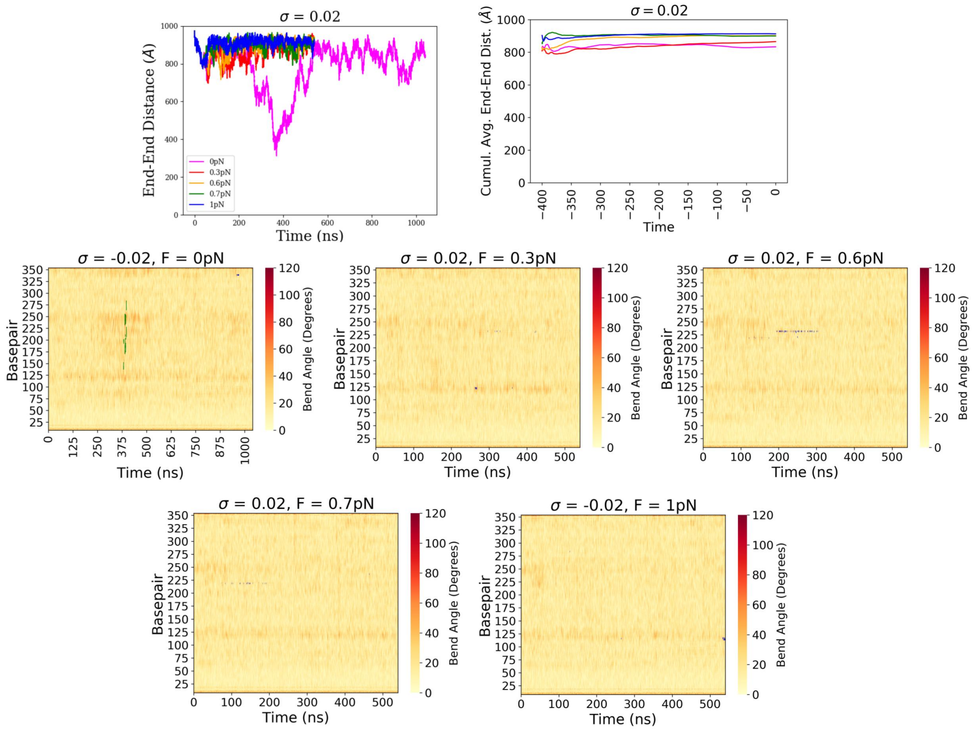


Figure A.7: Details of the dynamics of all σ = 0.02 systems in the hatcurve sequence.

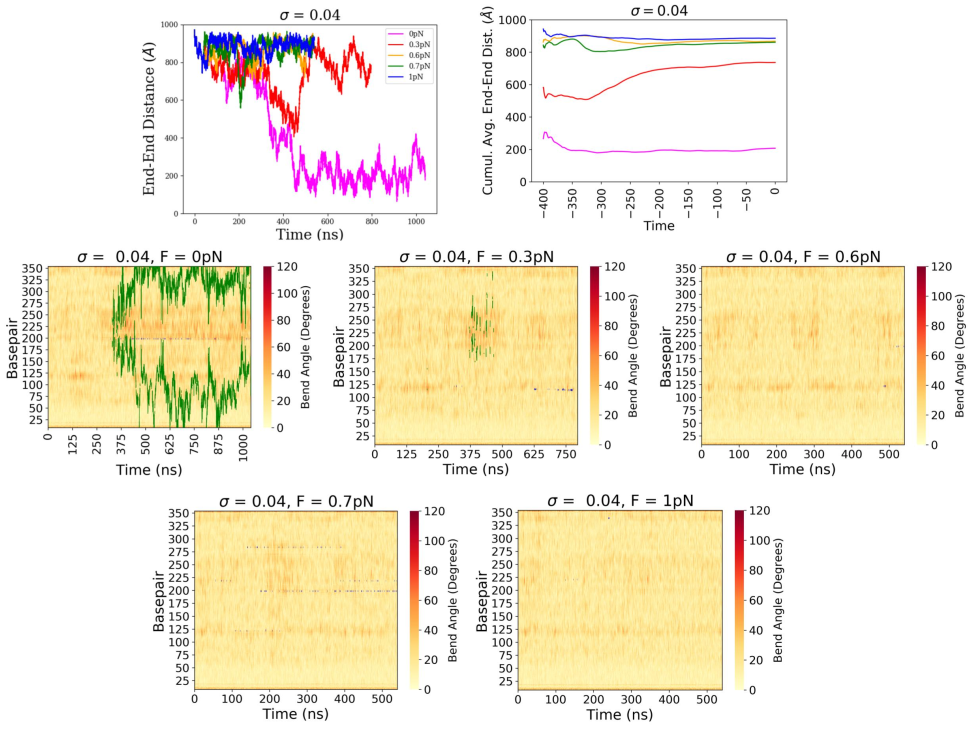


Figure A.8: Details of the dynamics of all σ = 0.04 systems in the hatcurve sequence.

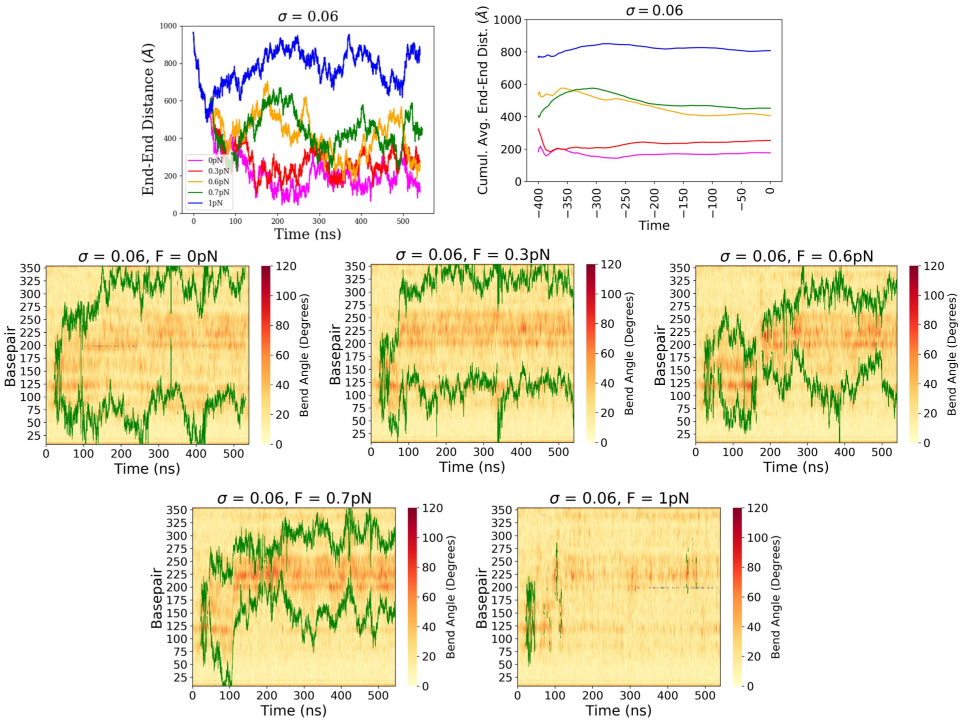


Figure A.9: Details of the dynamics of all σ = 0.06 systems in the hatcurve sequence.

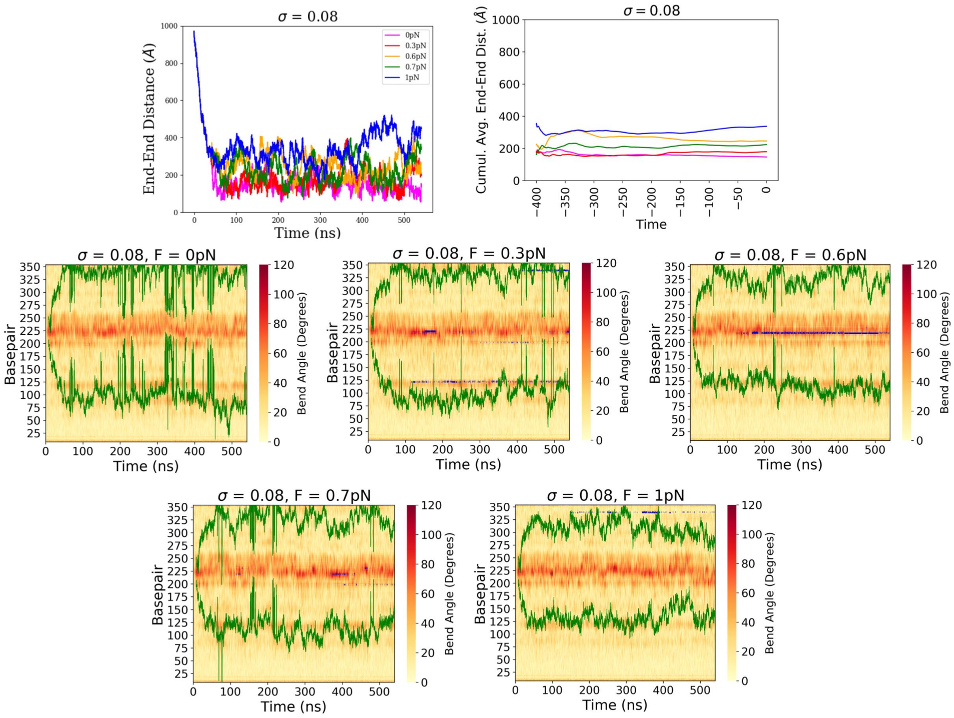


Figure A.10: Details of the dynamics of all σ = 0.08 systems in the hatcurve sequence.

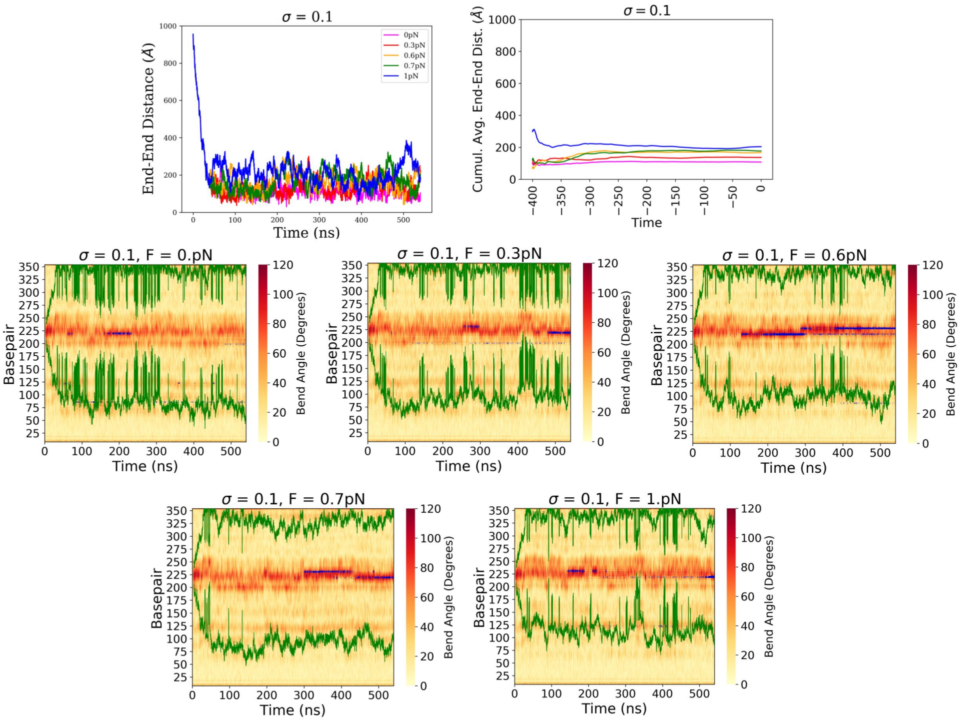


Figure A.11: Details of the dynamics of all $\sigma = 0.1$ systems in the hatcurve sequence.

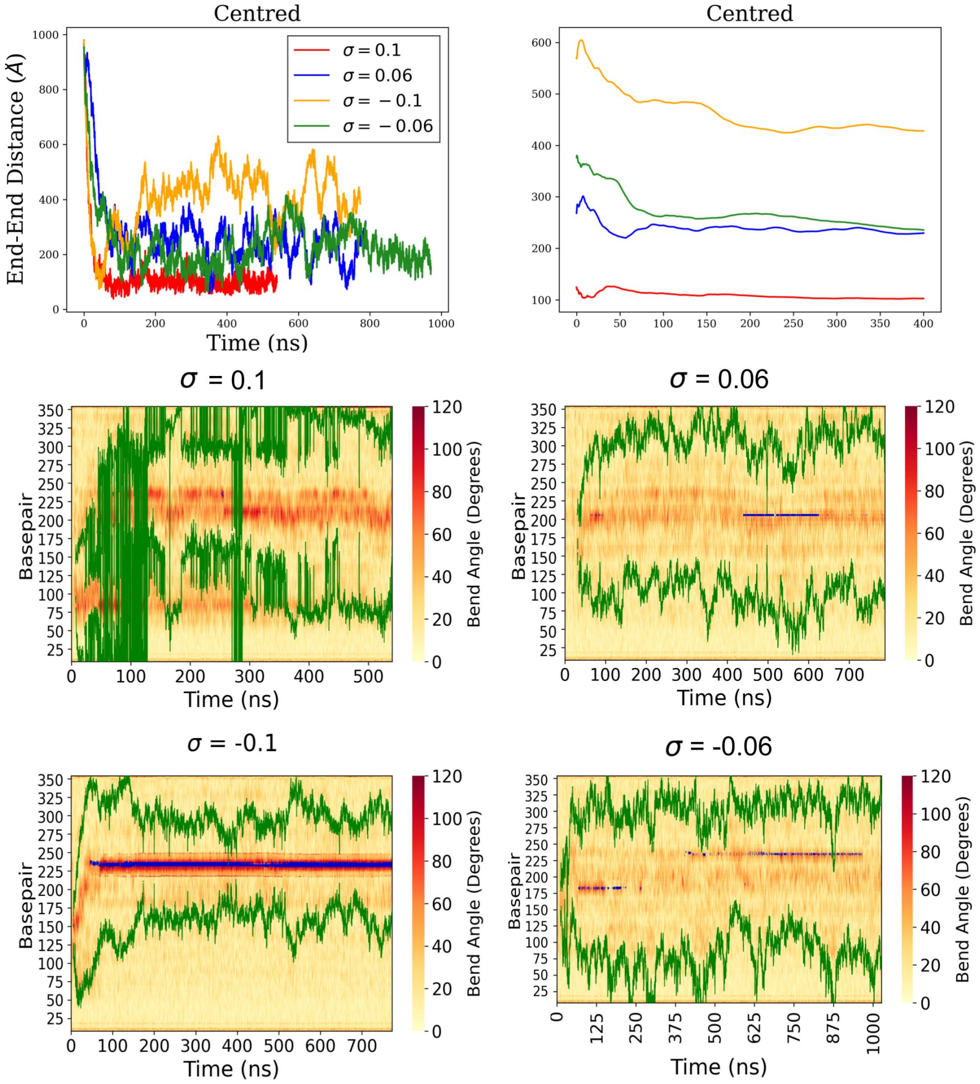


Figure A.12: Detailed dynamics of the ‘centred’ sequence, visualised in figure 4.3.

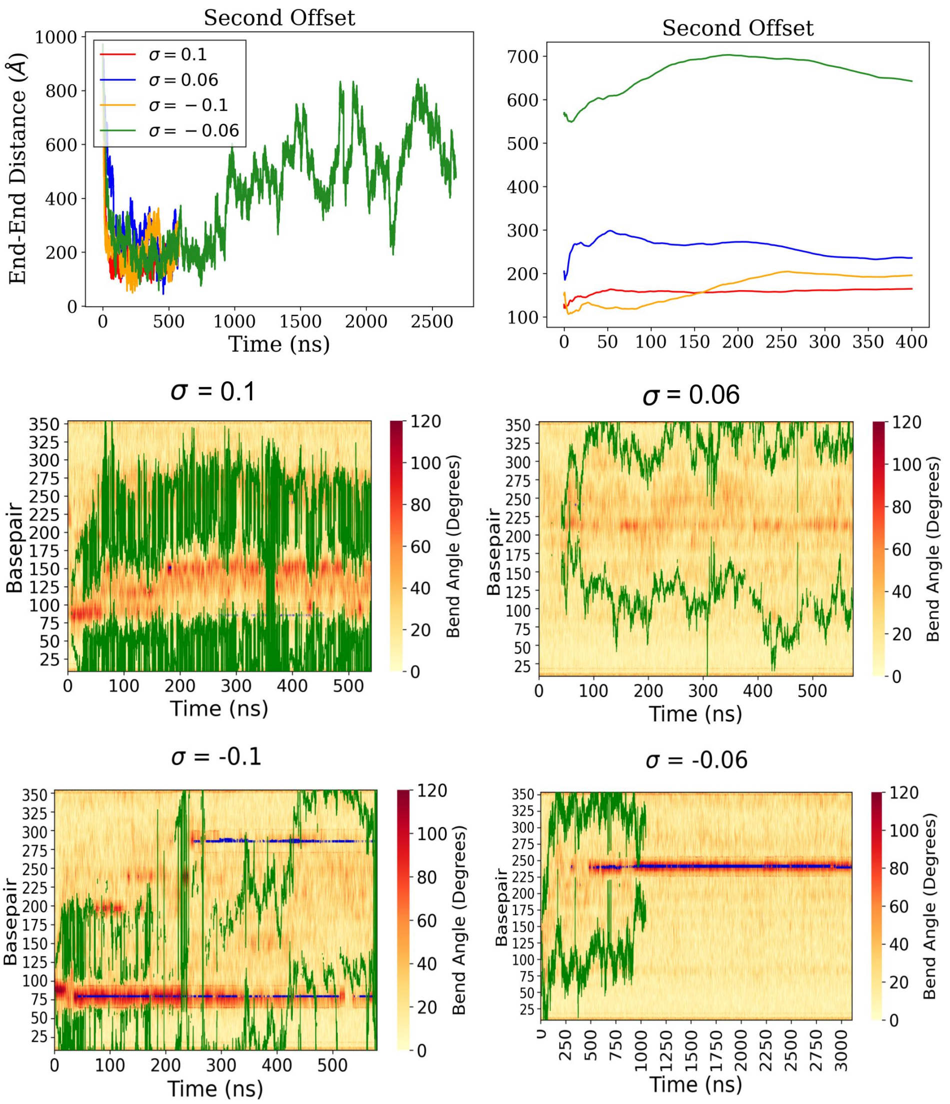


Figure A.13: Detailed dynamics of the ‘offset’ sequence, visualised in figure 4.3.

Appendix B

Code Listings

B.1 Python code to find bubbles within criteria

```
    """Python code to find bubbles from a ptraj nastruct output
Criteria for bubble is described in main text"""

import numpy as np
import pandas as pd


def group_consecutives(vals, step=1):
    """Return list of consecutive lists of numbers from vals (number list)."""
    run = []
    result = [run]
    expect = None
    for v in vals:
        if (v == expect) or (expect is None):
            run.append(v)
        else:
            run = [v]
            result.append(run)
        expect = v + step
    return result


def find_num_stds(array):
    avg = np.nanmean(array)
    std = np.nanstd(array)
    nums = np.absolute(np.divide(array - avg, std))
    return nums


min_bubble_size = 3
min_deg_std = 2
# Prepare bubble data

fname_dist = "file"
print("Reading " + fname_dist)
df_dist = pd.read_csv(fname_dist, delim_whitespace=True)
```

```
df_dist = df_dist.loc[:, ["#Frame", "Base1", "Propeller", "HB",
 "Opening", "Buckle"]]
base_max = df_dist.Base1.max()
df_dist = df_dist[
    df_dist.Base1 < base_max
]  # remove last row of each frame to make same size as bpstep dataframe
df_dist["Prop_stds"] = find_num_stds(df_dist.Propeller.to_numpy())
df_dist["Opstds"] = find_num_stds(df_dist.Opening.to_numpy())
df_dist["Buckstds"] = find_num_stds(df_dist.Buckle.to_numpy())
df_dist["Deg_of_bub"] = (
    df_dist["Prop_stds"] + df_dist["Opstds"] + df_dist["Buckstds"]
) / 3.0

df_bubs = df_dist[(df_dist["Deg_of_bub"] > 2) & (df_dist["HB"] == 0)]

bubbles_frames = (
    []
)  # Will store the frames in which the bubbles are stored,
#will be same size as array that stores start and end of bubbles
x_bubbles = []
for frame in df_bubs["#Frame"].unique():
    temp_df = df_bubs[df_bubs["#Frame"] == frame]
    temp_list = group_consecutives(temp_df["Base1"].to_numpy())
    for group in temp_list:
        groupsize = len(group)
        if groupsize < min_bubble_size:
            continue
        elif groupsize >= min_bubble_size:
            x_bubbles.append([group[0], group[-1]])
            bubbles_frames.append(frame)

xb = np.asarray(x_bubbles)
if xb.size == 0:
    print("No bubbles with size greater than" + str(min_bubble_size))
    df_SE = pd.DataFrame({"Frame": bubbles_frames, "Start": xb, "End": xb})
else:
    df_SE = pd.DataFrame({"Frame": bubbles_frames, "Start": xb[:, 0],
    "End": xb[:, 1]})
outfile_name = "Second_Offset_N006"
outfiles_loc = "/scratch/mb2055/New_seq_analysis/bublocs/"
outfile = outfiles_loc + outfile_name
df_SE.to_csv(outfile, sep="\t", index=False)
```

B.2 Bubble tracking python code

```
""" Take bubble loc files, finds bubbles that have a lifetime greater than
1ns and outputs them as a dataframe in the format
[frame],[start bp],[end bp]"""
import numpy as np
import pandas as pd


def group_consecutives(vals, step=1):
    """Return list of consecutive lists of numbers from vals
    (number list)."""
    run = []
    result = [run]
    expect = None
    for v in vals:
        if (v == expect) or (expect is None):
            run.append(v)
        else:
            run = [v]
            result.append(run)
        expect = v + step
    return result


class Bubble:
    def __init__(self):
        # This list will store the base pairs within this bubble object
        #at each frame
        self.bp = []
        # This list will store the frames in which the bubble is seen
        #(these will be continuous)
        self.frames = []
        # This will store the most recent frame in which this bubble has
        #been seen
        self.MR_frame = None
        self.firstframe = None

    # this function will append a given set of bases to the bp list
    #after they have been verified as part of the bubble
    # Only needs to store the start and end of the bubble
    def append_bp(self, list_bases):
        self.bp.append(list_bases)

    def append_frame(self, frame):
        self.frames.append(frame)
        self.MR_frame = frame

    def check_firstframe(self, frame):
        if self.firstframe is None:
            self.firstframe = frame
            self.MR_frame = frame

    def check_newframe(self, frame):
```

```
        if frame == self.frames[-1]:
            return False
        else:
            return True

    # Function to check whether two groups of numbers overlap
    def check_overlap(self, start1, end1, start2, end2):
        # print("Checking overlap for " + str(start1) + " to " + str(end1) +
        # " and " + str(start2) + " to " + str(end2))
        if max(start1, start2) - min(end1, end2) <= 0:
            # print("Overlap found")
            return True
        else:
            # print("No overlap found")
            return False

    # Function to check whether the input frame diectly follows the previous
    def check_consecutive(self, frame):
        if frame == self.frames[-1] + 1:
            return True
        else:
            return False

    # This function will be called in the main program,
    #returning true if a given bubble is part of the current object
    def is_in_bubble(self, bps, frame):
        if (
            self.check_overlap(self.bp[-1][0], self.bp[-1][-1], bps[0],
            bps[-1])
            and self.check_consecutive(frame)
            and self.check_newframe(frame)
        ):
            return True
        else:
            return False

    def firstframe_midpoint(self):
        return self.bp[0][-1] - np.floor((self.bp[0][-1] -
        self.bp[0][0]) / 2.0).astype(int)

    def initial_size(self):
        return self.bp[0][-1] - self.bp[0][0] + 1

    def lifetime(self):
        return len(self.frames)

    def output_info(self):
        bubble_start = self.bp[-1][0]
        bubble_end = self.bp[-1][-1]
        print(
            " Bubble Start: "
            + str(bubble_start)
            + " Bubble End: "
            + str(bubble_end)
```

```
            + " First Frame: "
            + str(self.firstframe)
            + " Last Frame: "
            + str(self.frames[-1])
            + " LifeTime: "
            + str(self.lifetime())
        )

    def avg_size(self):
        arr = np.array(self.bp)
        sizes = arr[:, -1] - arr[:, 0] + 1
        avg = np.average(sizes)
        SD = np.std(sizes)
        return avg, SD


def find_long_bubbles(filename):
    df = pd.read_csv(filename, delim_whitespace=True)
    df["Size"] = df["End"] - df["Start"] + 1
    firstframe = df.Frame.min()
    print("First frame in which a bubble is seen: ", firstframe)
    # indexes bubbles by their frame and size
    df = df.set_index(["Frame", "Size"])
    # Will store all bubbl#es in the system for all times
    list_of_bubbles = []
    # Will store all bubbles present in the previous frame
    prev_frame_bubbles = []
    # Loops over all frames
    for frame, mini_df in df.groupby(level=0):
        # No need to track bubbles in the first frame as their are no
        #members of the bubble class yet
        if frame == firstframe:
            for index, row in mini_df.iterrows():
                curr_bub = Bubble()
                bases = np.array([int(row.Start), int(row.End)])
                curr_bub.append_bp(bases)
                curr_bub.append_frame(frame)
                curr_bub.check_firstframe(frame)
                prev_frame_bubbles.append(curr_bub)
                list_of_bubbles.append(curr_bub)
        else:
            # Dummy list for temp storage of  bubble objects
            temp_storage = []
            # Loop over each row in the current frame
            for index, row in mini_df.iterrows():
                bases = np.array([int(row.Start), int(row.End)])
                # There has to be a way to concatentate these
                #3 loops in to 1, but cba finding it
                if any(bub.is_in_bubble(bases, frame) for bub in
                prev_frame_bubbles):
                    for bub in prev_frame_bubbles:
                        if bub.is_in_bubble(bases, frame):
                            bub.append_bp(bases)
                            bub.append_frame(frame)
```

```
                            temp_storage.append(bub)
                            break
                else:
                    curr_bub = Bubble()
                    bases = np.array([int(row.Start), int(row.End)])
                    curr_bub.append_bp(bases)
                    curr_bub.append_frame(frame)
                    curr_bub.check_firstframe(frame)
                    temp_storage.append(curr_bub)
                    list_of_bubbles.append(curr_bub)
            prev_frame_bubbles = temp_storage

    list_of_frames = []
    list_of_starts = []
    list_of_ends = []
    count = 0
    for bubble in list_of_bubbles:
        if bubble.lifetime() < 50:
            continue
        else:
            count += 1
            list_of_frames.append(bubble.frames)
            list_of_starts.append(np.array(bubble.bp)[:, 0])
            list_of_ends.append(np.array(bubble.bp)[:, 1])

    if count == 0:
        print("No bubble with lifetime greater than 1ns")
        list_of_frames = []
        list_of_starts = []
        list_of_ends = []
    else:
        list_of_frames = np.concatenate(list_of_frames).ravel()
        list_of_starts = np.concatenate(list_of_starts).ravel()
        list_of_ends = np.concatenate(list_of_ends).ravel()

    df_final = pd.DataFrame(
        {"Frame": list_of_frames, "Start": list_of_starts,
        "End": list_of_ends})
    df_final.sort_values(by=["Frame"], inplace=True)
    return df_final


# Second arg gives number of frames so that last 400ns can be found
def avg_sizes(filename, numframes):
    df = pd.read_csv(filename, delim_whitespace=True)
    # This is cheating, but should be OK - replaces DF with only last 400ns
    # In theory this ignores bubbles with a lifetime greater than 1ns
    # that vanish within 1ns of last 400, shouldn't matter
    df
    df["Size"] = df["End"] - df["Start"] + 1
    firstframe = df.Frame.min()
    print("First frame in which a bubble is seen: ", firstframe)
    # indexes bubbles by their frame and size
    df = df.set_index(["Frame", "Size"])
```

```
        # Will store all bubbl#es in the system for all times
        list_of_bubbles = []
        # Will store all bubbles present in the previous frame
        prev_frame_bubbles = []
        # Loops over all frames
        for frame, mini_df in df.groupby(level=0):
            # No need to track bubbles in the first frame as their are no
            #members of the bubble class yet
            if frame == firstframe:
                for index, row in mini_df.iterrows():
                    curr_bub = Bubble()
                    bases = np.array([int(row.Start), int(row.End)])
                    curr_bub.append_bp(bases)
                    curr_bub.append_frame(frame)
                    curr_bub.check_firstframe(frame)
                    prev_frame_bubbles.append(curr_bub)
                    list_of_bubbles.append(curr_bub)
            else:
                # Dummy list for temp storage of  bubble objects
                temp_storage = []
                # Loop over each row in the current frame
                for index, row in mini_df.iterrows():
                    bases = np.array([int(row.Start), int(row.End)])
                    # There has to be a way to concatentate these 3 loops
                    #in to 1, but cba finding it
                    if any(bub.is_in_bubble(bases, frame) for bub in
                    prev_frame_bubbles):
                        for bub in prev_frame_bubbles:
                            if bub.is_in_bubble(bases, frame):
                                bub.append_bp(bases)
                                bub.append_frame(frame)
                                temp_storage.append(bub)
                                break
                    else:
                        curr_bub = Bubble()
                        bases = np.array([int(row.Start), int(row.End)])
                        curr_bub.append_bp(bases)
                        curr_bub.append_frame(frame)
                        curr_bub.check_firstframe(frame)
                        temp_storage.append(curr_bub)
                        list_of_bubbles.append(curr_bub)
                prev_frame_bubbles = temp_storage
        print("here")
        list_of_avgs = []
        list_of_sds = []
        count = 0
        for bubble in list_of_bubbles:
            # CHANGE THIS NUMBER TO CHANGE MINIMUM LIFETIME (50 FRAMES = 1NS)
            if bubble.lifetime() < 50:
                continue
            else:
                count += 1
                avg, SD = bubble.avg_size()
                list_of_avgs.append(avg)
```

```
                list_of_sds.append(SD)

        if count == 0:
            print("No bubble with lifetime greater than 1ns")
            list_of_avgs = []
            list_of_sds = []

        df_final = pd.DataFrame({"Avg": list_of_avgs, "SD": list_of_sds})
        return df_final
```

B.3 Plectoneme finder python code

```
"""Code to read in projected.xyz from SerraLINE and find the first crossing point
outputs a txt file with Frame-Start-End"""
import numpy as np
import pandas as pd
from numpy import linalg as LA


# WILL ONLY WORK FOR 1D ARRAYS
# ENUM_START MARKS FIRST COUNTER (NOT ENUM_START+1)
def enumerate2D(array1, array2, enum_start=0):
    assert array1.shape == array2.shape, "Error - dimensions."
    for indexes, data in enumerate(array1):
        yield indexes + enum_start, data, array2[indexes]


seq_length = 362
# Read in xyz data from WrLINE, igorning blank rows and those with NaN
filename = "file"
titles = ["name", "x", "y", "z"]
df_coords = pd.read_csv(filename, delim_whitespace=True, names=titles)
df_coords.dropna(inplace=True)
df_coords.drop(columns=["name"], inplace=True)
# Convert dataframe to numpy array
coord_array = df_coords.to_numpy()
num_frames = int(coord_array.shape[0] / seq_length)
print("Total frames read = " + str(num_frames))
start_basepair = 0
min_separation = 40  # Min num of base pairs that can make up a plectoneme
max_dist = 3.0  # Max dist (nm) that the first and last bp in the plectoneme
# are separated by in the x-y plane
coord_array = coord_array.reshape(
    num_frames, seq_length, 3
)  # Reshape coordinate array to be frame-by-frame
start_end = []  # Will store [Frame, start of plect, end of plect]
for frame_num, frame in enumerate(coord_array[:], 1):
    for bp_number, coord in enumerate(
        frame[start_basepair : seq_length - min_separation], start_basepair
    ):
        found = False
        vects = np.subtract(frame[bp_number + min_separation :, :2], coord[:2])
```

```
        dists = LA.norm(vects, axis=1)
        for (
            end_counter,
            val,
        ) in reversed(list(enumerate(dists, min_separation))):
            if val <= max_dist:
                start_end.append([frame_num, bp_number, bp_number + end_counter])
                found = True
                break
        # Once the first bp fulfilling the criteria is found there is no
        #need to continue (we never see more than one plectoneme)
        if found:
            break

print("Plectonemes found")
start_end_df = pd.DataFrame.from_records(start_end,
columns=["Frame", "Start", "End"])
outfile = "out.txt"
start_end_df.to_csv(outfile, sep="\t", index=False)
```

B.4 Modified igor pro code

```
#pragma rtGlobals=3   // Use modern global access method and strict wave access.

Function PlectonemeCode(Swave, Twave, Wwave, Dwave, CwaveRR, CwaveTT,Pen40,
Pen48,Pen56,Pen64,Pen72,Pen80,Pen88,Pen96,Pen104,Pen112)

wave Swave    //the sequence of DNA we are considering
wave Twave, Wwave, Dwave, CwaveRR, CwaveTT
//Twist, Wedge, and Direction parameters; Covariance parameters for
//flexibility (roll-roll and tilt-tilt)
wave Pen40, Pen48, Pen56, Pen64, Pen72, Pen80, Pen88, Pen96, Pen104, Pen112
variable mm, ii, Letter1, Letter2, index, alpha_n, beta_n, omDiv2_n

variable SeqLength=dimsize(Swave,0)
variable rise=0.339
print "ASS"

make/d/o/n=(SeqLength,4) DNApath, DNApathMajorGroove
//These 2 paths will trace the center and the major groove
make/d/o/n=(SeqLength,2, 2) BasepairCovariance, LocalCovariance
//covariance matrices to estimate local stoffness along the sequence
make/d/o/n=(2, 2) BendRot, Covariance, LocalCov
//more matrices for stiffness calculation
DNApath=0
make/d/o/n=(SeqLength) CurvatureSequence
make/d/o/n=(SeqLength) Sequence_phase, Sequence_angle_energy, Sequence_angle_exp,
EndEffects
//Sequence_angle_exp will eventually store the weight assigned
//to a plectoneme at each position on the DNA
CurvatureSequence=0;
Sequence_phase=0
```

```

make/d/o/n=(4,4) T_n, Romega_n, Q_n, Rzplus, Rx, Rzminus, Minverse_tot, M_tot
make/d/o/n=4 StartPos, StartPosMG
StartPos={0,0,0,1}
StartPosMG={1,0,0,1}
make/d/o/n=4 OldPos

DNApath[0][]= StartPos[q]

OldPos=StartPos

Minverse_tot={{1,0,0,0},{0,1,0,0},{0,0,1,0},{0,0,0,1}}
M_tot=Minverse_tot
T_n={{1,0,0,0},{0,1,0,0},{0,0,1,-rise/2},{0,0,0,1}}

//This loop finds the 3D path of the relaxed DNA

Letter1=0;
For (ii=1; ii<SeqLength; ii+=1)
  //print Swave[ii]
  //If  ((Swave[ii])==char2num("A"))
  //   Letter2=0
  //Elseif ((Swave[ii])==char2num("C"))
  //   Letter2=1
  //Elseif ((Swave[ii])==char2num("G"))
  //   Letter2=2
  //Else
  //   Letter2=3
  //EndIf
  index = Swave[ii]
    //The index defines the current dinucleotide, AA=0, AC=1,...TT=15

  Sequence_phase[ii]=Sequence_phase[ii-1]+Twave[index]
    //This is used to measure how far around the DNA the major groove has
    rotated relative to the first base pair
  BendRot={{cos(Sequence_phase[ii]), sin(Sequence_phase[ii])},
    {-sin(Sequence_phase[ii]),cos(Sequence_phase[ii])}}  //Rotation matrix
  Covariance={{CwaveRR[index], 0},{0,CwaveTT[index]}}
    //The covariance matrix for the current basepair, expressed i
    n the coordinates of the current basepair
  MatrixOp/o CovRot = ( BendRot x Covariance x BendRot^t)
    //Rotating the covariance matrix so it will line up with its neighbors
  BasepairCovariance[ii][][]=CovRot[q][r]
    //Rotated covariance matrix at position is recorded

  //these next steps are outlined in the paper I sent
  omDiv2_n=Twave[index]/2

  Romega_n={{cos(omDiv2_n),sin(omDiv2_n),0,0},{-sin(omDiv2_n),
    cos(omDiv2_n),0,0},{0,0,1,0},{0,0,0,1}}
  alpha_n=Wwave[index]

  beta_n=Dwave[index]-pi/2

```

```
  Rzplus={{cos(beta_n),sin(beta_n),0,0},{-sin(beta_n),cos(beta_n),0,0},
     {0,0,1,0},{0,0,0,1}}
  Rx={{1,0,0,0},{0,cos(-alpha_n),sin(-alpha_n),0},{0,-sin(-alpha_n),
     cos(-alpha_n),0},{0,0,0,1}}
  Rzminus={{cos(-beta_n),sin(-beta_n),0,0},{-sin(-beta_n),cos(-beta_n),0,0}
     ,{0,0,1,0},{0,0,0,1}}

  MatrixOp/o Q_n = ( Rzminus x Rx x Rzplus )
  MatrixOp/o Minverse_n = Inv( T_n x Romega_n x Q_n x Romega_n x T_n)
  MatrixOp/o Minverse_new = (  Minverse_n x Minverse_tot)

  Minverse_tot=Minverse_new //Updating the total tranformation matrix

  MatrixOp/o CurrentPos = ( Minverse_tot^t x StartPos )
     //Calculate the coordinates of the current basepair

  MatrixOp/o CurrentPosMG = ( Minverse_tot^t x StartPosMG )

  DNApath[ii][]= CurrentPos[q]
  DNApathMajorGroove[ii][] = CurrentPosMG[q]

  Letter1=Letter2
Endfor
print DNApath
print "path calculated"

//Make the curvature calculation

variable CurveWindow  //must be even
Sequence_angle_energy=0
Sequence_angle_exp=0
Variable CircFrac=0.667
//We assume the plectoneme tip makes a 240 degree arc before joining the
//bulk plectoneme region
Variable BindLength=450
//experimentally, ~450 nt are bound to the surface at each end of the DNA
Variable AvePlecLength=0
Variable EnergyOffset
Variable Z1, Z2, Z3, Z4, Z5, Z6, Z7
//These are used to assign Boltzmann weights to bending in 8 possible directions
Variable Cnorm, E_base

Variable tanlength=10
//must be even. This is number of basepairs used to calculate the
//local tangent vectors.
Variable VectorMag, CosCurve, SinCurve


For (CurveWindow=40; CurveWindow<120; CurveWindow+=8)
  print CurveWindow
  LocalCovariance=BasepairCovariance
  smooth/B/DIM=0 (CurveWindow+1), LocalCovariance
     //Find covariance matrix over the curvature window

```

```
make/d/o/n=(SeqLength,3) TanVector, NormVector, CurveVector
make/d/o/n=(SeqLength) CurveMag, CurvePhase, HalfCurveMag, HalfCurvePhase
TanVector=0
NormVector=DNApathMajorGroove-DNApath
  //identifies the normal vector alligned with the major groove
CurveVector=0
make/d/o/n=3 CurrentTan, CurrentCurve, tP, tM, CurveCross, CurrentNorm
  //tangent vectors

// find the tan vectors over tanlength
For (ii=tanlength/2; ii<SeqLength-tanlength/2; ii+=1)
  CurrentTan[]=DNApath[ii+tanlength/2][p]-DNApath[ii-tanlength/2][p]
  VectorMag=sqrt(MatrixDot(CurrentTan, CurrentTan))
  TanVector[ii][]=CurrentTan[q]/VectorMag
      //Normalizes tangent vector to unit length
EndFor

// find the curvature vectors and values over curvewindow
For (ii=CurveWindow/2; ii<SeqLength-CurveWindow/2; ii+=1)
  tP[]=TanVector[ii+CurveWindow/2][p]         //plus tan vector
  tM[]=TanVector[ii-CurveWindow/2][p]         //minus tan vector
  CurrentCurve ={tP[1]*tM[2]-tP[2]*tM[1],tP[2]*tM[0]-tP[0]*tM[2],tP[0]*tM[1]
      -tP[1]*tM[0]}    //Cross product
  CurveVector[ii][]=CurrentCurve[q]
      //curvature vector is recorded at this position
//  CurveMag[ii]=acos(MatrixDot(tP, tM))
  CurveMag[ii]=asin(sqrt(MatrixDot(CurrentCurve, CurrentCurve)))
  CurrentCurve/=CurveMag[ii]
      //normalize the curvature vector to track direction

  //Calculates the phase angle of the curvature relative to
      //major groove at start of DNA
  CurrentTan[]=TanVector[ii][p]
  CurrentNorm[]=NormVector[ii][p]
  CosCurve=MatrixDot(CurrentCurve, CurrentNorm)
  tP=CurrentNorm  /
      /Only using these variables as placeholders b/c of short names,
      //otherwise cross product calculation makes a long line of code
  tM=CurrentCurve
  CurveCross={tP[1]*tM[2]-tP[2]*tM[1],tP[2]*tM[0]-tP[0]*tM[2],tP[0]*tM[1]
      -tP[1]*tM[0]}
  SinCurve=MatrixDot(CurveCross, CurrentTan)
  CurvePhase[ii]=mod(atan2(SinCurve,CosCurve)+Sequence_phase[ii],2*pi)
      //curvature phase is recorded.
EndFor

//EnergyOffset=25-CurveWindow*0.334*3/4.06
  //ORIGINAL - NOT USED IN THIS VERSION

For (ii=BindLength; ii<SeqLength-BindLength; ii+=1)
  if (CurveWindow == 40)
    EnergyOffset = Pen40[ii]
  Elseif (CurveWindow == 48)
    EnergyOffset = Pen48[ii]
```

```
    Elseif (CurveWindow == 56)
      EnergyOffset = Pen56[ii]
    Elseif (CurveWindow == 64)
      EnergyOffset = Pen64[ii]
    Elseif (CurveWindow == 72)
      EnergyOffset = Pen72[ii]
    Elseif (CurveWindow == 80)
      EnergyOffset = Pen80[ii]
    Elseif (CurveWindow == 88)
      EnergyOffset = Pen88[ii]
    Elseif (CurveWindow == 96)
      EnergyOffset = Pen96[ii]
    Elseif (CurveWindow == 104)
      EnergyOffset = Pen104[ii]
    Else
      EnergyOffset = Pen112[ii]
    EndIf
    EnergyOffset = EnergyOffset*-1.0
    Covariance=LocalCovariance[ii][p][q]
    BendRot={{cos(CurvePhase[ii]), sin(CurvePhase[ii])},
        {-sin(CurvePhase[ii]),cos(CurvePhase[ii])}}
    MatrixOp/o CovRot = ( BendRot x Covariance x BendRot^t)
        //local covariance matrix alligned to major groove
    BendRot={{cos(pi/4), sin(pi/4)},{-sin(pi/4),cos(pi/4)}}
    MatrixOp/o CovRot45 = ( BendRot x Covariance x BendRot^t)
        //local covariance matrix alligned 45deg to major groove
    Cnorm=CurveMag[ii]/(2*pi*CircFrac)

    E_base=CircFrac^2*3000/Curvewindow
    Z1=exp(-E_base/CovRot[0][0]*((1-Cnorm)^2)+EnergyOffset)
        //Bend in direction of curve
    Z2=exp(-E_base/CovRot[0][0]*((1+Cnorm)^2)+EnergyOffset)
        //Bend against curve
    Z3=exp(-E_base/CovRot[1][1]*(1-(Cnorm)^2)+EnergyOffset)
        //Bend perpendicular to curve
    Z4=exp(-E_base/CovRot45[0][0]*(sqrt(Cnorm^2/2+1)-Cnorm/sqrt(2))^2
        +EnergyOffset)
        //Bend at 45, 135, 225, and 315 degrees
    Z5=exp(-E_base/CovRot45[0][0]*(sqrt(Cnorm^2/2+1)+Cnorm/sqrt(2))^2
        +EnergyOffset)
    Z6=exp(-E_base/CovRot45[1][1]*(sqrt(Cnorm^2/2+1)-Cnorm/sqrt(2))^2
        +EnergyOffset)
    Z7=exp(-E_base/CovRot45[1][1]*(sqrt(Cnorm^2/2+1)+Cnorm/sqrt(2))^2
        +EnergyOffset)

    Sequence_angle_exp[ii]+=Z1+Z2+2*Z3+Z4+Z5+Z6+Z7
  endfor

endfor

Sequence_angle_energy=-ln(Sequence_angle_exp)
//Backing out the implied energy landscape from the summed Boltzmann weights

Duplicate/O Sequence_angle_exp,Sequence_angle_exp_smth
```

```
Smooth/E=2/F/B=32 300, Sequence_angle_exp_smth
//this command applies an approximate Gaussian smooth, though in reality
//it is 64 sequential boxcar smoothing operations
wavestats/q Sequence_angle_exp_smth
Sequence_angle_exp_smth/=V_avg
For (mm = 1; mm<SeqLength; mm+=1)

  print Sequence_angle_exp_smth[mm]
endfor


End
```

B.5 Energy penalty code

```
    """Python code to compute energy penalties for a given sequence length
outputs in a dataframe designed to be read directly by the modified
igor pro code"""

import numpy as np
import pandas as pd

# All contsants ignore the effects of salt conc.
c = 100  # Tors modulus in kbTnm
p = 45  # persistence length of dsDNA in nm
kbT = 4.114  # kbt at 298k
lamb = 0.01  # writhe per base pair
rise = 0.334  # DNA rise in nm
bind_length = 0  # Number of base pairs bound to bead/surface
#must be same as bind length in dekker code


def dens_to_alpha(sigma, len_seq):
    """Function to convert SC dens to number of turns"""
    rest_twist = 34.62
    lk0 = (rest_twist * len_seq) / 360
    return sigma * lk0


# Calculates extension in a zero torque system with given force
def extension_zero(force):
    term1 = (p * force) / kbT
    term2 = 1 / 32
    ovr = (term1 - term2) ** (-1 / 2)
    return 1 - ovr * 0.5


# Effective modulus Cs
def calc_eff_mod(force):
    outbrack = c / (4 * p * kbT)
    inbrack = (kbT / (p * force)) ** (1 / 2)
    return c * (1 - outbrack * inbrack)
```

```


# Calculate the pre-buckling energy
def prebuck_energy(F, N):
    Cs = calc_eff_mod(F)
    return 0.5 * (Cs / L0) * 2 * 2 * np.pi * np.pi * N * N


# Calculate the post-buckling energy
def postbuck_energy(N, Nb, LT, F):
    Cs = calc_eff_mod(F)
    term1 = 0.5 * (Cs / L0) * 2 * 2 * np.pi * np.pi * (N - Nb) * (N - Nb)
    term2 = ((Nb / lamb) + LT) * rise * extension_zero(F) * F
    return term1 + term2


# Estimate the ideal plectoneme size
# Could find analytically, but this finds to within 10bp so not worth the effort
def opt_plectsize(alpha, F, LT):
    nbs = np.arange(0, alpha - 0.1, 0.01)
    post_ens = postbuck_energy(alpha, nbs, LT, F)
    energy_min = np.min(post_ens)
    index_min = np.argmin(post_ens)
    plectsize_optimal = nbs[index_min]
    opt_plectsize_numbases = plectsize_optimal / lamb
    return opt_plectsize_numbases, energy_min


# Energy penalty of forming plect in given location
# Defined by the maximum possible plectoneme size for that location
def energy_penalty(location, optsize, energy_min, LT):
    maxsize_location = min(
        location - bind_length - LT / 2,
        length_seq - bind_length - LT / 2 - location
    )
    if optsize / 2 < maxsize_location:
        plect_size = optsize
    else:
        plect_size = maxsize_location * 2
    tot = postbuck_energy(alpha, plect_size * lamb, tipsize, F)
    pen = tot - energy_min
    if pen >= 0:
        return pen
    else:
        return 0


seqfile = "362_num"
with open(seqfile) as f:
    for count, line in enumerate(f):
        pass

length_seq = 300  # Set manually due to dummy base pairs
print("Counted " + str(length_seq) + " base pairs")
```

```python
sig = 0.08
# Minus one as tip is accounted for searately
alpha = dens_to_alpha(sig, length_seq) - 1
print("Sigma = " + str(sig) + ", equating to " + str(alpha) +
" additional turns.")
nbs = np.arange(0.5, 500)
L0 = length_seq * rise
F = 0.01  # force in pN

locations = np.arange(bind_length, length_seq - bind_length)
penalties = {}  # will store energy pens for all tip sizes
tip_sizes = np.arange(40, 120, 8)
for tipsize in tip_sizes:
    key = "Pen" + str(tipsize)
    enpens = []
    for i in range(bind_length):
        enpens.append(10000.0)
    cs = calc_eff_mod(F)
    opt_plectsize_numbases, en_min = opt_plectsize(alpha, F, tipsize)
    print("optimum Plectoneme size = " + str(opt_plectsize_numbases))
    for loc in locations:
        tot = energy_penalty(loc, opt_plectsize_numbases, en_min, tipsize)
        enpens.append(tot)

    for i in range(bind_length):
        enpens.append(10000.0)
    penalties[key] = enpens

df = pd.DataFrame(penalties)
filename = "orig_pens_" + str(sig) + "_" + str(F) + "pN.dat"
print(filename)
df.to_csv(filename, index=False)
```

Bibliography

[1] Mendel G 1865 *Verhandlungen des Naturforschenden Vereines in Brünn* 3–47

[2] Miescher F 1897 *Vogel, Vol1* 5–32

[3] Kossel A 1891 *Archive Physiological* 181–186

[4] Levene P 1919 *Journal of Biological Chemistry* **40** 415–424

[5] Avery O T, Macleod C M and McCarty M 1944 *Journal of Experimental Medicine* **79** 137–158

[6] ELSON D and CHARGAFF E 1952 *Experientia* **8** 143–145

[7] Franklin R E and Gosling R G 1953 *Nature* **171** 740–741

[8] Watson J D and Crick F H C 1953 *Nature* **171** 737–738

[9] Hoogsteen K 1963 *Acta Crystallographica* **16** 907–916

[10] CRICK F H 1958 *Symposia of the Society of Experimental Biology* **12** 138–163

[11] Dunham I, Kundaje A, Aldred S F, Collins P J, Davis C A, Doyle F, Epstein C B, Frietze S, Harrow J, Kaul R, Khatun J, Lajoie B R, Landt S G, Lee B K, Pauli F, Rosenbloom K R, Sabo P, Safi A, Sanyal A, Shoresh N, Simon J M, Song L, Trinklein N D, Altshuler R C, Birney E, Brown J B, Cheng C, Djebali S, Dong X, Ernst J, Furey T S, Gerstein M, Giardine B, Greven M, Hardison R C, Harris R S, Herrero J, Hoffman M M, Iyer S, Kellis M, Kheradpour P, Lassmann T, Li Q, Lin X, Marinov G K, Merkel A, Mortazavi A, Parker S C J, Reddy T E, Rozowsky J, Schlesinger F, Thurman R E, Wang J, Ward L D, Whitfield T W, Wilder S P, Wu W, Xi H S, Yip K Y, Zhuang J, Bernstein B E, Green E D, Gunter C, Snyder M, Pazin M J, Lowdon R F, Dillon L A L, Adams L B, Kelly C J, Zhang J, Wexler J R, Good P J, Feingold E A, Crawford G E, Dekker J, Elnitski L, Farnham P J, Giddings M C, Gingeras T R, Guigó R, Hubbard T J, Kent W J, Lieb J D, Margulies E H, Myers R M, Stamatoyannopoulos J A, Tenenbaum S A, Weng Z, White K P, Wold B, Yu Y, Wrobel J, Risk B A, Gunawardena H P, Kuiper H C, Maier C W, Xie L, Chen X, Mikkelsen T S, Gillespie S, Goren A, Ram O, Zhang X, Wang L, Issner R, Coyne M J, Durham T, Ku M, Truong T, Eaton M L, Dobin A, Tanzer A, Lagarde J, Lin W, Xue C, Williams B A, Zaleski C, Röder M, Kokocinski F, Abdelhamid R F, Alioto T, Antoshechkin I, Baer M T, Batut P, Bell I, Bell K, Chakrabortty S, Chrast J, Curado J, Derrien T, Drenkow J, Dumais E, Dumais J, Duttagupta R, Fastuca M, Fejes-Toth K, Ferreira P, Foissac S, Fullwood M J, Gao H, Gonzalez D, Gordon A, Howald C, Jha S, Johnson R, Kapranov P, King B, Kingswood C, Li G, Luo O J, Park E, Preall J B, Presaud K, Ribeca P, Robyr D, Ruan X, Sammeth M, Sandhu K S, Schaeffer L, See L H, Shahab A, Skancke J, Suzuki A M, Takahashi H, Tilgner H, Trout D, Walters N, Wang H, Hayashizaki Y, Reymond A, Antonarakis S E, Hannon G J, Ruan Y, Carninci P, Sloan C A, Learned K, Malladi V S, Wong M C, Barber G P, Cline M S, Dreszer T R, Heitner

S G, Karolchik D, Kirkup V M, Meyer L R, Long J C, Maddren M, Raney B J, Grasfeder L L, Giresi P G, Battenhouse A, Sheffield N C, Showers K A, London D, Bhinge A A, Shestak C, Schaner M R, Ki Kim S, Zhang Z Z, Mieczkowski P A, Mieczkowska J O, Liu Z, McDaniell R M, Ni Y, Rashid N U, Kim M J, Adar S, Zhang Z, Wang T, Winter D, Keefe D, Iyer V R, Zheng M, Wang P, Gertz J, Vielmetter J, Partridge E, Varley K E and Gasper C 2012 *Nature* **489** 57–74

[12] Ludwig M Z 2002 *Current Opinion in Genetics and Development* **12** 634–639

[13] Arnone M I and Davidson E H 1997 *Development* **124** 1851–1864

[14] Bartel D P 2018 *Cell* **173** 20–51

[15] Biologydictionarynet Editors 2016 Nucleotide biology dictionary URL `https://biologydictionary.net/nucleotide/`

[16] Whelan D R, Hiscox T J, Rood J I, Bambery K R, McNaughton D and Wood B R 2014 *Journal of the Royal Society Interface* **11** 20140454

[17] Flatters D, Young M, Beveridge D L and Lavery R 1997 *Journal of Biomolecular Structural Dynamics* **14** 757–765

[18] Wang A H, Quigley G J, Kolpak F J, Crawford J L, van Boom J H, van der Marel G and Rich A 1979 *Nature* **282** 680–686

[19] Lee M, Kim S H and Hong S C 2010 *Proceedings of the National Academy of Sciences* **107** 4985–4990

[20] Wittig B, Dorbic T and Rich A 1991 *Proceedings of the National Academy of Sciences of the United States of America* **88** 2259–2263

[21] Herbert A 2020 *European Journal of Human Genetics* **28** 114–117

[22] Lu X J and Olson W 2008 *Nature protocols* **3** 1213–27

[23] Lu X J, El Hassan M A and Hunter C A 1997 *Journal of Molecular Biology* **273** 668–680

[24] Lu X J, El Hassan M and Hunter C 1997 *Journal of Molecular Biology* **273** 681–691

[25] Chua E Y, Vasudevan D, Davey G E, Wu B and Davey C A 2012 *Nucleic Acids Research* **40** 6338–6352

[26] Sheinin M Y and Wang M D 2009 *Physical Chemistry Chemical Physics* **11** 4800–4803

[27] Bauer W R, Crick F H and White J H 1980 *Scientific American* **243** 100–113

[28] Corless S and Gilbert N 2016 *Biophysical Reviews* **8** 245–248

[29] Burns H and Minchin S 1994 *Nucleic Acids Research* **22** 3840–3845

[30] Ma J and Wang M D 2016 *Biophysical reviews* **8** 75–87

[31] Champoux J J 2001 *Annual Review of Biochemistry* **70** 369–413

[32] Nudler E 2012 *Cell* **149** 1438–1445

[33] Ding Y, Manzo C, Fulcrand G, Leng F, Dunlap D and Finzi L 2014 *Proceedings of the National Academy of Sciences* **111** 15402–15407

[34] Dorman C J and Dorman M J 2016 *Biophysical Review* **8** 89–100

[35] Herrero-Ruiz A, Martínez-García P M, Terrón-Bautista J, Millán-Zambrano G, Lieberman J A, Jimeno-González S and Cortés-Ledesma F 2021 *Cell Reports* **35** 108977

[36] Sobetzko P 2016 *Nucleic Acids Research* **44** 1514–1524

[37] Fulcrand G, Dages S, Zhi X, Chapagain P, Gerstman B S, Dunlap D and Leng F 2016 *Scientific Reports* **6** 19243

[38] Catanese D J, Fogg J M, Schrock D E, Gilbert B E and Zechiedrich L 2012 *Gene Therapy* **19** 94–100

[39] Hassan S, Keshavarz-Moore E and Ward J 2016 *Biotechnology Bioengineering* **113** 2064–2071

[40] Berger M A and Prior C 2006 *Journal of Physics A: Mathematical and General* **39** 8321

[41] Fuller F B 1971 *Proceedings of the National Academy of Sciences of the United States of America* **68** 815–819

[42] White J H 1969 *American Journal of Mathematics* **91** 693–728

[43] Călugăreanu G 1961 *Czechoslovak Mathematical Journal* **11** 588–625

[44] Sutthibutpong T, Harris S and Noy A 2015 *Journal of Chemical Theory and Computation* **11** 150507162320005

[45] Irobalieva R N, Fogg J M, Catanese D J, Catanese D J, Sutthibutpong T, Chen M, Barker A K, Ludtke S J, Harris S A, Schmid M F, Chiu W and Zechiedrich L 2015 *Nature Communications* **6** 8440

[46] Wlodawer A, Minor W, Dauter Z and Jaskolski M 2013 *The FEBS Journal* **280** 5705–5736

[47] Zhong E D, Lerer A K, Davis J H and Berger B 2021 Exploring generative atomic models in cryo-em reconstruction *pre-print*

[48] Murata K and Wolf M 2018 *Biochimica et Biophysica Acta (BBA) - General Subjects* **1862** 324–334

[49] Cheng Y, Grigorieff N, Penczek P A and Walz T 2015 *Cell* **161** 438–449

[50] Frueh D P, Goodrich A C, Mishra S H and Nichols S R 2013 *Current Opinions in Structural Biology* **23** 734–739

[51] Granéli A, Yeykal C, Krishnamurthy P and Greene E 2006 *Langmuir : the ACS journal of surfaces and colloids* **22** 292–9

[52] Collins B E, Ye L F, Duzdevich D and Greene E C 2014 *Methods in Cell Biology* **123** 217–234

[53] Dna curtains `http://www.thegreenelab.com/dnacurtains` accessed: 2022-08-18

[54] Nelson P C, Zurla C, Brogioli D, Beausang J F, Finzi L and Dunlap D 2006 *Journal of Physical Chemistry B* **110** 17260–17267

[55] Binnig G, Quate C F and Gerber C 1986 *Physical Review Letters* **56**(9) 930–933

[56] McPherson A, Malkin A, Kuznetsov Y and Plomp M 2001 *Acta crystallographica. Section D, Biological crystallography* **57** 1053–60

[57] Dufrêne Y F 2014 *mBio* **5** e01363–14

[58] Yoshua S, Watson G, Howard J A L, Velasco-Berrelleza V, Leake M and Noy A 2021 *Nucleic Acids Research* **49** 8684–8698

[59] Pyne A L B, Noy A, Main K H S, Velasco-Berrelleza V, Piperakis M M, Mitchenall L A, Cugliandolo F M, Beton J G, Stevenson C E M, Hoogenboom B W, Bates A D, Maxwell A and Harris S A 2021 *Nature Communications* **12** 1053

[60] Blumberg S, Gajraj A, Pennington M W and Meiners J C 2005 *Biophysical Journal* **89** 1272–1281

[61] Schafer D A, Gelles J, Sheetz M P and Landick R 1991 *Nature* **352** 444–448

[62] Bustamante C, Bryant Z and Smith S B 2003 *Nature* **421** 423–427

[63] Meng H, Bosman J, van der Heijden T and van Noort J 2014 *Biophysical Journal* **106** 1174–1181

[64] Strick T, Allemand J, Bensimon D, Bensimon A and Croquette V 1996 *Science* **271** 1835–7

[65] Brutzer H, Luzzietti N, Klaue D and Seidel R 2010 *Biophysical Journal* **98** 1267–1276

[66] Tempestini A, Cassina V, Brogioli D, Ziano R, Erba S, Giovannoni R, Cerrito M G, Salerno D and Mantegazza F 2012 *Nucleic Acids Research* **41** 2009–2019

[67] Charvin G, Bensimon D and Croquette V 2003 *Proceedings of the National Academy of Sciences* **100** 9820–9825

[68] Eeftens J, Bisht S, Kerssemakers J, Kschonsak M, Haering C and Dekker C 2017 *The EMBO Journal* **36** e201797596

[69] Revyakin A, Ebright R H and Strick T R 2004 *Proceedings of the National Academy of Sciences* **101** 4776–4780

[70] Bustamante C J, Chemla Y R, Liu S and Wang M D 2021 *Nature Reviews Methods Primers* **1** 25

[71] Nieminen T A, Knöner G, Heckenberg N R and Rubinsztein-Dunlop H 2007 Physics of optical tweezers *Laser Manipulation of Cells and Tissues* (*Methods in Cell Biology* vol 82) (Academic Press) 207-236

[72] Chemla Y R 2016 *Biopolymers* **105** 704–714

[73] Moffitt J R, Chemla Y R, Smith S B and Bustamante C 2008 *Annual Review of Biochemistry* **77** 205–228

[74] Gross P, Farge G, Peterman E J and Wuite G J 2010 *Methods Enzymol* **475** 427–453

[75] Heller I, Hoekstra T P, King G A, Peterman E J and Wuite G J 2014 *Chemical Reviews* **114** 3087–3119

[76] King G A, Burla F, Peterman E J G and Wuite G J L 2019 *Proceedings of the National Academy of Sciences* **116** 26534–26539

[77] Deufel C, Forth S, Simmons C R, Dejgosha S and Wang M D 2007 *Nature Methods* **4** 223–225

[78] La Porta A and Wang M D 2004 *Physical Review Letters* **92**(19) 190801

[79] Rittman M, Gilroy E, Koohy H, Rodger A and Richards A 2009 *Science progress* **92** 163–204

[80] Kratky O and Porod G 1949 *Recueil des Travaux Chimiques des Pays-Bas* **68** 1106–1122

[81] Doi M and Edwards S 1986 *The theory of polymer dynamics* (Oxford: Oxford University Press)

[82] Howard J 2001 *Mechanics of Motor Proteins and the Cytoskeleton* (Sunderland, MA: Sinauer Associates)

[83] Kumar S and Li M S 2010 *Physics Reports-review Section of Physics Letters* **486** 1–74

[84] Andrews S S and Arkin A P 2007 *Biophysical Journal* **93** 1872–1884

[85] Andrews S 2014 *Physical biology* **11** 011001

[86] Landau L and Lifshitz E 1980 Chapter xii - fluctuations *Statistical Physics (Third Edition)* (Oxford: Butterworth-Heinemann) 333-400

[87] Marko J F and Siggia E D 1995 *Macromolecules* **28** 8759–8770

[88] Moroz J D and Nelson P 1997 *Proceedings of the National Academy of Sciences* **94** 14418–14422

[89] Marko J F 2007 *Physical Review E, Statistical, Nonlinear, and Soft Matter Physics* **76** 021926

[90] Jacobson D R and Saleh O A 2016 *Nucleic Acids Research* **45** 1596–1605

[91] Marko J and Neukirch S 2013 *Physical Review E, Statistical, Nonlinear, and Soft Matter Physics* **88** 062722

[92] Brahmachari S, Dittmore A, Takagi Y, Neuman K C and Marko J F 2018 *Physical Review E, Statistical, Nonlinear, and Soft Matter Physics* **97**(2) 022416

[93] Benham C J and Bi C 2004 *Journal of Computational Biology* **11** 519–543

[94] Zhabinskaya D, Madden S and Benham C 2014 *Bioinformatics* **31** 421–422

[95] Cieplak P and Kollman P A 1988 *Journal of the American Chemical Society* **110** 3734–3739

[96] Steger G 1994 *Nucleic Acids Research* **22** 2760–2768

[97] Benham C J 1992 *Journal of Molecular Biology* **225** 835–847

[98] Ak P and Benham C J 2005 *PLoS Computational Biology* **1** e7

[99] Potaman V N, Bissler J J, Hashem V I, Oussatcheva E A, Lu L, Shlyakhtenko L S, Lyubchenko Y L, Matsuura T, Ashizawa T, Leffak M, Benham C J and Sinden R R 2003 *Journal of Molecular Biology* **326** 1095–1111

[100] Kim S H, Ganji M, Kim E, van der Torre J, Abbondanzieri E and Dekker C 2018 *eLife* **7** e36557

[101] Mitchell J S, Glowacki J, Grandchamp A E, Manning R S and Maddocks J H 2017 *Journal of Chemical Theory and Computation* **13** 1539–1555

[102] Balasubramanian S, Xu F and Olson W 2009 *Biophysical journal* **96** 2245–60

[103] Buckle A, Brackley C A, Boyle S, Marenduzzo D and Gilbert N 2018 *Molecular Cell* **72** 786–797.e11

[104] Smrek J, Garamella J, Robertson-Anderson R and Michieletto D 2021 *Science Advances* **7** eabf9260

[105] Knotts T A, Rathore N, Schwartz D C and de Pablo J J 2007 *Journal of Chemical Physics* **126** 084901

[106] Sambriski E J, Schwartz D C and de Pablo J J 2009 *Biophysical Journal* **96** 1675–1690

[107] Savelyev A and Papoian G A 2010 *Proceedings of the National Academy of Sciences* **107** 20340–20345

[108] Cao Q, Zuo C, Ma Y, Li L and Zhang Z 2011 *Soft Matter* **7** 506–514

[109] Farr S E, Woods E J, Joseph J A, Garaizar A and Collepardo-Guevara R 2021 *Nature Communications* **12** 2883

[110] Doye J P, Ouldridge T E, Louis A A, Romano F, Šulc P, Matek C, Snodin B E, Rovigatti L, Schreck J S, Harrison R M and Smith W P 2013 *Physical Chemistry Chemical Physics* **15** 20395–20414

[111] Matek C, Ouldridge T E, Doye J P and Louis A A 2015 *Science Reports* **5** 7655

[112] Ouldridge T E, Louis A A and Doye J P 2011 *Journal of Chemical Physics* **134** 085101

[113] Snodin B E, Randisi F, Mosayebi M, Šulc P, Schreck J S, Romano F, Ouldridge T E, Tsukanov R, Nir E, Louis A A and Doye J P 2015 *Journal of Chemical Physics* **142** 234901

[114] Doye J P K, Fowler H, Prešern D, Bohlin J, Rovigatti L, Romano F, Šulc P, Kui Wong C, Louis A A, Schreck J S, Engel M C, Matthies M, Benson E, Poppleton E and Snodin B E K 2020 *arXiv e-prints* arXiv:2004.05052

[115] Schreck J S, Ouldridge T E, Romano F, Šulc P, Shaw L P, Louis A A and Doye J P 2015 *Nucleic Acids Research* **43** 6181–6190

[116] Desai P R, Brahmachari S, Marko J F, Das S and Neuman K C 2020 *Nucleic Acids Research* **48** 10713–10725

[117] Mitchell J S, Laughton C A and Harris S A 2011 *Nucleic Acids Research* **39** 3928–3938

[118] Sutthibutpong T, Noy A and Harris S 2016 *Methods in Molecular Biology* **1431** 195–219

[119] Roe D R and Brooks B R 2020 *The Journal of Chemical Physics* **153** 054123

[120] Macke T and Case D 1998 *ACS Symposium Series* **682** 379–393

[121] Case D A, Cheatham T E, Darden T, Gohlke H, Luo R, Merz K M, Onufriev A, Simmerling C, Wang B and Woods R J 2005 *Journal of Computational Chemistry* **26** 1668–1688

[122] Lennard-Jones J 1924 *Proceedings of the Royal Society London A* **106** 463–477

[123] Rutkai G, Thol M, Span R and Vrabec J 2017 *Molecular Physics* **115** 1104–1121

[124] Son J H, Ahn H S and Cha J 2017 Lennard-jones potential field-based swarm systems for aggregation and obstacle avoidance *2017 17th International Conference on Control,Automation and Systems (ICCAS)* 1068-1072

[125] Cornell W D, Cieplak P, Bayly C I, Gould I R, Merz K M, Ferguson D M, Spellmeyer D C, Fox T, Caldwell J W and Kollman P A 1995 *Journal of the American Chemical Society* **117** 5179–5197

[126] Taylor J and Triggle D 2006 *Comprehensive medicinal chemistry II* (Elsevier Science) 87-118

[127] Liu Y P, Kim K, Berne B J, Friesner R A and Rick S W 1998 *The Journal of Chemical Physics* **108** 4739–4755

[128] Khoury G, Thompson J, Smadbeck J, Kieslich C and Floudas C 2013 *Journal of chemical theory and computation* **9** 5653–5674

[129] Nguyen H, Pérez A, Bermeo S and Simmerling C 2015 *Journal Chemical Theory and Computation* **11** 3714–3728

[130] Li J, Topaloglu R and Ghosh S 2021 *IEEE Transactions on Quantum Engineering* **PP** 1–1

[131] Gkeka P, Stoltz G, Barati Farimani A, Belkacemi Z, Ceriotti M, Chodera J D, Dinner A R, Ferguson A L, Maillet J B, Minoux H, Peter C, Pietrucci F, Silveira A, Tkatchenko A, Trstanova Z, Wiewiora R and Lelièvre T 2020 *Journal of Chemical Theory and Computation* **16** 4757–4775

[132] Jorgensen W, Chandrasekhar J, Madura J, Impey R and Klein M 1983 *Journal of Chemical Physics* **79** 926–935

[133] Mark P and Nilsson L 2001 *Journal of Physical Chemistry A* **105**

[134] Still W C, Tempczyk A, Hawley R C and Hendrickson T 1990 *Journal of the American Chemical Society* **112** 6127–6129

[135] Davis M E and McCammon J A 1990 *Chemical Reviews* **90** 509–521

[136] Boltzmann L 1868 *Studien über das Gleichgewicht der lebendigen Kraft zwischen bewegten materiellen Punkten* (*Cambridge Library Collection - Physical Sciences* vol 1) (Cambridge University Press) 49-96

[137] Holst M J and Saied F 1993 *Journal of Computational Chemistry* **14** 105–113

[138] Kirkwood J G 1934 *The Journal of Chemical Physics* **2** 351–361

[139] Tanford C and Kirkwood J G 1957 *Journal of the American Chemical Society* **79** 5333–5339

[140] Onufriev A V and Case D A 2019 *Annual Review of Biophysics* **48** 275–296

[141] Still W C, Tempczyk A, Hawley R C and Hendrickson T 1990 *Journal of the American Chemical Society* **112** 6127–6129

[142] Onufriev A, Bashford D and Case D 2000 *Journal of Physical Chemistry B* **104** 3712–3720

[143] Hawkins G, Cramer C and Truhlar D 1996 *Journal of Physical Chemistry* **100** 19824–19839

[144] Mongan J, Simmerling C, McCammon J A, Case D A and Onufriev A 2007 *Journal of Chemical Theory Computation* **3** 156–169

[145] Nguyen H, Roe D R and Simmerling C 2013 *Journal of Chemical Theory and Computation* **9** 2020–2034

[146] Choe J I and Kim B 2000 *Bulletin of the Korean Chemical Society* **21** 419–424

[147] Ryckaert J P, Ciccotti G and Berendsen H J 1977 *Journal of Computational Physics* **23** 327–341

[148] Berendsen H, Postma J, van Gunsteren W, DiNola A and Haak J 1984 *The Journal of Chemical Physics* **81** 3684

[149] Langevin P 1908 *Comptes rendus de l'Académie des Sciences* **146** 530–533

[150] Anandakrishnan R, Drozdetski A, Walker R and Onufriev A 2015 *Biophysical Journal* **108** 1153–1164

[151] Watson G, Velasco-Berrelleza V and Noy A 2022 *Chromosome Architecture: Methods and Protocols* **2476** 95–109

[152] Srinivasan J, Trevathan M W, Beroza P and Case D A 1999 *Theoretical Chemistry Accounts* **101** 426–434

[153] Onufriev A, Bashford D and Case D A 2004 *Proteins* **55** 383–394

[154] Henriksen N, Fenley A and Gilson M 2015 *Journal of Chemical Theory and Computation* **11** 150807142433009

[155] Altan-Bonnet G, Libchaber A and Krichevsky O 2003 *Phys Rev Lett* **90** 138101

[156] Fei J and Ha T 2013 *Proceedings of the National Academy of Sciences of the United States of America* **110** 17173–17174

[157] Roe D R and Cheatham T E 2013 *Journal of Chemical Theory and Computation* **9** 3084–3095

[158] Velasco Berrelleza V 2022 *Study of the flexibility of DNA using molecular dynamics simulations* Ph.D. thesis University of York York, United Kingdom

[159] Watson G D 2021 *Atomistic simulation of interactions between DNA and integration host factor* Ph.D. thesis University of York York, United Kingdom

[160] Baranello L, Levens D, Gupta A and Kouzine F 2012 *Biochimica et biophysica acta* **1819** 632–8

[161] Dorman C J 2019 *BMC Molecular and Cell Biology* **20** 26

[162] Bauer W R and Benham C J 1993 *Journal of Molecular Biology* **234** 1184–1196

[163] Sarkar R and Rybenkov V V 2016 *Frontiers in Physics* **4**

[164] Agarwal R and Duderstadt K E 2020 *Nature Communications* **11** 4714

[165] Sinden R, Carlson J and Pettijohn D 1980 *Cell* **21** 773—783

[166] Muskhelishvili G and Travers A 2016 *Biophysical reviews* **8** 5–22

[167] Quail T, Golfier S, Elsner M, Ishihara K, Murugesan V, Renger R, Jülicher F and Brugués J 2021 *Nature Physics* **17** 1007–1012

[168] Zhou Z, Miller H, Baumann C and Leake M 2016 *Biophysical Journal* **110** 499a

[169] Shepherd J W, Greenall R J, Probert M, Noy A and Leake M 2020 *Nucleic Acids Research* **48** 1748–1763

[170] Adamcik J, Jeon J H, Karczewski K J, Metzler R and Dietler G 2012 *Soft Matter* **8** 8651–8658

[171] Fogg J M, Judge A K, Stricker E, Chan H L and Zechiedrich L 2021 *Nature Communications* **12** 5683

[172] Cortini R, Lee D and Kornyshev A 2012 *Journal of physics. Condensed matter* **24** 162203

[173] Roth E, Azaria A, Girshevitz O, Bitler A and Garini Y 2018 *Nano Letters* **18** 6703–6709

[174] Crothers D M and Spatz H C 1971 *Biopolymers* **10** 1949–1972

[175] Marko J and Neukirch S 2012 *Physical review. E, Statistical, nonlinear, and soft matter physics* **85** 011908

[176] van Loenhout M T J, de Grunt M V and Dekker C 2012 *Science* **338** 94–97

[177] Wan B and Yu J 2022 *Biophysical Journal* **121** 658–669

[178] Sharma R, Patelli A S, De Bruin L and Maddocks J H 2023 *Journal of Molecular Biology* 167978

[179] De Bruin L and Maddocks J H 2018 *Nucleic Acids Research* **46** 5–10

[180] Gore J, Bryant Z, Nöllmann M, Le M U, Cozzarelli N R and Bustamante C 2006 *Nature* **442** 836–839

[181] Bryant Z, Stone M D, Gore J, Smith S B, Cozzarelli N R and Bustamante C 2003 *Nature* **424** 338–341

[182] Zakharova S S, Jesse W, Backendorf C, Egelhaaf S U, Lapp A and van der Maarel J R 2002 *Biophysical Journal* **83** 1106–1118

[183] Boles T C, White J H and Cozzarelli N R 1990 *Journal of Molecular Biology* **213** 931–951

[184] Ubbink J and Odijk T 1999 *Biophysical Journal* **76** 2502–2519

[185] van Loenhout M T J, de Grunt M V and Dekker C 2012 *Science* **338** 94–97

[186] Kolbeck P J, Vanderlinden W, Gemmecker G, Gebhardt C, Lehmann M, Lak A, Nicolaus T, Cordes T and Lipfert J 2021 *Nucleic Acids Research* **49** 5143–5158

[187] Lim W, Randisi F, Doye J P K and Louis A A 2022 *Nucleic Acids Research* **50** 2480–2492

[188] Kim E, Gonzalez A M, Pradhan B, van der Torre J and Dekker C 2022 *Nature Structural & Molecular Biology* **29** 719–727

[189] Eastman P, Swails J, Chodera J D, McGibbon R T, Zhao Y, Beauchamp K A, Wang L P, Simmonett A C, Harrigan M P, Stern C D, Wiewiora R P, Brooks B R and Pande V S 2017 *PLoS Comput Biol* **13** e1005659