A Molecular Dynamics Investigation of the Interactions between DNA and Other Biological Molecules

Bor-yior Yee

Doctor of Philosophy

University of York
Physics

December 2019

Abstract

We used molecular dynamics simulations to investigate interactions between DNA and three antineoplastic drugs from the anthracycline family, viz. daunomycin, doxorubicin and idarubicin. This encompassed three important aspects of DNA/drug interactions, viz. conformational perturbations, dynamics and energetics.

First, we investigated the structural perturbations caused by intercalation of the drugs into DNA. We found, using the software PyralleX which simulates X-ray diffraction patterns, that the DNA tends to change into an intermediary conformation between canonical forms. Daunomycin, among the three drugs, caused the greatest conformational shift in the DNA. Structural perturbations were shared with the base pairs adjacent to the intercalation sites.

Second, we studied the effects of groove-binding on the supercoiling behaviour of closed-circular DNA using the coarse-grained force field SIRAH. In the case without drugs, we saw an accelerating upward trend in the supercoiling rate with the salinity of the solution. However, with the drugs, supercoiling was found to retard in hypernatremic environments. Anthracyclines were found to form multilayer complex systems within themselves, which were capable of bridging across two segments of DNA and stabilising the DNA structure.

Third, we calculated the free energy changes associated with the intercalation of anthracyclines into DNA, using hybrid coarse-grained/all-atom models for simulation and the novel "extended-system adaptive biasing force" method for analysis. The free energy changes of intercalation of daunomycin and doxorubicin were calculated theoretically to be (-7.27 ± 0.23) kcal mol^{-1} and (-8.61 ± 0.33) kcal mol^{-1} respectively, which are in close agreement with previous experimental data. It was found that the calculated free energy change of idarubicin's intercalation is (-7.75 ± 0.17) kcal mol^{-1}, i.e. between those of the previous two drugs. This work has demonstrated a new way of evaluating free energy changes of interactions, which could help in speeding up time-consuming drug discovery processes.

Contents

List of Tables

List of Figures

Acknowledgements

"The whole concept of the self-made man or woman is a myth. None of us can make it alone," said Arnold Schwarzenegger, the film star and politician, once at an interview. This is true indeed, especially for someone travelling on such a long journey (or fighting a long battle as some might call it). I am lucky enough to have crossed paths with so many good mentors, companions and comrades at different points in this journey, and I would like to express my deepest and heartfelt gratitude to them here.

First of all, I would like to say thank-you to my fantastic supervisor, Dr. Robert Greenall, without whom none of what follows in this thesis would have happened. Firstly, thanks for all the freedom you have given me to do what I wanted to in these four years. You have taught me that science works hand-in-hand with imagination, and that it is good sometimes to think (if not crazily) out of the box. Secondly, thanks also for pulling me back from time to time when those thoughts have really gone crazily far from being practical. Thirdly, thank you for your ever-so-helpful feedback, especially on my written documents [1]. Not only do your comments shape me into a better scientific writer, they make me a better scientist day by day. Thank you so much Robert, and happy retirement!

Secondly, I would like to thank the lovely people at the Physics Department at York who have kindly provided me with help in any form — from teaching me in courses, to constantly showing care by asking how things are going, or even just bumping into me at coffee but somehow managing to steer conversations into deeply physics-y ones. I would like to thank especially Profs. Matt Probert and Rex Godby, and Dr. Phil Hasnip, for the numerous constructive conversations we had, and for the opportunities they have given me. I have learned so much working as a teaching assistant in their modules. I would also like to thank Jacob and Ed for helping me out countless times with computer problems, and to Jack for being so resourceful and so helpful with questions about biophysics. Lastly, I would like to thank Manuel, Andrea and Marta for all the wonderful memories we created together.

I would like to thank everyone from the Newman's Group at St. Wilfrid's Church, for creating such a nice community and making York far more than a place where I study. I want to specially thank Rachel, Irene and George for being such great companions, who have always been there for me and with me through ups and downs [2], and Molly and Phillip for being excellent social (and hiking) organisers who have made my PhD journey a much healthier one, both mentally and physically.

Last but not least, the most important thank-you should be reserved for the most important people — my parents. Thank you for raising me up, for your constant support and for your love. Without the two of you I would not even have existed, let alone seeing all these miracles happen in my life.

[1] I now truly understand why so many people in the Department said that Robert sometimes wrote longer comments than students their works...

[2] "Wit beyond measure is man's greatest treasure" as the motto says, but being Ravenclaws isn't just about wits!

Declarations

I declare that the work presented in this thesis, except where otherwise stated, is based on my own research and has not been submitted previously for a degree in this or any other university. All sources have been acknowledged as references.

Signed

Bor-yior Yee

Chapter 1

Introduction

1.1 DNA as a biological molecule

1.1.1 A brief early history (1869-1953) of DNA

DNA, an acronym for *deoxyribonucleic acid*, is a biological molecule which exists in the nuclei of living cells. It was first discovered and isolated in 1869 by the Swiss physician and biologist Friedrich Miescher [59, 60, 202]. However, at the time of the discovery, the great significance of this molecule to all life-forms had been ignored, even by the discoverer himself. In effect, for nearly a century, people had thought that the entity which bears genetic information was some sort of protein.

Amongst the pioneers in the discovery of DNA, especially those in the first half of the 20th century, the name Phoebus Levene should be mentioned. A physician-turned chemist, he made several important observations from his studies. For example, he was the first person to discover the "phosphate-sugar-base" order of the nucleic acid components whose name, "nucleotide", was also coined by himself. Moreover, he was also the first person to discover the sugars in RNA (ribose, discovered in 1909) and in DNA (deoxyribose, discovered in 1929). One of the most significant propositions he had made, was the so-called "tetranucleotide hypothesis" in which he hypothesised that a strand of nucleic acid (henceforth "NA"), regardless of DNA or RNA, must be composed of a repetitive nucleotide sequence of length exactly four. Albeit it was proven to be wrong, though much later, his hypothesis had laid a cornerstone in our knowledge of the existence of four types of nucleobases in all NAs.

In 1928, Griffith carried out an experiment [119] by applying two strains of *Streptococcus pneumoniae* — the bacterium which is the major cause of pneumonia, one of smooth and virulent type III-S and the other rough and non-virulent type II-R, in mice. He divided his experiment into four similar tests. In the first two sets of tests, he injected pure bacterial solution into the mice. The mice which were infected by the virulent type were all killed eventually, whereas those by the non-virulent type survived. Griffith noticed that whilst the virulent strain has a self-protection mechanism by wrapping itself in a polysaccharide capsule (hence the name "*smooth*"), the non-virulent type does not (hence "*rough*"). Without

the self-protection mechanism, the rough strain was defeated by the immune system of the mice.

After carrying out the control experiments, Griffith killed the smooth strain bacteria by heat and applied them to living mice. Because the bacteria were all dead prior to the injection, they did not have any effects on the mice, and the test subjects all remained alive. He then further added the remainder of his heat-killed bacteria into the II-R type, and infected the mice with the mixture. To his surprise, all the mice in this batch were killed. He concluded that the type II-R bacteria, under the influence of the dead III-S, had been transformed into their counterpart according to a "transforming principle" which is inherent in III-S type but not II-R.

This "transforming principle" was eventually unveiled in 1944, when Avery, MacLeod and McCarty performed a ground-breaking experiment [9]. They killed, with heat, the types II-R and III-S pneumococci, the same strains Griffith had used, and extracted some components using saline. They were able to narrow down the potential cause (i.e. the active portion extracted) of such transformation to only the polysaccharides, some lipids, some proteins, RNA and DNA. Using biochemical methods, namely attempting to break down these components using suitable enzymes, only deoxyribonucleodepolymerase (an enzyme which could decompose DNA) but not trypsin, chymotrypsin or ribonuclease (protein- or RNA-breakers) could prevent the transformation from happening.

Through this experiment, Avery and his colleagues hypothesised that DNA is the substance which causes the transformation of bacteria. They further generalised this idea and suggested that not only bacteria, but also viruses and higher organisms may have DNA as their hereditary material [9].

Less than a decade later, in 1952, Hershey and Chase confirmed Avery's hypothesis that DNA, not protein, is the genetic material in viruses, through a series of experiments they conducted [135]. Hershey and Chase used the T2 bacteriophage as their subject, and labelled the protein shell of some phages with the radioactive sulphur-35 (^{35}S) isotope, and the DNA contents of the remaining phages with phosphorus-37 (^{37}P) isotope, with the reason being the presence of these two elements in the respective components. The labeled viruses were then allowed to infect bacteria. The infected bacteria and the progeny (i.e. the phages) were separated by agitation in a blender. It was found out that the group of progeny with ^{35}S remained labelled, whereas that with ^{37}P became clean. Since it was known, prior to this experiment, that viruses infect their targets by lysing their cells and injecting their own genetic material, Hershey and Chase concluded from their study, that DNA, not proteins, is the genetic material of viruses.

Nearly at the same time, Chargaff and his colleagues conducted a set of experiments on DNA in the sperm of sea urchins [39]. They shone UV light on their samples and noted the absorption. They discovered, through the computation of the molar ratios of adenine (A) to guanine (G) and of thymine (T) to cytosine (C), that in three of their four samples, the ratios were very similar. However, for the ratios of A:T and C:G, all the four samples had both numbers very close to unity. They concluded that in a double-stranded DNA (dsDNA), the

abundance of A should be the same as that of T, and similarly for C and G. This equality is later known as the Chargaff's first parity rule.

However, despite these in-depth studies of DNA in this past century, allowing people to understand more about the functionality of the molecule, there were still a lot of questions which remained unanswered, with the true molecular structure being one (and the most fundamental) of them. Prior to this point, it was known already, that the DNA contains a certain amount of carbon, oxygen, hydrogen, nitrogen and phosphorus atoms, and their relative abundance within the molecule was known to a reasonable accuracy. This fact was exploited by Avery and his colleagues, in their experiments, to identify the DNA from the mixture of compounds. Moreover, from the extraction of the bacterial DNA using alcohol, it was found that in its pure form, the DNA is a fibrous strand. This, plus the Chargaff's rule, are the knowledge the scientific world had in general about the structure of the DNA as of 1952 — bits and pieces of important information, but no-one had put them into the holistic picture just yet.

The breakthrough came in the next year, 1953, when Watson and Crick created a chemical model [289] and proposed that a possible structure of DNA is a stack of two firmly bound nucleosides [1] and that it exists as an anti-parallel double helix. Since the proposition was done through physical, i.e. cardboard, modelling, their observation had to be rigidly tested. One name ought to be mentioned here specifically — Rosalind Franklin, whose work has critical importance on a significant proportion of this project. Franklin had been performing groundbreaking experiments using X-ray fibre diffraction methods on DNA structures, and had produced the "Photo 51" (lower right of Fig. 5.4) which turned out to be one of the most well-known experimental images in modern natural sciences. This photo, which had been kept inside Franklin's much detailed manuscript for more than two years before being published [95] together with Watson and Crick [289] and Wilkins [305], depicted the true structure of DNA as a dimer of anti-parallel strands with the nucleosides bridging between the backbones, and had vastly influenced Watson and Crick's molecular model [191].

In 1962, Watson, Crick and Wilkins were nominated for and awarded the Nobel Prize in Physiology or Medicine "for their discoveries concerning the molecular structure of nucleic acids and its significance for information transfer in living material" [270].

1.1.2 The function of nucleic acids

It was discovered earlier, that the DNA is a macromolecule which is responsible for the storage and transmission of genetic information. This is in fact, fundamentally, how proteins are constructed, and how the instructions for the development and functioning of living organisms are transmitted as a consequence. Moreover, another type of nucleic acid, namely the *ribonucleic acid* (or RNA), aids the functionality of DNA by converting the codes in DNA into the amino acid sequences of proteins.

The coding in nucleic acids is done through the sequences of the bases in the nucleic strand

[1] Watson and Crick integrated the discovery by Chargaff with the existing knowledge, and proposed that the pair of bases which are bound can be A-T or C-G, and they should be bound by hydrogen bonds.

(in the O5'⟶O3' sense). In 1961, Crick, Brenner, Barnett and Watts-Tobin carried out an experiment [57] which demonstrated for the first time that exactly three bases ("triplet") of DNA code correspond to one amino acid in the protein (for example, "GCA" or "GCC" in the DNA correspond to an alanine monomer being inserted by the RNA). In 1964, Nirenberg and Leder's experiment [169] showed that there exists a specificity of the "transfer RNA" (or tRNA) in the binding of ribosomes promoted by triplets. This suggests that the codon (i.e. a "unit" of the genetic code) consists of three nucleobases.

1.2 The structure of DNA

Many biological molecules are highly complex in nature. For example, a strand of DNA, say in human chromosome 13, consists of more than 114 million base pairs and is about 3 centimetres long when stretched out [106,171]. Therefore, in order to describe the structure of them accurately, especially such polymeric ones as DNA and proteins, we break them down into four different levels of details.

The primary structure of a polymeric molecule includes the most basic constituents, i.e. molecular *residues*, which can be standalone molecules *per se* but combine to form the ultimate structure. The secondary structure depicts the interactions between these residues to form the monomers which give the shape to the polymeric structure. The tertiary structure is the polymerised 3-dimensional structure of the molecule itself, i.e. what the molecule looks like macroscopically. Last but not least, the quarternary structure describes how two or more of the molecules interact with each other to form a complex.

1.2.1 Primary and secondary structures — nucleotides and double helices

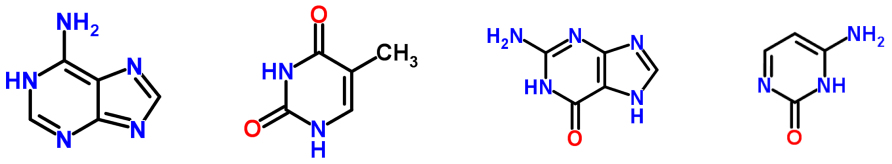

Figure 1.1: Nucleobases. From left to right: A, T, G, C. (Image obtained from ChemSpider.)

DNA is a long polymeric chain with "nucleotides" as its monomers, i.e. the basic unit. A nucleotide is a group of residues of a base ("nucleobase"), a 5-membered furanose sugar ring ("deoxyribose") and a phosphate; the subgroup of the nucleobase plus the deoxyribose within a nucleotide is known as a nucleoside. There are four kinds of nucleotides, namely deoxyadenylate (dAMP), deoxythymidylate (dTMP), deoxyguanylate (dGMP) and deoxycytydilate (dCMP). The only difference among these nucleotides lies in the constituent bases, which are adenine (A), thymine (T), guanine (G) and cytosine (C) correspondingly; the deoxyribose and phosphate groups are always the same for all nucleotides.

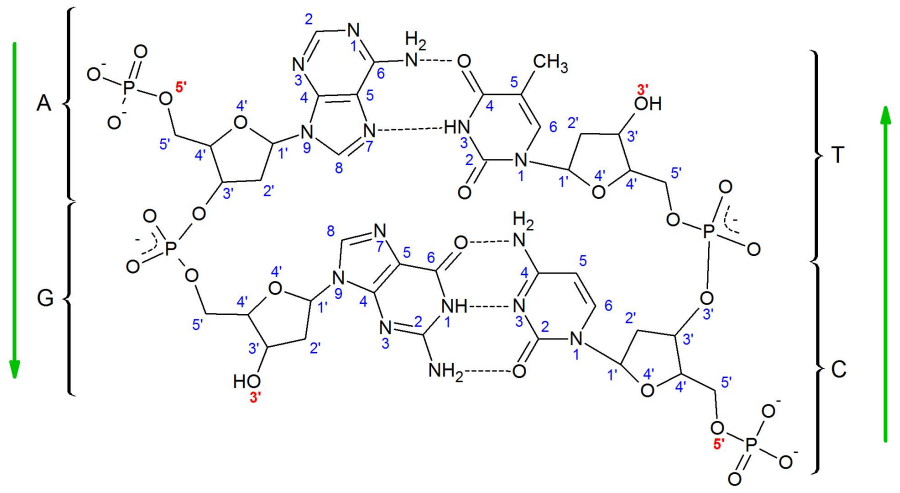

Figure 1.2: Polymerisation of nucleotides and the 5′ ⟶ 3′ directionality of DNA.

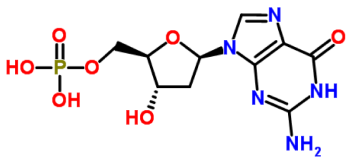

Figure 1.3: A nucleotide (dGMP). From left to right: phosphate, deoxyribose, nucleobase. (Image obtained from ChemSpider.)

Polymerisation of a polymeric chain starts with the dimerisation process. In the context of DNA, The dimerisation is done via a hydrolytic reaction where one of the nucleotides gives up the hydroxyl (–OH) group on the 3′ position of the sugar ring, while the double-bonded oxygen atom on the phosphate group in the other nucleotide shares one of the bonds with the now positive site on the ring. Moreover, due to the high electronegativity of the phosphate group, the two hydrogens in the hydroxyl groups of the phosphate are readily released into the aqueous environment, leaving the phosphate linkage with a bare negative charge. Therefore, the dimerisation of nucleotide results in a *dinucleotide* with a net negative charge, and the by-product of an H_3O^+ molecular ion which is the cause of acidity, hence the name "nucleic *acid*".

As shown in Fig. 1.1, two of the nucleobases (T and C) are six-membered rings whilst the remaining two (A and G) are fused aromatic rings. T and C are known collectively as the pyrimidines and A and G as purines. Here, the Chargaff's first parity rule can be more precisely reiterated as "the ratio between pyrimidines and purines in a DNA strand is in close proximity to unity". The molecular model made by Watson and Crick in 1953, did not only predict the tertiary structure of DNA as a double helix, but also helped proved Chargaff's rule by postulating that the forces which hold the two strands together are the same as those which hold the nucleobases in place: hydrogen bonds — three between G and C, two between A and T.

Chargaff's rule only applies to the ratio between pyrimidines and purines but not specifically any of the bases. Of course, one can argue that pairing can be equally done between

A and C, as well as between G and T, maintaining the 1:1 ratio between pyrimidine and purines. However, Watson and Crick's model suggests that by pairing A with T and G with C, the maximum number of hydrogen bonds can be attained, hence maximising the stability of the overall structure. Furthermore, since all the bases contain a lot of hydrogen atoms, as well as highly electronegative atoms such as nitrogen or oxygen, it is natural to think that there can be more than one way of forming hydrogen bonds. In 1963, Hoogsteen reported the discovery of different hydrogen bonding formation sites within base pairs [140]. However, this model depicts only two hydrogen bonds between C and G rather than three as in the Watson-Crick system, and hence is less energetically favourable. Moreover, according to the Hoogsteen rule the purinic nucleobases (A and G) must flip such that the six-membered ring, rather than the five-membered one, is attached to the furanose. This alters how the two bases within the same pair are oriented, and hence the overall double-helical structure of the DNA must change accordingly [140]. As a result, the "Hoogsteen base pair" is only seen in crystals and its occurrence is extremely rare.

1.2.2 Tertiary structure — conformations and polymorphism

DNA, whilst usually depicted as a double helix, is one of the most complex molecules in a biological organism — be it a highly developed eukaryote or as simple as a prokaryote — alongside proteins. The reason behind the complexity, structurally, is because the DNA is soft matter, meaning that it changes its shape according to the environment. For example, when the DNA segment is exposed in an ionic (i.e. salty) environment it prefers the A-form, but when both humidity and salt concentration are low it would change into the tightly wound C conformation, etc. [238, 306] Due to the flexibility of the molecule, the aforementioned named forms are only indicative of their appearances in crystalline structures, and the molecule constantly changes its appearance in aqueous environments.

Though always drawn as a straight helical staircase, the depiction is but a tip of an iceberg in the holistic picture of a DNA molecule, and has the name *oligomer*, meaning "a few (*oligo-*) parts (*-mer*)". In real life, short segments of such straight-chained DNA (scDNA) cannot exist for a long period of time, because the base pairs on either ends of the chain are only held tight by the phosphate linkage on one side and with temperature effects, the end pairs would start to melt and thus initiate a domino effect on the subsequent inner base pairs, resulting in the melting of the entire molecule eventually. Khandelwal *et al.* [156] formulated an estimation of the melting temperatures of arbitrary lengths and sequences of DNA by means of curve-fitting to experimental data and found that

$$T_m = 7.35E + 17.34 \ln L + 4.96 \ln C + 0.89 \ln D - 25.42 \tag{1.1}$$

where T_m is the theoretical melting point (in centigrades), E is the per-base DNA strength parameter, L is the number of base pairs, C is the concentration of the solution and D is the total nucleotide strand concentration. As an example, for a 15-bp short oligomer in 0.22M salt solution and $D = 2 \times 10^{-6}$, it was calculated that $T_m = 65.04°\text{C}$ which is less than 1% off from experimental results [225].

In cell nuclei, the lengths of the DNA strands are typically 10^5 to 10^8 base pairs long and

they are wound in such a way that the "head" of the molecule is covalently bonded with its own "tail", and as such are closed circles which are named "closed circular DNA" (ccDNA). This, of course, effectively solves the problem of melting. However, with the total length adding up to a metre in a mammalian cell [245], packing all genes into the nucleus of about 6 microns diameter [3] in their fully relaxed and perfectly circular conformation would not be viable. In order to attain the optimal packaging, the strands would wound against its own helical axis to form a supercoiled structure, just as the old-fashioned telephone wires would do. The functions and theories behind such behaviour will be discussed in fuller details in the upcoming chapters.

While the aforementioned mechanism occurs naturally in relatively simpler cells such as prokaryotic cells and viruses, it is much oversimplified for complex eukaryotic cells, whose long supercoiled DNA strands are further wrapped around proteins called histones, which are roughly spherical in shape, to produce even more effective packing. However, it does not mean that loose ccDNA does not occur in eukaryotic cells. In fact, this so-called extrachromosomal circular DNA (eccDNA) was first discovered in 1972 and later found to be in cells of all tested organisms, including roundworms and humans [256]. Among the subtypes of eccDNA, there is one called the "double minute", which is a small fragment of eccDNA, and has been detected in many types of tumours, such as breast, lung, lymphoma and neuroblastoma. As such, it is associated with gene amplification due to chromothripsis processes as tumours grow [261, 308].

Watson and Crick mentioned in their famous 1953 Nature paper [289] that in a DNA, "both chains follow right-handed helices". However, such claims is by no means specific, and does not depict the holistic picture of nucleic acid conformations. It was discovered, through the X-ray fibre diffraction studies of Franklin [94, 95] around the same time, that there were at least two rather different conformations of the DNA, later known as the A- and B-forms, which would qualify for the description of Watson and Crick.

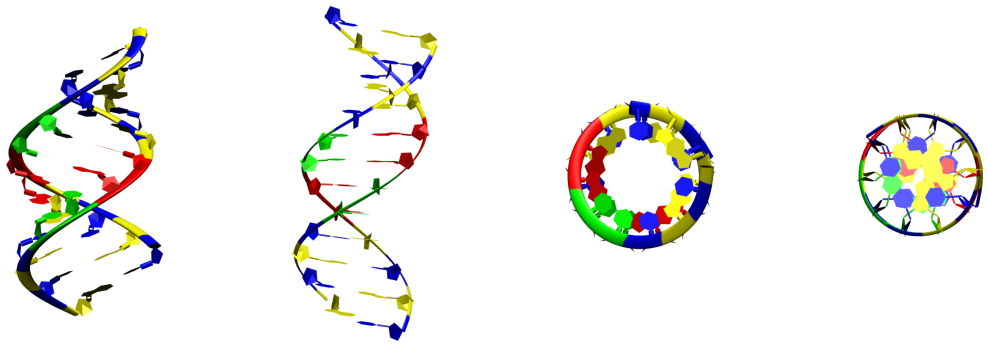

Figure 1.4: A- and B-forms of the Drew-Dickerson dodecamer, in side and top-down views. Representations drawn to the same scale.

Fig. 1.4 shows the A- and B-forms of the Drew-Dickerson dodecamer [77] which has the sequence of d(CGCGAATTGCGC). We can see that for the A-form, the helical structure is slightly more tightly wound then the B-form, in that it is shorter and fatter in appearance.

Form	Pitch (Å)	Rise (Å)	Turn (°)	Relative humidity	Ionic environment
A	28.2	2.56	32.7	< 85%	medium, not Li^+
B	33.8	3.38	36.0	> 92%	high Na^+ or high Li^+
C	31.0	3.32	38.6	< 66%	low
D	24.3	3.03	45.0	< 92%	heavy ions best
S / Z	43.2	3.60	−30.0	< 85%	high

Table 1.1: DNA conformations with respective helical parameters, obtained from X-ray fibre diffraction studies. Reproduced from [238, 306, 309].

Moreover, another measure of the tightness of helical twisting is the extent to which the nucleobases tilt. It is obvious that whilst the bases in B-DNA are largely parallel to each other and maintain the perpendicularity with the helical axis throughout, those in A-DNA are heavily tilted. As such, if we inspect the molecule from a top-down perspective, because of the tilting of the bases, the coverage of the central "cavity" becomes much less in the A-form than in the B-form.

While the two forms described above are the two most observed forms, they are not all the conformations a DNA can take, as the DNA is a soft matter and hence is very flexible structurally. Table 1.1 shows the currently identified general forms of the DNA in different environments. From this we could see why the A- or B-form is the most dominant conformation found in living cells — In normal conditions, the cell is highly hydrated (hence favouring the B-form) with the salt level maintained at a certain level through homeostasis (which favours the A-form).

The C-form is a rather loosely wound form of DNA, in that there are only about 9.3 bases per period, and is only favoured specifically in a relatively desiccated environment with low ionic concentration. On the other hand, while the environmental requirements of the D-form much resembles those of A and B, its ionic requirement of the presence of heavy ions makes it very rare to be found as heavy ions only exist in trace amounts in most living cells. Structurally speaking, the D-form is the most loosely wound conformation among the four mentioned, having only eight bases per period. In terms of sequence specificity, it was discovered that the D-form only exists for some sequences like $[d(AT)]_2$, but not others [176]. Last but not least, the traditional conformations for the right-handed helix are not always true, in that the DNA can also have left-handed helicity. The S-form or the Z-form is an example of a left-haded DNA [2]. Similar to the D-form, the S-form only exists in the specific sequence of $[d(CG)]_2$ [138]. In terms of the structure of the S-form, it is the most tightly wound conformation of those mentioned above, in that it has 12 ($= \frac{360^\circ}{30^\circ}$) bases per period. Moreover, it is also the most elongated conformation of all, as each period spans 43.2Å, which is nearly a third longer than the longest of the others (the B-form, 33.8Å).

Finally, the numbers quoted above are data collected from X-ray *fibre* diffraction experi-

[2]The S- and Z-forms are two names used interchangeably for the left-handed structure by different authors. The S-form is a polymeric conformation obtained from fibre diffraction, whereas the Z-form is an oligonucleotide conformation obtained from X-ray crystallography studies. The author of this thesis is aware that, in some biophysical communities, the "S-DNA" is also used in conjunction with the "S-ladder form", which is an extended and unwound DNA under external tension, derived from numerical modelling [47, 255]. It should be emphasised that the "S-DNA" (or "S-form") hereof does *not* mean the S-ladder form, and the discussion of the S-ladder DNA is beyond the scope of this work.

ments, where the samples used are assumed to be sufficiently long and have near-perfect periodicity, hence the numbers are averages over all the nucleotides. However, from non-fibre diffraction studies where relatively short oligonucleotides are used, it was discovered that the variations in those structural parameters can fluctuate a lot on a per base-pair basis [73,74].

1.3 DNA functionality and mutations

1.3.1 DNA functionality

The functionality of the DNA as a biological molecule comes in many aspects, but the most important ones include the storage of genetic information, transcription and replication.

Genetic information storage Not the entirety of the DNA molecule carries genetic information, and the proportion of it which does so (a.k.a. *genes*) varies with the organism [230]. For a gene, its main function is to store genetic information as codons, i.e. chunks of three successive nucleotides. These codons are instructions as to what amino acids are to be produced. As there are four types of nucleotides and a codon consists of three nucleotides, there are $4^3 = 64$ available combinations of codons for the 20 types of protenogenic amino acids [3]. Of the 64 codons, three are also used as initiators or terminators which labels the start and end of the production of a protein chain [180].

Transcription Transcription is the initiating step of the gene expression process, in which the information encoded in the genetic material is used in the synthesis of a gene product (proteins). Transcription by itself, entails the copy of a short segment of the coding DNA into RNA by the enzyme RNA polymerase, and it can be divided into three key stages: initiation, elongation and termination. During initiation, the RNA polymerase first binds to the DNA template ("promoter") and melts the subsequent 12 to 14 base pairs into a so-called *transcription bubble*. It then polymerises two of the loose ribonucleoside triphosphate fragments to form a dimer. The dimer is extended during the elongation process (Fig. 1.5), by means of the sliding of the polymerase along the DNA, into a chain of growing RNA ("*nascent* RNA"). Lastly, as the polymerase hits the stop codon of the DNA, it releases the produced RNA strand and dissociates itself from the DNA. Transcription is a vital process in the cell cycle as the produced RNAs serve as templates for further syntheses of proteins, without which life cannot happen.

Replication DNA replication is a process where a molecule of DNA is *cloned* to form a new and identical molecule. This is a crucial procedure in cell division [3], for a newly formed cell must have the same set of genetic materials as its mother cell. Like transcription, the initiation kick starts the process when initiator proteins gather around and form

[3] Protenogenic amino acids are those which are polymerised to form proteins. There are actually 22 protenogenic amino acids but only 20 are in the standard genetic code. [4]

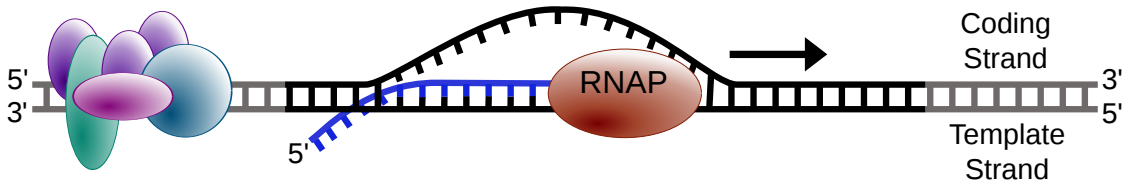

Figure 1.5: Schematic diagram showing the elongation stage of DNA transcription. Image taken from [303].

a complex around the initiation point ("origin") where the DNA would be unwound [19]. Once the strands are unwound and separated ("replication fork"), the enzyme primase adds the primers onto the two strands ("leading" and "lagging" depending on the directionality) of the template DNA and the enzymes helicase and ligase work in this process to ensure the size of the replication fork remains constant and the smooth sliding of the fork. It is noteworthy, that although the A-T and G-C pairing is dominant and other pairing schemes are possible, there are other forbidden combinations which upon creating would be deemed erroneous. These errors do occur in replication processes despite the precision of the enzymes' work. If a mismatch is created, the DNA polymerase which facilitates the polymerisation of the primers and the Okazaki fragments [4] would be forbidden to carry on so that the proofreading procedure can remove the mismatched nucleotides.

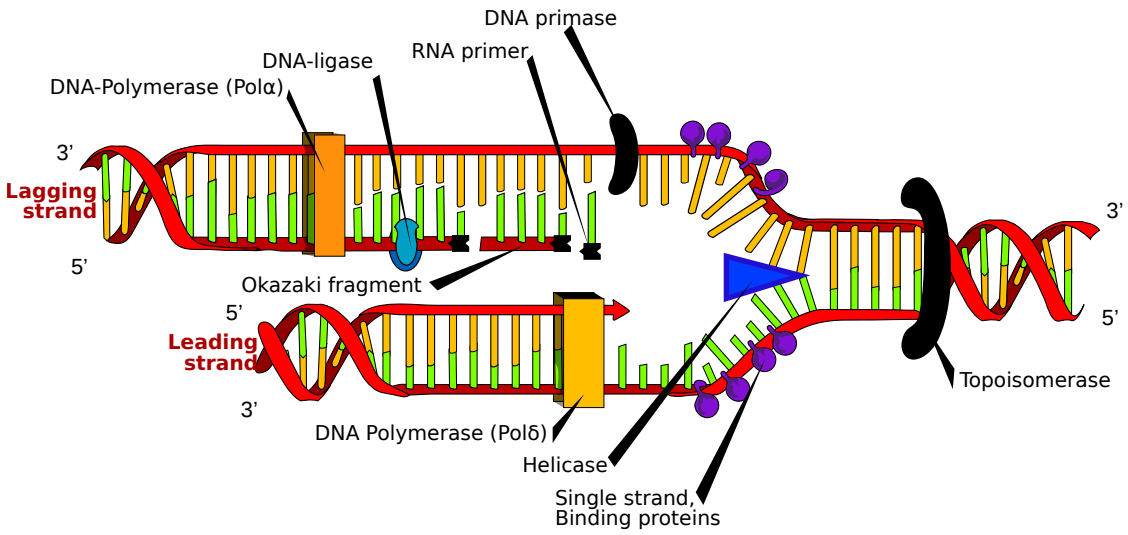

Figure 1.6: Schematic diagram showing the replication of DNA. Image taken from [301].

[4] Discovered by Okazaki *et al.* [219, 264] in 1968, Okazaki fragments are short discontinuous fragments of nucleotides which are paired with the separated strands to create the lagging strand. The fragments are then joined by the enzyme DNA ligase (cf. Fig. 1.6).

1.3.2 DNA-small molecule interactions

The DNA is a polyanion, meaning that it carries multiple negative charges. The origin of the charges is the phosphate residues on the backbone which bridge between two consecutive base pairs. Since each phosphate group carries a -2 charge and the 5′-end of the strand does *not* have a phosphate group, the total charge of an N-base pair segment is $-2(N-1)e$, with $e \approx 1.609 \times 10^{-19}$C being the standard proton charge. Furthermore, the prefactor 2 accounts for the two phosphodiester backbones of the molecule.

Being electronically charged implies the DNA readily interacts with other charged molecules through electromagnetic interactions, or with polar molecules through van der Waals' forces. This subsection will be dedicated to the explanation of DNA interactions with water, followed by those with other molecules which exist in the vicinity of the DNA.

DNA-water interactions Water plays a central role in biological systems, and on average, about 60% of our body is composed of it [288]. The most useful aspect of water does not come from it being an excellent lubricant, but from the fact that it is highly polar and can make large number of hydrogen bonds with neighbouring molecules. A single water molecule can have as many as four hydrogen bonds, where the oxygen (or its two pairs of lone-pair electrons) serves as the donor of two bonds and the two hydrogen atoms as the receptor of the remainder. Thus, water plays a crucial role in the maintenance of structural rigidity in biological systems, especially that of DNA.

Not only do water molecules interact with DNA themselves, their interactions are often associated with other ions, especially metal ions which carry *positive* charge(s); some of these ions include sodium (Na^+) and magnesium (Mg^{2+}). The ability of water to form metal-aquo complexes with metal ions, $[Na(H_2O)_6]^+$ and $[Mg(H_2O)_6]^{2+}$ for the two aforementioned examples [240], makes the DNA structure more rigid. This is achieved since the binding of water onto the ions provides a screening effect which disperses the charge(s) from the central ion to the surrounding water, in turn increasing the effective electrostatic range of the complex.

DNA-small molecule recognition The interactions between DNA and other external small molecules (especially organic ones) can be classified into two large categories, namely covalent-bonding and noncovalent-binding. Covalent-bonding interactions are carried out through chemical reactions of the external molecule with the DNA, resulting in changes in the composition of the DNA, whereas noncovalent-binding interactions are done physically without a resultant alteration in the chemical composition of the DNA. An example of a covalent-bonding reaction is the substitution of a nucleobase by the base analog 5-bromouracil. On the other hand, the docking of the molecules spermine [144, 254, 263] and spermidine [309] in the major and minor grooves of the DNA is a good example of noncovalent-binding interaction. One of the obvious features of noncovalent-binding interactions, which their covalent-bonding counterparts do not have, is the maintenance of relative high mobility of the reactant after interaction.

1.3.3 DNA damage and mutations

Errors during the replication process discussed above are only the tip of an iceberg: there are many different ways in which DNA or chromosomes could be damaged. That the DNA can be damaged was not discovered as a by-product of some pathological studies of late, but had been predicted by Erwin Schrödinger in his famous writing *What is Life?* in 1944 [252] where he questioned whether the seemingly simple structures of genes could withstand long periods of influence of heat motion, and demonstrated even earlier by Timoféeff-Ressovsky, Zimmer and Delbrück [272] in their work on X-ray effects on chromosomes. These studies were published well before the correct composition and topology of DNA were discovered in 1953.

There are two types of DNA damage, *viz.* endogenous and environmental [99]. Endogenous damage is that induced by natural intracellular activities or the direct interaction of the DNA with molecules which are within the same cell. For instance, the nucleobase cytocine can be deaminated (i.e. losing the amino groups) spontaneously due to the pH or temperature alteration in the cell [178]. Another example of endogenous damage is that induced by the so-called reactive oxygen species [67], i.e. chemical compounds which have reactive oxygen atoms. For instance, the highly reactive hydroxyl radical ·OH [28] can readily attack the double bonds in the nucleobases and form the unfavourable formamidopyridine (FaPy) products with pyridine-type bases [41]. Last but not least, the earlier example of incorporation of wrong bases during replication is another example of endogenous damage.

On the other hand, environmental damage is that which occurs due to the interaction or reaction of the DNA with extracellular agents. These agents can be naturally existing molecules, synthesised molecules, pollutants, or even molecules which were produced as by-products of metabolism. These molecules react with the DNA in different ways. For example, psoralen and some of its derivatives can intercalate into DNA due to their planar structure, and may form covalent adducts with pyrimidines when exposed to UV-A radiation [132]. Another example of mutagenic molecules is 5-bromouracil (5BrU), which looks very similar to uracil (methylated thymine), a base which occurs naturally in RNA but not DNA. 5BrU can attack the thymines and substitute itself in the site. Moreover, upon exposure to radiation of a specific wavelength, the bromine drops out and yields uracil [179]. Some other molecules may break the strand(s) of their target DNA, with a particularly interesting class of them being the topoisomerase inhibitors. Topoisomerases are enzymes which nick and close DNA strands during transcription and replication processes. Their inhibitors can bind to the DNA and when the topoisomerase opens up the strand(s) they can form a cleavable intermediate complex with the enzyme, preventing them from detaching from the broken DNA and moving on [86,100], resulting in the DNA being unable to be replicated. This feature makes these topoisomerase inhibitors very toxic to growing cells, and these chemicals are hence extensively used in cancer therapy [86,146,279].

Though one would normally associate extracellular agents as chemicals not originated from cells, they are *not* necessarily confined to molecules or free radicals, but can also be radiation, which has been proven to induce a large variety of lesions in DNA [93,112,147,177,242, 269,284–286]. In this work, however, we shall be focussing on DNA-molecule interactions

and so further discussion of radiation-induced mutations is out of scope here.

Before we continue with the discussion, there is a need for the clarification of some terminology. In layman terms, the words *change* and *mutation* may be used interchangeably depending on the context, but in cell biology they have very different definitions. A mutation is not only any change, but a *heritable* change in the sequence of of an organism's genome [99]. Simply put, a normal change (e.g. damage) to a DNA molecule may be detected and corrected by the self-repairing system within a few generations of replication, but a mutation would not be rectified and the change is carried over to the "offspring" and hence is permanent.

1.3.4 DNA intercalation and its impact on DNA functionality

As we have touched upon briefly, some molecules like psoralen can intercalate into DNA, but the definition of intercalation is yet to be clarified.

An intercalation compound, by the Oxford Dictionary of Chemistry [61], is defined as:

> "A type of compound in which atoms, ions, or molecules are trapped between layers in a crystal lattice. There is no formal chemical bonding between the host crystal and the trapped molecules (see also clathrate). Such compounds are formed by lamellar solids and are often nonstoichiometric; examples are graphitic oxide (graphite-oxygen) and the mineral muscovite."

In the context of biological entities, intercalation means the insertion of a molecule into another molecule. More specifically, in the case of DNA, intercalation points to the insertion of a molecule (typically aromatic, hence planar), or a part of a molecule with such features, in between two successive base pairs of a target DNA.

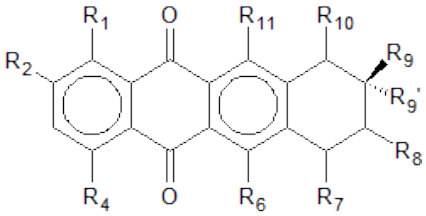

Figure 1.7: Basic structure of anthracycline aglycone. Reproduced from [243], drawn using [70].

The process of intercalation was proposed by Lerman in 1961 [174], with respect to acridines, dyes used extensively in experiments to tag DNA molecules, which can dock in the DNA by means of intercalation. This process results in the increase in the distance between the interstitial base pairs, and was first accurately measured by Waring [287] to be 3.4Å, which is roughly the same as the value for the inter-base pair rise [5] [210, 245]. Moreover, since the helical structure of the DNA is held relatively tight due to the $\pi - \pi^*$ molecular orbital stacking between atoms of two successive base pairs, the DNA has to unwind so that the base pairs could relax to allow the intercalator to enter. The extent of such unwinding depends largely on the size and structure of the intercalator, but is typically within the range of 15° to 25° which is rather large, considering that the twist of a canonical DNA is only about 32° to 36° [210].

[5] The rise is the axial distance between two successive base pairs in the DNA.

Such changes to the DNA structure can prove devastating to the functionality of DNA [22, 146, 160, 226, 239]. One of the aspects of the damage to DNA function lies in the impairment in the decipherability of it by messenger RNA (mRNA). mRNA, which has a similar structure to a single-stranded DNA, is a macromolecule which reads and decodes the genetic information from DNA. This makes a big difference in transcription, as the correct amino acids can be created only if the codons are correct. Secondly, these intercalators are often topoisomerase inhibitors as well, meaning that they would prevent the DNA topoisomerases from functioning properly, which ultimately causes cells to die since they cannot replicate.

1.4 Anthracycline drugs

Anthracycline is a class of antibiotic drugs which have been used as anticancer chemotherapeutic agents for more than 40 years to date. These drugs were first extracted from the bacteria *Streptomyces* [101], which produce the anthracyclines daunomycin and doxorubicin [6].

Structurally speaking, all anthracyclines have the form of anthracycline aglycones, also known as anthracyclinones, whose basic structures are all 7,8,9,10-tetrahydro-5,12-naphthacene quinone [243], which implies that they all have a partially-aromatic four-fused ring system (Fig. 1.7). While throughout the years nearly two thousand derivatives have been found and synthesised, only a few of them are currently used for clinical treatments [84]. Table 1.2 lists four of the most commonly used anthracycline drugs and their usage in chemotherapy.

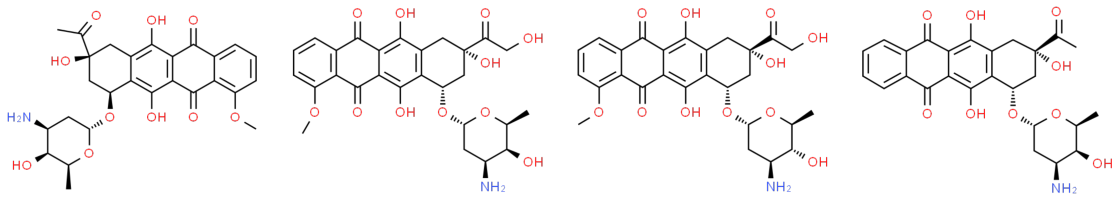

Figure 1.8: Structures of some anthracyclines: (Left-to-right) Daunomycin, doxorubicin, epirubicin, idarubicin. Images from [215–218].

Anthracycline type	Activity
Daunomycin	Acute myeloid leukaemia, acute lymphocytic leukaemia, chronic myelogenous leukaemia, Kaposi's sarcoma
Doxorubicin	Carcinomas (breast, lung, ovarian, liver and thyroid), leukaemias, lymphomas
Epirubicin	Cancers (breast, ovarian, gastric and lung), lymphomas
Idarubicin	Acute myeloid leukaemia

Table 1.2: Commonly used anthracyclines and their anti-tumour activities. Adopted from [40].

Just like many other environmental and biological chemicals, anthracyclines interact with the DNA via a multitude of mechanisms and pathways. Some of the proposed mechanisms are of particular pharmaceutical significance [107], and will be discussed here.

Inhibition of DNA replication Since their discovery in the 1960s, anthracyclines have been known to be potential DNA intercalators, since they have the planar naphthacene chromophore. This interaction with the DNA has been linked to its inhibitive effects on DNA replication [31,72], as it directly induces *frameshift mutations* on the target DNA. Moreover, since anthracyclines are generally rather readily absorbed into cells and are preferentially concentrated in the cell nucleus [124], the prevalence of such interaction is made even higher. However, there is another pathway by which anthracyclines inhibits DNA replications, which is the impairment of DNA polymerase activity [110, 113, 114, 232, 268, 310]. Since both pathways would lead to growth inhibition of DNA, vast studies have been conducted to investigate the contributions. However, many of the results were contradictory with one another, and a unified conclusion has not yet been reached. Gewirtz proposed that the growth-inhibitory effects may relate to a cytostatic and transient component of drug actions. This means that these effects are likely to be in the local cellular level (by slowing or stopping proliferation of cells) rather than in the global level as to actively be lethal for the entire tumour [107].

Free radical generation As have been touched on in the previous section, one of the sources of DNA damage is free radicals. It was discovered that the two molecular oxygens in the quinone of the aglycone, being highly electronegative, act as strong electron receptors in certain conditions [13,258], so that superoxides, hydroxyl radicals and peroxides are formed [12, 13, 17, 30, 258], in turn inducing damage to the DNA target. However, though it is a known fact that free radicals are produced from anthracyclines, its responsibility for the drugs' cyto- and cardiotoxicity still remains unknown to pharmacologists [107].

Topoisomerase poisoning Topoisomerase poisoning is, thus far, the most accepted explanation for anthracyclines' pharmaceutical mechanisms. Topoisomerases are enzymes which are responsible for the breakage of double stranded DNA, the alteration of the torsions of the DNA and the religation of the DNA, and topoisomerase poisoning is the impairment of these functions. In the context of anthracyclines, the poisoning process goes hand in hand with the intercalative properties. The drugs first intercalate into the target DNA and further form extremely stable DNA-anthracycline-Topo2a ternary complex [23], due to which the topoisomerase IIa (Topo2a) cannot religate the broken DNA strands and cannot retract from the DNA. Subsequently, the self-repair mechanism of the DNA would be triggered to repair this growth arrest [40]; and if such fix should fail, programmed cell death would be initiated [205].

1.5 Outline of research

The work presented in this thesis makes use of molecular dynamics and various peripheral techniques to probe the dynamical and energetic properties of interactions between DNA and three anthracycline drugs, *viz.* daunomycin, doxorubicin and idarubicin. The overall layout of the thesis will be as follows.

Chapter 2 will deal with the theoretical considerations of the techniques employed throughout this project. Molecular modelling methods will first be reviewed, followed by a detailed discussion about the AMBER force field. Then an in-depth explanation of coarse-graining methods and the SIRAH coarse-grained force field, which is used extensively in this work, will be presented. After that, a detailed account of molecular dynamics methods, including energy minimisation, simulation ensembles and environment regulatory methods (thermostats and barostats) will be made. We also give a brief explanation of various computational solvent models used in this project. Lastly, the chapter will be concluded with the discussion of the theories behind three most important aspects of this work, *viz.* free energy evaluation, ribbon theory and X-ray diffraction.

In Chapter 3, we will present the preliminary work done for the preparation of the major part of this project, using the methods of targeted molecular dynamics and accelerated molecular dynamics. Discussion will be made regarding the methodology and results, with particular details given to the caveats and drawbacks of these methods, which ultimately led to the use of other more suitable methods which will be presented in Chapters 4 to 7.

In Chapter 4, we will present the study on the effects induced to straight-chain DNA segments due to intercalation of the aforementioned anthracycline drugs. Firstly, simulation protocols adopted in this chapter will be explained. Then we will analyse the simulation outputs using both static and dynamical methods. Static method includes the evaluation of the per-base pair root-mean-square deviation of DNA structure with respect to canonical conformations, whereas the dynamical method used a real-time tracking of a selection of structural parameters to trace the more sophisticated structural perturbations.

In Chapter 5, we will give a brief introduction to the PYRALLEX program (an X-ray diffraction pattern simulator), supported by a collection of sample outputs. Then, we will present a novel parameter for data comparison. Finally, results obtained using molecular dynamics simulation in Chapter 4 will be re-analysed using PYRALLEX with detailed discussion to justify the structural perturbation induced in different sequences of DNA by intercalation.

In Chapter 6, we will study the non-intercalative interactions between a 160 base pair covalently closed circular DNA and the three drugs, in different ionic environments. We first present a novel method for the evaluation of salinity and solvent dielectric constants. Then the system creation process will be explained, with focus specifically on the transformation from an all-atom model to a coarse-grained model. Moreover, we will elucidate the method used in determining the rate of conformational change which is at the heart of the study in this chapter. Lastly, a detailed discussion will be made regarding the simulations done with the DNA-drug complex systems, each of which being simulated in seven different salinities. Particular attention will be paid to the topological transformation, the DNA-drug interaction modes, the writhe-twist partitioning and the rate of supercoiling in each of the

systems.

Chapter 7 will be dedicated to the study of energetics of DNA-drug intercalations. We will first present a general scheme of computational simulation for the work, then explain the use of an all-atom / coarse-grained method employed in this work. Detailed results and discussion will then be given for the free energies obtained from calculations for each drug type-base step combination. From these we will determine, for each of the drugs, the overall free energy change of intercalation (hence the probability of intercalation) and the likelihood of intercalation in specific mode (hence sequence-specificity of intercalation).

Lastly, a brief summary of this work and the aspiration for potential future work will be given as an epilogue in Chapter 8.

Chapter 2

Theoretical Considerations

DNA reacts and interacts with other molecules of all sizes, from macromolecules such as proteins to the smallest monatomic ions like sodium or magnesium; these interactions induce *changes* to the structure of the DNA. Whilst some molecules attach themselves onto different parts of the DNA and form strong covalent bonds with it, others interact with it through other modes, electrostatic interactions for instance. These interactions, many of which involve quantum effects, happen within the regime of picoseconds to nanoseconds, and are thus extremely difficult to visualise using experimental techniques. Even with the newest technology in high-speed lab cameras or single-molecule microscopes (or spectroscopes), the trade-off between spatial and temporal resolutions is still unavoidable. Molecular simulation is a computational technique to effectively envisage molecular interactions *atomistically* in terms of spatial resolutions, and down to the range of femtoseconds (i.e. atomic timescale) temporally.

Originally, molecular simulation was used to model the system of interest, from the visualisation of the movements of individual constituent particles to the calculation of different energies and forces in the system. As this technique is highly customisable, more and more functionalities and features have been added to cope with more challenging calculations such as the free energy change of interactions, which in turn determines the reaction rates.

This chapter is dedicated to the explanation of underpinning theory of molecular simulations which are used in this project. We will start from the basics of molecular simulation methods, the AMBER force field and the SIRAH coarse-graining method which are used throughout the project, and finish with an in-depth discussion on theories of free energy calculations.

2.1 Molecular modelling

The modelling of a molecular system can be broadly divided into two aspects, *viz. static* and *dynamic*. Static modelling includes the calculation of the energies (various potential energies) in the system at a point (hence also known as single-point energy) of the system and the energy-minimisation of systems. On the other hand, dynamic modelling entails the motions of the particles within the system at a non-zero temperature. Hence another

term for dynamic modelling, more formally, is molecular dynamics (MD for short). In this section, methods for both static modelling and MD are discussed.

2.1.1 *Ab initio* methods

Originated from its Latin root, *ab initio* means "from the beginning". In the molecular modelling sense, *ab initio* methods determine the properties of the molecule(s) of interest from the first principles of quantum mechanics, by solving the many-body (time-independent) Schrödinger equation [109]

$$\left[-\sum_i \frac{\hbar^2}{2m_e}\nabla_i^2 - \sum_I \frac{\hbar^2}{2M_e}\nabla_I^2 + \frac{1}{2}\frac{e^2}{4\pi\varepsilon_0}\sum_{i\neq j}\frac{1}{\|\boldsymbol{r}_i - \boldsymbol{r}_j\|} + \frac{1}{2}\frac{e^2}{4\pi\varepsilon_0}\sum_{I\neq J}\frac{Z_I Z_J}{\|\boldsymbol{R}_I - \boldsymbol{R}_J\|} - \frac{e^2}{4\pi\varepsilon_0}\sum_{i,I}\frac{Z_I}{\|\boldsymbol{r}_I - \boldsymbol{R}_I\|}\right]\Psi = E_{\text{tot}}\Psi, \quad (2.1)$$

where capitalised index I denote the I-th atomic *nucleus*, and the small-case subscript i denote the i-th *electron*. $\|\boldsymbol{a}\| \equiv \sqrt{\sum_{i=1}^{n} a_i{}^2}$ denotes the *norm* of the vector $\boldsymbol{a}$ with n being the dimensionality of the vector. Whilst Eq. 2.1 may look nicely symmetric, both for the electronic and nuclear terms, it is solvable *exactly* only for one-body systems, even with Born-Oppenheimer approximation which states that to an acceptable accuracy the electronic and nuclear terms can be segregated into two separate equations. The solution of the two-body version of the equation is hard to obtain, and analytical solutions for three-or-more-body versions of it are nonexistent. Hence, approximation methods have been devised such that solutions may be obtained numerically with a cut in the computational cost without the necessity of largely compromising the accuracy. Two of the most widely used approximations are the Hartree and the Hartree-Fock theories.

The Hartree theory [127, 128] is built on the approximation that the Hamiltonian of the many-particle system acts separately on each of the constituent particles [291] and the resultant many-body wavefunction is merely a product of the constituent single-particle wavefunctions. While this method ensures the uniqueness and orthogonality between wavefunctions, it violates the generalised Pauli exclusion principle for fermions and hence is deemed an incomplete theory. It is considered to be a crude theory from another aspect, which is the inclusion of the unphysical interaction of the particle with its own charge density (i.e. self-interaction).

The Hartree-Fock theory is a theory modified from Hartree's original hypothesis (which he called self-consistent field method). It was then corrected by considering the Slater determinant of the single-particle wavefunctions rather than merely the product of them. This correction was proven to fulfil the antisymmetry property of the wavefunctions. The new joint theory was later on reformulated and published in 1935 [129].

Whilst both Hartree and Hartree-Fock theories are based on the direct application of variational principles which in turn are approximations *per se*, in 1965 Kohn and Sham developed the then-novel Density Functional Theory (DFT) [158] which is an exact method analytically. The core concept of the aforementioned theory is that, the total energy of a system

can be expressed as a *functional* of the density $n(\boldsymbol{r})$, and it reads mathematically:

$$E = \int d\boldsymbol{r}\, n(\boldsymbol{r}) V_n(\boldsymbol{r}) - \sum_i \left[\int d\boldsymbol{r}\, \varphi_i^*(\boldsymbol{r}) \frac{\boldsymbol{\nabla}^2}{2} \varphi_i(\boldsymbol{r}) \right] + \frac{1}{2} \iint d\boldsymbol{r}\, d\boldsymbol{r'}\, \frac{n(\boldsymbol{r}) n(\boldsymbol{r'})}{\|\boldsymbol{r} - \boldsymbol{r'}\|} + E_{\mathrm{xc}}[n], \tag{2.2}$$

with which the ground-state wavefunctions, hence also the particle density, can be obtained using self-consistent methods until the calculated energy and wavefunctions have both converged. Since *ab initio* methods had been of limited use in this project further discussion of their theories is beyond the scope of this thesis.

2.1.2 Semi-empirical methods

Whilst DFT is capable of predicting the highly complicated wavefunctions and thus molecular structures, it also has a rather severe drawback, which is its power-law scaling. This means that the time required to perform a full DFT calculation goes up with a certain power of the number of atoms in the system. This makes MD simulations with full DFT extremely restricted (only systems with a small number of atoms can be simulated). Semi-empirical methods are quantum chemical methods which solve this problem by making simulations more practical.

Semi-empirical methods, in short, are approximations. The quantum component of the methods follows closely the Hartree-Fock theory described above, but the most computationally costly terms are reduced into parameters which are fitted from existing experimental results.

2.1.3 Empirical methods and force fields

So long as there is a need for the calculation of the exact wavefunctions of particles in the system, quantum methods have to be used and the self-consistent methods have to be employed. These methods usually involve large loops in the program codes which increase the runtime drastically. Those methods described above, including also the semi-empirical methods, are then deemed unhelpful for large ensembles such as those in biological systems.

Empirical methods are ones which take more approximations to simplify the representation of the systems even further. Rather than a nucleus plus an electron cloud, an atom in empirical methods is represented as a soft sphere, which interacts with its neighbours via so-called "force fields". A force field (FF), in the quantum chemical sense, is a simple form of force acting on the particles in the system following fundamental physical laws. For example, the Lennard-Jones (LJ) potential is a popular potential to be used (differentiated, cf. $\boldsymbol{F} = -\boldsymbol{\nabla} V$) to give the distance-dependent forces. The atom-specific parameters, for instance, σ (the interatomic separation at which the potential is zero) and ε (the bonding energy between the atoms) in the standard LJ potential, however, are determined directly from experimental data.

One feature of FFs is the expression of the parameters in terms of *atom types* rather than elements. The concept behind atom types is that even atoms of the same element have dif-

ferent types of hybridisation when they form bonds in different circumstances, giving rise to the fundamental difference in bond strengths and equilibrium values from one another. For example, all carbon atoms in a pentane chain are sp^3 hybridised so the bonds are sp^3 carbon-sp^3 carbon bonds. On the other hand, those in a benzene ring are all sp^2 hybridised and the bonds between them are hence of the sp^2 carbon-sp^2 carbon type. These two bond types have different lengths and strengths albeit they both link between two carbon atoms.

2.2 AMBER force fields

The Assisted Model Building with Energy Refinement (AMBER) [51,292] package provides a set of force fields which was developed since the 1970s and was released as a MD simulation package first in 2002 by Kollman *et al.*. The AMBER FF comprises several terms, which read

$$\begin{aligned} V_{\text{total}} &= V_{\text{bonds}} + V_{\text{angles}} + V_{\text{torsions}} + V_{\text{ES}} + V_{\text{H-bonds}} \\ &= \sum_{\text{bonds}} K_b (b_{ij} - b_{ij}^0)^2 + \sum_{\text{angles}} K_\theta (\theta_{ijk} - \theta_{ijk}^0)^2 + \sum_{\text{torsions}} \frac{V_n}{2} \left[1 + \cos(n\tau_{ijkl} - \gamma)\right] \\ &\quad + \sum_i \sum_{j>i} \left(\frac{A_{ij}}{{R_{ij}}^{12}} - \frac{B_{ij}}{{R_{ij}}^{6}} + \frac{q_i q_j}{\varepsilon R_{ij}} \right) + \sum_{\text{H-bonds}} \left(\frac{C_{ij}}{{R_{ij}}^{12}} - \frac{D_{ij}}{{R_{ij}}^{10}} \right). \end{aligned} \tag{2.3}$$

This section is dedicated to explaining each of the terms in detail, and will conclude with a brief history of the evolution of the force fields.

2.2.1 Bond lengths and angles terms

The first term in Eq. 2.3 is the bond length term which represents the bond energy between all covalently bonded pairs of atoms (indexed i and j in the equation). Both terms resemble the classical Hookean potential because small deviations about energy minimum in these degrees of freedom can be fitted to a quadratic form according to Hopfinger and Pearlstein [143], thus

$$V_{\text{bonds}} = \sum_{\text{bonds}} K_b (b_{ij} - b_{ij}^0)^2 \tag{2.3a}$$

where K_b is the type-specific force constant of the bond between atoms i and j, with an instantaneous bond length b_{ij} and an equilibrium length of b_{ij}^0. A problem which immediately arises from this is that by disregarding the molecular orbital shapes, the program can only tell how far two atoms are apart but has no way of telling whether they are bonded or not. A reasonable solution to tackle this problem is to set up a cut-off distance beyond which the bond is taken as being cleaved. Using the same logic as above, because of the lack of knowledge in electron cloud distributions [1], the bond type (covalent or ionic) or the bond order is not explicitly determined.

One of the limitations which comes directly from the use of *atom types* is that the types of the

[1] Traditionally, FFs have fixed "orbital" shapes, i.e. bond shape parameters, but polarisable models had been developed to account for the anomalies encountered when atoms of high charges or high polarisabilities are involved. However, since the shapes of bonds in these atoms are calculated on the fly, the performance of the code is compromised and hence these new models are not used in this work.

atoms are preset as the model is built and cannot be changed in the course of simulation. For example, if a hydrogen atom is determined by quantum mechanical FF parameterisation software (e.g. ANTECHAMBER in the AMBERTOOLS suite described below) as a "HO" type, i.e. a hydrogen on a hydroxylic oxygen, even if it gains enough energy to break the H-O bond and form a new bond with another atom later, that atom can only be another hydroxylic oxygen.

Likewise, the angles term, reading

$$V_{\text{angles}} = \sum_{\text{angles}} K_\theta (\theta_{ijk} - \theta^0_{ijk})^2, \tag{2.3b}$$

works in the same fashion as in the bond length term, with the only difference being that the quantity of interest here is the angle formed by the three atoms i, j and k.

2.2.2 Torsional term

The torsional term in AMBER FF reads

$$V_{\text{torsional}} = \sum_{\text{torsions}} \frac{V_n}{2} \left[1 + \cos(n\tau_{ijkl} - \gamma)\right] \tag{2.3c}$$

where n is the periodicity of the function, and V_n is the potential energy barrier encountered by the 4-atom subsystem when rotating the atomic planes (the red and blue virtual planes formed by atoms $i-j-k$ and $j-k-l$, as shown in Fig. 2.1) around the $j-k$ axis.

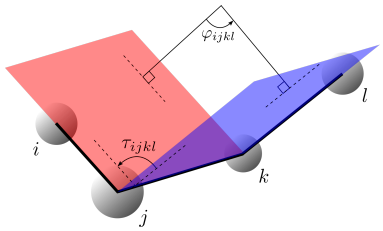

Figure 2.1: Torsional (τ_{ijkl}) and dihedral (φ_{ijkl}) angles.

The form of this term is different from those of the previous ones, in that rather than Hookean-like, the torsional term is a *periodic* function with a constant, positive offset.

It is not difficult to understand why the torsional term takes the sinusoidal form, if one notices the rotational symmetry. Moreover, such form ensures that the first maximum of the torsional energy must be at the specific angle $\tau_{ijkl} = \frac{\gamma}{n}$, which is dependent on the phase angle, γ. The phase angle arises from the fact that different atom pairs have different equilibrium bond lengths (i.e. $b^0_{ij} \neq b^0_{kl}$) and hence the coplanarity of the four atoms (if and only if $\tau_{ijkl} = 0$) would unlikely imply a maximum repulsion.

While some people may use the terms "torsional" and "dihedral" interchangeably (for example, in the User's Guide of the NAMD package [20]), it is noteworthy that such convention is inapplicable to this system, since the two angles are, in fact, supplementary to one another, i.e. $\tau_{ijkl} \equiv \pi - \varphi_{ijkl}$, where φ_{ijkl} is the dihedral, defined as the angle formed by the perpendiculars of the two planes. Apparently, due to this property, the torsional term can also be expressed in terms of φ_{ijkl}, hence

$$V_{\text{dihedral}} = \sum_{\text{dihed.}} \frac{V_n}{2} \left[1 + (-1)^n \cos(n\varphi_{ijkl} + \gamma)\right].$$

However, because of the alternating nature of the $(-1)^n$ prefactor, the whole summation is

now *conditionally* convergent, making computation very challenging[2].

Furthermore, since the torsional term is a 4-body term, which is strongly coupled to the lower order terms (e.g. bond length — 2-body, and bond angle — 3-body), the parameters are difficult to extract. Historically, Weiner *et al.* [292] used first a limited set of parameters for the atom types found in proteins and nucleic acids, which were fine-tuned to agree with experimental results. This was later on extended to atom types in a vast data bank of other molecules as well. Later in the 1990s, the derivation of the parameters was improved by employing *ab initio* methods such as MP2/6-31G* [51]. Nevertheless, regardless of the method with which the parameters are determined, adjustments are made constantly to refine the accuracy of the term [51, 292].

2.2.3 Electrostatic term

The electrostatic term V_{ES} can be further broken down into two components, *viz.* Coulomb and Lennard-Jones.

Coulomb component For an FF-based calculation, since there is little knowledge of an explicit representation of the molecular orbitals, the electrostatic field around an atom is usually represented by an atom-centred multipole expansion [36]. The general idea behind this is that, the higher the order of the expansion, the more accurate the approximation would be compared with the full quantum mechanical picture. Nonetheless, intuitively, the order of the expansion is directly related to the computational cost and hence it is once again a trade-off between overall precision and the cost. It has been shown that, for large biological systems such as the DNA and proteins, concision is valued over precision in that the use of even only the monopole is enough [238]; this is also the reason why in AMBER FF the first order of the Coulomb's law is adopted, hence

$$V_{\mathrm{Coulomb}} = \sum_{j>i} \frac{q_i q_j}{4\pi\varepsilon r_{ij}} \tag{2.3d}$$

where q_i and q_j are the partial point charges on the atoms i and j; ε is a system-specific dielectric constant[3] and r_{ij} is the distance between the atoms i and j. The partial (non-integral) charges on the atoms in a molecule arise due to the difference in electronegativities of the species, which induces an intra-molecular dipole moment, causing the electron clouds to deviate from ideal symmetric shapes, in turn giving rise to an unequal distribution of charges within the molecule [168].

The partial charges in Eq. 2.3d are calculated theoretically. This is done via a full quantum-mechanical energy-minimisation of the system, followed by the determination of the distribution of the molecular wavefunctions at the ground state. A somewhat primitive

[2] A conditionally convergent series is one whose convergence depends on the order of summation. Consider the series $I = \sum_n^\infty (-1)^n$ which has infinitely many ways of expansion, e.g. $I = (1+1+\cdots)+(-1-1-\cdots)$ or $I = (1+1+(-1))+(1+1+(-1))+\cdots$. Only one way of grouping, namely $I = (1+(-1))+(1+(-1))+\cdots$, converges to the finite solution of zero, whilst the others diverge either to $+\infty$ or $-\infty$.

[3] In general, ε should be a tensor $\underline{\underline{\varepsilon}}$, especially in the case of an anisotropic medium. On the other hand, in an isotropic medium, the tensor is reduced to $\varepsilon = \frac{1}{3}\mathrm{Tr}(\underline{\underline{\varepsilon}})$ where Tr denotes the trace of the tensor.

way of achieving this is known as the Mulliken population analysis [209], where the occupancies of the molecular orbitals are calculated as a linear combination of the orbitals (i.e. the basis sets). However, this method brings along two major problems. Firstly, Mulliken divides the off-diagonal terms in the density matrix (as an outer product of the two orbitals involved in a bond) equally, which is proven to overestimate the charge separation within a bond, giving rise to the exaggeration of the partial charges of atoms. Secondly and more importantly, the charge assignment using Mulliken analysis is extremely sensitive to the choice of basis sets. Since in the Mulliken scheme, all electrons are assigned individual atoms, the method poses no limit of convergence to the basis set, and the "exact" solution thus depends on how the limit is approached. As a result, the use of different basis sets usually gives rise to drastically different results [296].

A slightly more superior scheme is the electrostatic potential (ESP) method. This method entails a full quantum-mechanical calculation of the equipotential surface around an atom, and a least square fit between the ESP surface produced by an assigned partial charge at the same position of the physical atom. The charges of the atoms are then tuned self-consistently to match with the ESP surface. This method has shown not only to replicate the multipole moments of the true system, but also to be not so drastically influenced by the choice of basis functions [15]. However, there is a major drawback of this method which is the computational cost of the full quantum-mechanical calculation of the ESP profile of the molecule. As elucidated in the previous section, full QM treatments are viable only up to the scale of hundreds of atoms even using the most powerful supercomputers. Hence, the ESP method cannot be directly applied in systems involving biological macromolecules which typically have thousands of atoms each.

To alleviate this problem, a new algorithm known as AM1-BCC had been devised [150–152]. This method is based on the previous semi-empirical charge method of AM1 [71], which was shown to be excellent in capturing such fundamental features as formal charges and electron delocalisation but rather poor in replicating the ESP produced by the HF/6-31G* level of quantum-mechanical theory[4]. Jakalian *et al.* [152] discovered the addition of an additive bond charge correction (BCC) would rectify this flaw and hence the name of AM1-BCC [150,151]. This novel method achieved quantum-mechanical accuracy with the cost of a semi-empirical algorithm. The same group then refined the parameters against empirical data from more than 2,700 organic molecules, making it even more robust [152]. In this work, we have used this AM1-BCC charge method to determine the charge distribution in the computational model of molecules.

Lennard-Jones component The Lennard-Jones term V_{LJ} captures both the short-ranged electron orbital repulsive force and the long-ranged, attractive London dispersive force. It reads:

$$V_{\mathrm{LJ}} = \sum_{j>i} \left(\frac{A_{ij}}{{R_{ij}}^{12}} - \frac{B_{ij}}{{R_{ij}}^{6}} \right) \tag{2.3e}$$

[4]The "HF/6-31G*" notation is known as Pople's notation of basis sets [75]: a split-valence basis set applied onto the Hartree-Fock theory, with the core orbitals comprising 6 contracted Gaussian functions (GTOs), the inner valence orbitals 3 and the outer valence orbitals 1. Polarisation functions for the *d*-orbital are added for non-hydrogen atoms.

where R_{ij} is the distance between atoms i and j, and the coefficients A_{ij} and B_{ij} are parameters specific to the atom types of i and j, closely related to the zero-potential distance σ_{ij} and the bond energy ε_{ij}.

The Lennard-Jones potential is a crucial quantity in atomic physics as it contains the information about the optimal separation a_{ij} between a pair of atoms. Manifestly, such a distance is exactly when V_{LJ} is at its minimum and hence the force between the two atoms is zero. In the form of Eq. 2.3e, if we differentiate the summand we get the force between the two interacting atoms i and j:

$$F_{ij} = -\frac{dV_{\text{LJ}}}{dR_{ij}} = 12\frac{A_{ij}}{{R_{ij}}^{13}} - 6\frac{B_{ij}}{{R_{ij}}^{7}} \tag{2.4}$$

Setting it to zero gives the optimal separation as

$$a_{ij} = \left(2\frac{A_{ij}}{B_{ij}}\right)^{\frac{1}{6}} \tag{2.5}$$

Hydrogen-bond term Last but not the least, the final term of the AMBER FF is the hydrogen-bond term. It reads:

$$V_{\text{H-bond}} = \sum_{\text{H-bond}} \left(\frac{C_{ij}}{{R_{ij}}^{12}} - \frac{D_{ij}}{{R_{ij}}^{10}}\right) \tag{2.3f}$$

where C_{ij} and D_{ij} are, like in the case of the Lennard-Jones potential term, coefficients specific to the atom types of the electron donor and receptor.

Hydrogen bonds (H-bonds) are extremely important in biological molecules, especially macromolecules like proteins and nucleic acids. In DNA, there are two hydrogen bonds in a cytosine-guanosine base pair and three in an adenine-thymine pair. Albeit weak (typically 1-5 kcal/mol) when standalone, the additivity of the H-bond gives the DNA a relatively rigid structure [294]. Not only do H-bonds prevent the structure from melting (i.e. the splitting apart of the two strands), they maintain the distances between bases within the same pair to about 1.95 to 1.99Å with standard errors in the regime of just 0.3 to 0.5Å [76, 245, 246, 281, 282]. This feature also plays a key role in the usual maintenance of the double helical structure, hence preventing the helix, in normal circumstances, from unwinding locally or forming bubbles which would then destabilise the overall structure and damage the functionality of the molecule.

Despite the significance of H-bonds in real systems as discussed before, there exists a major caveat in the implementation of them in theoretical models, *viz.* their *directionalities*. In theory, any H-bond must consist of a hydrogen atom as the *donor* and a relatively electronegative atom, such as oxygen and nitrogen, as the *receptor*. Traditionally in physical chemistry, one would use an arrow which points from the donor to the receptor to represent an H-bond. It has been shown, that not only the bonds themselves but their directionalities play important roles in biological processes such as protein folding [68,96,97] and ligand-binding specificity [25]. The simple 12-10 Lennard-Jones formalism of the bond, just as in the case of the 12-6 force described above, does *not* hold any information about the directionalities

of the H-bonds. This, then, results in the high discrepancies in the directional preferences during simulation when compared with *ab initio* calculations [206].

For this reason, Eq. 2.3f had been used only in the early generations of the AMBERFFs, but taken out in later generations. Instead of this term, a correction term had been added to the FF for specifically the MNDO-type Hamiltonians of PM6 and AM1, the latter of which has been used in this project [159]. This correction takes a hydrogen bond as a charge-independent interatomic term between a hydrogen atom and another atom eligible for being a receptor and weights it using

$$f_{\text{bond}} = 1 - \frac{1}{1+\exp\left(-60\left(\frac{r_{XH}}{1.2}-1\right)\right)} \tag{2.6}$$

which is a y-reflected logistic sigmoid function, with r_{XH} being the distance between the hydrogen atom and the receptor of the bond. This weighting is used in the determination of the optimal geometry of the H-bond. A similar function

$$f_{\text{damp}} = \left(\frac{1}{1+\exp\left(-100\left(\frac{r_{XH}}{2.4}-1\right)\right)}\right)\left(1-\frac{1}{1+\exp\left(-10\left(\frac{r_{XH}}{7.0}-1\right)\right)}\right) \tag{2.7}$$

was also added to account for the damping effect of short- and long-ranged interactions. This correction has been added to recent versions of the SQM tool in the AMBERTOOLS toolbox, which performs all essential quantum-mechanical calculations during the calculation of force fields for new molecules. It is suggested that, apart from the fact that this correction still does not account for the directionality of the H-bond, an effective cutoff must be added to the geometric term in the correction, in order to be viable for use in MD simulations. Therefore, it is recommended that, this correction should be used only for single-point energy calculations and energy minimisations only.

2.3 Coarse-grained force fields and SIRAH

Ideally speaking, simulations should be able to replicate reality fully, given the level of detail one chooses to perform the simulations is realistic, *viz.* simulating *all* the particles down to the subatomic level, which would require of course the use of full quantum-mechanical formulations — without even the use of Born-Oppenheimer approximation because despite nucleons are roughly 1,837 times heavier than electrons they *do* move in real life! However, as we have seen earlier, this assumption *per se*, is unrealistic, since not even the most powerful supercomputer, as far as the current technology goes, would allow molecular dynamics simulation of such detail to be carried out in realistic timescales (nanoseconds to microseconds upwards). Hence, there was the advent of semi- or even fully empirical methods discussed in the previous section. However, as technology advances, simulations involving larger and larger systems (both in terms of the physical size and the number of atoms) had been realised, and the balance between computational cost and precision had become ever more important. Two immediate questions one can ask, then, is firstly, what information we want to obtain from these "mega-sized" simulations, and secondly, whether we

care about *all* the detail down to atomic level. More often than not, if we perform such a "mega-sized" simulation, the aim would be to study the macroscopic phenomena, where the microscopic details become less of a concern. For instance, when we simulate a long sequence (typically > 150-bp) of a closed-circular DNA to study its supercoiling behaviour we probably would not pay too much attention to how the furanose (sugar) of each base puckers. This provides a case for the further cutting-down of computational cost by cleverly grouping constituent atoms together and thus reducing the total number of particles in the system. For DNA simulations, the simplification can be done in several levels: a lower level can be achieved by joining just the functional groups intuitively, for example, the backbone phosphate being a group and the furanose being another, and so forth. A higher level can be achieved by simplifying the system even more drastically, say, putting everything on the backbone as a group and the whole nucleobase as another. The simulation package of oxDNA [223, 224, 276] is an example of high-level simplification. On the contrary, the package SIRAH [64, 186], which has been used extensively in this work, is an example of low-level simplification. The rest of the section will be dedicated to the explanation of how SIRAH attempts to simplify biomolecular systems.

2.3.1 SIRAH force field

SIRAH, an acronym for **S**outhamerican **I**nitiative for a **R**apid and **A**ccurate **H**amiltonian, is a force field developed by the Biomolecular Simulations Group at the Institut Pasteur de Montevideo in Uruguay, aiming to simplify systems involving proteins and/or DNA.

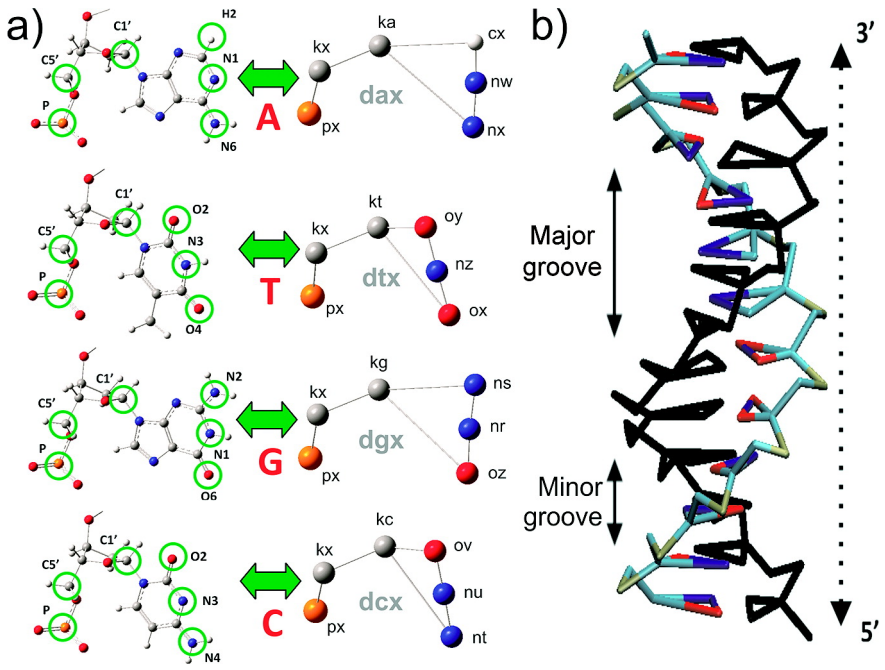

Figure 2.2: SIRAH mapping between all-atom and coarse-grained models. Taken from [62].

For the simplification of DNA systems, SIRAH transforms the vast majority of nucleotides into 6 superatoms: one for the backbone, one for the C5′ linkage between the backbone and

the nucleoside, one for the innermost C1′ atom of the furanose group, and lastly, three for the three heavy central atoms (i.e. excluding hydrogens) responsible for both the hydrogen bonds and dynamics of the nucleobases in the all-atom picture. The only exception applies to the 5′-ends of oligomeric strands, where they lack the "px" superatom (see Fig. 2.2) as conventionally the 5′-ends do *not* possess the phosphate group.

SIRAH, albeit called a force field, is actually a parametrisation for the use in local package environments, *viz.* AMBER or GROMACS. This means that rather than having its own formulation, it fits experimental data (e.g. atomic masses, partial charges, etc.) for the parameters within these two existing packages, without the use of any parametrisation algorithms. As such, SIRAH does not apply artificial constraints to fix secondary structures, thus making it an unbiased force field [186]. Moreover, SIRAH was originally designed for use in multiscale simulations, which means atomistic, semi-classical, or even quantum-mechanical calculations can be performed in conjunction with SIRAH in the same simulation framework [65, 187–190].

The parametrisation of the SIRAH model is similar to the atomistic AMBER FFs, in that the Hookean formulation for bonding and angular (three-body) terms are retained. However, the torsional term in SIRAH is determined by means of Fourier expansions rather than direct summations. Furthermore, the Lennard-Jones terms are calculated using the usual 12-6 potential, with the atom-type pair parameters obtained using the Lorentz-Berthelot combination rules [186].

Lastly, SIRAH does not only have coarse-grained models for amino acids and nucleic acids, it has its own coarse-grained model for solvent environments as well, which will be discussed in detail in Sec. 2.6.

2.3.2 Caveats concerning use of coarse-grained force fields

Due to the simpification of the system by coarse-graining models, it is inevitable that some of the intrinsic features in the simulated system have to be sacrificed as a trade-off for lower computational costs. One of the most obvious examples of such features is the individuality of the atoms. Using SIRAH, which groups about six atoms into one superatom (cf. Fig. 2.2), as an example, we can see that whilst the phosphate group should have a C_{3v} trigonal pyramidal symmetry in all-atom representations, it is coarse-grained in SIRAH into a single spherical blob ("px"), which has a totally different point group symmetry *per se*, and the protrusions from the side oxygen atoms are lost.

This has profound effects in several of the quantities calculated during simulations using CG methods. An immediate implication comes from the diminishing of the friction in the atomic level. Consider the previous example of the phosphate group, where the coarse-graining not only removes the ruggedness of the potential landscape, but also decreases the effective surface area of the potential surfaces. This results in the reduction in the effective friction between atoms, in turn increasing the diffusivity of the simulated system. Since the diffusivity (or the diffusion coefficient) has a unit of *area per unit time*, it becomes a good measure of the *effective timescale* of phase-space sampling. For example, with the MARTINI CG force field [196], it was found that the diffusion coefficient is 2 to 10 times higher using

the MARTINI force field than using traditional all-atom ones, and so the phase-space sampling is accelerated by a similar factor as well. This means that, not only is longer simulation time made possible by the cut in the computational costs, for each unit of time simulated the amount of phase-space sampled is also several times higher.

Unfortunately, the evaluation of the exact amount of speed-up by a CG model is not an easy task, due to the multitude of factors contributing to the it. These factors could be broadly divided into three categories, *viz.* solvent-solvent, solvent-solute and solute-solute interactions, and hence are highly specific to the system being simulated.

2.4 Molecular modelling methods

The calculation of the force field $V_{\text{total}}(\boldsymbol{r};t)$ of a system gives us a good idea of what the overall potential energy at a specific point $\boldsymbol{r}$ is, at a frozen spot in time t. Then based on principles of fundamental physics, we know that there must be a configuration where the total energy is the lowest, such that system is the most mechanically stable. An immediate question which follows is, then, how does one sample the configuration space in order to attain the most stable configuration? For sure, we can sample the whole space point by point, but the feasibility of such an attempt depends largely on the size of the system, or more precisely, the total number of degrees of freedom of the system. With this, we can see that this approach works only for small systems and will suffer as the number of atoms grows, as the number of degrees of freedom increases rapidly with the size of the system. Furthermore, even if such a method works, it would be considered rather uneconomical as the amount of labour put in is vast whilst the achievement is little. Therefore, numerous cleverer methods had been devised and some of the more widely used schemes are listed and described below.

2.4.1 Energy minimisation

Generically speaking, energy minimisation is a class of *static* processes which takes the current (initial) positions of the atoms in a system and moves them in a fashion such that the total potential energy of the system is reduced. Whilst these methods can mostly ensure that the endpoint potential energy is lower than the initial value, the attainment of global minimum is not guaranteed, as minimisation algorithms are local and have no knowledge of the full configuration space. This implies that the initial configuration fed into the minimiser is a major factor in the performance of the minimisation process.

Minimisation is a branch of mathematical problems known as optimisation, and there are multiple ways of performing multivariate optimisation, both analytically and numerically [50, 168]. We list below the major methods implemented in AMBER and NAMD, the packages which have been used throughout this work.

Steepest descent The steepest descent method entails, first, the computation of the gradient of the potential energy of individual atoms which gives the direction along which the potential rises the most. Then, as its name implies, the atoms are moved along the *opposite*

direction of the gradient. A classical analogy would be releasing a ball on a slope and the ball will naturally find the steepest downhill direction and roll down along that direction. In mathematical notation, the steepest descent method can be written as

$$\boldsymbol{r}_i^{k+1} = \boldsymbol{r}_i^k - \alpha \boldsymbol{\nabla} V\left(\boldsymbol{r}_i^k\right) \tag{2.8}$$

where k is the step number, $\boldsymbol{r}_i^k$ is the position vector of the i-th atom at the k-th step, and $\alpha \in \mathbb{R}^+$ is a constant that governs how big a step the minimiser takes. The choice of α directly affects the performance of the minimiser. If α is too large, the optimal point to make a turn will be missed and thus the convergence will be slow. Conversely, if α is too small, although the optimal turning point will generally always be found, more steps are required to reach that point, thus slowing down the convergence again. An optimal value of α which maximises the convergence rate, however, does exist; and it is one at which each minimisation step corresponds to a turning point. One way of getting the value is through quadratic interpolation [50], but we shall omit the explanation of it as it is out of the scope of this work.

Albeit proven to be robust and always convergent, the steepest descent method has some grave drawbacks; one of them is the incapability of dealing with systems with highly asymmetrical energy profiles. For example, in a two-dimensional system where the potential contour is highly elliptical with a large semi-major axis (i.e. long and narrow), the path of descent tends to oscillate and over-correct itself, and corrected errors in previous steps are re-introduced into the system in subsequent steps [168].

This oscillatory behaviour of the steepest descent method can be alleviated by using the method described below, which is the conjugate gradients method.

Conjugate gradients As briefly mentioned before, the conjugate gradients (CG) method can solve some problems intrinsic to the steepest descent method, and so is considered to be superior. The essence of this method is the use of orthogonal gradients and conjugate directions of travel.

In the CG algorithm, it is customary to define a new variable $\boldsymbol{g}^k = \boldsymbol{\nabla} V^k$ for cleaner notation. Then, for steps $k \geq 2$ the direction of travel for atom i is updated according to the equation

$$\boldsymbol{v}_i^k = -\boldsymbol{g}_i^k + \gamma_i^k \boldsymbol{v}_i^{k-1} \tag{2.9}$$

where $\gamma^k = \dfrac{\boldsymbol{g}^k \cdot \boldsymbol{g}^k}{\boldsymbol{g}^{k-1} \cdot \boldsymbol{g}^{k-1}}$ is a self-updating constant which prevents the same over-correction problem described above. The conjugacy of the directions and the orthogonality between gradients are enforced by the relations

$$\boldsymbol{g}^k \cdot \boldsymbol{g}^l = 0 \tag{2.10}$$

$$\boldsymbol{v}^k \underline{\underline{H}}^{kl} \boldsymbol{v}^l = 0 \tag{2.11}$$

$$\boldsymbol{g}^k \cdot \boldsymbol{v}^l = 0 \tag{2.12}$$

$\forall k \neq l$, where $\underline{\underline{H}}^{kl} = \dfrac{\partial^2 V}{\partial \boldsymbol{r}^k \partial \boldsymbol{r}^l}$ denotes the Hessian matrix of the system. For the first step $k = 1$,

since the direction $\boldsymbol{v}^0$ is ill-defined by nature, it is customary to set it to zero, and Eq. 2.9 is then reduced to Eq. 2.8, which is the steepest descent, by noting that $\boldsymbol{v}^k \sim \boldsymbol{r}^k - \boldsymbol{r}^{k-1}$ [168].

Velocity quenching While the method of conjugate gradients nicely eliminates the problem of conditional oscillatory behaviours of the steepest descent, it is at times still sub-optimal, since from Eq. 2.9 we see that the new direction of travel $\boldsymbol{v}^k$ depends on the old direction $\boldsymbol{v}^{k-1}$, which implies that the factor of "inertia" is taken into account as γ^k is always positive. Hence there are times when the particle takes multiple up-slope moves, and the time taken to optimise the system is thus lengthened.

Velocity quenching [236] is a method which applies upon the original CG, but adds another constraint to the choice of the step size factor γ^k — if at some point $\boldsymbol{g}^k \cdot \boldsymbol{v}^k > 0$, set γ^k to zero rather than the previous definition, and thus the direction taken next step will be that of the steepest descent[5]. The reason behind this is that, though $\boldsymbol{g}^k \cdot \boldsymbol{v}^k > 0$ is unavoidable *for a single step only* as the algorithm does not know the landscape ahead, further waste of time in moving farther uphill can be prevented.

2.4.2 Simulated annealing

Annealing, originally a technical term used in metallurgy, is a process where a blacksmith heats up a metal to a certain temperature and cools it back down either slowly at room temperature or rapidly by dipping in a cool water bath (i.e. quenching). The theory behind this process is that by heating the metal above the re-crystallisation temperature the atoms in the metal migrate thus eliminating original dislocations and grain structures, hence altering such intrinsic properties as ductility and hardness of the material. The metal re-crystallises with the new properties as it cools [260, 297].

The same principle can be applied to simulated systems, where the molecular structures are heated up to a high temperature rapidly and the slowly cooled to 0K (absolute zero). This cycle is typically repeated multiple times just as in traditional blacksmithing. In the case of molecular modelling, simulated annealing [162] (hereby "annealing" for simplicity) randomises the structure, allowing potential energy barriers to be surmounted and hence more of configuration space can be explored, resulting in a higher probability of attaining more stable states than using methods mentioned above. The reason for the need to perform the annealing cycle multiple times is that, since the simulation period is typically very short (in the regime of hundreds of picoseconds per cycle), the configuration space explored would be limited. By doing it multiple times, the sampling can then be expanded. It has been shown in this work, that the total energy of the system decreases exponentially with the number of annealing cycles, and can go down for a further 25% upon performing 15 cycles of annealing after even an over-converged minimisation.

Since annealing involves physical movements of atoms and hence is a *dynamical* process, the underlying theory will be explained in the next section.

[5] This abrupt "U-turn" is permitted and does *not* violate the law of conservation of momenta, as the "moves" are actually static re-positioning of atoms and no real movement is involved.

2.5 Molecular dynamics

As seen in the previous section, the use of static methods to determine certain structures, especially the optimal structure, is very limited. A good way of alleviating this problem is using the method of molecular dynamics (MD). The philosophy behind MD simulations is to allow a particular system to evolve with time naturally. There are two main classes of MD, *viz.* classical MD (CMD) and quantum MD (QMD). Since only CMD is used in this work, we will explain only it in detail.

2.5.1 Classical molecular dynamics

The core of a CMD algorithm lies in the integration of the Newton's laws of motion, which gives as the resultant the trajectories of particles in a particular system and how they evolve in time. The three laws of motion according to Newton read [212] (translated by Andrew Motte [207]):

1. Every body perseveres in its state of rest, or of uniform motion in a right line, unless it is compelled to change that state by forces impressed thereon.
2. The alteration of motion is ever proportional to the motive force impressed; and is made in the direction of the right line in which that force is impressed.
3. To every action there is always opposed an equal reaction; or the mutual actions of two bodies upon each other are always equal, and directed to contrary parts.

The trajectory is obtained by integrating the equation of motion directly derived from the second law of motion:

$$\ddot{\boldsymbol{r}}_i = \frac{d^2\boldsymbol{r}_i}{dt^2} = \frac{1}{m_i}\boldsymbol{F}_i \tag{2.13}$$

where $\boldsymbol{r}_i$ is the position of the i-th atom, $\boldsymbol{F}_i$ the external force acted upon it, and m_i the mass of it. The force is, by the assumption that the system is conservative, the negative gradient of the potential, hence

$$\boldsymbol{F}_i = -\boldsymbol{\nabla}_i V(\boldsymbol{r}_i)\,. \tag{2.14}$$

In the case of CMD, the potential V is pre-determined as force fields, in the way described in the previous section.

2.5.2 Integration schemes

It is an art to choose a suitable integrator to be used to obtain the trajectory, as there are numerous schemes which could be chosen. Popular examples of such schemes include Euler [81], leapfrog [136] and velocity Verlet [267, 280]. Each of them have advantages and disadvantages over one another [277]. This subsection is dedicated to give a brief review of how they work, their advantages and disadvantages.

Euler method The Euler method is the most straightforward and the most intuitive integrator among the four listed above, in that it is the reformulation of the derivative:

$$\begin{aligned} \left.\frac{dy}{dx}\right|_{y=y_i} &= \lim_{\Delta x \to 0} \frac{y_i - y_{i-1}}{\Delta x} \\ \Longrightarrow y_i &\approx y_{i-1} + \frac{dy}{dx} \cdot \Delta x \end{aligned} \tag{2.15}$$

where Δx is the step length in the independent variable of the problem. This matches the definition of Riemann summation [233] that

$$\int_{x_i}^{x_f} f(x)dx \approx \lim_{\Delta x \to 0} \sum_{i=1}^{N} f(x_i)\Delta x_i \tag{2.16}$$

if one equates the integrand (summand) in Eq. 2.16 with the derivative in Eq. 2.15. Despite its straightforward definition and simple implementation, the Euler method suffers in two aspects, *viz.* errors and stability. The Euler integration scheme is a first-order method, meaning that the global error scales linearly with the step size Δx whereas the local error scales quadratically with the step size. Moreover, the stability of this method is poor as well, as the stability region is rather narrow in some cases, and is even ill-defined for a lot of other differential equations (e.g. the oscillatory equation $y' = -y$). The stability of an integrator, usually a function of the step size, is the ability of it in *not* magnifying its numerical errors in the course of iterations. The ill-definition of the stability region implies that the global error will approach infinity with the increase in iterations, regardless of the step size.

Leapfrog method The leapfrog method, as its name implies, uses a half-step point as a pivot to integrate, more accurately than the Euler method, a given first-order differential equation. The algorithm reads

$$x_{i+1} = x_i + v_{i+\frac{1}{2}}\Delta t \tag{2.17}$$

$$v_{i+\frac{1}{2}} = v_{i-\frac{1}{2}} + a_i \Delta t \tag{2.18}$$

$$v_i = \frac{1}{2}\left(v_{i+\frac{1}{2}} + v_{i-\frac{1}{2}}\right), \tag{2.19}$$

and a_{i+1} is determined directly using the known form of force. Note the interleaving of the update of the velocity and of the position. This feature allows the precision of both the velocity and position to be improved at the same time, whilst maintaining the total number of steps required to perform the full calculation. Furthermore, the leapfrog method is a second-order scheme, i.e. the global error scales quadratically with Δt. Another feature which makes this method far superior to Euler is its stability, as it does not have as many ill-defined stability regions as Euler. The leapfrog is conditionally stable for the aforementioned oscillatory equation, so long as $\Delta t \leq \frac{2}{\omega}$ and is a constant [24].

However, there are two great disadvantages to this method, *viz.* the ill-definition of the initial velocity and the asynchrony of the velocity and the position. Firstly, if one inspects Eqs. 2.18 and 2.19, it is obvious that in order to determine $v_{\frac{1}{2}}$ after the initialisation, $v_{-\frac{1}{2}}$ is required although it is ill-defined. This is usually alleviated by asserting that $v_{\frac{1}{2}} = \frac{1}{\Delta t}(x_1 - x_0)$,

which unfortunately causes the problem of having to estimate x_1 which is yet to be calculated and depends on $v_{\frac{1}{2}}$ itself. This implies the overall accuracy of the method depends directly on how the x_1 predictor is defined. Secondly, from Eqs. 2.17 and 2.19, it is clear that the velocity and the position are out of synchronisation, implying that the velocity-dependent kinetic energy and the position-dependent potential energy are out-of-sync as well. Hence the evaluation of the total energy of the system is *always* wrong using leapfrog.

Velocity Verlet The velocity Verlet method is a scheme which appears very similar to the leapfrog. In fact, some authors regard the two, alongside some others, as a standalone class of integrators called the leapfrog-type integrators [237]. The algorithm reads

$$x_{i+1} = x_i + v_i \Delta t + \frac{1}{2} a_i \Delta t^2 \tag{2.20}$$

$$v_{i+1} = v_i + \frac{1}{2} (a_i + a_{i+1}) \Delta t. \tag{2.21}$$

The reason for the velocity Verlet to be leapfrog-like, is that the algorithm is indeed based on the interleaved half-step method. The half-step procedures are made implicit in most MD implementations given normal *unconstrained* atomic interactions, i.e. interatomic forces, are velocity-independent [304] and hence holonomic [111, 117, 275]. One of the advantages of velocity Verlet over leapfrog is the synchronicity of position and velocity which ensures the conservation of total energy. Variations of the velocity Verlet exist, e.g. Beeman's algorithm [16], but they are usually of higher order than the original Verlet, hence more computationally expensive and so less widely used.

2.5.2.1 Optimal choice of timesteps for MD simulations

We have previously touched on different integration schemes used by different MD simulation packages. Moreover, we have learnt that for an order-N integrator, the global error propagates as the N-th power of the timestep Δt, whereas the local error propagates as the $(N+1)$-th power of Δt. Then a natural follow-up question would be whether there is an optimal choice of Δt. This short subsection is dedicated to the discussion of the criteria in choosing reasonable step sizes for MD simulations.

Manifestly, we could make such an argument that "shrinking Δt as much as possible would be a good idea" as this is mathematically equivalent to taking the limit in the Riemann summation (Eq. 2.16) and we would effectively approach the continuum limit. However, this is plainly impractical as the calculation time is a function of the number of steps in simulations which is inversely proportional to the step size, given a constant simulation time.

If we cannot solve the problem by reducing Δt, can we at least estimate the maximum value of it so that the simulations would still remain stable? The answer to this is yes, and Choe *et al.* [44] derived expressions for the calculation of the quantity for different systems. It was discovered that for simple force field models (with only Hookean terms, cf. Eq. 2.3), the maximum viable Δt goes up as the reduced masses of the constituent atoms. Although in general, practical force fields like AMBER are more complicated and involve many more

terms, a similar prediction can still be drawn, that larger values Δt can be used for heavier systems. For instance, by using the SIRAH force field, whose masses of the superatoms are set to 50 atomic units [186], i.e. about 20 to 50 times those of the masses of real atoms, it should be safe to tune up the value of Δt by about four to seven times (i.e. up to about 4fs to 7fs).

However, there is also one very important caveat concerning the choice of MD time steps. The scheme for the choice of step sizes only works for *non*-constrained MD simulations. This is because the stability of the system is intrinsically linked to the thermal fluctuation during simulations [87] which is inversely proportional to the square root of the number of degrees of freedom in the system [98] (to be discussed in detail in the next subsection). This means that for systems under constraints, because of the reduction in the number of degrees of freedom, the temperature fluctuation would be more prevalent. As a result, the timestep must be small enough to prevent the simulations from becoming unstable.

2.5.3 Thermodynamic ensembles

2.5.3.1 Microcanonical ensemble (NVE)

In a classical MD simulation, we assume the system of interest to be isolated from the outer world, i.e. it exchanges neither particles nor energy with the extra-system environment. In this case, by just integrating the Newton's equation of motion, the *total* energy of the system must be constant, given the stability of the integrator.

Moreover, it is useful to consider the simulated ensemble to be in a virtual box, i.e. the simulation cell. The size of such a box plays a crucial role in the correct representation of the real system, as the boundary of the box brings about some interesting physics, if not artefacts. Here, a few questions can be raised. Firstly, if the box is of a finite size, what does a particle do when it hits the boundary? Secondly, if the wall is physical, should it be a hard wall or a soft one? On the other hand, if the wall is virtual, what happens if a particle drifts out of bound? The answer is that if the wall should be physical, it should be *soft*, for the reason that if the wall is hard there will be an injection of momentum and energy into the system as a particle bounces off the wall, which clearly would violate physics. If the wall is virtual, the particle which left the box must come back in from the other side of the box in order to conserve number of particles. In this case, the boundary condition is said to be periodic as the primitive cell is virtually cloned indefinitely in all 3 dimensions. Such an ensemble, conserving the number of particles N, the system volume V, and the total energy E, is called a microcanonical ensemble or an NVE ensemble.

One of the issues of an NVE ensemble is that the temperature of the system is *not* constant, for the total energy does *not* have injective (one-to-one) mapping with the temperature, and hence the temperature would fluctuate rather wildly due to the definition of the instantaneous thermodynamic temperature [98]

$$T(t) = \sum_{i=1}^{N} \frac{m_i \boldsymbol{v}_i{}^2(t)}{k_B N_f} = \frac{1}{3} \sum_{i=1}^{N} \frac{m_i \boldsymbol{v}_i{}^2(t)}{k_B (N-1)} \tag{2.22}$$

where k_B is the Boltzmann constant and the $N_f = 3(N-1)$ factor accounts for the total number of degrees of freedom in a N-body holonomic system. From this we see that for an NVE calculation to be physical, the initial configuration of the system must be physical and should preferably be well energy-minimised. This is because if the energy minimisation is performed poorly there may still be atomic clashing at the start of the simulation. The extremely high potential energy stored in the initial configuration would be partially converted into kinetic energy, causing the temperature to remain high while fluctuating wildly [6].

2.5.3.2 Canonical ensemble (NVT)

More often than not, truly physical systems would tend to maintain their temperatures rather than their total energies, since they normally interact with the outer world which could be seen as a heat bath. This happens in most of living creatures as they have to maintain their homeostasis, and a sharp fluctuation in the body temperature (or cell temperature for prokaryotes) disrupts greatly the vital cell processes such as diffusion and osmosis. Hence ectothermic (cold-blooded) an animal may be, its cell temperature should still be roughly constant in a short period of time (simulation timescale, ns to μs).

In light of this, we can impose a constraint on the temperature, instead of the total energy, of the system when we perform simulations. In this case, the ensemble is called a canonical ensemble or an NVT ensemble. The maintenance of the temperature is done via a thermostat. Some of the popular thermostats include Andersen [5], Berendsen [18], Langevin [37, 164] and Nosé-Hoover [141, 142, 213, 214]. Since in this work we have used the Langevin thermostat, the rest of the subsection shall be dedicated to the discussion of it.

The Langevin thermostat is a specific implementation of Langevin dynamics which has the generic equation for the i-th atom in the ensemble

$$\frac{d^2\boldsymbol{x}_i}{dt^2} = \boldsymbol{F}_i - \gamma m_i \frac{d\boldsymbol{x}_i}{dt} + \boldsymbol{R}_i(t) \tag{2.23}$$

where $\boldsymbol{x}_i$ is the position of the atom, $\boldsymbol{F}_i$ is the interaction (force) acting on the atom and γ is a damping constant. $\boldsymbol{R}$ is a time-dependent zero-mean stochastic term following a Gaussian distribution which, when used as a thermostat, fulfils the correlation relationship

$$\left\langle \boldsymbol{R}_i(t)\boldsymbol{R}_j(t') \right\rangle = \sqrt{2m_i\gamma k_B T}\delta_{ij}\delta(t-t') \tag{2.24}$$

where T is the temperature of the heat bath, and the Kronecker delta δ_{ij} enforces that the system is *self*-correlated, i.e. particle motions should not correlate with each other. Langevin dynamics is an approach to model the stochastic behaviour of particles in fluids at a particular temperature, and is hence related to Brownian motion in some specific cases (in particular, when γ is very large). As a result, the γ variable in Eq. 2.23 can be viewed as the viscosity of the solvent environment which contributes mostly to the dissipative friction of the motions.

[6] The relative error of temperature is proportional to ${N_f}^{-1/2}$ [98], so the higher the ensemble temperature, the higher absolute fluctuation of it would be.

2.5.3.3 Isothermal-isobaric (NPT)

Thus far, all the ensembles we have discussed are constant in volume, but there is a rather grave drawback regarding constant-volume simulations especially for some biological systems [7]. With an NVE ensemble, since the temperature fluctuation is large, the fluctuation of the particle velocities is big as well. From kinetic theory we know that the velocities of particles bombarding a wall give rise to the pressure exerted on it. This implies that the fluctuation of pressure in the system using NVE ensemble must be also large, which is not realistic in biological systems. On the other hand, an NVT ensemble poses even more of a problem as simulations are usually started at a relatively low temperature and gradually heated up to a desired temperature through the thermostat. With the volume of the simulation box kept constant, the pressure inside must build up gradually with the temperature. This is clearly not biological either, as high pressure could rupture cell membranes or blood vessels.

For this reason, it is more reasonable in simulations of biomolecular systems to constrain both temperature and pressure at the same time. Such ensembles are called isothermal-isobaric or NPT ensembles. Much like thermostats, a selection of barostats – algorithms which keeps pressure balance – has been devised, with examples including Andersen [5], Berendsen [18] and Langevin piston [85]. The Andersen method is based on an extended system whereas the Berendsen scheme is based on a weak coupling with an external "pressure bath". Since NAMD, the simulation package we have used in this work, uses the Langevin piston method, we will elaborate its algorithm below.

The Langevin piston, like the Langevin thermostat, obeys the Langevin dynamics for stochastic dynamical systems. The three key equations for the piston are

$$\frac{d\boldsymbol{x}_i}{dt} = \frac{1}{m_i}\boldsymbol{p}_i + \frac{1}{3V}\frac{dV}{dt}\boldsymbol{x}_i \tag{2.25}$$

$$\frac{d\boldsymbol{p}_i}{dt} = \boldsymbol{f}_i - \frac{1}{3V}\frac{dV}{dt}\boldsymbol{p}_i \tag{2.26}$$

$$\frac{d^2V}{dt^2} = \frac{1}{W}\left[P(t) - P_{\text{ext}}\right] - \zeta\frac{dV}{dt} + R_p(t) \tag{2.27}$$

where $\boldsymbol{x}_i, \boldsymbol{p}_i, \boldsymbol{f}_i$ are the position, momentum and force of the i-th atom, V, W, P_{ext} and ζ are the system volume, "mass" of the piston, the imposed pressure and the collision frequency respectively. R_p, as in the case of the Langevin thermostat, is a zero-mean stochastic term with the self-correlation

$$\left\langle R_p(t)R_p(t')\right\rangle = \frac{2\zeta k_B T}{W}\delta(t-t'). \tag{2.28}$$

A special case of the Langevin piston happens when $\zeta = 0$, corresponding to a zero-friction system, where the Andersen barostat is reproduced [85].

[7] Some cells (the vast majority of prokaryotes, algae, fungi and plants) have rigid cell walls so the volume of such cells can be regarded constant in simulation timescales. Animal cells only have flexible cell membranes but not cell walls, so the volume is a variable.

2.6 Solvent models

2.6.1 Explicit water models

As has been mentioned briefly in the last chapter, water plays a crucial role in the composition of our body as it makes up a significant proportion of it. Moreover the nucleoplasm, i.e. the sap in the nucleus where DNA is dissolved, consists mostly of water. Therefore, in order to accurately simulate the dynamics and interactions involving DNA, one must take into account the effect of the surrounding water, i.e. the solvent. One way of doing so is to treat the water as individual molecules — hence the name "*explicit* solvent model".

The general method of simulating a system in explicit water is to surround the system (originally *in vacuo*) by a box of water molecules, which in AMBER can be either cuboid or truncated octahedron. The system thus created is taken as a periodic cell and replicated infinitely in all three dimensions, so that if a molecule drifts out of the primary box it would come back into the box from the other side due to the periodicity.

Whilst the explicit model should be accurate because it accounts for the interactions of individual water molecules with the molecules of the system (and with one another), there is also a great limitation of the implementation of this model due to the unavoidable high computation cost it causes. Assuming that the density of the whole solvated system is a constant, the number of water molecules in the primary cell increases roughly proportionally as the volume, i.e. *third* power of the box width. With this, it is not difficult to see that the number of solvent atoms in the box alone could easily exceed 10,000, which adds a heavy burden to computation.

However, the factor which affects the computational cost does not merely depend on the total number of *real* atoms — the model of water is another crucial factor. One of these models is the family of TIP (an acronym for *t*ransferable *i*ntermolecular *p*otential) models.

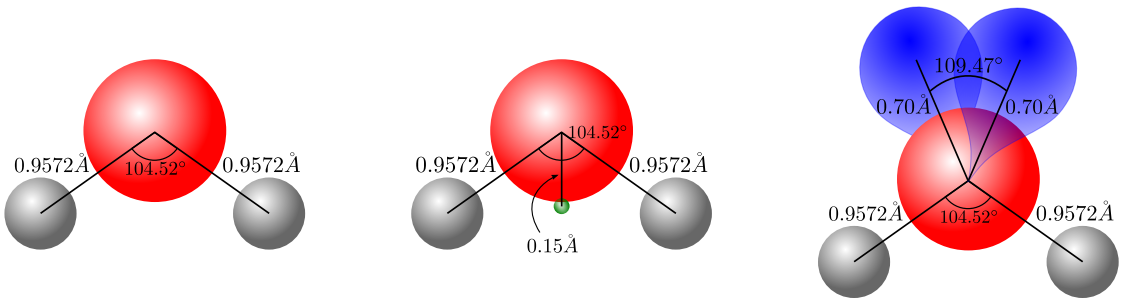

Figure 2.3: Graphical representation of TIP3P (*left*), TIP4P (*middle*) and TIP5P (*right*). Bond lengths *not* drawn to scale.

TIP*n*P models The TIP3P model [154] is one of the three most commonly used TIP models. The 3P in the name implies that it is a 3-site model, meaning that there are three points of interactions for the molecule, *viz.* the oxygen and the two hydrogen atoms. The bond lengths ($r_{OH} = 0.9572$) and the bond angle ($\angle HOH = 104.52°$) were determined empirically

from experimental data. Moreover, the charges on the hydrogen and oxygen atoms are $+0.417e$ and $-0.834e$ respectively [38, 121], giving the dipole moment of the molecule as $2.347D$, deviating rather far away from the experimental result of $2.95D$ [38]. Also, the temperature of maximum density (TMD) which is well known to be about 4°C, becomes −91°C as simulated using TIP3P which is one of the worst in all existing models. This is due to the oversimplification of the model and the resulting failure in representing the hydrogen bonding between adjacent molecules.

The TIP4P model [154], designed at the same time as its 3-site counterpart, is another commonly used model. The main difference between this and TIP3P is the shift of the charge centre of the oxygen atom. In the TIP3P model, the charge centre of the oxygen atom is the atom itself, but in the TIP4P model, the charge centre is displaced along the bisector of ∠HOH by 0.15Å and the oxygen atom is made neutral. With this change, the TMD is improved greatly to −25°C but the dipole moment is further decreased to $2.18D$ due to the reduced distance between the hydrogen atoms and the charge centre for the oxygen (cf. $\boldsymbol{\mu} = \sum_i Q_i \boldsymbol{r}_i$).

Last but not least, the TIP5P model [193], which did not come out until 2000, is a relatively new model which consists of 5 interaction sites. Rather than putting the charge directly on the oxygen or displacing it entirely down along the ∠HOH angle bisector, it tries to simulate the two lone pairs L (each pair is simplistically represented by a point charge) by equally distributing the negative charge between the Ls. The two Ls, 0.7Å away from the oxygen, subtend with the oxygen an angle of ∠LOL = 109.47°, with the LOL plane perpendicular to the HOH plane. With this model, the partial charges on the hydrogen atoms and the lone pair centres are greatly reduced to $+0.241e$ and $-0.241e$ respectively, but with no further penalties on the dipole moment since the separation between the charge centres is increased. Since it is the only model, among all those mentioned, which correctly represents the geometry of the molecule, it is able to form correct hydrogen bonds with nearby molecules and hence produces the 4°C TMD accurately. However, despite being much more accurate than the two previous models, TIP5P makes simulation much more computationally demanding as the two lone pairs are usually represented by two virtual massless atoms. Assuming the number of water molecules in the simulation cell is far greater than that of the solute being simulated, by using TIP5P the total number of atoms in the whole system goes up by roughly two-thirds, which would slow down the simulation much because of the power-law scaling of MD simulations.

Coarse-grained solvation model — WT4 As briefly mentioned in the previous section, SIRAH force field offers a coarse-grained solvation model, known as WT4 (an acronym for "WAT-FOUR"), for use in MD simulations. This model takes 11 traditional TIP3P or SPC water molecules and re-parametrise them as four tetrahedrally interlinked superatom beads. The model was tested in physiological temperatures ranging from 278K to 328K and was shown to accurately reproduce macroscopic properties of water within this region of temperatures [63].

2.6.2 Implicit water models

In this subsection we describe the implicit solvent models, starting with the definition of solvation free energy, then carrying on to the theoretical backgrounds of the Poisson-Boltzmann and the generalised Born models, which lead ultimately to the discussion about the limitations of adopting them.

Solvation free energy In an MD simulation, one of the key components of the result which one would first investigate is the total energy of the system. It is a crucial indicator of whether the simulation is physical or not, or whether there has been some anomalies happening during the simulation. For instance, if, in a long simulation run, the total energy increases monotonically, it could indicate that the system is blowing up as the system temperature increases with the energy. On the other hand, if the energy profile fluctuates continuously and wildly, it could imply that the system is not well equilibrated.

The total energy of a *solvated* system can be written as

$$E_{\text{tot}} = E_{\text{vac}} + \Delta G_{\text{solv}} \tag{2.29}$$

where E_{vac} is the energy of the system *in vacuo*, i.e. in its gas phase. ΔG_{solv} is known as the *solvation free energy*, which describes the energy of transferring the molecule from vacuum into solvent. Moreover, the free energy can be further divided into two parts which could be added up linearly,

$$\Delta G_{\text{solv}} = \Delta G_{\text{el}} + \Delta G_{\text{nonpolar}} \tag{2.30}$$

where $\Delta G_{\text{nonpolar}}$ is the energy of solvating the molecule from which all the atomic partial charges are neutralised, and ΔG_{el} is the energy required to remove the charges (and replacement of them afterwards) so that the non-polar solvation can take place.

The evaluation of the free energy changes of an evolving system is always the pièce-de-resistance in simulations, and that of the solvation free energy is not an exception. One of the reasons behind this is that the pairwise electrostatic force obeys inverse-square law, making it a long-ranged force which diverges upon infinite summation. To accurately evaluate the solvation free energy, certain analytical approximations have to be taken, such as the Poisson-Boltzmann (PB) and the generalised Born (GB) models.

Poisson-Boltzmann model The Poisson-Boltzmann model is a preliminary theory which ultimately leads to the GB model which is more widely used but is still worth mentioning here. It is based on the crude assumption that that solvent is a *continuum* and a *linear-response dielectric*. Hence in the absence of any mobile ions, the electrostatic potential $\phi(\boldsymbol{r})$ created by an arbitrary charge density $\rho(\boldsymbol{r})$ can be readily obtained from the Poisson equation

$$\boldsymbol{\nabla}[\varepsilon(\boldsymbol{r}) \cdot \boldsymbol{\nabla}\phi(\boldsymbol{r})] = -4\pi\rho(\boldsymbol{r}) \tag{2.31}$$

where ε is the position-dependent dielectric constant [8]. This equation only has analytical solutions for very specific symmetries such as spherical, and a set of approximations has to

[8] We have used Gaussian units here.

be imposed in order to efficiently evaluate it — the most fundamental one being the setup of a virtual "dielectric boundary" and consider only two values for $\varepsilon(\boldsymbol{r})$, *viz.* ε_{in} for dielectric within the boundary and ε_{out} for that outside. Such a boundary is defined as a sphere whose radius is the solute's radius of rotation, with a shell of a finite thickness to allow for the dynamics of the solute. This simplification is built on the assumption that $\varepsilon(\boldsymbol{r})$ drops quickly with the increase of $\boldsymbol{r}$ from the surface of the solute. This assumption makes the calculation much less computationally costly, as the nonlinearity of the problem is reduced to a finite size for the space within the boundary only. On the other hand, since by the definition of the dielectric boundary, there should be no free charges beyond the dielectric boundary, Eq. 2.31 becomes a *homogeneous* second-order equation which is much easier to solve than an *in*homogeneous one. Looking from another perspective, there should be manifestly many places even within the dielectric boundary which do not have an effective charge, tests need to be performed on-the-fly to determine the suitable integrator to be used. However, since by definition the space beyond the boundary must *not* have charges, time could be saved from the checking for charges on grids by setting up the boundary.

Furthermore, the charge distribution is produced approximately by the sum of *point* charges, hence $\rho_f(\boldsymbol{r}) = \sum_i q_i \delta(\boldsymbol{r} - \boldsymbol{r}_i)$, and the total charge density is then

$$\rho(\boldsymbol{r}) = \rho_f(\boldsymbol{r}) + |e| \sum_j n_j z_j \exp\left(-\frac{\phi(\boldsymbol{r})\,|e|\,z_j}{k_B T}\right) \tag{2.32}$$

where n_j and z_j are the bulk density and charge of each ion species j, and $|e|$ is the elementary charge. Substituting this into Eq. 2.31, we have

$$\boldsymbol{\nabla}[\varepsilon(\boldsymbol{r}) \cdot \boldsymbol{\nabla}\phi(\boldsymbol{r})] = -4\pi\left(\rho_f(\boldsymbol{r}) + |e| \sum_j n_j z_j \exp\left(-\frac{\phi(\boldsymbol{r})\,|e|\,z_j}{k_B T}\right)\right) \tag{2.33}$$

which is called the non-linear Poisson-Boltzamann equation (NLPB), which, because of its non-linearity, produces potentials which are non-associative. It is known to be even harder to solve than Eq. 2.31 for not having any analytical solution even for a spherical ρ_f. Hence, it is essential that this equation is linearised before being evaluated; and the linearlisation can be done by substituting the second term for $\kappa^2 \varepsilon(\boldsymbol{r})\phi(\boldsymbol{r})$. Here, the κ factor is called the Debye-Hückel screening parameter [69] which is carefully chosen to suit the system. With the knowledge of the potential, the electronic part of the solvation free energy can be readily evaluated through the familiar expression from classical electrostatics

$$\Delta G_{\text{el}} = \frac{1}{2} \sum_i \left(\phi(\boldsymbol{r}) - \phi_{\text{vac}}(\boldsymbol{r})\right) \tag{2.34}$$

Generalised Born model The generalised Born model is a method of further approximating the PB by recasting the equation using Green function [149]:

$$\begin{aligned} \boldsymbol{\nabla}[\varepsilon(\boldsymbol{r}) \cdot \boldsymbol{\nabla}\Gamma(\boldsymbol{r}_i, \boldsymbol{r}_j)] &= -4\pi\delta\left(\phi(\boldsymbol{r}_i) - \phi(\boldsymbol{r}_j)\right) \\ \Longrightarrow \Gamma(\boldsymbol{r}_i, \boldsymbol{r}_j) &= \frac{1}{\varepsilon_{\text{in}}\left|(\boldsymbol{r}_i) - \phi(\boldsymbol{r}_j)\right|} + F(\boldsymbol{r}_i, \boldsymbol{r}_j) \end{aligned}$$

where $\Gamma(\boldsymbol{r}_i, \boldsymbol{r}_j)$ is the Green function and $F(\boldsymbol{r}_i, \boldsymbol{r}_j)$ is a function corresponding to the field due to polarisation charges induced at the boundary (hence it must satisfy the Laplace equation $\boldsymbol{\nabla}^2 F = 0$). With this, the expression for the solvation free energy (electronic part) is simplified to

$$\Delta G_{\text{el}} = \frac{1}{2} \sum_i \sum_j F(\boldsymbol{r}_i, \boldsymbol{r}_j) q_i q_j . \tag{2.35}$$

Here, again, $F(\boldsymbol{r}_i, \boldsymbol{r}_j)$ can take only spherical symmetry and it has the form [257]

$$F(\boldsymbol{r}_i, \boldsymbol{r}_j) = -\left(\frac{1}{\varepsilon_{\text{in}}} - \frac{1}{\varepsilon_{\text{out}}}\right) \frac{1}{A} \sum_{l=0}^{\infty} \frac{t_{ij}^l P_l(\cos\theta)}{1 + \frac{l}{l+1}\beta} \tag{2.36}$$

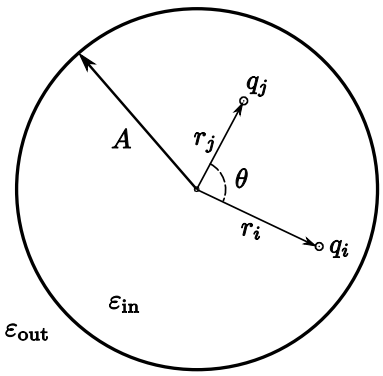

Figure 2.4: Illustration for Eq. 2.36. Reproduced from [220].

where $t_{ij} = \frac{r_i r_j}{A^2}$, r_i is the position of the i-th atom relative to the centroid of the sphere and $P_l(\cos\theta)$ is a Legendre polynomial with θ being the angle subtended by the i-th and j-th atoms and the system centroid. Moreover, A is the radius of gyration of the solute and $\beta = \frac{\varepsilon_{\text{in}}}{\varepsilon_{\text{out}}}$ is the ratio between the dielectric constants inside and outside of the sphere. Note that there can be two cases for Eq. 2.36, *viz.* $i = j$ and $i \neq j$. When $i = j$, i.e. the atoms of interest are the same and the term becomes a *self-interaction field* term. However, when $i \neq j$, the term corresponds to the contribution from normal interatomic interactions.

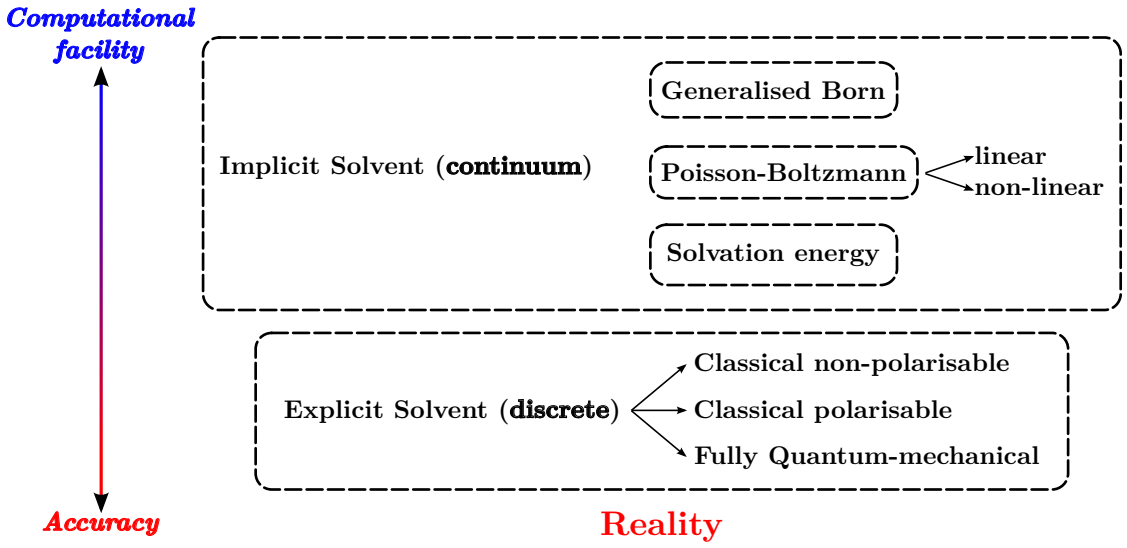

Figure 2.5: Hierarchy of representations of solvent effects in molecular modelling. Reproduced from [220].

Merits and limitations of implicit solvents In the previous subsection, we have discussed the explicit solvent models, how they work to simulate the real water molecule (despite none of them being able to reproduce all the fundamental physical quantities equally satis-

factorily) and their drawbacks. One way of alleviating the problems they cause is through the use of the so-called implicit solvent models, also known as implicit solvent framework [21, 55, 108, 139, 185, 192, 220, 249]. Some of the benefits of using this new model over the explicit solvent models include:

1. Lower direct CPU costs for large systems
2. Enhanced sampling of conformational space through the reduction (or turning-off) of solvent viscosity.
3. More straightforward estimation of free energies
4. Instantaneous dielectric response from solvent (which in the explicit model needs equilibration)
5. Noise reduction of the energy landscape in which local maxima and minima arise from the inter-particle gaps

Regardless of the PB or the GB theories, approximations were made in them in order to make the MD simulation computationally cheaper by drastically cutting down the number of atoms inside the system. The higher the level of a theory is in the hierarchy (see Fig. 7.1), the more approximations would be applied compared with those inferior to it. However, this also implies that the general accuracy of simulations using higher-level theories would be lower than those using lower-level theories.

There is one feature which exists in all the implicit solvent models, which is the "discrete-to-continuum" approximation as their cornerstone. Upon applying such an approximation, several of the fundamental features and effects arising from discrete molecules, such as the size of the water molecules and the tight-binding of the molecules, are totally eliminated. This can be disastrous for simulations focusing on the conformational stability of particular systems. Moreover, the "discrete-to-continuum" approximation also implies that the whole theory is switched to a mean-field one, which means that the sharp and discrete electrostatic interactions (such as hydrogen bonds and induced dipole moments) are smoothed and averaged out.

Another potential implication of using the implicit models comes from their mean-field and linear response approximations. A generic characteristic of such approximations is the neglect of correlation between counterions especially multivalent ones such as magnesium (Mg^{2+}). This can prove devastating to simulations of highly charged systems like DNA as the major interaction between the counterions and the main body cannot be modelled correctly. Furthermore, one should not forget that the analytical form of the form factor $F(\boldsymbol{r}_i, \boldsymbol{r}_j)$ in Eq. 2.36 only works for highly spherical local symmetries. This implies that the less spherical the system is locally, the larger the deviation would be produced using the GB method. In fact, it has been shown in previous studies [220], that while the computation time was greatly reduced, the error of the GB model with respective to the explicit solvent models) is considerably larger than that of the PB, even if the "perfect" effective Born radii [222] are used.

2.7 Free energy calculations

The energy of any system, by the definition of the Oxford Dictionary of Physics [1], is the measure of its ability to do work. The *free* energy of a *thermal* system, however, is the energy released or absorbed in a *reversible* thermal process under certain constraints. Thermodynamically, there are two important free energies, *viz.* Helmholtz free energy F and Gibbs free energy G, with the definitions [241]

$$F = U - TS \tag{2.37}$$

$$G = H - TS \tag{2.38}$$

where U, T and S are the internal energy, temperature and entropy of the system respectively. $H = U + PV$ is known as the enthalpy with P being the pressure acting *on* the system [9] and V being the volume of the system. From this, it is obvious that the only difference between Eqs. 2.37 and 2.38 lies in the PV term. This difference makes the Helmholtz free energy suitable for an NVT ensemble, whereas the Gibbs free energy is applicable to NPT ensembles. In biochemistry, when one refers to "free energy", the Gibbs free energy is usually implied. This is because physiological systems and normal laboratory settings are mostly isobaric rather than isovolumetric [104].

Just like all other forms of energies, the absolute magnitude of the quantity is meaningless, whilst the *difference* of two values tells the whole story. In the case of free energy, the change of the system free energy, $\Delta G \equiv G(\Sigma_2) - G(\Sigma_1)$, when it transits from state Σ_1 to another state Σ_2, denotes the minimum energy needed for the transition to happen [1]. Another piece of information one can obtain from the free energy is the probability of the system to be in a particular state Σ, as the two are directly related through the relationship [115]

$$p(\Sigma) = \frac{1}{Z} \exp\left(-\frac{G(\Sigma)}{k_B T}\right) \tag{2.39}$$

where $Z = \sum_i \exp\left(-\frac{G(\Sigma_i)}{k_B T}\right)$ is the partition function as a sum over energy contributions from *all* possible states. Again, despite the nice look of Eq. 2.39, the calculation of it is largely impractical due to the difficulty in evaluation in the partition function. In real life, seldom are states of systems discrete like in the classical case of a spin-$\frac{1}{2}$ dipole in a magnetic field [194], but are mostly continuous where the partition function for an NPT ensemble would become the high-dimensional phase-space integral [104]

$$Z = \frac{1}{h^{3N} N!} \int \cdots \int dV d\boldsymbol{p}_1 \ldots d\boldsymbol{p}_N d\boldsymbol{x}_1 \ldots d\boldsymbol{x}_N \exp\left(-\frac{1}{k_B T}\left(H(\boldsymbol{p}_1 \ldots \boldsymbol{p}_N, \boldsymbol{x}_1 \ldots \boldsymbol{x}_N) + PV\right)\right). \tag{2.40}$$

where h is the Planck constant, $\boldsymbol{p}_i$ and $\boldsymbol{x}_i$ are the momentum and position of the i-th atom respectively.

For a biological system consisting of typically $\sim 10^4$ to 10^6 atoms, Eq. 2.40 is hopeless to be solved even with the use of supercomputers. Unfortunately, even if it is solvable the

[9] Some texts take P as the pressure acting *by* the system, in which case the equation becomes $H = U - PV$.

information obtained would be of very limited usage, as Z is bound to be very large for an unconstrained system and the probability to be in a particular state has to be vanishingly small unless it is an extremely dominant mode. However, some of the states can be grouped together to make the calculation more viable. For instance, in the study of the likelihood of a drug binding to DNA, we are only interested in knowing whether the drug has bound to the DNA but not how it binds. As a result, such complicated systems can be drastically reduced into binary systems where the *ratio* between the probabilities of the two states is simply

$$\frac{p(\Sigma_2)}{p(\Sigma_1)} = \exp\left(-\frac{G(\Sigma_2) - G(\Sigma_1)}{k_B T}\right) = \exp\left(-\frac{\Delta G}{k_B T}\right) \tag{2.41}$$

which maps immediately onto the definition of the equilibrium constant of a *reversible* chemical reaction A + B $\leftrightharpoons$ AB

$$K_c \sim \frac{[AB]_{\text{eq}}}{[A]_{\text{eq}}\,[B]_{\text{eq}}} = \frac{p_{AB}}{p_A p_B} \tag{2.42}$$

where $[\dots]_{\text{eq}}$ denotes the equilibrium concentration of a species and $p_{\dots}$ is the probability of being in a particular state (see Appendix B.1 for the derivation). With this we arrive at the equation

$$K_c = \exp\left(-\frac{\Delta G}{k_B T}\right) \Longleftrightarrow \Delta G = -k_B T \ln K_c\,. \tag{2.43}$$

Hence, we see here once again that, for binding-type interactions, the free energy change is directly linked to the likelihood of a ligand binding to a receptor — the more negative the free energy change is the higher the probability of binding. This is particularly important in pharmaceutical applications as the higher the probability of a drug binding to the DNA, the more it is potentially effective in delivering its actions.

Now that the relationship between binding likelihood and the associated free energy change is established, we have to deal with the practicality, i.e. actually calculating ΔG. This task, no matter how much we simplify the problem as discussed before, remains still the most difficult to perform, as ΔG has a component of entropy which is not a calculable state variable, and in order to assess it precisely, a sufficient phase space of the system must be sampled. In order to achieve this, several approaches had been devised, with examples including (Gaussian) accelerated MD [200, 201], metadynamics [14, 163], umbrella sampling [273], alchemical transformations [43] and adaptive biasing force (ABF) [66, 133]. The rest of the section is dedicated to the detailed discussion about the ABF method as it is used extensively in this project.

2.7.1 The adaptive biasing force (ABF) method

The adaptive biasing force method is a method based solely on geometrical transformations in a system [42], which aims to alleviate the inefficient sampling of rugged energy landscapes using direct Boltzmann sampling methods [48].

A structural or conformational change in a system may not be just a single alteration in an atom's position, and more often than not it entails the change in the orientation of a whole set of atoms. Moreover, the changes themselves vary from system to system; they may be an

opening or closing of an angle, or splitting apart of two groups of atoms, or a combination of both. These angles or distances are often referred to as "transition coordinates", for they denote the directionality of the transition happening in the system. In some situations they are known as "collective variables (CV)" as well [88, 134].

The ABF method is built upon the fact that the free energy change in a transition is the minimum energy (potential) input required for the transition to occur and there should then be a corresponding average force driving the transition, which is equal to the negative gradient of the potential. Therefore such a potential is known as the *potential of mean force* (PMF) [48]. Now since this PMF is the potential associated with the transition, it is obviously expressed in terms of the transition coordinates or the CV's and therefore does *not* map directly to the potential energy in the real space, and certain transformations have to be made to convert it to the free energy change in the real space. Such transformations can be thus made:

$$\exp\left(+\beta\Delta G_{\xi}\right) = \frac{\int d\xi \exp\left(-\beta w(\xi)\right)}{\int d\xi \exp\left(-\beta(w(\xi)+u_{\xi})\right)} \tag{2.44}$$

where $w(\xi)$ is the PMF in the ξ-space and $u_{\xi} = \frac{1}{2}k_{\xi}\left(\xi-\xi_0\right)^2$ is a harmonic restraining force on the CV with the spring constant k_{ξ}.

It was derived by Woo *et al.* [307] that the equilibrium binding constant K_{eq} can be expressed as

$$K_{eq} = \frac{\int_{\text{site}} d\mathbf{1} \int d\boldsymbol{X} e^{-\beta w}}{\int_{\text{bulk}} d\mathbf{1}\, \delta(\boldsymbol{r}_1 - \boldsymbol{r}_1^*) \int d\boldsymbol{X} e^{-\beta w}} \tag{2.45}$$

where $\mathbf{1}$ and $\boldsymbol{X}$ are the state variables of the ligand and the receptor (which, in the context of this work, are the drug molecule and the DNA) respectively [10]. $\boldsymbol{r}_1$ is the centre-of-mass of the ligand and $\boldsymbol{r}_1^*$ is an arbitrary point (far away from the receptor) in the solvent environment. Here, the subscripts "site" and "bulk" denote the intercalation site and the solvent environment respectively. Hence, basically, the numerator counts the number of microstates which correspond to complete intercalation of *only one* ligand, whereas the denominator counts the number of microstates corresponding to cases where the ligand does not interact with the receptor at all.

Using the same logic as the chain rule in differential calculus, Woo [307] asserted that Eq. 2.45 can be expanded as

$$\begin{aligned} K_{eq} = {} & \frac{\int_{\text{site}} d\mathbf{1} \int d\boldsymbol{X} e^{-\beta w}}{\int_{\text{site}} d\mathbf{1} \int d\boldsymbol{X} e^{-\beta[w+u_c]}} \times \frac{\int_{\text{site}} d\mathbf{1} \int d\boldsymbol{X} e^{-\beta[w+u_c+u_o]}}{\int_{\text{site}} d\mathbf{1} \int d\boldsymbol{X} e^{-\beta[w+u_c+u_o+u_a]}} \\ & \times \frac{\int_{\text{site}} d\mathbf{1} \int d\boldsymbol{X} e^{-\beta[w+u_c+u_o+u_a]}}{\int_{\text{bulk}} d\mathbf{1}\, \delta(\boldsymbol{r}_1 - \boldsymbol{r}_1^*) \int d\boldsymbol{X} e^{-\beta[w+u_c+u_o]}} \times \frac{\int_{\text{bulk}} d\mathbf{1}\, \delta(\boldsymbol{r}_1 - \boldsymbol{r}_1^*) \int d\boldsymbol{X} e^{-\beta[w+u_c+u_o]}}{\int_{\text{bulk}} d\mathbf{1}\, \delta(\boldsymbol{r}_1 - \boldsymbol{r}_1^*) \int d\boldsymbol{X} e^{-\beta[w+u_c]}} \\ & \times \frac{\int_{\text{bulk}} d\mathbf{1}\, \delta(\boldsymbol{r}_1 - \boldsymbol{r}_1^*) \int d\boldsymbol{X} e^{-\beta[w+u_c]}}{\int_{\text{bulk}} d\mathbf{1}\, \delta(\boldsymbol{r}_1 - \boldsymbol{r}_1^*) \int d\boldsymbol{X} e^{-\beta w}} \end{aligned} \tag{2.46}$$

[10] We have used "1" here to denote the first ligand molecule. We have made the assumed that for every receptor only one ligand molecule is allowed to bind. See Sec. B.1 for rigorous derivation.

where u_c, u_o and u_a are the external forces applied on the system to restrain the ligand *c*onformation, *o*rientation and *a*ngular position with respect to the receptor. Here, u_c is a function of the root-mean-square deviation of the ligand, u_o is a function of the Euler angles (Θ_1, Φ_1, Ψ_1) and u_a is a function of the angles (θ_1, ϕ_1). The normal and Euler angles and distance are defined in the manner as shown in Fig. 2.6, using three atom groups from each of the ligand and the receptor.

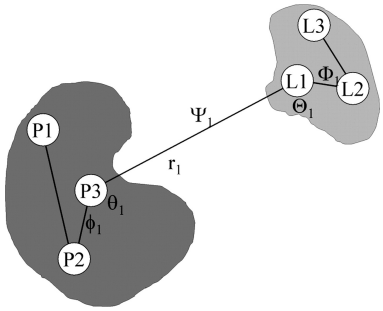

Figure 2.6: Definitions of reference angles and distance between ligand and receptor. Image from [307].

Now, consider the expectation value of an arbitrary function $f(x)$ [170],

$$\langle f(x) \rangle = \int f(x) p(x) dx \qquad (2.47)$$

for any normalised probability density function (p.d.f.) $p(x)$. Obviously, this equation can be extended to a function of any dimensions, i.e. $\varphi(\boldsymbol{X})$ with $\boldsymbol{X} = \{x_1, \ldots, x_n\} \in \mathbb{R}^n$, and the p.d.f., now $\pi(\boldsymbol{X})$ instead of $p(x)$, does not necessarily have to be normalised either, in which case the general form of the expectation value becomes

$$\langle \varphi(\boldsymbol{X}) \rangle = \frac{\int \cdots \int \varphi(\boldsymbol{X}) \pi(\boldsymbol{X}) \, d\boldsymbol{X}}{\int \cdots \int \pi(\boldsymbol{X}) \, d\boldsymbol{X}}, \qquad (2.48)$$

where the new denominator, which did not appear in Eq. 2.47, serves as the normalisation constant for the non-normalised p.d.f.. Comparing Eqs. 2.48 and 2.46, we see that five of the six terms in Eq. 2.46, *viz.* those with both "site" or both "bulk" integrals on the numerator and the denominator, resemble the form of Eq. 2.48. For example, the first fraction

$$\frac{\int_{\text{site}} d\mathbf{1} \int d\boldsymbol{X} e^{-\beta w}}{\int_{\text{site}} d\mathbf{1} \int d\boldsymbol{X} e^{-\beta [w + u_c]}} \equiv \frac{1}{\left\langle e^{-\beta u_c} \right\rangle}$$

with $e^{-\beta w}$ being the p.d.f.. Similarly, for the second fraction,

$$\frac{\int_{\text{site}} d\mathbf{1} \int d\boldsymbol{X} e^{-\beta [w + u_c]}}{\int_{\text{site}} d\mathbf{1} \int d\boldsymbol{X} e^{-\beta [w + u_c + u_o]}} \equiv \frac{1}{\left\langle e^{-\beta u_o} \right\rangle}$$

whose p.d.f. is now $e^{-\beta [w + u_c]}$. Inductively, it follows for all the other similar terms, that the fractions represent the expectation values (or the inverse) of the exponentials of the respective constraints. Moreover, since they all match with the form of Eq. 2.44, the two examples above can be written as $e^{+\beta \Delta G_c^{\text{site}}}$ and $e^{+\beta \Delta G_o^{\text{site}}}$ respectively. This then reveals the possibility of a very convenient way of computing the angular components of K_{eq}, which is the sequential assessments of components with the appropriate constraints added upon one another at each assessment like matryoshki, the Russian nested dolls [123, 307].

Now, for the fourth fraction in Eq. 2.46, since the numerator is a "site" integral whereas the denominator is a "bulk" integral, the fraction does *not* map directly onto the expectation

value discussion above, and thus needs special attention. Woo *et al.* proved that the term can be expressed as a product, which reads [307]

$$\frac{\int_{\text{site}} d\mathbf{1} \int d\boldsymbol{X} e^{-\beta[w+u_c+u_o+u_a]}}{\int_{\text{bulk}} d\mathbf{1}\, \delta(\boldsymbol{r}_1 - \boldsymbol{r}_1^*) \int d\boldsymbol{X} e^{-\beta[w+u_c+u_o]}} = S^* I^* \tag{2.49}$$

where

$$\begin{cases} S^* = (r^*)^2 \displaystyle\int_0^{\pi} d\theta \sin\theta \int_0^{2\pi} d\phi\, e^{-\beta u_a(\theta,\phi)} \\ I^* = \displaystyle\int_{\text{site}} dr\, e^{-\beta[w(r)-w(r^*)]} \end{cases} \tag{2.50}$$

with w being the PMF calculated in the presence of all constraints u_c, u_o and u_a. Eq. 2.49 then basically denotes the likelihood of binding whilst keeping the conformations and orientations both of the ligand and of the receptor, by means of the constraints u_c and u_o.

Another way of understanding Eq. 2.49 is via Eq. 2.50. We see that the S^* integral is a double polar integral which does not have an r-dependent term, whereas the I^* term is a single radial integral which depends only on r but not the angles. This in turn implies that the S^* term calculates the total free energy integrated on the surface of the sphere with radius r^* where the integrand maps out the free energy landscape of the sphere. The I^* integral, however, takes care of the radial part of the binding energy, and thus calculates the PMF when the ligand is pulled away *radially* from the receptor to the surface of the sphere.

For the fifth fraction in Eq. 2.46, Woo *et al.* asserted that it can be calculated analytically for any system [307], and the analytical form of the integral reads [122]

$$\begin{aligned} \frac{\int_{\text{bulk}} d\mathbf{1}\, \delta(\boldsymbol{r}_1 - \boldsymbol{r}_1^*) \int d\boldsymbol{X} e^{-\beta[w+u_c+u_o]}}{\int_{\text{bulk}} d\mathbf{1}\, \delta(\boldsymbol{r}_1 - \boldsymbol{r}_1^*) \int d\boldsymbol{X} e^{-\beta[w+u_c]}} &= \frac{1}{8\pi^2} \int_0^{\pi} d\Theta \sin\Theta \int_0^{2\pi} d\Phi \int_0^{2\pi} d\Psi\, e^{-\beta u_o} \\ &= \frac{1}{8\pi^2} \int_0^{\pi} d\Theta\, e^{-\beta u_\Theta} \sin\Theta \int_0^{2\pi} d\Phi\, e^{-\beta u_\Phi} \int_0^{2\pi} d\Psi\, e^{-\beta u_\Psi} \end{aligned} \tag{2.51}$$

where $u_o = \frac{1}{2}\left[k_\Theta(\Theta_0-\Theta)^2 + k_\Phi(\Phi_0-\Phi)^2 + k_\Psi(\Psi_0-\Psi)^2\right] = u_\Theta + u_\Phi + u_\Psi$.

Lastly, since the constraining forces u_o and $u_a = u_\theta + u_\phi = \frac{1}{2}\left[k_\theta(\theta_0-\theta)^2 + k_\phi(\phi_0-\phi)^2\right]$ are high dimensional, it is difficult to keep track of the contributions of each of the components (i.e. CVs). Luckily, when put in the exponential, the CVs are separable as they are mutually exclusive to one another, and our "matryoshka model" can be used on a per-CV basis. For instance, for the second fraction (i.e. the $e^{+\beta\Delta G_o^{\text{site}}}$ term in Eq. 2.46),

$$\begin{aligned} \frac{\int_{\text{site}} d\mathbf{1} \int d\boldsymbol{X} e^{-\beta[w+u_c]}}{\int_{\text{site}} d\mathbf{1} \int d\boldsymbol{X} e^{-\beta[w+u_c+u_o]}} &\equiv \frac{\int_{\text{site}} d\mathbf{1} \int d\boldsymbol{X} e^{-\beta[w+u_c]}}{\int_{\text{site}} d\mathbf{1} \int d\boldsymbol{X} e^{-\beta[w+u_c+(u_\Theta+u_\Phi+u_\Psi)]}} \\ &= \frac{\int_{\text{site}} d\mathbf{1} \int d\boldsymbol{X} e^{-\beta[w+u_c]}}{\int_{\text{site}} d\mathbf{1} \int d\boldsymbol{X} e^{-\beta[w+u_c+u_\Theta]}} \times \frac{\int_{\text{site}} d\mathbf{1} \int d\boldsymbol{X} e^{-\beta[w+u_c+u_\Theta]}}{\int_{\text{site}} d\mathbf{1} \int d\boldsymbol{X} e^{-\beta[w+u_c+(u_\Theta+u_\Phi)]}} \\ &\times \frac{\int_{\text{site}} d\mathbf{1} \int d\boldsymbol{X} e^{-\beta[w+u_c+(u_\Theta+u_\Phi)]}}{\int_{\text{site}} d\mathbf{1} \int d\boldsymbol{X} e^{-\beta[w+u_c+(u_\Theta+u_\Phi+u_\Psi)]}} \\ &= e^{+\beta\Delta G_\Theta^{\text{site}}} \times e^{+\beta\Delta G_\Phi^{\text{site}}} \times e^{+\beta\Delta G_\Psi^{\text{site}}}. \end{aligned} \tag{2.52}$$

Another advantage of "stacking the matryoshki" CV-by-CV over adding constraints in bulk, is the convergence of values, i.e. the efficiency of the calculations. For example, if τ_Θ, τ_Φ and τ_Ψ are the time (in terms of simulation steps) required for each of the terms in Eq. 2.52 to converge [11], then the total time required for attaining convergence of Eq. 2.52 will be trivially $\tau_\Theta + \tau_\Phi + \tau_\Psi$, as the three terms are evaluated separately. However, if the constraints are applied simultaneously as u_o, the total time for convergence would be the product $\tau_\Theta \tau_\Phi \tau_\Psi$, which is much larger than $\tau_\Theta + \tau_\Phi + \tau_\Psi$.

Combining all the terms above, we have for the overall equilibrium binding constant

$$K_{eq} = S^* I^* e^{-\beta\left[\Delta G_c^{\text{bulk}} + \Delta G_o^{\text{bulk}} - \Delta G_c^{\text{site}} - \Delta G_o^{\text{site}} - \Delta G_a^{\text{site}}\right]}. \tag{2.53}$$

Last but not least, since the product $S^* I^*$ has the units of Å^3 and the exponential is dimensionless by definition, the binding constant thus calculated also has the units of Å^3. In order to convert to the more familiar *per molar* (= mol^{-1} dm^3) in biomolecular sciences, we divide the result by a factor of 1660.54 (= $\frac{10^{27}}{N_A}$, $N_A \sim 6.022 \times 10^{23}$ = Avogadro constant).

2.7.2 Extended-system ABF (eABF) method

What has been described in the previous subsection when we discussed the ABF method, is the step-by-step procedure to obtain the equilibrium constant and hence the free energy change of the interaction. Basically, the discussion above is all about the relationship among $w(\xi)$, u_ξ and ΔG_ξ. In this subsection, we will discuss in details about $w(\xi)$, which is the most important piece of information needed in the calculation, yet whose method of evaluation remains unexplained.

In traditional ABF implementations [134], the CVs of interest are taken to be functions of the real spatial coordinates of atoms, hence $z = \xi(\boldsymbol{q})$ for $\boldsymbol{q} \subseteq \mathbb{R}^{3N}$, with the coordinates subject to an external constraint of the closed form $\sigma_k(\boldsymbol{q}) = 0$. It was then proven [48, 66, 173] that such system provides, in a long run, a uniform distribution of z (or ξ) and gives an unbiased estimate of $w(z)$, the PMF [175], such that

$$w(z) = -k_B T \ln \rho(z) \tag{2.54}$$

Moreover, there are two important conditions which must be fulfilled in the traditional ABF method, *viz.*

$$\begin{cases} \boldsymbol{\nu} \otimes (\boldsymbol{\nabla}\xi) = \underline{\underline{\mathbf{I}}}^{3N} \\ \boldsymbol{\nu} \otimes (\boldsymbol{\nabla}\sigma_k) = \underline{\underline{\mathbf{0}}} \end{cases} \tag{2.55}$$

where $\boldsymbol{\nu}$ is any arbitrary vector field (also known as the "inverse gradient" [134]), $\underline{\underline{\mathbf{I}}}^{3N}$ is the $(3N \times 3N)$ identity matrix and $\otimes$ is the outer product operator. Lelièvre *et al.* [172] asserted that by using an invertible autocorrelation matrix $\underline{\underline{\mathbf{G}}} = (\boldsymbol{\nabla}\xi) \otimes (\boldsymbol{\nabla}\xi)$, one of the solutions for $\boldsymbol{\nu}$ could then be $\boldsymbol{\nu} = \underline{\underline{\mathbf{G}}}^{-1} \cdot (\boldsymbol{\nabla}\xi)$. However, albeit mathematically valid and robust, the evaluation of the $\boldsymbol{\nu}$ field quickly becomes impractical due to the size of the system and the

[11] The scaling factor for free energy calculation in terms of the system size and the phase-space dimensions is very complicated to derive. For the same system and same phase-space dimensions, it is much easier to communicate using the absolute convergence time as the common language.

difficulty in inverting $\underline{\underline{\mathbf{G}}}$.

For the free energy profile $w(z)$, things are even nastier as the preliminary calculation goes only as far as the *gradient* of w and an integrator has to be implemented to retrieve w. Furthermore, the evaluation of $\boldsymbol{\nabla} w$ *per se* involves not only $\boldsymbol{v}$, which is already nearly impossible to obtain, but its *divergence* as well. The whole term then reads [45, 175]

$$\boldsymbol{\nabla} w(z) = \left\langle \boldsymbol{F}^{\xi} \right\rangle_{\xi(q)=z} = \left\langle \boldsymbol{\nabla} V \cdot \boldsymbol{v} - k_B T \boldsymbol{\nabla} \cdot \boldsymbol{v} \right\rangle_{\xi(q)=z} \tag{2.56}$$

where $\boldsymbol{F}$ is the instantaneous collective force, and $V = V(\boldsymbol{q})$ is the system total potential energy.

Lesage *et al.* [175] reported an alleviation to this problem, by coupling the reaction coordinates to a fictitious and non-physical degree of freedom λ, with an associated "mass" m_λ. The additional degree of freedom can be thought of as an extended system in the Car-Parrinello dynamics sense [33, 148], and hence the method is named "extended-system ABF". Such coupling is done via a fictitious potential $V'(\boldsymbol{q}, \lambda) = \frac{k}{2}(\xi(\boldsymbol{q}) - \lambda)^2$, where k is the spring constant for the coupling, then the entirety of the eABF algorithm would be transformed from the q-dependent reaction coordinate $\xi(\boldsymbol{q})$ into purely λ-dependent extended coordinate $\xi^{\text{ext}}(\boldsymbol{q}, \lambda) = \lambda$.

With this, the total *extended* potential of the system (*excluding* biasing force) becomes

$$V^{\text{ext}}(\boldsymbol{q}, \lambda; k) = V(\boldsymbol{q}) + \frac{k}{2}(\xi(\boldsymbol{q}) - \lambda)^2 \tag{2.57}$$

and the associated probability density of distribution of λ reads

$$\begin{aligned}
\rho(\lambda; k) &\sim \int d\boldsymbol{q} \exp\left(-\beta V^{\text{ext}}(\boldsymbol{q}, \lambda; k)\right) \\
&= \int d\boldsymbol{q} \exp\left(-\beta V(\boldsymbol{q})\right) \exp\left(-\beta \frac{k}{2}(\xi(\boldsymbol{q}) - \lambda)^2\right) \\
&= \int dz\, \rho(z) \exp\left(-\beta \frac{k}{2}(z - \lambda)^2\right)
\end{aligned} \tag{2.58}$$

which can be written as $\rho(z, \lambda; k)$ to emphasise that it is a *joint* distribution.

In the case where a biasing force is applied, an extra term has to be added to Eq. 2.57 to account for the energy associated with the extra force, hence

$$\tilde{V}^{\text{ext}}(\boldsymbol{q}, \lambda; k) = V(\boldsymbol{q}) + \frac{k}{2}(\xi(\boldsymbol{q}) - \lambda)^2 - w(\lambda; k). \tag{2.59}$$

Following the same logic as above, the associated distribution, now with the biasing force added, becomes,

$$\begin{aligned}
\tilde{\rho}(\boldsymbol{q}, \lambda; k) &\sim \exp\left(-\beta \tilde{V}^{\text{ext}}(\boldsymbol{q}, \lambda; k)\right) \\
&= \exp\left(-\beta V(\boldsymbol{q})\right) \exp\left(-\beta \frac{k}{2}(\xi(\boldsymbol{q}) - \lambda)^2\right) \exp\left(\beta w(\lambda; k)\right) \\
\tilde{\rho}(z, \lambda; k) &= \int dz\, \tilde{\rho}(\boldsymbol{q}, \lambda; k)\, \delta(\xi(\boldsymbol{q}) - z)
\end{aligned}$$

$$= \rho(z) \exp\left(-\beta \frac{k}{2} (z-\lambda)^2\right) \exp\left(\beta w(\lambda; k)\right) \tag{2.60}$$

with which the biased marginal distribution of z alone can be obtained by integrating with respect to λ. Hence,

$$\tilde{\rho}(z) = \rho(z) \int d\lambda \exp\left(-\beta \frac{k}{2} (z-\lambda)^2\right) \frac{1}{\rho(\lambda; k)} \tag{2.61}$$

Finally, the slope of the free energy landscape can be estimated using the so-called "corrected z-averaged restraint" (CZAR) [175], which reads

$$\frac{dw(z)}{dz} = -\frac{1}{\beta} \frac{d \ln \tilde{\rho}(z)}{dz} + k \left(\langle \lambda \rangle_z - z\right) \tag{2.62}$$

where $\langle \lambda_z \rangle_z$ is the conditional average of λ at a given value of z. It is noted in Lesage *et al.* [175] that the CZAR estimator (Eq. 2.62) can be trivially extended to arbitrary dimensions:

$$\boldsymbol{\nabla} w(\boldsymbol{z}) = -\frac{1}{\beta} \boldsymbol{\nabla} \left(\ln \tilde{\rho}(\boldsymbol{z})\right) + k \left(\langle \boldsymbol{\lambda} \rangle_{\boldsymbol{z}} - \boldsymbol{z}\right). \tag{2.63}$$

However, since only scalar constraints are applied in this work, further discussion regarding higher dimensional methods is out of the scope of this thesis.

Chapter 3

Preliminary investigations on MD simulation of DNA intercalation

3.1 Introduction — intercalation as rare events

In a previous discussion in Chapter 1, we have introduced the intercalation process as one of the many ways through which the DNA is mutated. We have also mentioned such mutations in the DNA affects the production of amino acids and subsequently proteins. Such changes in the composition of proteins may be devastating to the body, as they may cause dysfunction in the regulation of cell process and cell reproduction.

However, in view of the average lifespan of human beings (about 75 years) and the fact that most cases of cancers emerge after the age of 40, we know that mutations cannot *successfully* happen very frequently. In fact, mutations of all causes do happen all the time but DNA has excellent repairing mechanisms which prevent the vast majority of such changes becoming permanent [99]. As a result, the germline mutation rate in humans is kept to a very low value of about 0.5×10^{-9} per base pair per year [248], whereas the rate of somatic mutations is about an order higher ($\sim 10^{-7}$) than that of germline mutation [1] [204].

In the context of computational simulation of cell processes using all-atom models, since a typical simulation timescale is in the regime of nanoseconds to microseconds, the mutations rates quoted above translate to $\sim 10^{-26}$ and $\sim 10^{-25}$ per base pair per ns, for germline and somatic mutations respectively. This implies that the probability of visualising this happening during a "normal" MD simulation is extremely thin (unless we simulate an extremely long DNA sequence, say $> 10^{20}$ bps, for several ms; or a normal ~ 100 bp sequence for $\sim 10^{23}$ ns [2], which are both impossible), hence the term "rare events" for such reactions.

Fortunately, there are a few algorithms which allow us to boost the probability of seeing such reactions happening within simulation timescales. Targeted MD (TMD) [250,251] and accelerated MD (aMD) [125] are two of the examples which were used in the early stages in this work. The rest of this Chapter will be dedicated to explaining these methods, with particular focus on the discussion of preliminary results and the limitations of them which

[1] Germline mutations are mutations which happen in germ cells which would develop into either sperms or ova, whereas somatic mutations are those which happen in normal tissue cells.

[2] This converts to nearly 3.2 million calendar years!

led to the choice of the method adopted in later chapters rather than further development of usage of the two aforementioned algorithms.

Throughout the work presented in this Chapter, we have used the drug ellipticine, which is one of the topoisomerase II inhibitors which can intercalate into the DNA. Ellipticine was used as a preliminary test case because it has a relatively simple structure with a four-fused aromatic ring structure which can easily be inserted between base pairs, but without side chains which induces steric hindrance which hampers the intercalative actions.

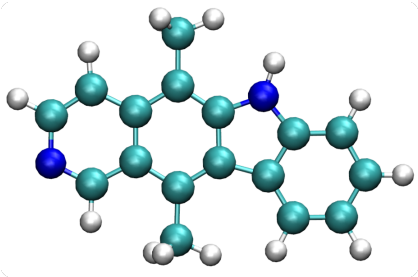

Figure 3.1: An ellipticine molecule.

3.2 The obtainment of force field parameters for ellipticine

Since the AMBER toolbox only includes parameters for biological molecules (e.g. amino acids, nucleic acids, etc.) and a selection of trace metal ions which exist in biological entities, any external molecules which a user wishes to add to their simulations must be treated before they can be used. Such treatments include the geometry optimisation of the molecules and the determination of their partial charges. This small section is dedicated to the explanation of the procedure we followed for all the external molecules, using ellipticine as an example.

In order to obtain the crude coordinates of the atoms in the ellipticine molecule, we first created the structure in AVOGADRO [126] using its SMILE code. The crude structure was then geometry-optimised using CASTEP [46, 137, 158, 228, 231]. The post-optimisation structure was used as the input for the tool ANTECHAMBER within the AMBERTOOLS16 toolbox. ANTECHAMBER was used to determine the charge distributions and the atom types. In particular, we calculated the partial charges on the atoms using the AM1-BCC theory, as elucidated before in Chapter 2 as the force field parameters were discussed.

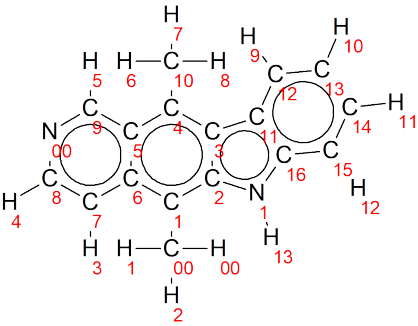

Figure 3.2: Numbering system for atoms in ellipticine.

As an example, Table 3.1 shows the partial charges determined using the aforementioned tool.

C15 (−0.156)	H12 (+0.139)	C16 (+0.031)	N1 (−0.219)	H13 (+0.230)	C14 (−0.099)
H11 (+0.134)	C13 (−0.159)	H10 (+0.134)	C12 (−0.079)	H9 (+0.137)	C11 (−0.075)
C3 (−0.062)	C4 (+0.005)	C10 (−0.188)	H6 (+0.091)	H7 (+0.089)	H8 (+0.093)
C2 (+0.052)	C1 (−0.110)	C00 (−0.168)	H00 (+0.075)	H1 (+0.084)	H2 (+0.090)
C6 (+0.021)	C5 (−0.110)	C9 (−0.039)	H5 (+0.158)	N00 (−0.146)	C8 (−0.078)
H4 (+0.158)	C7 (−0.175)	H3 (+0.143)			

Table 3.1: AM1-BCC partial charges of atoms in ellipticine (in units of e), calculated using ANTECHAMBER with a tolerance of $10^{-3}e$.

3.3 Constrained MD simulation

In order to make an extremely rare event such as intercalation happen, there are two routes we can take:

1. *Force* the system to make the desired transition;
2. Smoothen the free energy landscape so that the probability of transition can be drastically improved.

Constrained MD algorithms take the first route and allow only few degrees of freedom in the transition whilst constraining the vast majority of the others. TMD is an example of constrained MD. Since TMD is primarily built on a more primitive model, known as "targeted energy minimisation" (TEM), we will explain the two models briefly with particular focus placed on TMD.

3.3.1 TEM and TMD

The method of "Targeted Energy Minimization" ("TEM") [80], first reported in 1992, is a computational method to guide an arbitrary initial structure to a specified final configuration. This is done through alternate atomic shifts and energy minimisation until the desired configuration is attained. It was first used to investigate the structural transition of insulin. The algorithm is explained briefly below:

For an N-atom system, vectors of $3N$ coordinates are determined ($\boldsymbol{I}$ for initial, $\boldsymbol{F}$ for final). Typically, from the practical point of view, the initial ($\boldsymbol{I}$) and the target ($\boldsymbol{F}$) states are created and energy-minimised seperately. For the application in our work, we will further explain the protocol being used in the next subsection. For a transformation performed in n steps, the next intermediate position vector $\boldsymbol{L}_1$ will be an n-th closer to $\boldsymbol{F}$. Hence, for each intermediate (transient) $\boldsymbol{L}_i$ (where $i = 1, 2, \ldots, n-1$), the distance from the target structure is

$$\|\boldsymbol{F} - \boldsymbol{L}_i\| = \frac{n-1}{n} \|\boldsymbol{F} - \boldsymbol{I}\| \tag{3.1}$$

Each step consists of two sub-steps, namely "enforced displacement" and "relaxation displacement". During enforced displacement, the shift is performed on the linear route between the current ($\boldsymbol{I}$ if initial, $\boldsymbol{L}_i^{\mathrm{EM}}$ otherwise) and the final state, until the next inner layer of the concentric arc is reached [3]. Then "relaxation displacement" is carried out by performing

[3] In this 2D representation, the intermediary states are represented by two-dimensional arcs, but in reality for a system with $3N$ particles, the states should be concentric hyperspheres in $3N$ dimensions.

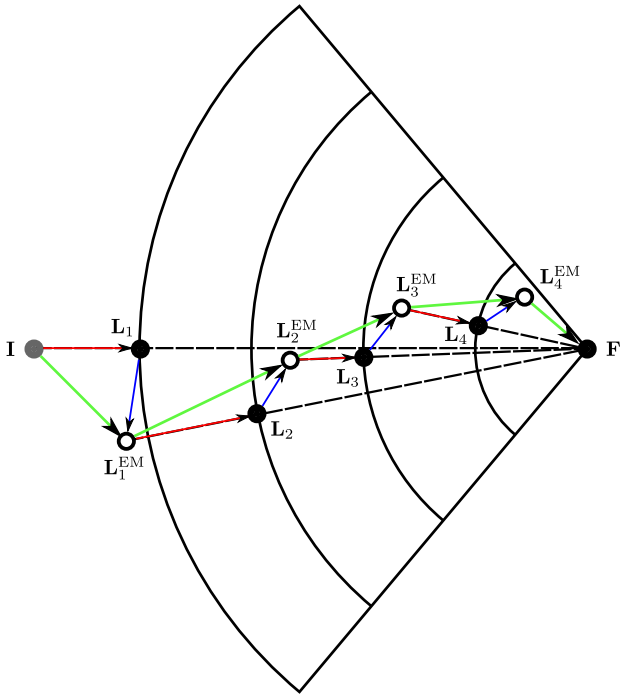

Figure 3.3: Schematic representation of a 5-step TEM on a 2D plane. Red arrows: enforced displacements. Blue arrows: minimisation displacements. Green arrows: resultant shift for each step. Reproduced from [80].

a conjugate gradient minimisation for all degrees of freedom, attaining the $\boldsymbol{L}_i^{\mathrm{EM}}$ transient state vectors (The hollow dots in Fig. 3.3. "EM" denotes *energy-minimised*). The new state vector after each enforced displacement is thus

$$\boldsymbol{L}_i = \boldsymbol{L}_{i-1}^{\mathrm{EM}} + \boldsymbol{t}_{i-1}\left(d_{i-1} - \frac{n-1}{n} d_{\mathrm{tot}}\right) \tag{3.2}$$

where d_{tot} is the root-mean-square deviation (RMSd) of atom positions between $\boldsymbol{F}$ and $\boldsymbol{I}$, d_{i-1} the RMSd from $\boldsymbol{L}_{i-1}^{\mathrm{EM}}$ to $\boldsymbol{F}$ and $\boldsymbol{t}_{i-1}$ the unit vector targeted from $\boldsymbol{L}_{i-1}^{\mathrm{EM}}$ to $\boldsymbol{F}$. In summary, the algorithm of TEM can be listed out as follows:

1. At $i = 0$, $\boldsymbol{I} = \boldsymbol{L}_0^{\mathrm{EM}}$
2. $i \longrightarrow i+1$
3. Calculate $\boldsymbol{L}_i$ from $\boldsymbol{L}_{i-1}^{\mathrm{EM}}$ using Eq. 3.2
4. Perform energy minimisation of $\boldsymbol{L}_i$ to obtain $\boldsymbol{L}_i^{\mathrm{EM}}$
5. If $i < n$ go to Step 2, otherwise end.

Hence the transition pathway (see Fig. 3.3, green path) from $\boldsymbol{I}$ to $\boldsymbol{F}$ using an n-step TEM is:

$$\boldsymbol{I} \rightarrow \boldsymbol{L}_1^{\mathrm{EM}} \rightarrow \boldsymbol{L}_2^{\mathrm{EM}} \rightarrow \cdots \rightarrow \boldsymbol{L}_{n-1}^{\mathrm{EM}} \rightarrow \boldsymbol{F}$$

Note that TEM is a totally static method in pulling a structure to a final (desired) configuration, as neither of the displacements within a step is done dynamically according to Newton's laws of motion — the enforced displacement is merely a manual re-positioning

of atoms according to mathematical results, which is by no means physical, and hence has to be minimised. Moreover, the "EM" structure of each step is obtained from the minimisation of the previous sub-step but not the previous step. Hence the transition from $\boldsymbol{I}$ to $\boldsymbol{F}$ is in fact piecewise and multilinear rather than a smooth pathway. In this sense, the physicality of the transition pathway plotted using TEM increases with the number of steps n.

The method of "targeted molecular dynamics" (TMD) [250, 251] is a relatively new idea emerging from the previous TEM method, by the same group, published only a few months after TEM. In contrast to TEM, TMD is a proper *dynamical* algorithm which determines the trajectories of atoms solely through the integration of the Newton's laws.

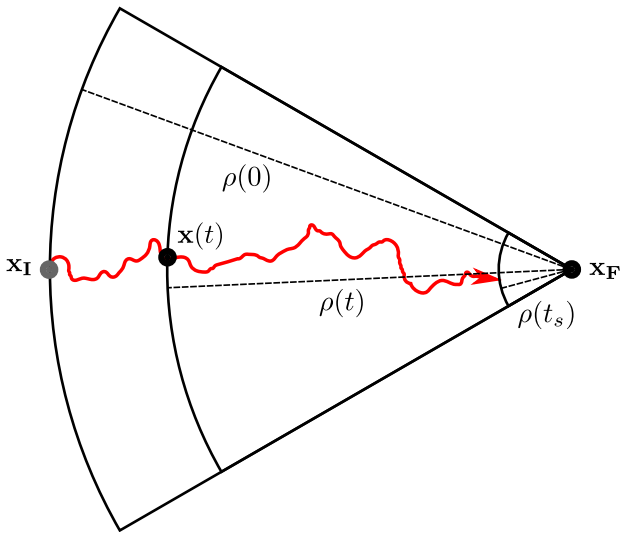

Figure 3.4: Schematic representation of TMD on a 2D plane. Red arrows: trajectory of molecule under TMD constraints. Reproduced from [250, 251].

Rather than having an energy minimisation procedure done every time step as in TEM, TMD takes a quasi-continuous approach. Just like TEM, the targeted RMSd for a single step is shrunken by an n-th of $\|\boldsymbol{x}_F - \boldsymbol{x}_I\|$, i.e. the total distance between the final and initial state vectors. A constraining force is imposed on all the degrees of freedom, with the functional form $\boldsymbol{F}^c = 2\lambda(\boldsymbol{x} - \boldsymbol{x}_F)$, where λ is a Lagrange parameter. Note that the direction of the force points *away* from the final vector. This is crucial as it steadily slows down the transition, eventually "soft-landing" the whole process, so as to prevent possible overkills and undesired fluctuations around the final structure (hence a waste of computational effort) [4].

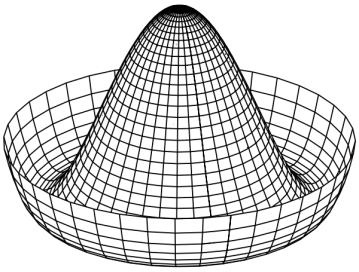

Figure 3.5: A Mexican hat shaped Higgs-like potential.

The most significant difference between TEM and TMD can be demonstrated in the following thought experiment. Assume a particle is inside a "Higgs-like" (Mexican hat shaped, 3D

[4] An analogy to this constraining force would be the dissipative (or damping) force in a harmonic motion (HM). The system undergoing a HM can be *critically damped* with the carefully-chosen constraining force, but *underdamped* or even undamped without it.

with cylindrical symmetry, cf. Fig. 3.5) potential and the goal of the experiment is to move it to the opposite side of the "hat". By using TEM, because it is a totally static method, the particle is first moved linearly up towards the top of the hat — since the shortest pathway (on the plane of projection) is a straight line. However because the local energy minimum occurs at the trough, the particle will be "rolled" (statically) back to the bottom upon minimisation. Eventually TEM will either roll the particle to its destination along the trough, or even more likely, fail to simulate such a system. However, in real life, (classical) dynamical systems are rarely in their ground states, and small potential barriers are surmountable given a suitable initial velocity or thrust, which could be given as the constraint force in TMD. In TMD, the particle will be pushed upwards toward the top of the hat and roll back down the other slope. Hence for systems which are prone to influences of external forces, TMD is much more physical than TEM.

The algorithm of TMD is listed as follows (cf. Fig. 3.4):

1. Establish the initial distance $\rho = \rho(0) = \|\boldsymbol{x}_F - \boldsymbol{x}_I\|$.
2. Choose initial coordinates $\boldsymbol{x}(0)$ and appropriate initial velocities
3. Integrate the equations of motion with the additional constraining force $\boldsymbol{F}^c$
4. Decrease ρ by a *constant* amount of $\Delta\rho = (\rho(0) - \rho(t_s))\,\dfrac{\Delta t}{t_s}$ after each time step Δt, where t_s is the total simulation time. Steps 3 and 4 are repeated until $t = t_s$.

Note that Fig. 3.4 is only a 2D representation of TMD. In reality, for an N-atom system the state vectors should reside on a $3N$-dimensional hypersphere. The "r-axis", i.e. the horizontal axis, is selected such that the initial configuration defines the origin and $\boldsymbol{x}_F$ lies horizontally, positively and collinearly with it. The vertical axis is called the "s_i-axis" and it accounts for one of the residual Cartesian coordinates. Hence, in our N-atom system, the hypersphere should have *one* r-axis and $(3N-1)$ s_i-axes.

While this kind of coordinate scheme can effectively ensure the forced transition happens in most cases, it also brings some serious drawbacks. First of all, by shrinking ρ continuously, TMD is able to keep the state vector on the surface of the hypersphere, at least at the end of every time step. However it does not constrain all the degrees of freedom at the same rate; it uses the *average* of all the coordinates and hence leaves a massive freedom for the system to transform in the unconstrained dimensions. In other words, no matter the state vector is at the "north pole" or the "equator" at a particular time step t, it is still valid and legitimate so long as it still lies on the surface of the hypersphere of radius $\rho(t)$. This implies that, since the distance is calculated as $\rho = \left(r^2 + \sum_i {s_i}^2\right)^{1/2}$, degenerate solutions are bound to exist for different sets of $(r, \{s_i\})$, $\forall \rho \neq 0$. Put simply, unless $\rho \equiv 0$ (which would imply a *unique* set of $(r, \{s_i\}) = (0, \{0\})$, TMD has little control over how far off a particular coordinate deviates.

Secondly, since the coordinate system of the atoms is intrinsically chosen to be Cartesian, the application of TMD is bound to produce translational transitions only, but not rotational ones. For example, it is extremely difficult to simulate the transition of a left-handed DNA helix into a right-handed one which involves the unwinding and re-winding of the helix and hence is an angular transition. It has been tested in this work, that TMD failed to unwind a left-handed B-form DNA and re-wind it as a canonical right-handed B-DNA.

Thirdly and lastly, as ρ shrinks over time, the system ideally converges to the final, i.e. designated, configuration. However, as explained in the previous paragraphs, while TMD can exert control over one or several coordinates, the more coordinates it controls the higher the deviations would the uncontrolled coordinates have. In simpler words, at best the final distance $\rho(t_s)$ can be diminished to a small but finite number, but ultimately $\rho = 0$ is unattainable.

3.3.2 DNA-ellipticine intercalation using TMD

To demonstrate TMD, we created a very simple system with an 8bp sequence of $d(ACTGACTG)_2$ and made two branches out of it. The first branch consists of an extra ellipticine molecule, manually placed at about 15Å from the centroid of the DNA molecule. The other branch consists of the same ellipticine molecule, but manually placed at a intercalated position, between the two base pairs in the middle. The two systems were energy-minimised separately using 500 steps of line minimisation. Then they were heated up to 300K and equilibrated with constraints imposed to maintain the relative distance between the external molecule and the DNA.

After the systems were equilibrated, Branch 2 (ellipticine pre-intercalated) was used as the reference system (i.e. the system with configuration $\boldsymbol{F}$ in the discussion above) and Branch 1 (i.e. the system with configuration $\boldsymbol{I}$ in the discussion above) was used to simulate the transition into intercalation, with a TMD bias applied. The general simulation protocol consists of a 1ns run *with* TMD, followed by another 1ns run with*out* TMD to let the system freely evolve. Three simulations runs were performed, each with a different Lagrange factor (or force constant) λ. λ was set to be 0.05, 0.10 and 0.15 kcal mol^{-1} $Å^{-1}$ for the three sets respectively.

In terms of general observation of the intercalation behaviour, we found that rather than the ellipticine ramming into the interstitial site, in turn forcing the two immediately adjacent base pairs to open up, the base pairs opened up first, letting the intercalator into the site. It was a surprising finding, as it was as though the bases *actively* weakened the original $\pi - \pi^*$ stacking interactions and formed intermediary bonds with the similar delocalised orbitals in the ellipticine.

3.3.3 Limitations and drawbacks of TMD

Although the above-said findings seem to be promising and prove that TMD is useful for the simulation of intercalations, TMD has a few important limitations which can be demonstrated from the individual behaviours of the three systems being simulated.

Figure 3.6 shows the root-mean-square deviation of the systems with respect to the reference structure $\boldsymbol{F}$ (Branch 2), with the colour coding explained in the caption of the figure. The way of reading the graph is as below.

In the graph, each "step" before $t = 1000$ps (e.g. the sharp drop of the red line from 1.95Å to about 1.4Å at around $t = 170$ps) indicates a stage in the intercalation process. For instance, for the red line, the plateau around $t = 2.5 - 170$ps corresponds to the period when

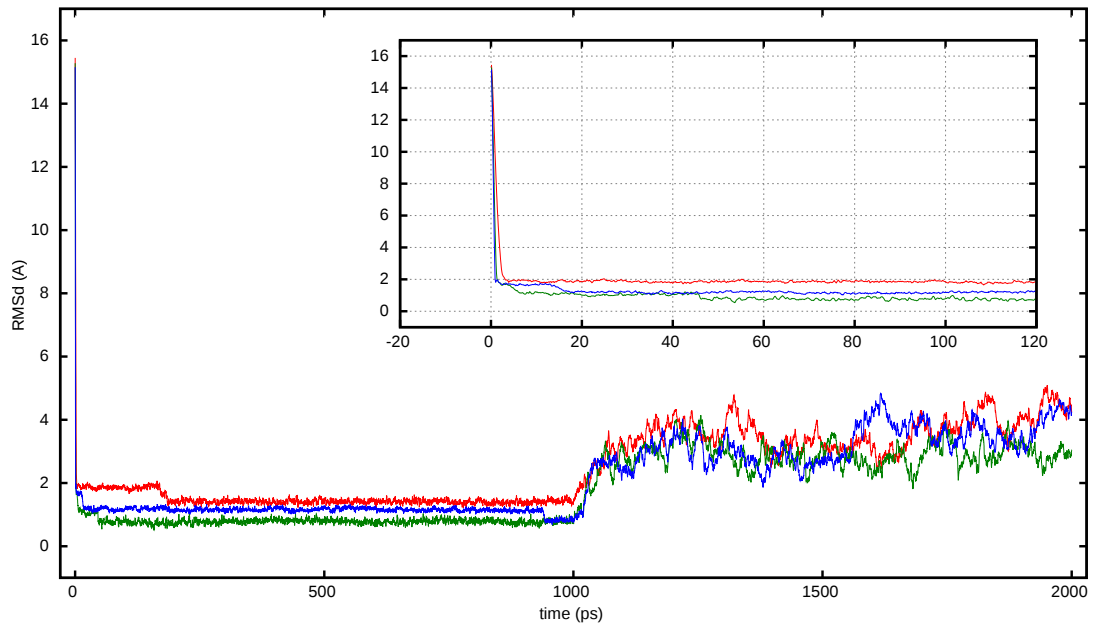

Figure 3.6: Instantaneous RMSd of system from the preset final structure. (Colour legend — Red: $\lambda = 0.05$, Green: $\lambda = 0.10$, Blue: $\lambda = 0.15$)

the intercalator stayed in the groove and aligned with the base pairs. The lower flat region which spans from $t = 185 - 1000$ps corresponds to the *intercalated* state. At $\lambda = 0.05$, TMD was successful in guiding the ellipticine to intercalate into the designated gap.

At $\lambda = 0.10$, although the molecule took much less time (only 17ps) to align itself with the base pairs, it went into the *wrong* gap initially. It was found from the simulation snapshots that, initially, rather than going into the gap directly, it drifted itself in between two nucleobases *within the same base pair*. It stayed in position for about 40ps before slipping back into the correct gap.

At $\lambda = 0.15$, since the force which led the molecule into the DNA was so strong, the entrance of it into the DNA was not smooth like the previous cases. It crashed into the DNA instead and quickly intercalated, again, in the *wrong* gap (the one directly adjacent to that designated) and stayed there for more than 850ps, until it swapped its position with the bases and corrected its position.

One of the most interesting yet striking results, which was visualised only at $\lambda = 0.15$, was that the positional swap between the base pair and the ellipticine was not only a translation swap (i.e. slipping against each other), but it also involved the flipping of one of the bases. This has not been observed in previous MD simulations of DNA systems, perhaps because in normal cases the planes of bps are held roughly parallel to each other through the $\pi - \pi^*$ stacking of the bases' molecular orbitals, which would be broken (fully or partially) due to the opening of the gaps upon intercalation.

The moral behind these findings is that, for every system there seems to be a "perfect" force constant which could directly and smoothly guide the system to the designated state, and a "threshold" beyond which could possibly cause a similar but wrong transition. This means

that tests have to be done very carefully on a per-system basis in order to obtain that value.

Taking the simulations performed in this section as an example, from the $\lambda = 0.05$ case we can see that intercalation is a multistep process, and that it is a *converging* one, in that the nearer to the intercalated state the fewer microstates there would be. Mukerjee *et al.* [208] suggested that between the groove-binding and the intercalated stages, there is an excited-state transition which is highly dependent on the orientations of the DNA and the ligand, which has been observed in this study. We assert that a high enough value of λ would provide extra energy to surmount the free energy barrier (i.e. excitation). On the other hand, if we chose too small a value for λ, then the intercalator would be stuck in the groove-binding state.

Moreover, since the TMD algorithm guides the system to transit to the designated configuration via the *straightest* possible path in the $3N$-dimensional hyperspace, the variations in the trajectory of particles due to stochastic thermal fluctuations (cf. Eqs. 2.23 and 2.24) would be rather low even with a different seed for the random number in the simulation. This implies that if we want to assess the energy changes associated with the transition by probing the external TMD force applied to drive the transition, what would be obtained is the value for the (near-)perfect trajectories, whereas the values for other trajectories which have contributions towards the overall energy change remain very much unexplored. In technical terms, we attribute it to the under-sampling of the phase-space. In order to facilitate a more efficient way of sampling the phase-space, we may use the methods described in the following section.

3.4 Energy landscape modification strategies

In the previous section, we have mentioned that there are two routes one may take to visualise rare events in simulations. We have also discussed in detail the application of constrained MD as one of the two routes, using TMD as an example. In this section, we will switch to the second strategy, which is the modification of free energy landscapes. The basic principle behind this strategy is that the rate of a (single-step) reaction is primarily determined by the energy barrier which the system needs to surmount in order to make the desired transition. Mathematically, it can be expressed via the Arrhenius' equation [7, 8]

$$k \sim \exp\left(-\frac{E_a}{k_B T}\right) \tag{3.3}$$

where k is the rate of reaction and E_a is the activation energy. Since E_a is a positive quantity, this implies that the higher the value of E_a, the lower would be the probability of surmounting the barrier, hence a longer period would be needed for the transition to happen on average. Energy landscape modification strategies are those which aim to boost the likelihood (hence, rate) of transitions by means of diminishing the E_a values. Because such alteration in the free energy landscape is global, the rate boost is not confined to specific transitions, but *all* possible transitions of the system. This means that the overall sampling of the phase-space can be greatly enhanced. We will dedicate the rest of this section to the discussion of accelerated MD and the use of this algorithm in our work.

3.4.1 Accelerated MD

Accelerated MD (aMD), as implemented in the NAMD package, is a tool which fills up deep free-energy troughs at run-time. Obviously, in order to assess which troughs to fill and which not to, one has to know the depths of *all* the valleys before making the decision. Whilst it may be viable for small systems such as those used in the test case [201], it quickly evolves into something impractical with the increase in the system size. For the systems studied in this work, it is simply impossible even with the use of sampling methods such as Monte Carlo.

To this end, aMD offers a much smarter solution, which is to make use of the various energies calculated on-the-fly (which is by default) during the MD simulation to determine whether a boost is to be made locally. Such a modification in the potential energy is made by trivially adding a correction term in the potential:

$$V^*(\boldsymbol{r}) = V(\boldsymbol{r}) + \Delta V(\boldsymbol{r}) \tag{3.4}$$

where, if the original potential is lower than a threshold value E, the correction term $\Delta V(\boldsymbol{r})$ takes the form

$$\Delta V(\boldsymbol{r}) \equiv \frac{(E - V(\boldsymbol{r}))^2}{\alpha + E - V(\boldsymbol{r})} \tag{3.5}$$

where $\alpha \in [0, +\infty)$ is a parameter which determines to what extent the potential is modified.

A special case occurs when $\alpha = 0$, where the corrected potential is reduced to the threshold value E. Another feature of aMD is that the minima in $V^*(\boldsymbol{r})$ must be the corresponding ones in $V(\boldsymbol{r})$ at the identical point in space. This in turn implies the overall landscape is preserved (unless $\alpha = 0$), and can be proven straightforwardly by considering the gradients of the two expressions above.

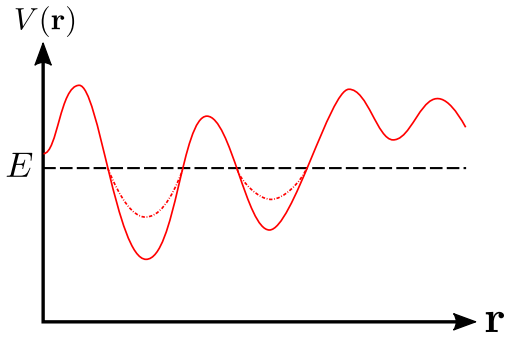

Figure 3.7: Schematic diagram of potential modification in aMD. Reproduced from [125]. Legend: solid red line – original potential profile; dashed red line – modified potential profile.

A neat fact about the aMD algorithm is that, the *boosted* probability density, $p^*(A)$, along any well-defined reaction coordinate $A(\boldsymbol{r})$, can be converted back to the *un*boosted density $p(A)$ through suitable reweighting [201] which reads

$$p(A_j) = p^*(A_j) \frac{\left\langle e^{\beta \Delta V(\boldsymbol{r})} \right\rangle_j}{\sum_{j=1}^{M} \left\langle e^{\beta \Delta V(\boldsymbol{r})} \right\rangle_j} \tag{3.6}$$

where j denotes the j-th bin, M is the total number of bins and $\langle \cdots \rangle_j$ means the canonical ensemble average found in the j-th bin.

3.4.2 DNA-ellipticine intercalation using aMD

To demonstrate the application of aMD in the context of DNA intercalation of drugs, we performed preliminary simulations using this method. This subsection will be dedicated to the discussion of the protocols and results of the calculations.

We first created a short 8 base-pair system with an arbitrary sequence of $d(ACGTACGT)_2$ and added 5 ellipticine molecules around the DNA. The system was then neutralised using 14 sodium ions and was solvated using a truncated octahedral shell (8Å thickness) of TIP3P waters.

In terms of simulation protocol, the system was first energy-minimised crudely using 500 steps of conjugate-gradient minimisation to ensure the absence of atomic collisions. Then it was heated slowly from 0K to 300K in 50ps, and was allowed to attain a steady-state in about 1ns. After the first nanosecond, modification parameters for the dihedral and total energies were calculated using the suggested values in Miao *et al.* [201]:

$$(E_{\text{dihed}}\,;\,\alpha_{\text{dihed}}) = \left(\overline{V_{\text{dihed}}} + 3.5N_{\text{res}}\,;\,0.7N_{\text{res}}\right) \tag{3.7}$$

$$(E_{\text{total}}\,;\,\alpha_{\text{total}}) = \left(\overline{V_{\text{total}}} + 0.175N_{\text{atom}}\,;\,0.175N_{\text{atom}}\right) \tag{3.8}$$

where N_{res} and N_{atom} are the number of *solute* residues and total number of atoms *in the system* respectively, and $\overline{V}$ are the mean potential energies obtained on-the-fly during the equilibration run. Once these parameters had been determined, we performed a further 50ns simulation, turning on the dual-boost (dihedral and total energy components) aMD with them.

Figure 3.8 shows four static snapshots of the simulation at different points of time. We observed that at $t \sim 2.1$ns, one of the ellipticine molecules started to approach from the major groove and eventually intercalated between two base pairs (subfigures 3.8b and c). More intriguingly, at $t \sim 4.2$ns, another ellipticine molecule did the same as the one before, but intercalated into exactly two gaps below where the first molecule intercalated. It was also seen that the receptor DNA was unwound and elongated locally, around the interstitial sites whenever the external molecules stayed intercalated. However, strikingly, while we have witnessed two events of intercalation happening at a very short interval, we also observed the de-intercalation of the second molecule at around $t \sim 14$ns.

Through this study, we proved that aMD was helpful in greatly enhancing the intercalation rate from $\lesssim 10^{-7}$ per bp per ns [5] to ~ 0.5 per bp per ns. Moreover, we have confirmed a previous claim that the general features of the energy landscape do *not* get altered. This was shown through the reproduction of the trajectory of ellipticine from the groove-binding state to the intercalated state in previous studies using TMD [250, 251].

[5] N.B. this number accounts the rate of *all* somatic mutation *combined*, and so should be taken as a ballpark figure only. In fact, with intercalation being only one of the many routes by which the DNA mutates, the true intercalation rate should be much lower than this number.

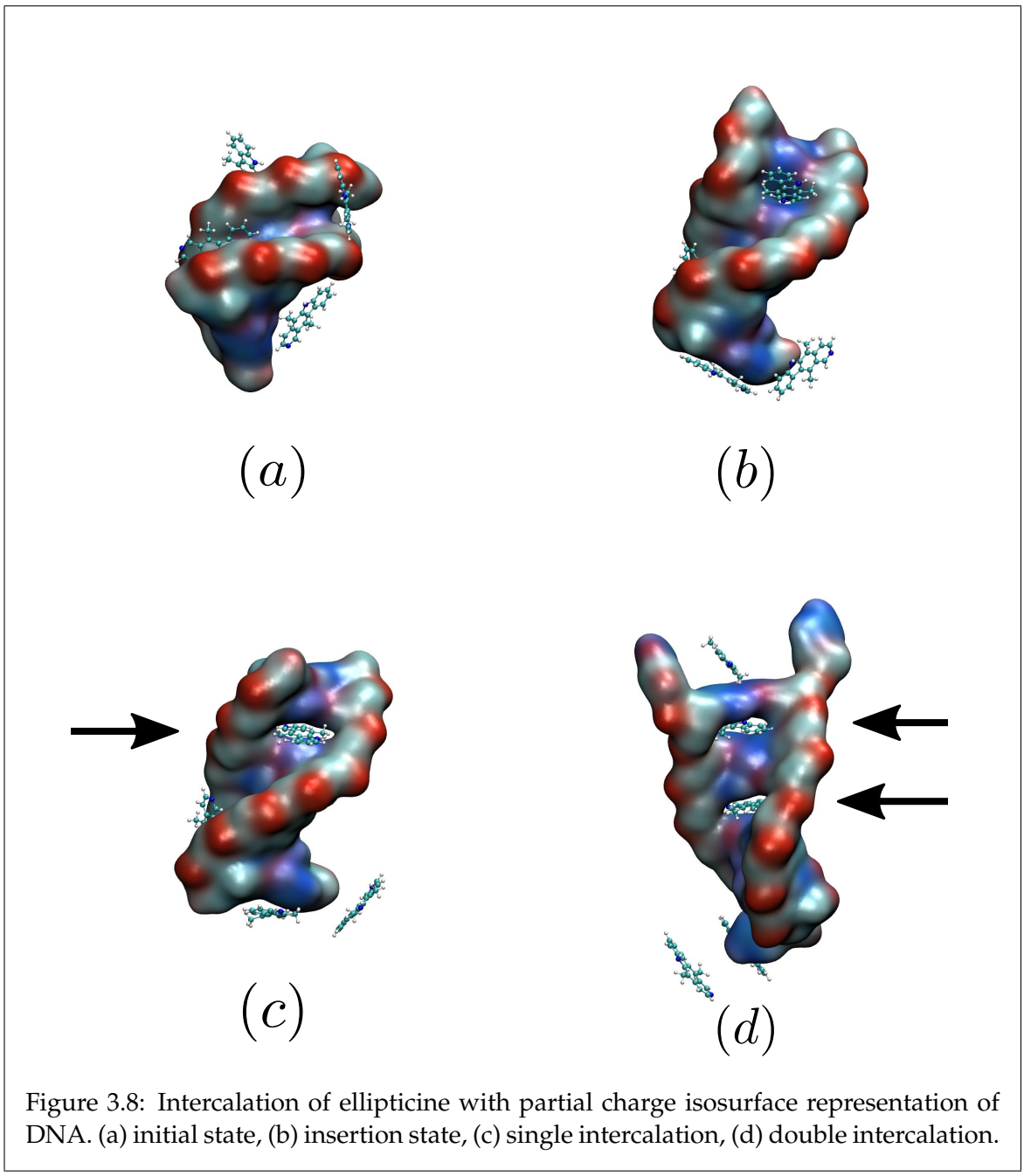

Figure 3.8: Intercalation of ellipticine with partial charge isosurface representation of DNA. (a) initial state, (b) insertion state, (c) single intercalation, (d) double intercalation.

3.4.3 Limitations and drawbacks of aMD

Despite the excellent results obtained from the study mentioned above, the method of aMD does have a number of rather grave drawbacks. This small subsection will be dedicated to the discussion of them.

First of all, aMD is an unbiased method, meaning that the user cannot specify one or more reaction coordinates for the energy boosts; the only options are the total or the dihedral energies, or both. This means that it would be extremely hard for one to assess the energetics associated with a particular transition pathway.

Secondly, although the reweighting formula (Eq. 3.6) look simple, it is practically very hard to evaluate, primarily due to the two ensemble averages. This is because in order to obtain a credible partition function the sample size must be large enough. This in turn suggests

that the length of simulation must be long, which is more viable for small systems than for large ones. Moreover, owing to some technicalities the reweighting needs to be done using the toolbox implemented in the NAMD package by the developers [201], which has not been parallelised and scales rather poorly with system size. For instance, we tried using this code on our system, and found out that the estimated runtime was more than 50 hours which was longer than the simulation runtime itself by far.

Thirdly, and perhaps most fatally, since there are infinitely many pathways a molecule can take to intercalate into the DNA, the energy (or probability) calculated for a single simulation run only corresponds to that particular pathway, and is hence by no means representative of all pathways added up. One of the possible solutions to this is the use of techniques such as umbrella sampling, in conjunction with aMD, then apply the Jarzynski's equality [153] to calculate the average free energy change. However, this requires multiple runs of simulations, which would lead to massive time requirement in the reweighting procedure as described above.

Last but not least, since the current version of aMD boosts only the total potential and the dihedral energies, the aforementioned set of parameters only work with DNA systems in explicit solvents, i.e. systems with large enough number of atoms and residues. We performed a test simulation on the same system described above but in *implicit* solvent environments. We discovered that whilst ellipticines still intercalate, it was due to the further weakening of bonds and interactions within the DNA as there are no water molecules around to restrict the melting.

3.5 Conclusion

In this Chapter, we have discussed the theory and the application of the methods of targeted MD and accelerated MD. Through the preliminary investigations of these methods using a DNA-ellipticine system as a test subject, we demonstrated that both of them are extremely powerful tools for enabling rare biological events to happen even at a simulation timescale.

We have also discussed the limitations and drawbacks of these tools. We found out that while targeted MD takes a direct approach in guiding an unlikely transition to occur, it does not allow system evolution via inter-molecular interactions and thus is much restricted in terms of phase-space sampling.

As an alternative of TMD, we also explored accelerated MD. We demonstrated via the simulation of the DNA-ellipticine system, that accelerated MD is capable of boosting the likelihood of the occurrence of intercalation of ellipticine by more than ten orders of magnitude. However, we also discovered that the backtracking of modification made to the free energy landscape is the pièce-de-resistance of the whole calculation, and is typically tens to hundreds of times more computationally demanding than the MD simulation itself. This implies the evaluation of the free energy change of a transition is very difficult, if not impractical, using accelerated MD. With this, we have also explored other methods of simulation of intercalations and evaluation of the associated free energy, which will be discussed in the next few chapters.

Chapter 4

Study of straight-chained DNA-drug intercalation complexes

4.1 Introduction

As explained in the previous chapters, DNA interacts with the surrounding environment and the molecules in the near vicinity via different modes. As a result of the diverse nature of interactions and the structural complexity of the molecule *per se*, the structural conformation of DNA is rather sensitive to the reactions with external agents.

In this Chapter, we study how the intercalation interactions between DNA and three anthracycline antibiotics, *viz.* daunomycin (DAU), doxorubicin (DOX) and idarubicin (IDA), affect the conformation of the DNA.

The chapter will be divided into two parts. Firstly the simulation protocols used throughout this Chapter will be explained in details. Finally, the Chapter will end with a discussion about the results obtained in this work.

4.2 Preparation of drug molecules

As mentioned in the discussion before, whenever we need to introduce external molecules into the simulation environment using AMBER, we have to determine the force field parameters and charge distributions for them beforehand. The three aforementioned anthracycline antibiotics which will be used for the rest of the work are no exceptions.

As before, we performed geometry optimisation of the structures using CASTEP first to obtain the ground-state structures of the molecules. Then the coordinates of the optimised structures were used as inputs for the software ANTECHAMBER to calculate the charge distributions. Tables 4.1, 4.2 and 4.3 below show the charge distribution determined from ANTECHAMBER calculations using the AM1-BCC theory.

C (−0.248)	H5 (+0.095)	H6 (+0.085)	H7 (+0.086)	C1 (+0.011)	H (+0.107)
O (−0.304)	C2 (−0.045)	H1 (+0.093)	O9 (−0.325)	H28 (+0.206)	C3 (−0.019)
H2 (+0.115)	N (−0.329)	H26 (+0.138)	H27 (+0.162)	C4 (−0.182)	H8 (+0.121)
H9 (+0.113)	C5 (+0.152)	H3 (+0.102)	O1 (−0.276)	C6 (+0.085)	H4 (+0.126)
C7 (−0.196)	H10 (+0.134)	H11 (+0.098)	C8 (+0.052)	C25 (+0.228)	O7 (−0.252)
C26 (−0.272)	H22 (+0.094)	H23 (+0.110)	H24 (+0.109)	O8 (−0.304)	H25 (+0.223)
C9 (−0.151)	H12 (+0.135)	H13 (+0.113)	C10 (−0.028)	C15 (+0.116)	O2 (−0.255)
H14 (+0.257)	C11 (−0.081)	C12 (+0.133)	O6 (−0.267)	H21 (+0.260)	C13 (−0.176)
C14 (−0.169)	C16 (+0.319)	O3 (−0.314)	C17 (−0.075)	C18 (−0.120)	H15 (+0.161)
C19 (−0.078)	H16 (+0.144)	C20 (−0.193)	H17 (+0.147)	C21 (+0.128)	O5 (−0.181)
C24 (−0.081)	H18 (+0.118)	H19 (+0.075)	H20 (+0.078)	C22 (−0.138)	C23 (+0.327)
O4 (−0.296)					

Table 4.1: AM1-BCC partial charges (in units of e) of atoms in daunomycin, calculated using ANTECHAMBER.

C (−0.249)	H5 (+0.095)	H6 (+0.085)	H7 (+0.088)	C1 (+0.011)	H (+0.108)
O (−0.305)	C2 (−0.045)	H1 (+0.091)	O10 (−0.325)	H28 (+0.206)	C3 (−0.018)
H2 (+0.115)	N (−0.329)	H26 (+0.139)	H27 (+0.162)	C4 (−0.181)	H8 (+0.121)
H9 (+0.110)	C5 (+0.152)	H3 (+0.104)	O1 (−0.281)	C6 (+0.085)	H4 (+0.128)
C7 (−0.196)	H10 (+0.127)	H11 (+0.116)	C8 (+0.053)	C25 (+0.206)	O7 (−0.294)
C26 (−0.055)	O8 (−0.320)	H24 (+0.214)	H22 (+0.107)	H23 (+0.140)	O9 (−0.332)
H25 (+0.235)	C9 (−0.151)	H12 (+0.129)	H13 (+0.136)	C10 (−0.026)	C15 (+0.118)
O2 (−0.255)	H14 (+0.258)	C11 (−0.083)	C12 (+0.133)	O6 (−0.269)	H21 (+0.260)
C13 (−0.175)	C14 (−0.168)	C16 (+0.320)	O3 (−0.314)	C17 (−0.075)	C18 (−0.120)
H15 (+0.162)	C19 (−0.078)	H16 (+0.144)	C20 (−0.193)	H17 (+0.147)	C21 (+0.128)
O5 (−0.181)	C24 (−0.081)	H18 (+0.118)	H19 (+0.074)	H20 (+0.078)	C22 (−0.138)
C23 (+0.327)	O4 (−0.296)				

Table 4.2: AM1-BCC partial charges (in units of e) of atoms in doxorubicin, calculated using ANTECHAMBER.

C (−0.248)	H5 (+0.096)	H6 (+0.084)	H7 (+0.086)	C1 (+0.011)	H (+0.108)
O (−0.305)	C2 (−0.045)	H1 (+0.093)	O8 (−0.325)	H26 (+0.206)	C3 (−0.019)
H2 (+0.116)	N (−0.328)	H24 (+0.138)	H25 (+0.162)	C4 (−0.181)	H8 (+0.121)
H9 (+0.113)	C5 (+0.153)	H3 (+0.101)	O1 (−0.277)	C6 (+0.085)	H4 (+0.125)
C7 (−0.197)	H10 (+0.135)	H11 (+0.098)	C8 (+0.052)	C24 (+0.228)	O6 (−0.251)
C25 (−0.272)	H20 (+0.094)	H21 (+0.111)	H22 (+0.110)	O7 (−0.304)	H23 (+0.223)
C9(−0.152)	H12 (+0.136)	H13 (+0.114)	C10 (−0.022)	C15 (+0.118)	O2 (−0.256)
H14 (+0.260)	C11 (−0.085)	C12 (+0.143)	O5 (−0.268)	H19 (+0.262)	C13 (−0.182)
C14 (−0.166)	C16 (+0.320)	O3 (−0.320)	C17 (−0.106)	C18 (−0.083)	H15 (+0.161)
C19 (−0.114)	H16 (+0.144)	C20 (−0.114)	H17 (+0.144)	C21 (−0.083)	H18 (+0.161)
C22 (−0.105)	C23 (+0.323)	O4 (−0.326)			

Table 4.3: AM1-BCC partial charges (in units of e) of atoms in idarubicin, calculated using ANTECHAMBER.

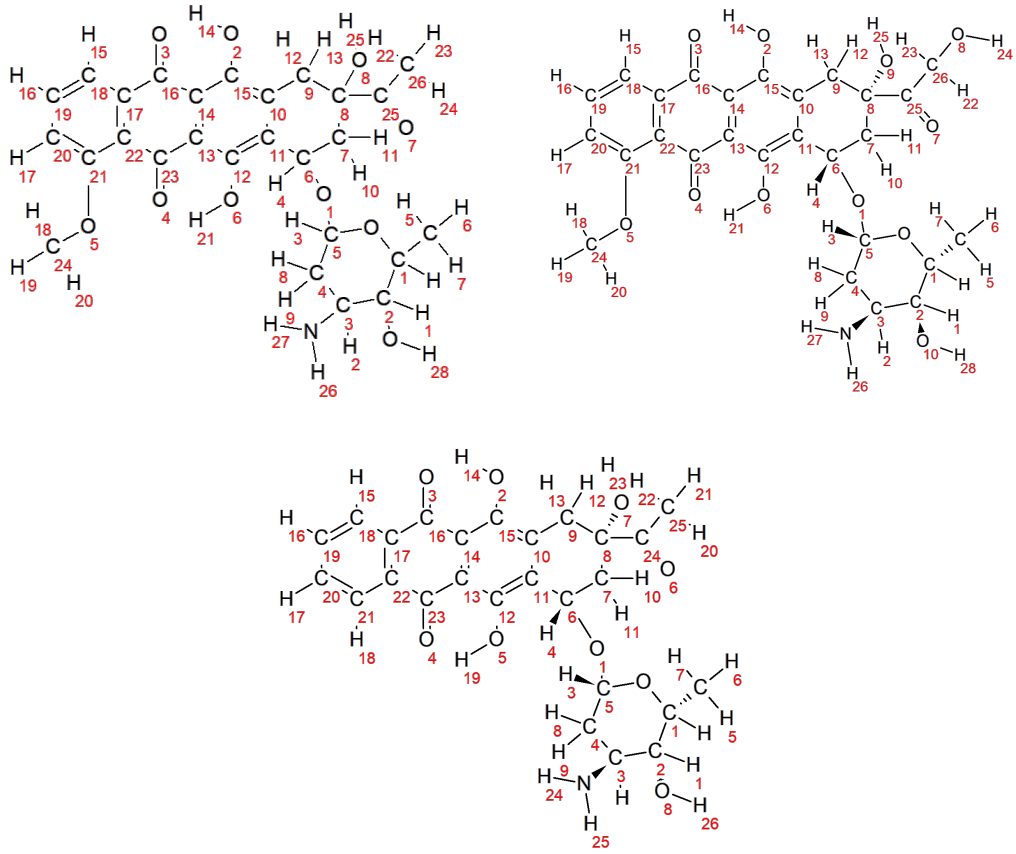

Figure 4.1: Numbering system for atoms in the anthracyclines. (Top left) Daunomycin, (Top right) Doxorubicin, (Bottom) Idarubicin.

As a quick verification of the values, we can use our knowledge of an element's electronegativity χ as a measure of its electron affinity. Firstly, in the Pauling scale [227], $\chi_{\mathrm{H}}(2.20) < \chi_{\mathrm{C}}(2.55) < \chi_{\mathrm{N}}(3.04) < \chi_{\mathrm{O}}(3.44)$ for the four elements existing in the three anthracyclines. This means that the electron cloud should always get pulled away from the hydrogen atoms, whereas for the oxygen atoms they always pull electrons from whatever atoms they bond with. In short, this implies that hydrogens should always have net positive charge whereas oxygens should always have net negative charge, which *is* indeed the case we observe from the tables.

Similar arguments applies to the nitrogen atoms. Since in anthracyclines, the nitrogens only bond with either carbons or hydrogens which both have much lower electronegativity, it is then anticipated that the partial charges on nitrogens in these molecules can only be negative, which again is observed from the data shown above.

4.3 Systems and simulation protocols

In this Chapter, three 72-base pair basal DNA systems have been used throughout the work. The sequences of these basal systems are dA_{72}, $\mathrm{d(AC)}_{72}$ and dC_{72} respectively [1]. The reason

[1] The subscript in this notation denote the *total number of base pairs*, rather than repetition counts in the usual polymer chemistry notations.

behind such a choice of basal systems is that, these three cases represent 0%, 50%, 100% GC-contents in the DNA, hence should give an indicative account for the general case.

The creation of the simulated systems included two steps, *viz.* conversion from NAB code to PDB coordinates and the intercalation of drug into the basal DNA using the XLEAP tool within the AMBERTOOLS16 toolbox. We hereby denote each of the system using the code format "XX-Y-zzzz", where "XX" is the sequence. For instance, "AA" means consecutive AA base steps, hence the dA_{72} oligomer. "Y" in the code means the starting conformation of the basal DNA. "zzzz" tells the type of the drug molecule intercalated (or a bare DNA system). As two examples of the systems used in this chapter, the $d(AC)_{72}$ B-form with a doxorubicin molecule pre-intercalated will have the code "AC-B-dox", whereas the dC_{72} A-form without anything intercalated will have the code "CC-A-bare".

In terms of the simulation, each of the system was first energy-minimised using 1,000 steps of line minimisation, in order to remove any unphysical atomic contacts. The semi-minimised systems then underwent 15 cycles of simulated annealing to attain the ground-state energy. Each of the annealing cycles consisted of a fast heating process (3ps) from 0K to 300K, then a 5ps cooking stage where temperature was maintained by means of Langevin heat bath, followed by a slow cooling process (15ps) back to 0K and another 5ps dissipation stage afterwards to let the residual heat in the system dissipate back to the heat bath at 0K. Finally, after all 15 cycles, the system was allowed another 10ps for the atoms to stop moving.

4.4 Data Analysis

The analysis of data obtained for this part of the work consists of two main parts, *viz.* static analysis and dynamical analysis. This section is dedicated to the explanation of them in details.

In the static analysis of data, the final frame from each of the simulation snapshots was post-processed using the AMBERTOOLS16 utility `cpptraj`. The post-production includes the stripping of the intercalator and a conversion of coordinates format from the binary `.dcd` to pdb. The coordinates of the DNA-drug complexes were then compared with those from the respective bare DNA counterparts. For instance, the coordinates of "AC-B-ida" would be compared with "AC-B-bare", et cetera. The comparisons were done using the "RMSD Visualization" toolkit within the VMD program [145].

Rather than performing an atom-by-atom RMSD calculation, which would give out a single value for each system, averaging out all the atomic contributions, the calculation in this work had been performed on a residue-by-residue, hence nucleotide-by-nucleotide, basis. The benefit of doing so over single value calculations is that, through residue-by-residue calculation, we can obtain much more useful information on how each nucleotide has changed. For example, in this work where each system consists of a 72-bp oligomer, an RMSD value would be given out for each of the 144 ($= 72 \times 2$) bases.

A python script has been written to read in the VMD-generated heatmap (RMSD) file and consolidate data into bp-based information. This is done via a further RMS calculation

between the RMSD value for a specific base and that for its complementary. Hence for the i-th base, the RMSD is set to be

$$\mathrm{RMSD}_i^{\mathrm{bp}} := \sqrt{\frac{1}{2}\left(\mathrm{RMSD}_i{}^2 + \mathrm{RMSD}_{2N-i+1}{}^2\right)} \tag{4.1}$$

where N is the total number of base pairs.

Apart from the base-pair-based RMSD calculation, real-time structural parameter calculations were carried out to probe the structural perturbations brought by the intercalation of different drug molecules. These structural parameters were calculated using the programs CURVES+ and CANAL. The graphs of a selection of parameters are presented in Appendix A. The parameters chosen include those which accounts for the groove dimensions, *viz.* groove widths and depths, and those which describe the nucleobases and the helical structure, *viz.* buckle, roll, shift, slide, and helical rise and twist.

4.5 Results and discussion

We now move on to the discussion of the study of the structural changes brought by the anthracycline drugs.

4.5.1 Bare DNA — energy-minimised

We first investigate the post-annealing RMSD of the base pairs with respect to canonical structures (cf. Fig. 4.2). We can see that for all the cases, irrespective of the original conformation, the RMSD values on the edges are in general higher than those in the middle region. This means the base pairs towards either ends of the DNA strand deviated more from the respective canonical (i.e. starting) structure during the annealing process, than those in the middle of the DNA. This is very much as predicted since the DNA is not closed on itself. The two termini of the DNA were held only by the phosphate backbone on *one* side but not *two* as the other non-terminal nucleotides were. Thus, the edge effects must be prevalent, and they contribute to the partial melting of the molecule.

We also observe that the A-start systems generally have larger range of RMSD across the bases than their respective B-start counterparts. We assert that this is because of the base-tilting in canonical A-form DNA but not in B-form, making the $\pi - \pi^*$ orbital stacking less efficiency in A-form than B-form. This in turn lowers the rigidity of the structure.

We now turn to the structural parameters, presented in Appendix A, which may give us insights into what caused the the RMSD graphs to appear so. Firstly, we look at the Fig. A.1 which gives the parameters for the AA-A-bare system. The subfigure (1,1) [2] (also Fig. 4.3, left) reveals that the widths of the major grooves had widened globally. However, the extent to which they expanded differ by their positions. In particular, we see rather clearly that there are rather regular alternating dark and light red fringes with a period of about 6 to 8 base pairs. We assert that this may be one of the major causes of the rippling pattern in the

[2] The indexing of subfigures follow the "(row, column)" convention — (3,1) means the third row and the first column, etc.

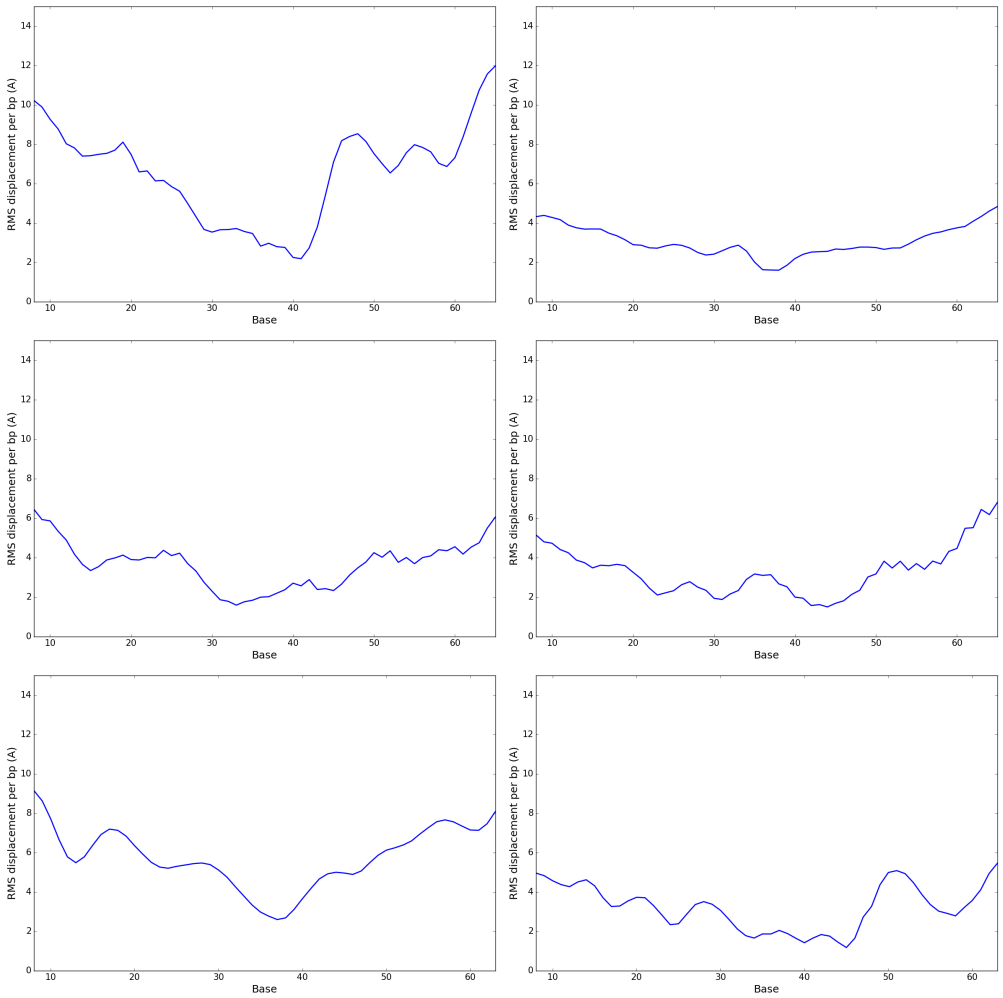

Figure 4.2: RMSD of post-minimisation w.r.t. canonical structures. (Left-to-right, top-to-bottom) AA-A-bare, AA-B-bare, AC-A-bare, AC-B-bare, CC-A-bare and CC-B-bare

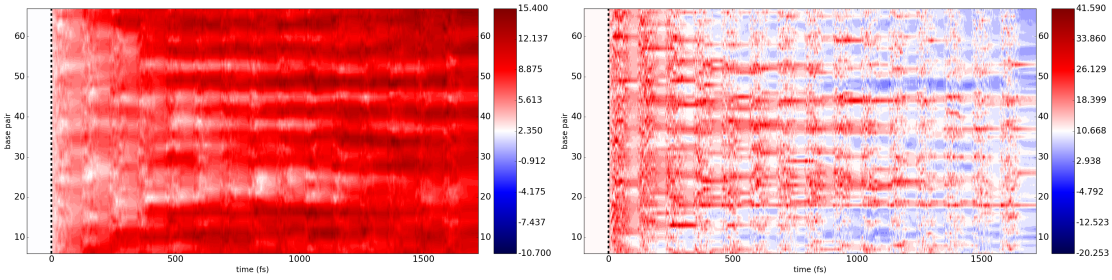

Figure 4.3: (left) Major groove width and (right) roll of AA-A-bare system.

corresponding RMSD graph, which have similar periodicity of the ripples. Similar things happen with subfigure (4,2) (also Fig. 4.3, right) which shows the time variation in the roll parameter: the roll parameter tells the extent of non-parallelarity between two successive base pairs in the groove-ward direction. The figure showed that for the vast majority of the bases, the roll parameter decreased from about 20° to about 5°. However, similar fringe

patterns emerge in this case as in the major width: some of the base pairs remained with relative high roll, and the separation among them (the red fringes in the subfigure) are also about 6 to 8 base pairs.

Finally, we note that there are some anomalies in some of the parameters of the AA-A-bare system. It can be observed that, since the fourth annealing cycle, high differences in the rise, the twist, and the shift between the pairs 18 and 19 have started to occur, and such differences have persisted until the end of the simulation. It was noticed in the simulation snapshots, that the heating stage prior to that might have provided enough energy to break some of the inter-base stacking between the two aforementioned bases and hence produced the abnormal slide. Moreover, due to the same reason, subsequently a kink is produced around these nucleotides, where the local curvature was measured to be as high as 8.75Å^{-1} in some temporal regions, as opposed to $\lesssim 1.5\text{Å}^{-1}$ in other non-kinked regions.

Compared with the AA-A-bare case, the AC-A-bare and CC-A-bare cases have lower spread of the RMSD values across the base pairs. The reason to which is, however, not difficult to see from the structural parameters of the two cases. Firstly, if we look at the same subfigures, (1,1) and (4,2), as we did in the AA-A-bare case, we see that the extent to which some of the major grooves opened up have diminished, which can be understood from the uniformity of the colours in the heat-map. Moreover, regarding the depths of the major grooves, we see, in the AA-A-bare case, that there are two broad regions where the depth drastically decreased by nearly a half. However, this is not observed in the other two A-start systems. Instead, the values of the major groove widths remained rather steady throughout the simulation.

We hereby also make a minor comment regarding the parameters for the B-start systems. Details will not be given fully as before since, firstly, B-start systems have relatively flat profiles of base pair RMSD; secondly, the parameters for the B-start systems appear rather similar across systems. It is observed in all the B-start systems, that the variations in the major groove widths during the annealing processes were higher than those in in the A-start systems. Nevertheless, the variations across bases at the post-annealing states show more consistency than the A-start systems. Moreover, the "red-to-blue" (or vice versa) colour changes in the heat-maps, which signifies that the structural parameters underwent drastic decrease (or increase), are much more prevalent in A-start systems than B-start ones. This means that in general the B-form seems to be more resilient to structural changes.

4.5.2 Daunomycin

In this subsection we discuss the structural perturbation induced by the intercalation of a daunomycin molecule into an inter-base pair site in the same DNA systems used in the previous subsections. The daunomycin was inserted into the system prior to the simulation and the intercalation site was chosen arbitrarily in the central region.

The change brought by the intercalative interaction can be probed when one measures the structural parameters. We have mentioned in earlier chapters that it is a well-studied behaviour that intercalation can cause the interstitial site to widen [174], and it is faithfully shown in the subfigures (3,1) of Fig. A.2 and subsequent corresponding figures in Appendix

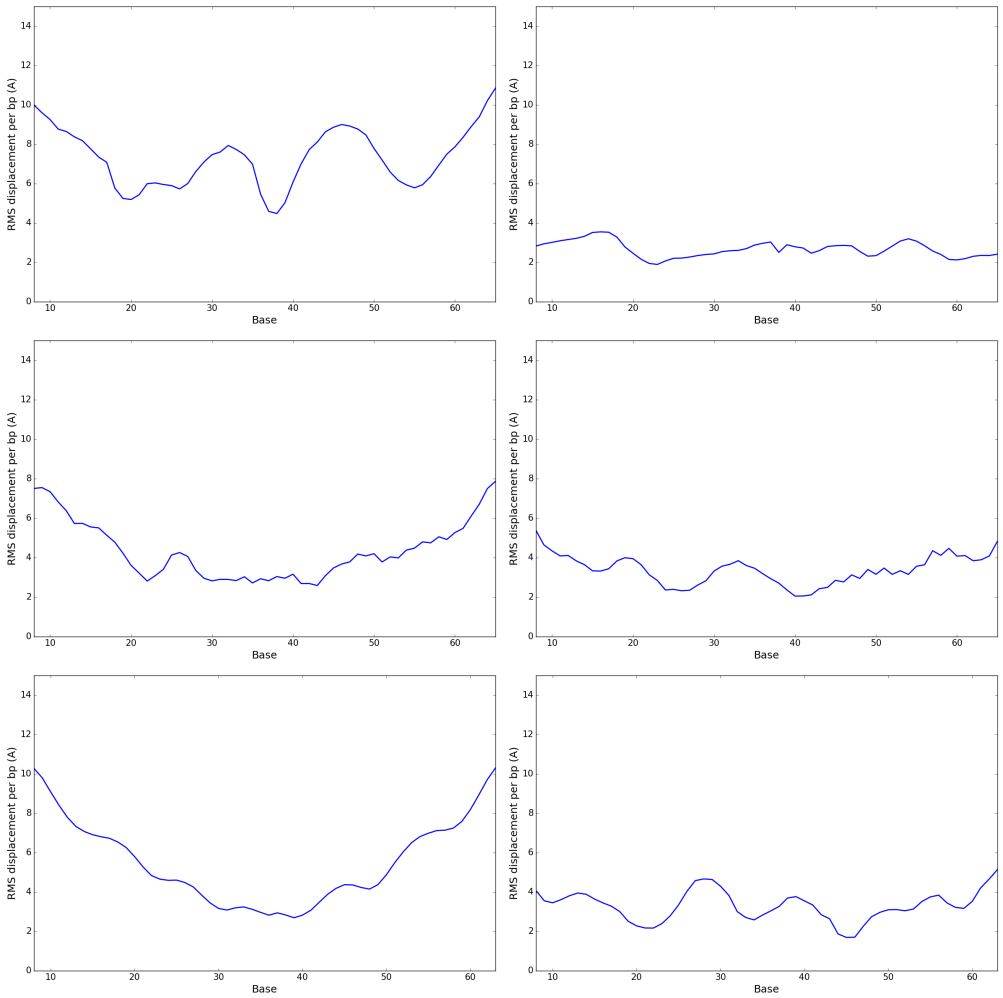

Figure 4.4: RMSD of post-minimisation w.r.t. canonical structures. (Left-to-right, top-to-bottom) AA-A-dau, AA-B-dau, AC-A-dau, AC-B-dau, CC-A-dau and CC-B-dau

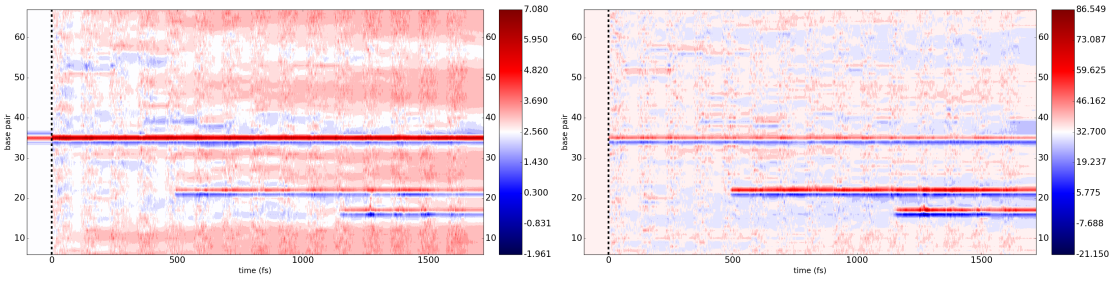

Figure 4.5: (left) Helical rise and (right) helical twist of AA-A-dau system.

A.

Let us inspect the subfigures (3,1) and (3,2) from Fig. A.2 (also corresponds to the left and right subfigures of Fig. 4.5) more closely. In the subfigure (3,1), we see that in the linear minimisation region, the vast majority of the region is white in colour, meaning that the rise is

about 2.56Å. However, at base pair 35 the colour becomes bright red and the two immediate adjacent pairs appear rather dark blue in colour. This means that during the minimisation the intercalation site has widened whereas the immediate previous and next gaps have narrowed. This phenomenon persisted throughout the minimisation period. Once the annealing process has started we observer that the red and blue colours of base pairs 35 and 34 have further intensified, whereas the blue band of pair 36 has nearly disappeared. This implies that through heating the interstitial site has further blown up whereas the natural squeezing of the adjacent gaps have re-partitioned. The fading of the blue bands around the central red strip can be seen even more prevalently in other systems, for example, AC-A-dau: this further shows that the DNA is in fact a soft matter, which would delocalise induced perturbations to preserve original structures.

An even more interesting phenomenon can be found from subfigure (3,2). In the minimisation region, we can see a nearly-uniform pale pinkish colour throughout the minimisation, which signifies that the twist of the system is rather uniformly $\gtrsim 32^\circ$ which is about the value expected for a canonical A-DNA. However, once the annealing cycle started a clear red band is seen at pair 35 and a blue band at pair 34. Moreover, the colour of the red band has faded with the annealing time whereas the blue band has continued to intensify towards the end of the annealing cycles. This means that during the annealing cycles, there is a clear unwinding of the helix around the interstitial cite (from 32° down to as low as $\sim 5^\circ$).

These two observations have not only proved previous experimental results of the blowing up of intercalation sites and the partial unwinding of the helical structure, but have also exposed the weakness of the linear minimisation method. From the vast discrepancies in the pre-annealing and post-annealing values and the relative stability of values during the minimisation periods, we can see that whilst linear minimisation may be fast-converging (converging in as few as 20 to 30 steps), the convergence is only limited to local energy minima; this minimisation method does not have tests for structures with even lower total energy. On the other hand, since annealing involves the partial reordering of the system particles, much wider phase space could be explored and thus configurations with lower energy can be attained.

4.5.3 Doxorubicin and idarubicin

In this subsection, we present the results from the studies of the doxorubicin-DNA and idarubicin-DNA complex systems. Unlike the previous case of daunomycin, since the behaviours of the doxorubicin and idarubicin systems resemble those of daunomycin systems, we only briefly discuss the general appearances of the RMSD profiles and structural parameters, and explain the difference between the doxorubicin and idarubicin cases with the daunomycin systems we have discussed before.

We first note that the profiles of the RMSD of base pairs for the doxorubicin (Fig. 4.6) and for the idarubicin system resemble those of the daunomycin systems. This is much anticipated as the three molecules have very similar structures and so the interactions they have with DNA are thought to be rather similar. Moreover, the DNA which interacted with the

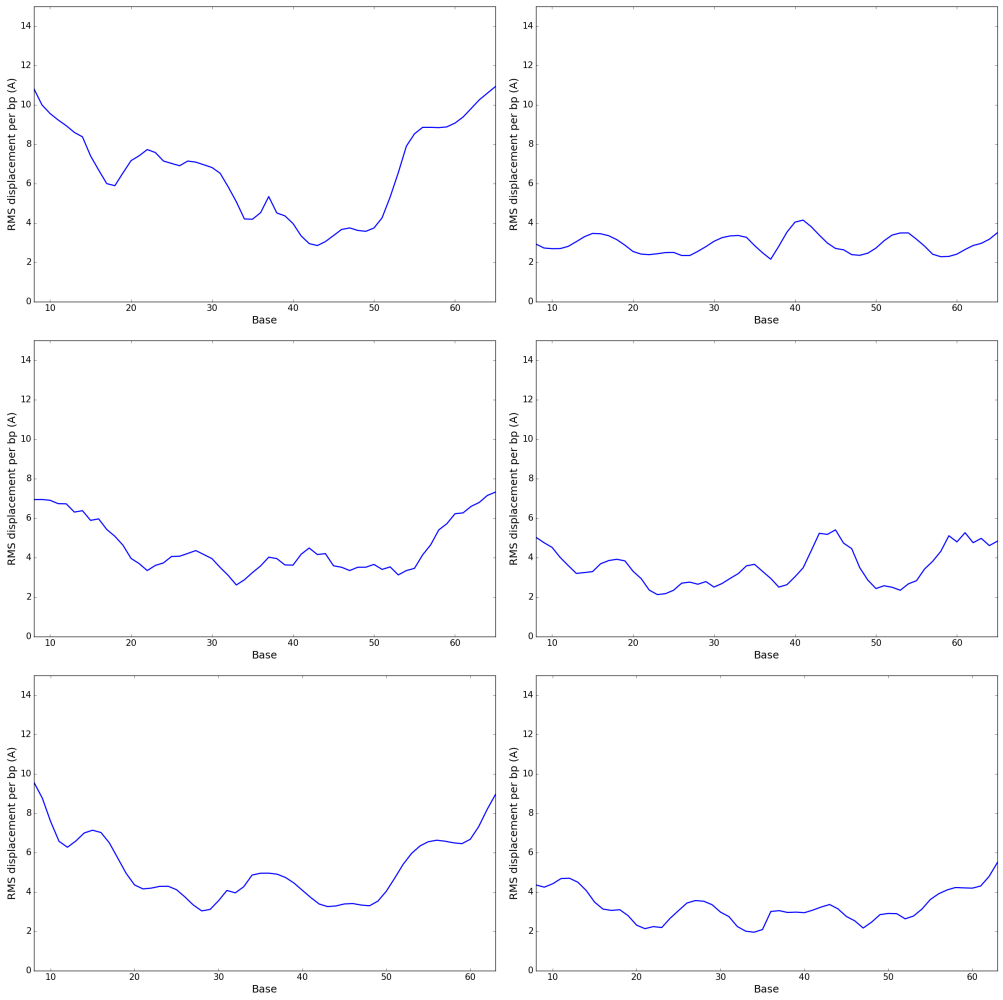

Figure 4.6: RMSD of post-minimisation w.r.t. canonical structures. (Left-to-right, top-to-bottom) AA-A-dox, AA-B-dox, AC-A-dox, AC-B-dox, CC-A-dox and CC-B-dox

molecules have its own rigidity which made it hard to deviate too much from one another.

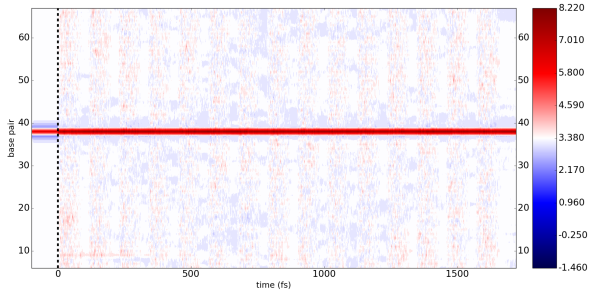

Figure 4.8: Helical rise of the AA-B-ida system.

Lastly, there is one important feature which occurs in all systems involving daunomycin, doxorubicin and idarubicin is the delocalisation of the structural perturbations. For instance, if we look at the subfigure (3,1) of Fig. A.8 (also presented here as Fig. 4.8), which accounts for the helical rise of the AA-B-ida system, we see that in the static energy-minimisation procedure, the minimiser predicted that the rise at the intercalation site would be roughly doubled whereas those at the adjacent sites would be roughly halved in order to compensate the loss in space. This, of course, is based on the assumption that the DNA is a rigid body and other parts of the macromolecule has little information of the local perturbation.

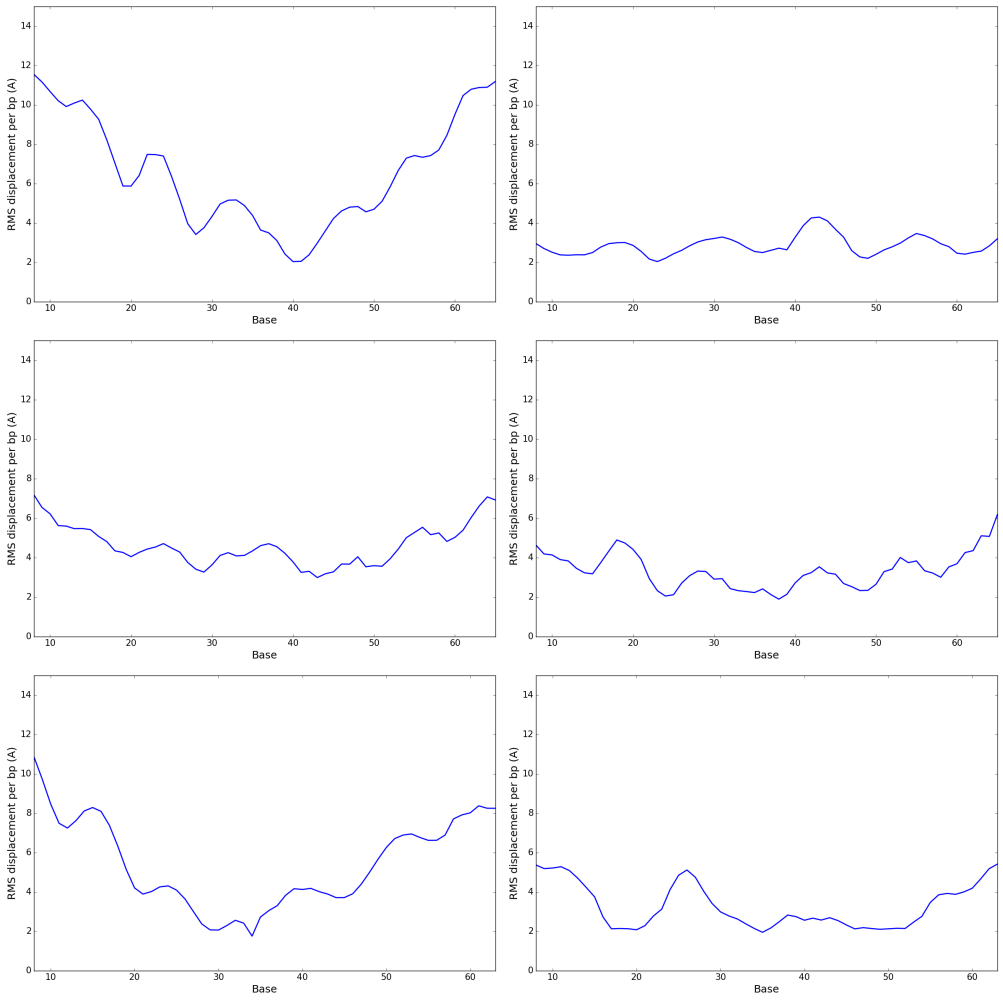

Figure 4.7: RMSD of post-minimisation w.r.t. canonical structures. (Left-to-right, top-to-bottom) AA-A-ida, AA-B-ida, AC-A-ida, AC-B-ida, CC-A-ida and CC-B-ida

However, once the dynamic process had started we could see clearly an intensified red band at the interstital site, whereas the blue bands at the *directly* adjacent sites have weakened and widened to cover several more sites on both sides. This not only echoes with our previous discussion on a similar phenomenon, but also confirms that the DNA is *not* a rigid body but a soft matter. Any *local* shape- or conformation-changing effects caused by external interactions with other molecules will likely be dispersed to other parts of the macromolecule, so that the impact to the function may be reduced.

We stress here that the structural parameters being used throughout this Chapter are only a few examples of those we have calculated. The full collection of parameters for all systems can be found in Appendix A.

4.6 Summary

In this Chapter, we have studied the structural perturbation caused by anthracycline drugs intercalations on DNA. We used traditional energy-minimisation techniques in conjunc-

tion with simulated annealing to obtain steady-state structures of anthracycline-DNA complexes. We performed static analysis on these structures by assessing the per-base pair RMSD with respect to the canonical forms. Lastly, we also developed a method in probing the time-evolution of the DNA's groove and base parameters, with the use of third-party software such as Curves+ and Canal.

From the static study using base pair RMSD, we discovered two major features concerning DNA and its intercalation complex. Firstly, in all cases, we saw that the RMSD values near the termini are in general higher than those near the middle. We assert that this is because of the edge effect, which further stresses the need to use a long enough sequence when performing simulations. However, we also observed that the difference between the values in the middle and those near the termini are larger in the A-start cases than in the B-start cases. This could be another signal of the B-form being a more structurally stable conformation than the A-form. Secondly, we noticed that the magnitudes of the RMSD values did not increase drastically with intercalation. We assert that this is because the RMSD was evaluated at the end of the simulations with the most stable structures, and the local effects have rippled outward to the farther ends and have thus weakened.

From the studies of the time-evolution of structural parameters, we have demonstrated the properties of the DNA as a soft matter. In particular, we observed from the heat-maps that although at the first instance when the intercalator is inserted the helical rise of the interstitial site was immediately increased by 3.4Å, and the helical twist around the site decreased by about 20° to 30°, once the complex was allowed to evolve freely with the surrounding environment, the effects were spread out at once. For instance, we observed that the "squishing" of the adjacent base pairs were fanned out to the bases further down the DNA strand. The same phenomenon happened to the untwisting local to the intercalation site and the compensating overtwisting of the adjacent base pairs. Once the system was allowed to evolve by itself, we observered that the extreme overtwisting of the next base pair seemed to disperse further down the DNA. However, since this compensation was unnatural and an abnormal stress would be induced somewhere in the DNA, it resulted in a nick in the DNA which would move around as the simulation time advanced.

Finally, whilst it is natural to question whether a specific starting conformation would lead to faster structural convergence during annealing processes, we assert that it is rather unlikely that a clear-cut yes-or-no conclusion can be made. This is because whilst the total energy of the system converges exponentially with the number of annealing cycles performed, this is by no means a good indicator for the convergence of the system structures, primarily due to the complexity of the structure of the DNA. For instance, in some of the systems, we saw that nicks and kinks suddenly emerged for several cycles and disappeared some time later, whereas the total energy followed the same exponential decreasing trend. Moreover, since the heating process at each cycle is effectively a re-randomisation of the structure, the inherent stochastic nature of the simulation would then prevent the systems from converging to identical structures.

Chapter 5

Alternative method for structural analysis of DNA-drug complexes

5.1 Introduction

In the previous Chapter, we have studied the structural perturbations caused by the intercalation of the anthracycline drugs daunomycin, doxorubicin and idarubicin into specific 72 base-pair DNA oligomers with different GC contents.

In particular, when we performed analyses on the data obtained from the MD simulations, we used *dynamical* methods to probe those structural perturbations — We have investigated the time-evolution of a selection of structural parameters specific to DNA.

However, we assert that there must be some alternative methods which would allow us to probe the structural changes without having to go through so many parameters which might, from time to time, be rather confusing. We found out that X-ray diffraction method, which has been used by scientists for more than a century now, is very helpful in this respect.

Initiated as a side project to this work, we have written a program called PYRALLEX, which allows user to input the structure (coordinates) of a system and then simulates the X-ray diffraction pattern of the structure. The rest of the Chapter will be dedicated to showcase how these simulations help probe differences between two samples.

The Chapter will first start with the explanation of basic theory of X-ray diffraction. After that, we will explain what PYRALLEX is and its functionalities through two sample outputs. Moreover, we will present a set of novel reliability factors which we devised for the quantitative comparison between two 2D images. Finally, the Chapter will end with the discussion of the 2D images obtained from the simulations done in the previous Chapter.

5.2 Theory of X-ray diffraction

X-rays are a class of electromagnetic radiation, with the typical wavelength in the range of 0.1Å to 100Å. It was discovered in 1912 by Max von Laue [78], that X-rays diffract off and

from crystals. This discovery has led to one of the most applied experimental methods used in the past century, especially in the determination of the detailed structures of crystals and macromolecules. In this section the theory of X-ray diffraction is discussed in details.

5.2.1 Source of radiation

X-ray, characterised by its relatively short wavelength, hence high frequency, is usually created by the bombardment of high energy electrons (~10 keV) at a metal sample. These electrons, having so high an energy, are capable of knocking electrons off from the target metal. The electrons knocked off could be from different shells in the atomic structure, for example, the innermost shell ("K shell") or the second shell ("L shell"), et cetera. The vacancy, or hole, then, would be filled up by another electron in the atom by falling down from an excited state. Such stabilisation of electron, according to quantum mechanics, must give out energy equal to the difference of the shell levels. Since this is an electromagnetic scattering event, the energy given out is hence in the form of a photon, whose energy is given by Einstein's formula [79, 291]

$$\Delta E = (E_n - E_m) = h\nu \tag{5.1}$$

where E are the energies of the respective electron shells, h is the Planck constant and ν is the frequency of the photon emitted. Sometimes it may be more convenient to express quantities in terms of wavelengths rather than frequencies, in which case by using the relation $c = \lambda\nu$, Eq. 5.1 becomes

$$\lambda = \frac{hc}{E_n - E_m} \tag{5.2}$$

where c is the speed of light.

Since there are many possible combinations of m and n in Eq. 5.1, crystallographers have devised a nomenclature for easier communication. In this nomenclature, each radiation is characterised by two letters, an English letter followed by a Greek letter. The English letter tells the final state of the electron, whereas the Greek letter tells the number of shells the electron has fallen. For instance, "K_α" would mean an $L \rightarrow K$ transition, whereas "L_β" indicates an $N \rightarrow L$ transition.

Two of the most commonly used radiation sources are the Cu K_α (copper) and the Mo K_α (molybdenum) series [32], with the wavelengths being 1.5418Å and 0.7107Å respectively [235]. Note that both of them are in the K_α series, i.e. the lowest possible transition, because the Fermi Golden Rule which governs the time-dependent amplitude (hence probability) of the transition reads [199]

$$P_{m\rightarrow n}(t) = 2\,|\langle m|V|n\rangle|^2\,\frac{1-\cos(\omega_{ks}t)}{(E_n - E_m)^2} \tag{5.3}$$

which implies that the probability diminishes as the second power of the energy difference between two shells. This means that the K_α series are the easiest ones to obtain and hence the most commonly used sources. Other metals which are used to generate X-rays include silver, palladium, rhodium, zinc, et cetera [235].

5.2.2 Geometry in X-ray crystallography

A typical experimental setup of an X-ray diffraction (XRD) chamber consists of a source of radiation, a goniometer (a clamp holding the sample which is also capable of rotating the sample in all directions) and a screen to record the diffraction pattern on the opposite side to the source. Then it follows that geometry plays a vital role in the formation of the diffraction pattern. This subsection is dedicated to explain the geometry in XRD.

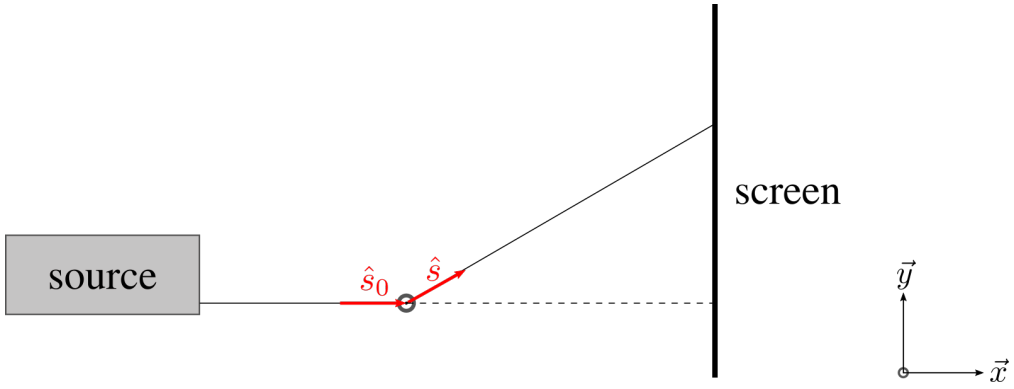

Figure 5.1: Scattering of radiation from charge at origin. Image adopted from [32].

We first consider the simplest case, where a single point charge is placed at the origin (Fig. 5.1). Let the unit vector (hence direction) of the incoming radiation be $\hat{\boldsymbol{s}}_0$ and that of the position of the detector be $\hat{\boldsymbol{s}}$. It is customary to let the angle of deflection be *twice* that of the scattering angle, hence

$$\hat{\boldsymbol{s}} \cdot \hat{\boldsymbol{s}}_0 \equiv \cos(2\theta). \tag{5.4}$$

Define the scattering vector $\boldsymbol{S}$ as

$$\boldsymbol{S} \equiv \frac{1}{\lambda}(\hat{\boldsymbol{s}} - \hat{\boldsymbol{s}}_0) \tag{5.5}$$

where λ is the wavelength of the incoming radiation. $\boldsymbol{S}$ bears the unit of inverse length, and it measures the number of cycles of radiation per unit length.

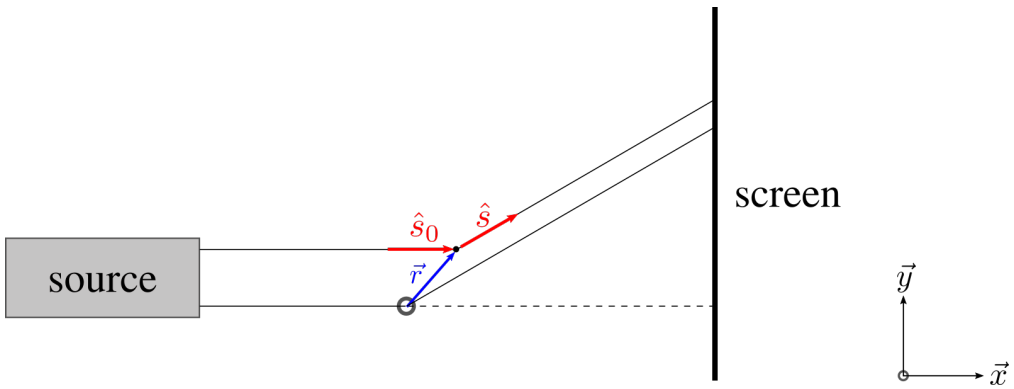

Figure 5.2: Scattering of radiation from charge displaced from origin. Image adopted from [32].

Now, assume the charge is displaced by the vector $\boldsymbol{r}$. Since the size (thus length scale) of the charge is several orders of magnitude smaller than the distance between the charge and the screen, the scattering angle θ still holds by Eq. 5.4. However, since the charge is displaced, the path traversed by the deflected ray has changed, and the path difference is simply $(\hat{\boldsymbol{s}} - \hat{\boldsymbol{s_0}}) \cdot \boldsymbol{r}$. In terms of the light wave, the path difference can be expressed in terms of number of extra cycles travelled by dividing this relation by the wavelength λ, hence,

$$\begin{aligned} \text{p.d.} &= \frac{1}{\lambda} (\hat{\boldsymbol{s}} - \hat{\boldsymbol{s_0}}) \cdot \boldsymbol{r} \\ &= \boldsymbol{S} \cdot \boldsymbol{r}. \end{aligned} \tag{5.6}$$

When expressed in number of extra cycles traversed, the path difference becomes the phase shift in the wave, hence inducing an extra factor of $e^{i2\pi \boldsymbol{S}\cdot\boldsymbol{r}}$ to the radiation scattered at the origin. Then, it follows trivially that for any arbitrary charge distribution $\rho(\boldsymbol{r})$, the radiation scattered onto $\boldsymbol{S}$, or the "form factor", can be generalised as

$$F(\boldsymbol{S}) = \iiint_{\text{all space}} d^3\boldsymbol{r}\, \rho(\boldsymbol{r}) e^{i2\pi \boldsymbol{S}\cdot\boldsymbol{r}} \tag{5.7}$$

which is, of course, the formal definition of the Fourier transform of the charge density.

For the completeness of the discussion, let us use the previous example to demonstrate the calculation of the form factor. For the point charge at the origin, its distribution can be expressed as $\rho(\boldsymbol{r}) = z\delta^3(\boldsymbol{r})$ where δ^3 is the three-dimensional Dirac delta function. Then Eq. 5.7 becomes

$$\begin{aligned} F(\boldsymbol{S}) &= z \iiint_{\text{all space}} d^3\boldsymbol{r}\, \delta^3(\boldsymbol{r}) e^{i2\pi \boldsymbol{S}\cdot\boldsymbol{r}} \\ &= z e^{i2\pi \boldsymbol{S}\cdot\boldsymbol{0}} \\ &\equiv z, \end{aligned}$$

which implies that the form factor is uniform everywhere on the detector screen, in turn meaning that there will not be any patterns formed.

5.2.3 X-ray diffraction of atoms

A real atom can be thought of as a fuzzy cloud of electrons surrounding a tiny core of positively charged nucleus. Whilst photons interact with both electrons and the nucleus, because the wavelength of X-ray is several orders of magnitudes longer than the radius of the nucleus, the diffraction is then mainly due to the interaction with the electron clouds, and hence the nucleus can be safely omitted in X-ray crystallography. As for the electron clouds, the most accurate means of determining the shape and spatial density would be the use of *ab initio* methods. However, since X-ray crystallography mainly deals with the positions of the atoms and the electron configurations do not matter too much, it is customary to adopt rather crude approximations for the electron distributions. There are many different models used by crystallographers, and the one used in this work is by approximating the

electron cloud as a Gaussian sphere [32], hence

$$\rho(r) \sim zNe^{-kr^2}, \tag{5.8}$$

where z is the total charge of the atom, k is a factor related to the width of the atom (thus atomic radius) and N is the normalisation factor (cf. Sec. B.2). The adoption of this form makes the calculation of the form factor much easier, as the Fourier transform of a Gaussian function is another Gaussian function, thus using Eq. 5.7 (see Sec. B.2 for full derivation), we have

$$f(S) = z \exp\left(-\frac{\pi^2 S^2}{k}\right) \tag{5.9}$$

where $S = \|\boldsymbol{S}\|$ is the magnitude of the scattering vector. Note that a different variable name f has been used here as opposed to the F used above. This is because the form factor here only accounts for an atom centred at the origin (the $\boldsymbol{r}$ in Eq. 5.7 only accounts for the displacement of the electron from the nucleus placed at origin). Moreover, the lack of dependence on the actual $\boldsymbol{S}$ but only the magnitude shows that, if a collimated beam is shone onto a spherical sample, it diffracts of the surface as a cone, as the magnitude of the scattering vector on the cone at the same depth must be the same.

For an atom n which is at an arbitrary position $\boldsymbol{r}_n \neq \boldsymbol{0}$, the same argument applies for the phase difference as has be discussed above, and so an extra exponential factor must be introduced, so that

$$f_n(\boldsymbol{S}) = f(S) e^{i2\pi \boldsymbol{S}\cdot\boldsymbol{r}_n} \tag{5.10}$$

Lastly, if the sample consists of more than one atom, the overall form factor of the system is simply the sum of the contributions from all the atoms, hence

$$F(\boldsymbol{S}) = \sum_n f_n(\boldsymbol{S}) e^{i2\pi \boldsymbol{S}\cdot\boldsymbol{r}_n}. \tag{5.11}$$

This overall form factor is also known as the spectral density of the diffraction pattern. The *intensity* of the pattern which appears on the screen is given by the square of the spectral density, thus

$$I(\boldsymbol{S}) = |F(\boldsymbol{S})|^2 = F^*(\boldsymbol{S}) F(\boldsymbol{S}) \tag{5.12}$$

5.2.4 X-ray diffraction from crystals

A crystal, in the solid-state physics and mineralogy point of view, is a structure where an arbitrary arrangement of atoms (the “base") is superimposed onto a regular and periodic grid (the “lattice") [91]. The lattice is a space-filling grid which is formed by imposing a periodic boundary condition (PBC) on the most fundamental component of the grid, known as the *primitive cell*, which is characterised by the cell vectors $(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$. These cell vectors tell the direction and the dimensions of the primitive cell. Hence the scalar triple product of them is the volume of the primitive cell. The PBC on the primitive cell enforces that the same component of the periodic cell occupies the same relative position spatially within

the extended cells. Hence, mathematically, the whole lattice can be expressed as

$$L(\boldsymbol{r}) = \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} \delta\left(\boldsymbol{r} - (m\boldsymbol{a} + n\boldsymbol{b} + l\boldsymbol{c})\right) \tag{5.13}$$

Now if one superimposes an atomic base onto such a lattice to create an atomic crystal, it is mathematically equivalent with the convolution between the base and the lattice, hence

$$\begin{aligned} \rho_{\text{lattice}}(\boldsymbol{u}) &= [\rho * L](\boldsymbol{u}) \\ &= \iiint_{\text{all space}} d^3\boldsymbol{r}\, \rho(\boldsymbol{r}) \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} \delta\left[\boldsymbol{u} - (\boldsymbol{r} - (m\boldsymbol{a} + n\boldsymbol{b} + l\boldsymbol{c}))\right] \end{aligned} \tag{5.14}$$

Then for the form factor of the entire infinite crystal, we can use the convolution theorem which states that the Fourier transformation of a convolution is the product of the Fourier transforms of the constituent factors [27, 130]. Hence,

$$F(\boldsymbol{S}) = \mathcal{F}\left[\rho(\boldsymbol{u})\right] \mathcal{F}\left[L(\boldsymbol{u})\right] \tag{5.15}$$

where the curly $\mathcal{F}$ denotes a Fourier transformation. Now, consider the Fourier transform of the lattice,

$$\begin{aligned} \mathcal{F}\left[L(\boldsymbol{u})\right] &= \iiint_{\text{all space}} d^3\boldsymbol{u} \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} \delta\left(\boldsymbol{u} - (m\boldsymbol{a} + n\boldsymbol{b} + l\boldsymbol{c})\right) e^{i2\pi \boldsymbol{S}\cdot\boldsymbol{u}} \\ &= \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} e^{i2\pi \boldsymbol{S}\cdot(m\boldsymbol{a}+n\boldsymbol{b}+l\boldsymbol{c})} \\ &= \left(\sum_{m=-\infty}^{\infty} e^{i2\pi m\boldsymbol{S}\cdot\boldsymbol{a}}\right)\left(\sum_{n=-\infty}^{\infty} e^{i2\pi n\boldsymbol{S}\cdot\boldsymbol{b}}\right)\left(\sum_{l=-\infty}^{\infty} e^{i2\pi l\boldsymbol{S}\cdot\boldsymbol{c}}\right). \end{aligned} \tag{5.16}$$

Note that the use of $\boldsymbol{u}$ or $\boldsymbol{r}$ here does not affect the result as they are dummy variables only. Moreover, since $\rho(\boldsymbol{r})$ represents the charge distribution in the primitive cell, the Fourier transform of it is the form factor contributed by the primitive cell alone.

Finally, Eq. 5.16 is still valid if the sample is finite, i.e. an $M \times N \times L$ supercell. The only modification needed to be made is the change in the summation limits. For instance, for the first factor [118, 157],

$$\begin{aligned} \sum_{m=-M}^{M} e^{i2\pi m\boldsymbol{S}\cdot\boldsymbol{a}} &= \sum_{m=-M}^{M} \operatorname{cis}\left(2\pi m\boldsymbol{S}\cdot\boldsymbol{a}\right) \\ &= 1 + \frac{1}{2}\sum_{m=1}^{M} \cos\left(2\pi m\boldsymbol{S}\cdot\boldsymbol{a}\right) \\ &= 1 + \frac{1}{2}\frac{\sin\frac{M+1}{2}\theta}{\sin\frac{\theta}{2}} \cos\frac{M+1}{2}\theta \end{aligned} \tag{5.17}$$

where $\theta = 2\pi\boldsymbol{S}\cdot\boldsymbol{a}$, $\operatorname{cis}\theta = \cos\theta + i\sin\theta$ and use has been made of the evenness of the cosine function. The other terms follows by substituting in the respective supercell dimensions for M.

5.3 The PYRALLEX program

PYRALLEX (pronounced [ˈpʰɪrəˌlɛks]) is a Python program for the simulation of X-ray diffraction patterns. Originally written as a side project, the program takes in a atom coordinates file in the castep `cell` format [46] and calculates the 2-dimensional diffraction pattern formed using the formalism described in Sec. 5.2.

5.3.1 Functionality of program

In view of the very many differing formats for the atom coordinates file, a converter tool has been written so that users can convert a raw input coordinates file (currently `pdb` and `xyz` formats are supported) into the castep `cell` format for use in PYRALLEX.

Users can customise the preset parameter file which contains the following input and output parameters:

- Input parameters
  - `task`: the task to be performed — X-ray (`xray`) or fibre diffraction (`fibre`)
  - `bessel_order`: the maximum order of Bessel function up to which summation is to be performed (if `task` = `fibre`)
  - `xr_source`: the source of X-ray radiation.
  - `supercell`: dimensions of the supercell ([1, 1, 1] if using only primitive cell)
  - `crystal_plane`: plane of crystal to be examined. PYRALLEX automatically rotates sample using Rodrigues' rotation formula [244].
  - `s0`: the unit wave vector of the input radiation
  - `screen_shape`: shape of the simulated screen (either `flat` or `cylindrical`)
  - `screen_dim`: dimensions of the simulated screen (in cm)
  - `max_two_theta`: maximum of diffraction angle 2θ for the automatic calculation of sample-screen distance
  - `bs_radius`: radius of backstop on simulated screen to prevent infinite intensity at centre
- Output parameters
  - `showspec`: turn on switch for the calculation of XRD spectrum
  - `showfig`: plot 2D XRD pattern on screen using `matplotlib`
  - `resolution`: resolution of output pattern in pixels per dimension
  - `log_output`: pseudo-gamma correction of output contrast
  - `log_power`: extent of gamma correction if `log_output = True`

The X-ray sources which are currently supported in PYRALLEX are listed with their respective wavelengths in Table 5.1

Source	X-ray series	Wavelength λ (in Å)
Silver	Ag Kα	0.5608
	Ag Kβ	0.4970
Palladium	Pd Kα	0.5869
	Pd Kβ	0.5205
Rhodium	Rh Kα	0.6147
	Rh Kβ	0.5456
Molybdenum	Mo Kα	0.7107
	Mo Kβ	0.6323
Zinc	Zn Kα	1.4364
	Zn Kβ	1.2952
Copper	Cu Kα	1.5418
	Cu Kβ	1.3922
Nickel	Ni Kα	1.6591
	Ni Kβ	1.5001
Cobalt	Co Kα	1.7905
	Co Kβ	1.6208
Iron	Fe Kα	1.9373
	Fe Kβ	1.7565
Manganese	Mn Kα	2.1031
	Mn Kβ	1.9102
Chromium	Cr Kα	2.2909
	Cr Kβ	2.0848
Titanium	Ti Kα	2.7496
	Ti Kβ	2.5138

Table 5.1: X-ray sources and wavelengths in PYRALLEX. Data taken from [235].

5.3.2 Sample outputs of PYRALLEX

Case I: Inorganic crystals Fig. 5.3 shows three simulations of finite-sized crystals of different structures in the (100) direction. The crystals include sodium chloride (body-centred cubic, BCC), copper (face-centred cubic, FCC) and diamond (diamond cubic). It is observed that the relative intensities of the major constructive interference points between in the reciprocal planes as suggested in the schematic diagram, i.e. the points with integral (hkl) numbers in the von Laue formalism [32], are faithfully reproduced.

Note that the lines in the diffraction patterns are where the reciprocal lattice dimensions are integers.

Case II: Double-stranded DNA Fig. 5.4 shows the simulated XRD patterns for canonical A-DNA and B-DNA respectively, using a Cu Kα series (1.5418Å) radiation. It can be seen that in the case of A-DNA, there is strong diffraction around the first two layers from the equator (the central horizontal line) and on layers 6 to 8. Moreover, the dumbbell shape of the diffraction pattern is symmetric about the meridian (the central vertical line). On the contrary, in the case of B-DNA, we can clearly see the characteristic X-shape spanning from layer −3 to layer +3. Furthermore, the reappearance of strong diffraction on layers 8 and 10 are noted as well. Finally, it is rather obvious that the “diamond" pattern, only occuring in B-DNA, which reveals the structure of the phosphate backbone [181–184], occurs in the

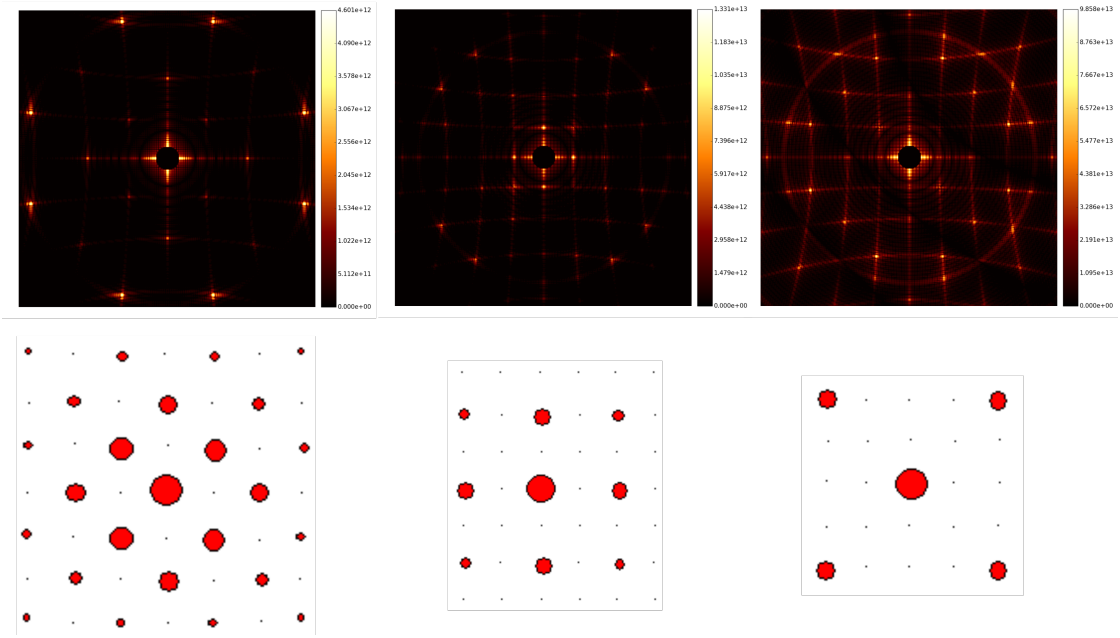

Figure 5.3: (upper) The simulated 2D diffraction pattern of sodium chloride, copper and diamond in (100) direction, using PYRALLEX. (lower) Schematic showing projected intensities of diffraction for the respective generic crystal structures. Images adopted from [92].

simulation of B-DNA too but not A-DNA. We note that all the characteristics mentioned above occur in the experimental data of Franklin and Gosling [94].

Last but not least, since the NAB program uses experimental data as basis in the creation of nucleic acid coordinates for simulations, the reciprocation of the axial rise difference between A-DNA and B-DNA should be reflected on the simulated 2D diffraction pattern if PYRALLEX is valid. This implies that since the periodicity in an A-DNA is about 28Å, and that in a B-DNA is about 34Å (hence, ~ 121% that of A-form) [210], the layer separation in the diffraction pattern of A-DNA should be about 121% that of B-DNA. In the simulated patterns, we observe that the layer 10 (i.e. the large meridian arc) of the B-DNA is at about a quarter between layers 8 and 9 of A-DNA. This implies that there are about 20% more layers in B-DNA than in A-DNA, which is of course equivalent to saying the interlayer separation in A-DNA is about 120% that in B-form, in turn proving that PYRALLEX is valid and is suitable for use in DNA studies.

5.4 Data Analysis

Analyses of systems were done via a simulated X-ray diffraction study using the program PYRALLEX. As in the analyses performed in the previous Chapter, the coordinate files from MD simulations and those of the canonical form DNA were used as raw inputs to PYRALLEX for the calculation of 2-dimensional diffraction patterns. Then the likenesses within pairs of diffraction patterns were compared using the modified van Hove reliability

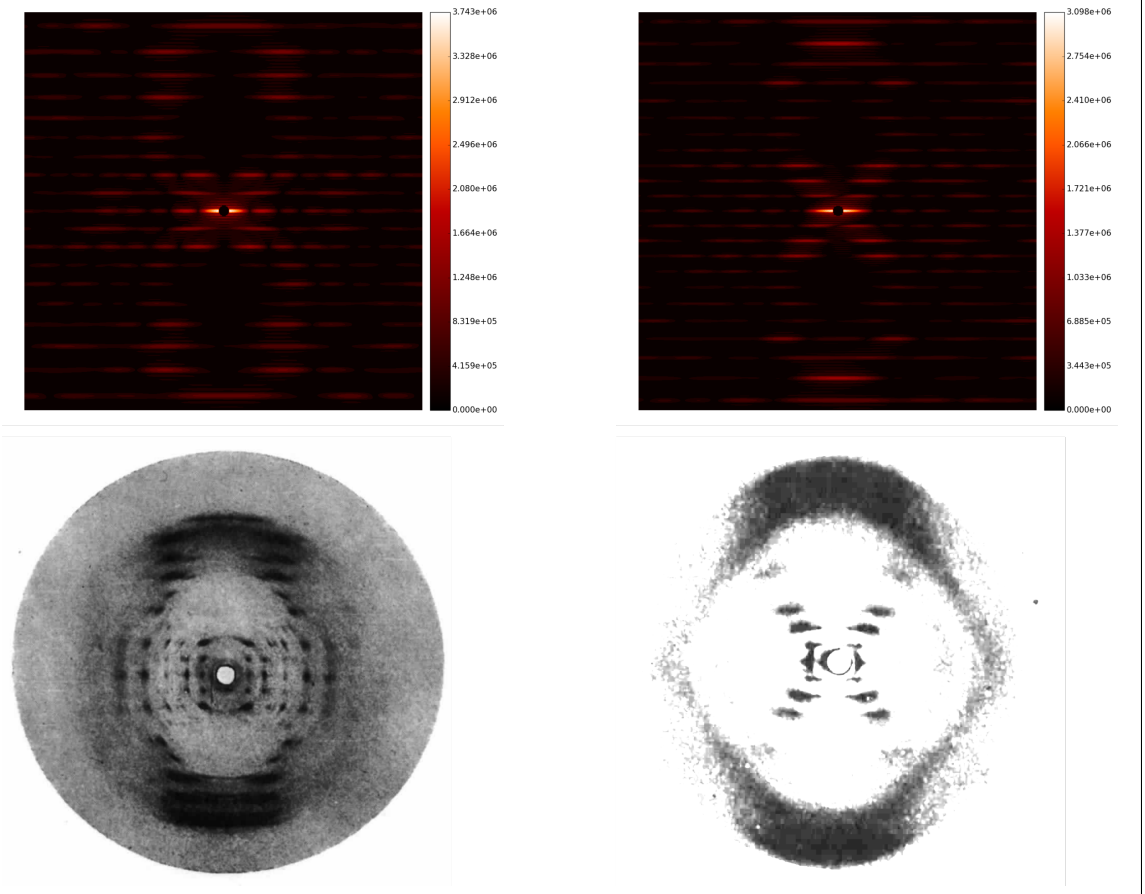

Figure 5.4: (upper) Simulated XRD pattern of canonical A-DNA and B-DNA. (lower) Experimental XRD pattern adapted from Franklin *et al.* [94].

factors (R-factors), which will be explained below.

van Hove R-factors The van Hove R-factors (VHRs) [278], first reported by van Hove *et al.*, is a set of five different R-factors which is used to determine the goodness of fit of an experimental crystallography output with calculations from theories. The five R-factors read [278]

$$R1 := \frac{\sum_n \left| I_n^{\mathrm{exp}} - c I_n^{\mathrm{th}} \right|}{\sum_n \left| I_n^{\mathrm{exp}} \right|}$$

$$R2 := \frac{\sum_n \left| I_n^{\mathrm{exp}} - c I_n^{\mathrm{th}} \right|^2}{\sum_n \left| I_n^{\mathrm{exp}} \right|^2}$$

$R3 :=$ "Fraction of E-range with slopes of different signs"

$$R4 := \frac{\sum_n \left| I'^{\mathrm{exp}}_n - c I'^{\mathrm{th}}_n \right|}{\sum_n \left| I'^{\mathrm{exp}}_n \right|}$$

$$R5 := \frac{\sum_n \left| I'^{\mathrm{exp}}_n - c I'^{\mathrm{th}}_n \right|^2}{\sum_n \left| I'^{\mathrm{exp}}_n \right|^2}$$

where I_n is the intensity of the n-th energy point on the intensity-energy spectrum in low energy electron diffraction (LEED) experiments, and I'_n is the associated derivative of the intensity at the same point on the spectrum. The superscripts "exp" and "th" denote experimental data and theoretical results respectively.

$c = \frac{\sum_n I_n^{\text{exp}}}{\sum_n I_n^{\text{th}}}$, being the ratio between the *total* intensity in experiment and that in theory, serves as the scaling factor. This is much needed as the intensity is in effect the total number of photons hitting the screen per unit time per unit area. Without scaling, the electromagnetic flux across the detector screen would be different for the cases of experiment and theory, and the comparison between the two would then be meaningless.

The van Hove R-factors can be broadly divided into two groups, *viz.* the first two and the last three. The first two, strictly speaking, were not devised by van Hove *et al.*, as they have been used traditionally by X-ray crystallographers [11, 83]. These two factors have profound significance in picking out the similarities in the positions, heights and widths of the spectral profile. However, as pointed out in [278], the two factors alone are not sufficient in depicting the characteristics of the spectral profile down to the micro-structural level, as they are incapable in distinguishing between smooth peaks and bumpy ones, so long as the macro-structures have the same height and width and have the same mean position.

The second group of the R-factors, i.e. the last three factors, were devised to alleviate the problems described above. These factors take into account also the slopes (hence curvature) of the profile. $R3$, whose real form was not disclosed in the original paper [278] but given in Awrejcewicz [10] as

$$R3 = \frac{N_+(I'^{\text{exp}})}{N_-(I'^{\text{exp}})} - \frac{N_+(I'^{\text{th}})}{N_-(I'^{\text{th}})} \tag{5.18}$$

where N_+ and N_- are the total number of points having positive and negative slopes respectively. This factor is useful in differentiating between split and non-split peaks [278]. The two factors $R4$ and $R5$, analogous to $R1$ and $R2$ in the first group, add sensitivity to the slope at each energy.

Last but not least, as suggested by Awrejcewicz [10], the five R-factors can be combined to form the total R-factor

$$R_T = \sqrt{\sum_{i=1}^{5} {R_i}^2}, \tag{5.19}$$

which gives the overall goodness-of-fit between experimental data and theoretical calculations.

Modified van Hove R-factors While the VHRs are more general than most other reliability factors [10], there is a grave shortcoming of them, which is the metricity of the terms. For example, if we express the scaled theoretical value as the experimental value plus the difference between the two, i.e. $cI^{\text{th}} := I^{\text{exp}} + \varepsilon$, then $R1 \longrightarrow \frac{\sum_n |\varepsilon_n|}{\sum_n |I_n^{\text{exp}}|}$, which has the range of $R1 \in \mathbb{R}_{\geq 0}$. The same logic can be applied to $R2$, $R4$ and $R5$ as well and thus they all have the same range.

The analysis in the metricity of $R3$ deserves more attention, as there is no parity due to

the lack of the absolute values as in the other terms. Consider the two extreme cases of monotonically increasing and monotonically decreasing intensity with respect to the input energies. If the data set is monotonically increasing, then $N_+ : N_- = +\infty$. On the contrary, if the data set is monotonically decreasing, then $N_+ : N_- = 0$. This implies that $R3$ can span all real numbers. This then has massive impact on the validity of the total R-factor R_T as the constituent terms are no longer in the same scale, and $R_T \in \mathbb{R}_{\geq 0}$ *per se*, which means there is no way of quantifying how much poorer a fit is if, say, $R_T = 10$, than if $R_T = 5$, and the value becomes totally meaningless.

Moreover, since the VHRs are originally designed for use in 1-dimensional spectral data, they cannot be directly applied to 2-dimensional data sets. In view of this, modifications to the original VHRs have been devised to alleviate the problems explained above and to extend the idea behind the VHRs to 2-dimensional.

We first consider the modifications made to $R1$ and $R2$ as the change is most straightforward. Inspired by the Pendry R-factors [229] which are very popular within the electron energy loss spectroscopy (EELS) community, a term of the scaled theoretical value has been added to the denominator, hence

$$MR1 := \frac{\sum_n \left| I_n^{\text{exp}} - cI_n^{\text{th}} \right|}{\sum_n \left| I_n^{\text{exp}} + cI_n^{\text{th}} \right|}$$
$$MR2 := \frac{\sum_n \left| I_n^{\text{exp}} - cI_n^{\text{th}} \right|^2}{\sum_n \left(\left| I_n^{\text{exp}} \right|^2 + \left| cI_n^{\text{th}} \right|^2 \right)}$$

which ensures that both factors have the range of $[0, 1]$.

The second group of terms which concerns the gradient of the input signals has proven to be rather tricky to alter, as the first derivative of a univariate function does not map directly onto the gradient of a multivariate function. Moreover, since the gradient of a multivariate function is a vector field, the second group of terms would render ill-defined as the division between two vectors is illegitimate. In light of this, we turn to the *second* derivative of the function, or more specifically, the Laplacian of the intensity, i.e. $\boldsymbol{\nabla}^2 I \equiv \dfrac{\partial^2 I}{\partial x^2} + \dfrac{\partial^2 I}{\partial y^2} \equiv \boldsymbol{\nabla} \cdot (\boldsymbol{\nabla} I)$.

The merit of using the Laplacian comes with the nature of it as the divergence of the gradient field — a local minimum (a large negative value) implies the tip of a peak, a local maximum (a large positive value) means the bottom of a trough, whereas a zero value can be attained if and only if the position is either on a flat terrain or is itself a saddle point [1]. Another useful information which could be obtained from the Laplacian is the steepness of the nearby landscape, as the steeper the landscape would give a larger magnitude of the Laplacian and vice versa.

With this, and by letting $L := \boldsymbol{\nabla}^2 I$, VHRs in the second group are changed to

$$MR3 := \left| \frac{N_+(L^{\text{exp}}) - N_+(L^{\text{th}})}{N_\tau} \right| \approx \left| \frac{N_-(L^{\text{exp}}) - N_-(L^{\text{th}})}{N_\tau} \right|$$

[1] That is to say, $\partial_x^2 I + \partial_y^2 I = 0$ iff $\partial_x^2 I = -\partial_y^2 I$ or $\partial_x^2 I = \partial_y^2 I = 0$.

$$MR4 := \frac{\sum_n \left| L_n^{\text{exp}} - cL_n^{\text{th}} \right|}{\sum_n \left(\left| L_n^{\text{exp}} \right| + \left| cL_n^{\text{th}} \right| \right)}$$

$$MR5 := \frac{\sum_n \left| L_n^{\text{exp}} - cL_n^{\text{th}} \right|^2}{\sum_n \left(\left| L_n^{\text{exp}} \right|^2 + \left| cL_n^{\text{th}} \right|^2 \right)}$$

where N_τ is the total number of samples points (pixels) on the screen. With this notation, the range of $MR3$ must be $[0,1]$. Note the difference between the denominators of $MR1$ and $MR4$. Although the two factors are analogous to one another, the splitting of the absolute value in $MR4$ is necessary in obtaining the maximum range of $\boldsymbol{\nabla}^2 I_n$, as $\boldsymbol{\nabla}^2 I_n$ can be negative.

Finally, with modification made to all R-factors so that now all of them are within zero and unity, sense can be made of the overall R-factor, which is changed into

$$MR_T := \sqrt{\frac{1}{5} \sum_{i=1}^{5} {MR_i}^2}, \tag{5.20}$$

i.e. the root-mean-square of all the MVHRs, which obviously have the range of $[0,1]$ as well.

5.5 Results and discussion

We now move on to the discussion of the study of the structural changes brought by the anthracycline drugs. It is worth presenting the results from the bare DNA oligomers as controls.

5.5.1 Bare DNA — canonical structures

Fig. 5.5 shows the simulated XRD patters of the A- and B-forms canonical structures. It can be clearly seen that while all of the A-DNA show the characteristic dumbbell pattern, whereas the B-DNA show the large cross in the middle, regardless of the real sequence. This is because the overall structure of the DNA is maintained roughly the same; the helical pitch and width are maintained. However, it can also be observed, especially for the peripheral fringes, that whilst the positions of the bright ones are very similar to each other across different base sequences, they have different diffraction intensities. One of the causes of this is the contributions from nucleobases, as the fundamental compositions of the bases are slightly different from one another. Another cause of the difference in the intensities is the orientation of the bases. Since in this study, single-crystal methods, rather fibre diffraction methods, have been employed, the diffraction patterns are prone to change in the relative positions of the atoms, as they are calculated directly from the physical positions of the particles rather than the cylindrically-averaged coordinates [116]. However, we assert such contributions should be negligible in this part, as all the images captured in Fig. 5.5 were taken in the same direction, i.e. that which is parallel to the (100)-plane of the primitive cells of the respective structures.

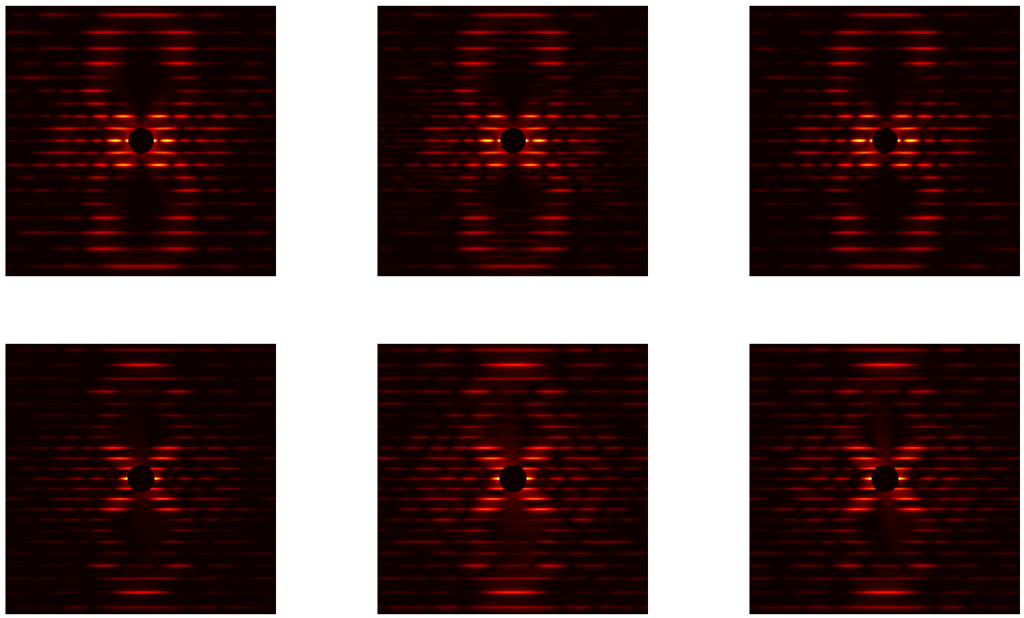

Figure 5.5: XRD patterns of (Left-to-right, top-to-bottom) AA-A-bare, AC-A-bare, CC-A-bare, AA-B-bare, AC-B-bare and CC-B-bare simulated using PYRALLEX.

	AA-A	AC-A	CC-A	AA-B	AC-B	CC-B
AA-A	0.0000	0.0187	0.0435	0.2633	0.2635	0.2618
AC-A		0.0000	0.0369	0.2623	0.2624	0.2607
CC-A			0.0000	0.2743	0.2742	0.2722
AA-B				0.0000	0.0206	0.0386
AC-B					0.0000	0.0325
CC-B						0.0000

Table 5.2: MR_T values for pairs of simulated X-ray images.

Table 5.2 shows the total MVHR of the diffraction patterns of different pairs of systems. It can be observed that, expectedly, when an image is compared with itself, $MR_T = 0$ as the intensity at each of the data point is the same. Moreover, we see that the MR_T values when the pair consists of two different canonical forms (i.e. the top-right quadrant) are much larger than those when the two compared systems are of the same form (i.e. the top-left and bottom-right quadrants). This is justifiable as VHRs and MVHRs measure the differences in two images, and the differences in the patterns for A- and B-forms are rather obvious from Figs. 5.4 and 5.5. Here we can use the rule-of-thumb in determining the goodness-of-fit from the R-values — the fit is near-perfect if $MR_T \leq 0.05$, $MR_T \approx 0.25$ implies a certain extent of disagreement between the compared pair and $MR_T \approx 0.35$ is the threshold above which the convergence between the two images is small [32]. Lastly, due to the parity in the MVHR calculation, the swapping between the test image and the reference image gives the same value, and hence the missing lower triangle in Table 5.2 can be obtained by transposing the current upper triangle.

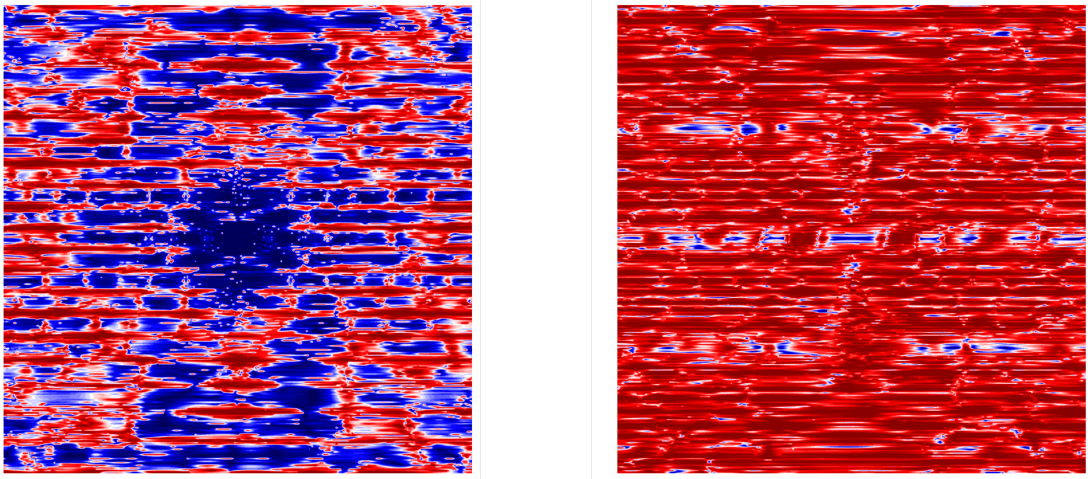

Figure 5.6: Heat maps for the comparison between (left) AA-A/AC-A; and (right) CC-A/AA-B

Although the method of calculating MR_T, the overall MVHR, is very useful in telling whether two output images are similar to each other, it cannot tell where on the images that fits are good, and where that fits are bad. We assert that the method of calculating individual MVHRs can be extended to produce a point-to-point heat map between the two images. Here, two of the terms, *viz.* $MR1$ and $MR4$ are used, thus

$$MR1_{(ij)} := \frac{\left|I^{\text{exp}}_{(ij)} - cI^{\text{th}}_{(ij)}\right|}{\left|I^{\text{exp}}_{(ij)} + cI^{\text{th}}_{(ij)}\right|}$$

$$MR4_{(ij)} := \frac{\left|L^{\text{exp}}_{(ij)} - cL^{\text{th}}_{(ij)}\right|}{\left|L^{\text{exp}}_{(ij)}\right| + \left|cL^{\text{th}}_{(ij)}\right|}$$

where (ij) denotes the (i,j)-th pixel on the output image. Similar to MR_T, an equivalent quantity is defined as the RMSD of $MR1_{(ij)}$ and $MR4_{(ij)}$, hence

$$MR_T^{(ij)} = \sqrt{\frac{1}{2}\left(MR1_{(ij)}{}^2 + MR4_{(ij)}{}^2\right)}$$

As an example, Fig. 5.6 shows the heat maps comparing the systems AA-A and AC-A, and those CC-A and AA-B. These two pairs are chosen as examples as they have the most extreme non-zero values in Table 5.2 — the pair AA-A/AC-A has the lowest MR_T whilst the pair CC-A/AA-B has the highest MR_T. The colouring in the heat maps are normalised to the range $[0,1]$ as the plotted quantity $MR_T^{(ij)}$ is, by nature, normalised to the same range. The deeper blue the colour denotes the smaller the $MR_T^{(ij)}$ value, hence higher similarity at the particular point. On a contrary, the deeper red the colour implies the larger the $MR_T^{(ij)}$ value, thus lower similarity at that pixel.

It can be clearly seen from the AA-A/AC-A pair, that the dark blue patches form a clear

dumbbell shape which resemble those of the individual AA-A and AC-A images. This means that not only are the positions of the bright spots on the main dumbbell structure similar in both images, even the relative intensity are similar too. Moreover, many of the "stray" layer fringes outside of the main dumbbell structure appear to be rather dark blue in colour. This signifies that secondary structures, i.e. the arrangements and alignment of the inner base pairs, are very similar within the pair.

On the other hand, the heat map for the CC-A/AA-B pair appears to be mostly in dark red, apart from only a few fringes, including the central fringe, being blue in colour. This implies that most of the secondary and tertiary structures are different. Furthermore, the reason behind the very dark blue colour of the central equatorial fringe is due to its position in the reciprocal space. The centre of the screen corresponds to the $(hkl) = (000)$ three-way conjunction in the reciprocal Bragg plane. This means that the surrounding fringe the corresponds to the first Brillouin zone of the crystal structure, which is the trivial solution in the von Laue conditions, which states that constructive interference of diffracted radiations takes place only at integral values of h, k and l. This solution implies that the central spot must have near-to-infinity relative intensity whereas anywhere in the first couple Brillouin zones would have intensity tens, if not hundreds, of orders of magnitudes higher than that of the dimmest spot. This in turn says that, the difference in the intensity in that region is nearly negligible comparing with the absolute intensity, hence the dark blue colour.

5.5.2 Bare DNA — energy-minimised

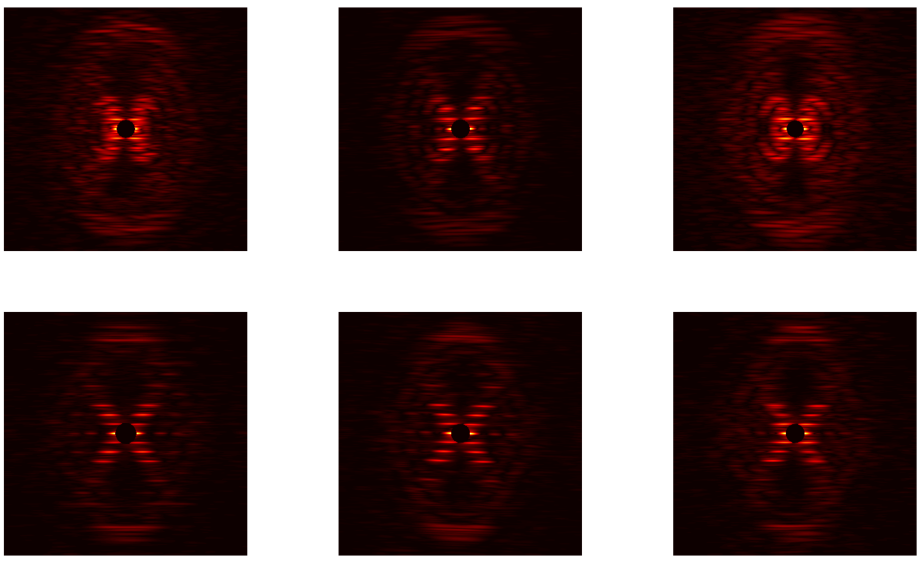

Figure 5.7: XRD patterns of (Left-to-right, top-to-bottom) energy-minimised AA-A-bare, AC-A-bare, CC-A-bare, AA-B-bare, AC-B-bare and CC-B-bare simulated using PyralleX.

From the XRD patterns (Fig. 5.7), we can see that for the A-start systems (top row), the central bright fringes all appear to have deviated into structures which vaguely resemble those

for both the canonical A- and B-forms. However, the extent to which the central “circular blob" turned into the X-shape depends largely on the base sequences. For example, for the diffraction pattern of the AA-A-bare system, it could be seen that the X-shape is rather obvious, with the bright fringes spanning the negative to positive three layers. However, it is observed that the layer 2 and 3 fringes are slightly bent. Similar phenomenon occurs in the CC-A-bare system, where the layer 3 fringes are bent into arcs, and the layer 1 and 2 fringes form a circular cloud being wrapped in this envelope. This could mean that while the CC-A-bare system has changed to somewhere between A and B, the fact that it retains much characteristics of A-form whereas the other systems (*viz.* AA-A-bare and AC-A-bare) look more like B-form implies the system having 100% GC content has higher resilience to structural perturbations than the those with lower GC contents.

On the other hand, for the B-start systems (second row of Fig. 5.7), we see that the major difference in the diffraction patterns across different structures is mostly in the secondary characteristics, rather than in the most prominent ones *viz.* the central X-shape and the bright layer 10 fringes. It can be observed that the “fuzziness" of the layer lines between layers 4 and 9 increases with the CG content. The clarity of the layer lines is a symbol for the regularity or periodicity of the system: the more fuzzy the lines are, the less regular is the structure. From the subfigures, we can see that while the layers are still rather distinguishable from one another in the AA-B-bare case, it is very difficult to tell which fringe is from which layer in the CC-B-bare case. Moreover, if we compare the width of the layer 1 fringe, it is obvious that it also increases with the CG content of the system — in the CC-B-bare case layer 1 is even longer than layer 2, which is typical in the A-DNA diffraction pattern.

5.5.3 Daunomycin

In this subsection we discuss the structural perturbation induced by the intercalation of a daunomycin molecule into an inter-base pair site in the same DNA systems used in the previous subsections. The daunomycin was inserted into the system prior to the simulation and the intercalation site was chosen arbitrarily in the central region.

Fig. 5.8 shows the simulated XRD pattern of the daunomycin-intercalated systems. By inspection, all the six subfigures appear very similar to those in Fig. 5.7; even the secondary features in the patterns look similar in the corresponding subfigures. This is because there is only one daunomycin molecule inserted, and its effect on the global structure is limited. Moreover, the symmetry of the DNA is already broken due to the energy minimisation, and further small perturbation would not be easily observable by macroscopic means.

	AA-A-dau	AA-B-dau	AC-A-dau	AC-B-dau	CC-A-dau	CC-B-dau
Canon. A	0.40899	0.38897	0.32310	0.36373	0.43294	0.38604
Canon. B	0.36950	0.34084	0.37888	0.28616	0.42717	0.29512

Table 5.3: MR_T values for daunomycin systems (compared with respective canonical forms).

Here, we present the MR_T values for the daunomycin systems (Table 5.3). Each row repre-

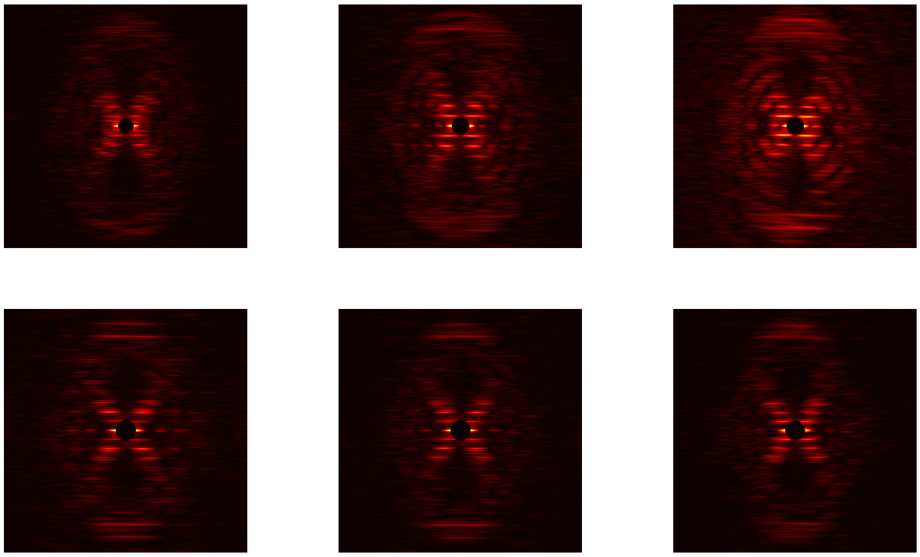

Figure 5.8: XRD patterns of (Left-to-right, top-to-bottom) energy-minimised AA-A-dau, AC-A-dau, CC-A-dau, AA-B-dau, AC-B-dau and CC-B-dau simulated using PYRALLEX.

sents the canonical form of DNA, whose base sequence is determined by that of the drug-DNA complex system. For instance, for the cell on the first row ("Canon. A") and third column ("AC-A-dau"), the value (0.32310) is the MR_T value when the fully minimised AC-A-dau system is compared with the canonical A-form of the sequence $d(AC)_{72}$.

From the table, we can see that all the 12 values are rather high, that even the smallest value (0.28616) is larger than any value in Table 5.2. This means that all the systems with daunomycin has deviated drastically from their respective canonical forms. What is more intriguing here is that, if we compare the pure sequences (AA and CC) with the mixed sequences (AC), we can see that for the pure sequences, the MR_T values for the canonical B-form comparisons are *lower* than those for the canonical A comparisons. However, for the mixed sequences, the forms are more-or-less preserved upon minimisation. For example, the MR_T value for the "AC-A-dau/canon. A" pair is more than 0.55 lower than that of the "AC-A-dau/canon. B" pair — the A-likeness is higher than the B-likeness. On the contrary, the MR_T value for the "AC-B-dau/canon. B" pair is nearly 0.8 lower than that of the "AC-B-dau/canon. A" pair, signifying that the tendency of the ACB system to stay in a more B-like structure is higher than its ACA counterpart to remain in the A-like structure.

5.5.4 Doxorubicin and idarubicin

	AA-A-dox	AA-B-dox	AC-A-dox	AC-B-dox	CC-A-dox	CC-B-dox
Canon. A	0.35537	0.40817	0.33261	0.40263	0.42329	0.34979
Canon. B	0.34011	0.37909	0.39529	0.38903	0.41451	0.33587

Table 5.4: MR_T values for doxorubicin systems (compared with respective canonical forms).

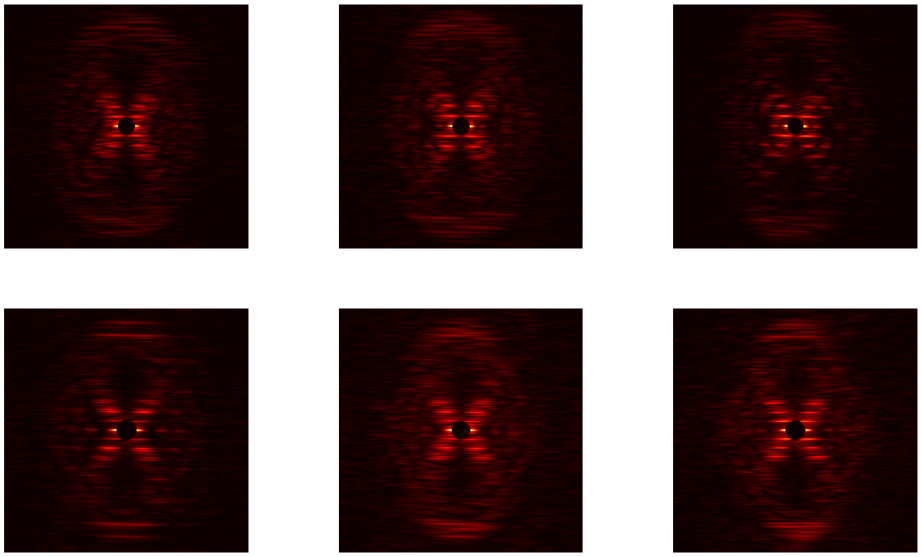

Figure 5.9: XRD patterns of (Left-to-right, top-to-bottom) energy-minimised AA-A-dox, AC-A-dox, CC-A-dox, AA-B-dox, AC-B-dox and CC-B-dox simulated using PYRALLEX.

In this subsection, we present the results from the studies of the doxorubicin-DNA complex systems. Unlike the previous case of daunomycin, since the XRD patterns for the doxorubicin and idarubicin systems resemble those of daunomycin systems, we only briefly discuss their general appearances and explain the difference between the doxorubicin and idarubicin cases with the daunomycin systems we have discussed before.

We first note that for the A-start doxorubicin systems, the bright fringes are much dimmer than those in the corresponding figures in the daunomycin systems. This is probably because the absolute brightness of the fringe at the equator is higher in the doxorubicin cases. As a result, the luminosity of the peripheral fringes are normalised to be much dimmer. Nevertheless, the patterns resemble closely those in the daunomycin case.

In Table 5.4, we present the MR_T values for doxorubicin systems. When compared with the numbers in the daunomycin case, we see rather clearly that spread of the values down columns have greatly reduced in general. For example, the difference between the pairs "AC-B-dau/canon. A" and "AC-B-dau/canon. B" (in Table 5.3) is about 0.077 whereas the corresponding pairs "AC-B-dox/canon. A" and "AC-B-dox/canon. B" in the doxorubicin case has a spread in values of only about 0.013. Moreover, the majority of MR_T values in the doxorubicin systems is higher than the corresponding ones in daunomycin systems. These imply that not only is the tendency of preserving the form in some subsystems lost, but also the general disruption to the structure brought by the intercalation of doxorubicin is more prevalent than that of daunomycin.

Finally, we present the results from the simulations of the idarubicin systems, with Fig. 5.10 showing the relevant XRD images and Table 5.5 showing the MR_T values for the systems.

We first discuss about the MR_T values. One of the interests regarding these numbers is,

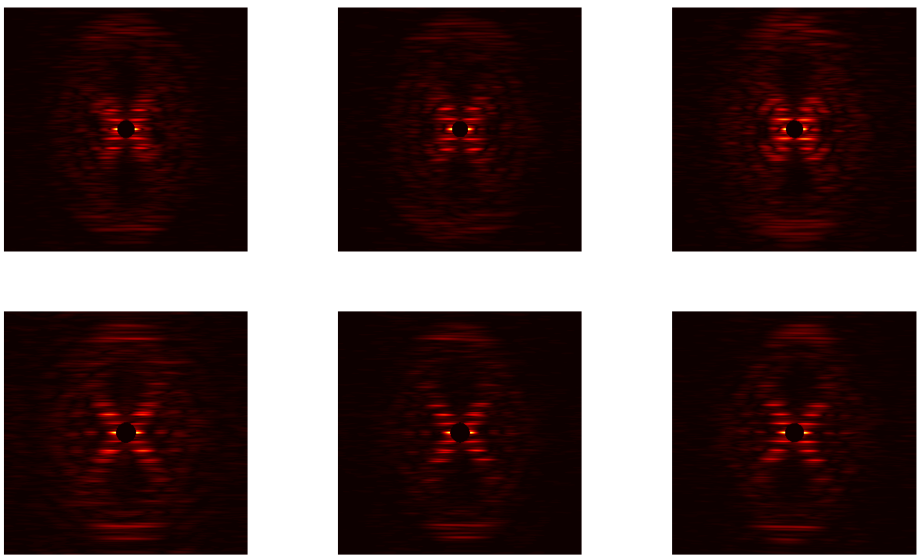

Figure 5.10: XRD patterns of (Left-to-right, top-to-bottom) energy-minimised AA-A-ida, AC-A-ida, CC-A-ida, AA-B-ida, AC-B-ida and CC-B-ida simulated using PYRALLEX.

	AA-A-ida	AA-B-ida	AC-A-ida	AC-B-ida	CC-A-ida	CC-B-ida
Canon. A	0.32132	0.36228	0.32256	0.36616	0.35026	0.47500
Canon. B	0.34841	0.30883	0.38067	0.33587	0.36380	0.39518

Table 5.5: MR_T values for idarubicin systems (compared with respective canonical forms).

again, the spread of values down a column. It can be observed that while some of the subsystems have rather narrow spread of less than 0.02 (for instance, the AA-A-ida subsystem), other have much wider spread (for example, the CC-B-ida subsystem, having a spread of nearly 0.08). This means that the tendency of the CC-B-ida system to take a near-canonical A-form is much more unlikely than to take a near canonical B-form.

Secondly, in terms of the absolute values of the MR_T, we note that the values for the idarubicin systems are slightly lower than those in both cases involving daunomycin and doxorubicin. This is a sign of this drug causing less structural disruption to the intercalated DNA than the previously two drugs. This observation can be supported by noting which cell down a column has *lower* MR_T value, as this value tells quantitatively how closely a system resembles a canonical conformation.

We can classify these values by two categories, *viz. matching* and *mismatching*. A *matching* situation is when the MR_T value for a pair where the starting form matches with the canonical form (say, A-start with the canonical A) and the value is lower than the other on the same column. On the contrary, a *mismatching* situation is the opposite: a lower MR_T value occurs when the starting form does *not* match with the canonical form (say, A-start with the canonical B). We note that for all the six subsystems with idarubicin, they all fall into the catergory of *matching* pairs. This means that whilst structural changes are induced by the intercalation of the drug, in no cases was the interaction able to change the structure

enough to attain one close to the counterpart conformation.

5.6 Summary

In this Chapter, we have revisited the structural perturbation caused by anthracycline drugs intercalations on DNA. We used different methods other than classical MD simulations to probe the changes and quantify them. In particular, we wrote the PYRALLEX program which simulates the X-ray diffraction of a structure. Using PYRALLEX and heatmaps, we showed that we can probe the structural differences between two similar structures by selectively pick out useful features. We also assert that, the technique of the two-dimensional modified van Hove R-factors, which was used in this work to compare two simulated images, can also be applied in the comparison between experimental and simulated images. In this case, one would need the intensity mapping of the experimental images and a global normalisation of the experimental results with the simulated (or vice versa) prior to the calculation.

From the structural studies using the PYRALLEX, we discovered that for the vast majority of the systems with anthracyclines, the conformation of the DNA converged to somewhere between the A-form and the B-form. This was shown by the appearance of both the characteristic XRD patterns of A- and B-forms simultaneously in the most stable states of the DNA-drug complexes after simulated annealing. However, as we also probed the relative likeness of those structures to the canonical conformations. We found that although the three drugs being studied only differ with each other by a small component in the side chains, they exhibit rather different behaviours on the DNA. For instance, daunomycin and doxorubicin seem to cause a structural transit to a more B-like conformation with little regard to the starting conformation. On the other hand, idarubicin produced more "matching pairs" in the modified van Hove R-factors, while keeping the R-factors lower than the other two drugs, indicating that it may stabilise the DNA to the original conformation.

Although initially PYRALLEX was written to simulate the X-ray diffraction pattern of a given structure, its potential usage could be extended beyond its initial intended purpose. For instance, numerical simulations can be performed to produce images prior to doing imaging experiments. In this way, structural experimentalists would only need to compare their results with those calculated using PYRALLEX. If the images agree with each other then it could massively save analysis time for experimentalists, as they only need to obtain the raw structural parameters from the PYRALLEX input, rather than having to reverse engineer the structures themselves.

Chapter 6

Simulation of covalently closed circular DNA

6.1 Introduction

As have been touched on in the previous chapters, covalently closed circular DNA (cccDNA) is a double-stranded DNA whose 5'-end is connected to its 3'-end via strong covalent bonds. It exists naturally in prokaryotes such as yeasts and bacteria, both of whose chromosomal DNA and plasmids are in this form. In higher organisms, DNA is usually wound around proteins (called *histones*), but since the packing of the molecule is very tight, it can be approximated as a circular ring of DNA.

Not only is the DNA tightly packed as mini-circles, it was found also that in physiological conditions, i.e. in the presence of water and salt and other biological molecules, the molecule is generally maintained in a homeostatically underwound state [54, 161, 253, 274]. This is not merely a physical result from the interactions among molecules, but also bears very important biological significance, as the "loosening" of the DNA structure by means of underwinding facilitates processes which need other molecules' access to the DNA [90], with transcription and replication being two of the most obvious examples of.

In this Chapter, we will study the effects of anthracycline drugs on the supercoiling behaviours of DNA. However, in order to quantify such behaviours we need to use the ribbon theory to aid our communication. Therefore we shall start the discussion with the basics of this theory.

6.2 Ribbon theory

Ribbon theory is a branch of mathematics, in the areas of topology and differential geometry, which deals with how simple and closed ribbons (strips which have the two ends connected together) [1] and interwoven curves behave. It is closely related to knot theory

[1] The rigid definition of a *mathematical* ribbon is a smooth curve with its defining vector depending on a continuous curve of arc-length s, where $s \in (a, b)$, and a smooth perpendicular which varies at each point on the curve. [26, 295]. In particular, a simple and closed ribbon must satisfy the conditions of 1) no self-intersection; and 2) continuity of derivative at a and b.

and has vast application in physics, especially in the fields of quantum field theory, general relativity and biophysics. It is also a very useful tool for the analysis of the supercoiling of circular DNA, which is at the heart of this work.

There are three crucial concepts in ribbon theory, *viz.* the link, the twist and the writhe, and they will be discussed in detail in this section.

6.2.1 Link

In topology, a link can be defined as a collection of non-intersecting knots being *linked* together by means of crossing over with one another. The simplest case is known as a Hopf link [234], which is formed by two components (two knots) crossing each other exactly once. It then follows that there are two quantities which could be used to quantify how two knots are interwound with each other. One of them is known as the *crossing number*, and the other the *linking number*.

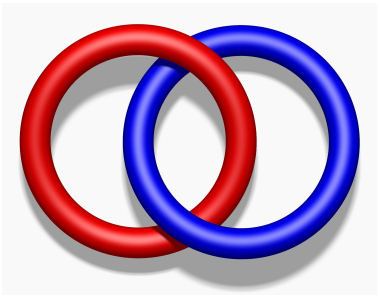

Figure 6.1: Hopf link being the simplest non-trivial link. Image taken from [298].

The crossing number $Cr(C_1, C_2)$ is an unsigned quantity (i.e. with no directionality) which tells the total number of crossovers between curves C_1 and C_2. For example, in Figure 6.1, we can see two crossovers between the red ring and the blue ring, and so the crossing number of the Hopf link is 2.

A very similar quantity is the linking number $Lk(C_1, C_2)$, which incorporates also the directionality of the crossovers, where the assignment of its sign follows the right-hand screw rule; it is also *always* an integer. For instance, in the leftmost diagram of Figure 6.2, if we align our right palm with the blue arrow, with fingers pointing at the direction of the arrowhead and curl around the shaft of the red arrow, the thumb points at the *same* direction as the red tip (i.e. upwards), and hence the sign of the associated linking number is positive. On the other hand, taking the third diagram as an example, if we curl our fingers in the same fashion as before, the thumb will point at the direction (i.e. leftwards) of the opposite end of the red arrow shaft: the associated linking number for this case will be negative. The total linking number of a knot system is then defined as [300]

$$Lk = \frac{1}{2}(n_1 + n_2 - n_3 - n_4) \tag{6.1}$$

where the n numbers are the total number of crossings in the left-to-right order in Figure 6.2. It then follows, if we assign the directionality of both circles in the Hopf link to be anticlockwise, the linking number is +1, since the upper crossover is the second case and the lower crossover the first case in Figure 6.2.

There are different methods of computing the linking number for an arbitrary link system,

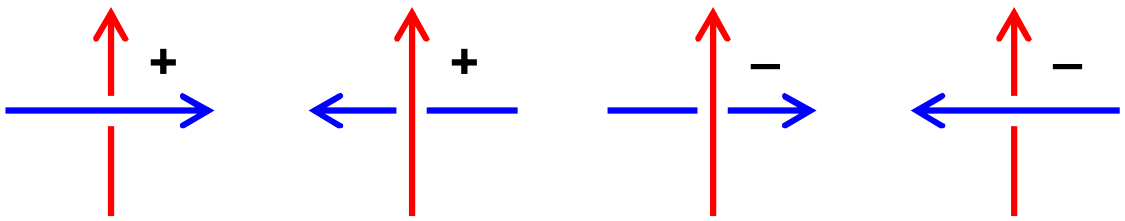

Figure 6.2: The four possible link crossing combinations. A broken shaft means the arrow is beneath the other arrow. Image taken from [300].

one of them is via the so-called Gauss linking integral [53]:

$$\begin{aligned} Lk(C_1,C_2) &= \frac{1}{4\pi}\oint_{C_1}\oint_{C_2}\frac{\boldsymbol{x}_1(s_1)-\boldsymbol{x}_2(s_2)}{|\boldsymbol{x}_1(s_1)-\boldsymbol{x}_2(s_2)|^3}\cdot(\boldsymbol{dx}_1\times\boldsymbol{dx}_2) \\ &= \frac{1}{4\pi}\oint_{C_1}\oint_{C_2}\frac{(\boldsymbol{t}_1(s_1)\times\boldsymbol{t}_2(s_2))\cdot[\boldsymbol{x}_1(s_1)-\boldsymbol{x}_2(s_2)]}{|\boldsymbol{x}_1(s_1)-\boldsymbol{x}_2(s_2)|^3}ds_2ds_1 \end{aligned} \tag{6.2}$$

where $\boldsymbol{x}_i(s_i)$ are the coordinates of a point on curve i in terms of the arc length s_i and $\boldsymbol{t}_i(s_i)\equiv\dfrac{d}{ds_i}\boldsymbol{x}_i(s_i)$ is the spatial derivative of the coordinates.

In DNA research, it is customary to choose C_1 as the helical axis and C_2 as one of the two strands [266]. Mathematically, it is convenient to define a quantity, called the *frame* $\mathcal{F}$, for the ribbon, with a formal definition of being a closed curve displaced along the normal of the ribbon by a small amount ε [197]. Hence, for every point on $\mathcal{F}$, the coordinate $\boldsymbol{y}(s)$ is defined as

$$\boldsymbol{y}(s)=\boldsymbol{x}(s)+\varepsilon\hat{\boldsymbol{n}}(s) \tag{6.3}$$

where $\boldsymbol{x}(s)$ is the basis point on the knot and $\hat{\boldsymbol{n}}(s)$ is a unit normal vector from the knot. It thus follows that a plasmid can be considered as a simple closed ribbon with the frame defined as one of the two strands, hence C_2 is a frame for C_1. Now, because the quantity denotes the linking properties between a ribbon and its own frame, the linking number is also known as the *self-linking number* SLk [155], hence

$$SLk(C_1,C_2=\mathcal{F}_{C_1})\equiv Lk(C_1,C_2) \tag{6.4}$$

6.2.2 Twist

Similar to the linking number, the twist tells how much a ribbon twists around its own axis. Then, for a knot C, the twist of the knot with respect to its frame $\mathcal{F}$ is defined as [155]

$$Tw(C)=\frac{1}{2\pi}\oint_C ds\,\epsilon_{\mu\nu\alpha}\frac{dx^\mu}{ds}n^\nu\frac{dn^\alpha}{ds} \tag{6.5}$$

where $\epsilon_{\mu\nu\alpha}$ is the Levi-Civita pseudotensor. Obviously, Eq. 6.5 can be readily re-written from the component form into the more accessible vectorial form [2]:

$$\begin{aligned} Tw(C) &= \frac{1}{2\pi}\oint_C ds \left[\frac{d\boldsymbol{x}}{ds}\times\hat{\boldsymbol{n}}\right]_\alpha \left[\frac{d\hat{\boldsymbol{n}}}{ds}\right]^\alpha \\ &= \frac{1}{2\pi}\oint_C ds \left(\frac{d\boldsymbol{x}}{ds}\times\hat{\boldsymbol{n}}\right)\cdot\frac{d\hat{\boldsymbol{n}}}{ds}. \end{aligned} \tag{6.6}$$

For application in DNA systems, we follow the definitions of C_1 and C_2 from above, i.e. C_1 is the helical axis of the DNA and C_2, as the frame, is one of the phosphate backbones. Then Eq. 6.6, meaning the twist of C_2 around C_1, becomes

$$Tw(C_2, C_1) = \frac{1}{2\pi}\oint_{C_1} ds \left(\frac{d\boldsymbol{x}_1}{ds}\times\hat{\boldsymbol{n}}\right)\cdot\frac{d\hat{\boldsymbol{n}}}{ds}, \tag{6.7}$$

thus having the same form as the definition in Swigon [266] where $\boldsymbol{n}$ was defined as $\boldsymbol{n} = \boldsymbol{x}_2(\sigma(s)) - \boldsymbol{x}_2(s)$ [3].

6.2.3 Writhe

Last but not least, the writhe is a concept which characterises the extent of chiral deformation of a single ribbon. In particular, it measures how much a closed ribbon deviates from being planar, implying that the writhe of a ribbon is exactly zero if and only if it is planar.

Much similar to the concept of the linking number, but only working on a single strand, the definition of the writhing number Wr of a ribbon C bears much resemblance to that of the linking number (Eq. 6.2):

$$\begin{aligned} Wr(C) &= \frac{1}{4\pi}\oint_C\oint_C \frac{\boldsymbol{x}_1-\boldsymbol{x}_2}{|\boldsymbol{x}_1-\boldsymbol{x}_2|^3}\cdot(\boldsymbol{dx}_1\times\boldsymbol{dx}_2) \\ &= \frac{1}{4\pi}\oint_C\oint_C \frac{(\boldsymbol{t}(s_1)\times\boldsymbol{t}(s_2))\cdot[\boldsymbol{x}(s_1)-\boldsymbol{x}(s_2)]}{|\boldsymbol{x}(s_1)-\boldsymbol{x}(s_2)|^3} ds_2 ds_1, \end{aligned} \tag{6.8}$$

where $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are two different points on ribbon C and the definitions of the rest of the variables are the same as those before.

The physical significance of Eq. 6.8 is that, the image of the ribbon (with directionality predefined) is first projected onto a flat plane and the *signed* crossings along *all* directions are summed up (hence the double line integrals) and averaged (hence the prefactor of $\frac{1}{4\pi} = \frac{1}{2}\cdot\frac{1}{2\pi}$).

[2] We have strictly followed the rule in tensor calculus that index contraction can only be performed between one *contra*variant component and one *co*variant component, hence the subscripts and superscripts.

[3] In the original article, the author did not specify the normal vector $\boldsymbol{n}$ to be a unit vector. However, it can be shown trivially, that the perturbation factor ε which appeared in Eq. 6.3 can be absorbed into $\hat{\boldsymbol{n}}$ such that $\boldsymbol{n} \equiv \varepsilon\hat{\boldsymbol{n}}$, and that ε is cancelled out in the vector transformation process.

6.2.4 Călugăreanu-White-Fuller theorem

Of the three quantities mentioned above, only the (self-)linking number is topologically invariant, meaning that this quantity is independent of homotopic deformations [266] [4]. On the other hand, the writhing and twist numbers are *not* topological invariants; these numbers change with *any* topological deformations.

Călugăreanu proved in 1961 [58], for the first time, that these three numbers of seemingly very different natures — not only their topological invariance, but also their integrality (SLk is always an integer but Tw and Wr are rarely integers [5]) — can be nicely linked together by a simple relationship

$$SLk(C_1, C_2) \equiv Wr(C_1) + Tw(C_2, C_1) \tag{6.9}$$

This theorem (the "Călugăreanu theorem") was first proposed for a closed three-dimensional system, but was extended to systems of higher dimensions by White [293] in 1969. Two years later, Fuller [102] focussed on the calculation of the writhing number, applied this theorem to an elastic rod system and produced speculations on the applicability to supercoiled DNA systems. The theorem is hence known fully as the "Călugăreanu-White-Fuller theorem" (hereby "CWF theorem"), and has been used extensively in the study of circular DNA systems, even by Crick [56] who had proposed the double helical structure.

The CFW theorem can be easily observed using a soft rubber tube or a coiled telephone wire, which are manufactured to have intrinsic material properties such as the tensile strength. If one holds the two ends of the tube together (hands on top of circle), joining them to form a circle while twisting the right hand forward (positive direction according to the right hand screw rule, hence overtwisting), the bottom of the circle can be observed to start spinning in the anticlockwise direction, and the tube becomes supercoiled. This is because as we induce an overtwist ($\Delta Tw > 0$), the linking number is increased by the same amount simultaneously. Note that Lk is *not* an invariant in this case since the tube, the topological object, is *not* closed, and hence a non-homotopic deformation would change its value. Now since the addition of twist would add a torsional stress to the system, the tube would supercoil itself in order to relieve this stress. In this process, the writhe would have the same direction as the original twist and the magnitude of the twisting number would decrease. In this process, the linking number is held constant as no external factors are currently causing the topological deformation, and such deformation is purely homotopic.

An immediate implication of the CFW theorem, when incorporating the fact that the self-linking number is geometrically and topologically invariant, is that in a homotopic deformation a change in the twist is compensated exactly by a change in the writhe in the opposite direction, and vice versa. Mathematically,

$$\begin{aligned} \Delta Wr(C_1) + \Delta Tw(C_2, C_1) &= 0 \\ \Delta Wr(C_1) &= -\Delta Tw(C_2, C_1). \end{aligned} \tag{6.10}$$

This redistribution of the twisting and writhing number is known as the *twist-writhe par-*

[4] A homotopic deformation is a change of shape of a topological object through a *continuous* path of homeomorphisms. In layman terms, such deformation does not require the breaking of the original topological object.

[5] The proof is beyond the scope of this work.

titioning [247]. This relationship is of crucial importance in the quantitative studies of circular DNA systems. This is because, as applied to DNA systems, it states that unless the backbones are cleaved (by DNA topoisomerases) [6], twisted (by gyrases) and religated (by ligases), the overall linking number of the closed DNA is kept constant.

Here, another important quantity, known as the *persistence length* L_p, should be mentioned. The persistence length of a material, in layman terms, is a measure of the rigidity or the stiffness of the material. More scientifically, it is the length at which the position of a segment of a material (or, molecules or residues in a polymeric chain) loses its correlation with that of another segment. That is to say, strings of lengths shorter than their persistence lengths can be modelled roughly as semi-rigid beams, whereas those longer than the respective persistence lengths have to be modelled using multi-dimensional random walker models [302].

Now, in the previous model of the rubber tube, since the tube has a persistence length of about 25cm [211], and the typical length L of the tube being used in the demonstration is about one or two metres, the L/L_p ratio is relatively low, and the tube is hence rather rigid. This is the reason why as soon as we induce a twist, the tube would start supercoiling almost immediately. On the other hand, as a thought experiment, if we could have a giant machine which does exactly the same as our hands did in the mini-experiment, but only on a circularised tube of, say, 50 metres, the *rate* of the supercoiling should be much slower due to the lower average torsional stress on the structure.

The latter case is exactly what happens in DNA systems. The persistence length of a double-stranded DNA is typically about 390Å [120] (which converts to about 115 bps, assuming an axial distance of 3.4Å per bp), and bacterial plasmid DNA is of a typical length of 1 to 200 kbps [271]. Hence the plasmid DNA usually has a very large L/L_p ratio and the topoisomerases and gyrases act as the mighty machinery in the previous thought experiment, and a slow supercoiling rate (comparing with atomic timescales) of the DNA thus results.

6.3 Methodology

6.3.1 Evaluation of salinity and solvent dielectric constants

In this chapter, since simulations are performed using the generalised Born implicit solvent (GBIS) model, a few parameters are known to contribute greatly to the credibility of any simulation outcomes, with two particularly important examples being salinity and the associated dielectric constant of the solvent. This is because as explained earlier in Section 2.6.2, implicit solvent schemes model the environment as a continuum with a preset dielectric constant.

The dielectric constant ε_r is a factor which modifies the electric permittivity ε of a medium. In vacuum, the permittivity is called the *permittivity of free space*, ε_0. But in other media, this value is higher, hence can be expressed as a multiple of ε_0, i.e. $\varepsilon = \varepsilon_r \varepsilon_0$. The generalised

[6] Whether one or both backbones are cleaved depends on the type of the topoisomerase.

Coulomb's Law for inter-charge forces

$$\|\boldsymbol{F}_{12}\| = \frac{1}{4\pi\varepsilon}\frac{q_1 q_2}{\|\boldsymbol{r}_{12}\|^2} = \frac{1}{4\pi\varepsilon_r\varepsilon_0}\frac{q_1 q_2}{\|\boldsymbol{r}_{12}\|^2} \tag{6.11}$$

then tells us that the force exerted on a charge q_2 by another charge q_1 is weakened by a factor of ε_r when in a non-vacuum medium, since $\varepsilon_r > 1$. This is called electrostatic screening effect and can be measured by means of the Debye length λ_D of the medium,

$$\lambda_D = \sqrt{\frac{\varepsilon_r\varepsilon_0 k_B T}{2\times 10^3 N_A e^2 I}}, \tag{6.12}$$

which is a measure of how far electrostatic effects reach in a screened system. Here, I is the *ionic strength* of the solvent, defined as [299]

$$I = \frac{1}{2}\sum_i c_i {z_i}^2 \tag{6.13}$$

where c_i and z_i are the molarity and charge of the i-th ionic species. In particular, the electric potential due to a point charge Q is given by

$$\Phi(\boldsymbol{r}) = \frac{Q}{4\pi\varepsilon r}\exp(-\frac{r}{\lambda_D}), \tag{6.14}$$

implying that for every λ_D away from the point source, the field is decreased by $1 - e^{-1}$, i.e. $\sim 63\%$.

Screening effect has profound influence on the supercoiling behaviour of cccDNA, as supercoiling is primarily done due to the long-ranged electrostatic effects. Hence, a change in the solvent screening strength, which in turn alters the Debye length of the environment, must change how an atom on one end of the DNA interacts with another on the opposite end.

The evaluation of the dielectric constant of a solvent may be trickier than it sounds, one of the reasons being that sophisticated instrumentation is essential in obtaining accurate results. Studies have been conducted for more than a century but it was not until 1948, that Hasted *et al.* performed the first systematic experimental study of the dielectric properties of salt-water solutions [131] and discovered that for dilute solutions of concentration c less than 1.5M the dielectric constant decreases linearly,

$$\varepsilon_r = \varepsilon_w - \alpha c \tag{6.15}$$

where ε_w is the dielectric constant of pure water and α is an ion-specific parameter called the total excess polarisation of the species [105]. Hasted *et al.* noted that the linearity gradually decreases as c increases beyond 1.5M, and reaches saturation at a certain high concentration. The theoretical framework of this experimental discovery was investigated by Gavish *et al.* in whose work [105] the functional form of ε is formulated as

$$\varepsilon(c) = \varepsilon_w - (\varepsilon_w - \varepsilon_{ms})\,\hat{L}\left(\frac{3\alpha}{\varepsilon_w - \varepsilon_{ms}}c\right). \tag{6.16}$$

Here, ε_{ms} is the limiting dielectric constant of the electrolyte (i.e. molten salt) and $\hat{L}(x) = \coth(x) - x^{-1}$ is the Langevin function. The parameters α and ε_{ms} are originally obtained by curve-fitting experimental data for electrolytes. It was noted in Buchner *et al.* [29] that both α and ε_{ms} are in fact temperature-dependent.

In this work, we have adopted the aforementioned model and have added temperature dependence of individual quantities to make the model even more robust.

For the temperature-dependence of pure water, we have directly adopted the model developed by Meissner *et al.* [198]:

$$\varepsilon_w(T) = \frac{3.70886 \times 10^4 - 82.168T}{421.854 + T} \tag{6.17}$$

where T is the temperature *in centigrade*. We fitted the parameters α and ε_{ms} with the experimental data from [29] to give

$$\begin{aligned} \alpha^{\mathrm{Na}}(T) &= 14.2254 e^{-0.0083T} \\ \varepsilon_{ms}^{\mathrm{Na}}(T) &= -0.01069T^2 + 1.0409T + 10.7323, \end{aligned}$$

with both R-squared values very close to unity for the temperature range of 5°C to 35°C. We assert that this fit can be extrapolated for use at a temperature not too far away from this range, say, the physiological temperature of 37°C (310K) which we have used throughout this work, as the dielectric properties should be smooth and so the trend should not change drastically with such a small perturbation in temperature. We stress here, once again, that this set of parameters only works for sodium chloride solutions but not those of other ions, as the parameters are specific to the ion species.

6.3.2 System creation and specifications

In order to simulate the cccDNA-drug complex system we first created the system. In this project, this was done using the NAB (acronym for Nucleic Acid Builder) program. The NAB is also a scripting language under the same name. When an NAB script is compiled and run, the NAB program extracts parameters from the script and feeds them into a sample C program within the AMBERTOOLS toolbox, which creates an all-atom double-stranded DNA-only system in the PDB format.

Since the original version of this sample program only allows for the creation of straight-chained DNA segments, appropriate changes were made in the raw code to include the bending (hence circularisation) of the segment. Moreover, alterations were also made so that an initial non-zero change of twist ΔTw can be induced at the creation step. Lastly, to increase the versatility of the program, subroutines have been added into the code so that random sequences can be tailor-made to the user's preferred GC content of the entire segment. For example, if a user wants a 400b circular segment, with 30% GC content, with an initial ΔTw of -5, the program will create a 400b random-sequenced circle with around 120 G or C pairs and having exactly 35 turns (cf. a canonical B-DNA has 10 bases per turn).

In this work, we have used a 160b circular segment of the randomly generated sequence

(with 50% projected GC content)

5′- TACGTAGCGAGCTATTGTTTCTAGGAGATAGATTCCGGCG
GAAACTTATCAAATCGAAACAAGCAGAAGGCTGCGAGGAT
TCGTCCCAAGACAATTAATCTCGGTGCGTTCCCGTTTCCT
CCTTTCTTAAAAGTTCGCGAAGTTGTTGGTTAAAGGCCGA **-3′**

This sequence has 43 A's, 44 T's, 34 C's and 39 G's, corresponding to an overall $45\frac{1}{8}$% GC content [7] which is close to the above-said projected value. Moreover, the initial ΔTw was set to be −2 so that there are exactly 14 turns in the 160b segment (Fig. 6.3).

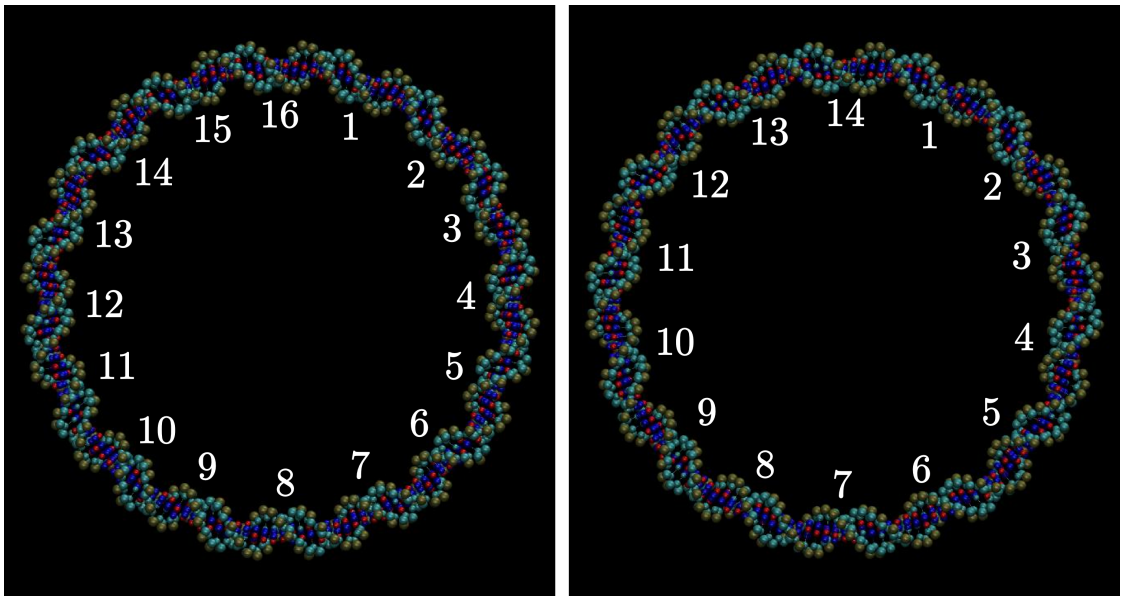

Figure 6.3: Coarse-grained representation of 160b circular DNA with turn counts. Left — Natural system with 16 turns. Right — System with 14 turns used in this work. Figures produced using VMD [145].

In order to speed up simulation and to allow for a longer simulation time, hybrid all-atom (AA) / coarse-grained (CG) systems have been adopted throughout this work, where the DNA segment was fully coarse-grained using the CGCONV utility within the SIRAH toolbox. The coarse-grained DNA was read into the XLEAP program and drug molecules (in AA model) were added to the system manually using the graphical user interface of XLEAP. For each of the systems, seven drug molecules, i.e. one molecule per two turns, were placed near the grooves of the DNA. We chose particularly this number as it is the closest integral factor of the number of turns (14).

In this chapter, the drug molecules chosen were daunomycin (DAU), doxorubicin (DOX) and idarubicin (IDA). Then the 5′ and 3′ ends of each of the two strands were bonded together so that the DNA would not relax back into an open-chain during the course of simulation. Moreover, systems were set up for the GBIS solvent model using the Onufriev-Bashford-Case model [221] for the effective radii for atoms.

[7] Since the complementary base for A is T and that for G is C (and vice versa) the GC content on the Crick strand is the same as that on the Watson strand.

Concentration (in M)	Dielectric constant
0.01538 (= 0.1c)	74.042
0.03076 (= 0.2c)	73.881
0.07690 (= 0.5c)	73.399
0.1538 (= 1.0c)	72.595
0.3076 (= 2.0c)	70.997
0.7690 (= 5.0c)	66.349
1.5380 (= 10.0c)	59.510

Table 6.1: Concentrations and corresponding dielectric constants of NaCl solution used in this work.

6.3.3 Energy minimisation and simulation protocols

Once the systems [8] were created, they were energy-minimised using NAMD. The minimisation process adopted in this chapter only consisted of a single static stage, in which systems were minimised under the conjugate gradient scheme for 5,000 steps at 0K. The purpose of this stage was to crudely remove any unphysical atomic contacts which would result in extremely high pairwise potential energy. Dynamic minimisation, i.e. simulated annealing, was not adopted in this part, due to two main reasons. Firstly, the systems would evolve during the MD simulation stage, and hence the conformation of the system would change drastically, and thus it is pointless to absolutely minimise the initial structure. Secondly, and much more importantly, an artificial and near-perfect circle with a non-zero initial ΔTw value is, by nature, carrying a lot of stress in the structure. An annealing of the initial system would hence render the conformation randomised and resulting in the loss of meaningfulness of the study in this chapter, which is the rate of relaxation of a stressed plasmid into a plectoneme.

For the MD simulations, each of the systems was first heated from 0K to 310K in 124 ps, then the temperature was kept at 310K for the entire duration of the simulation, i.e. 100 ns. Since the vast majority of the particles being used in the simulations were coarse-grained "super-atoms", we have chosen a rather long temporal step size of 4 fs per step, hence each simulation consisted of 25 million steps. We have used the Langevin thermostat in this work to regulate the system temperature, and the damping factor (collision frequency) was chosen to be 3.0 ps^{-1}. Moreover, to facilitate the long-ranged electrostatic interactions, the electrostatic cutoff radius and the radius of pair-listing were set to 150Å and 200Å respectively [9].

Since one of the main focuses of this chapter is to study the effect of salinity on drug-DNA interactions, each of the DNA-drug complex systems was split into seven different cases, which were simulated using the same protocol as described above, but in different concentrations of NaCl. The concentrations used throughout this chapter, and their corresponding

[8] By a system we mean an ensemble consisting of the 160b plasmid and 7 drug molecules as described above.

[9] A pair-list is a list of atoms within a shell of a certain thickness beyond the cutoff radius. In this work, the thickness was (200 – 150)Å= 50Å . The significance of such a list lies in the enhancement of computational efficiency. The pair-list is computed once per few calculation cycles, and atoms inside the list are the candidates of those which could potentially enter into the cutoff radius, hence having interactions. Having such a list could prevent unnecessary searches for atoms beyond cutoff.

dielectric constants (calculated using Eq. 6.16) are listed in Table 6.1.

Another feature in the MD protocol is the *forced* hydrogen bonds. In this work, we constrained the interatomic distances between corresponding super-atoms within the same base pair to their SIRAH preset value, which range from 2.81Å to 2.90Å according to the species, by means of a strong spring constant of 10 kcal mol^{-1} Å^{-2}. This is because whilst *natural* inter-bp hydrogen bonds are strong, they are still weaker than the overall stress. It was shown in early tests of this work, that the DNA would partially melt at regions with the highest stress to form bubbles as a means of relieving the stress.

Last but not least, since the simulations in this chapter were all performed using the GBIS scheme, i.e. without a finite-sized simulation box, the small molecules have the potential of drifting off from the DNA, causing the extent of interaction to decline over time. To alleviate this problem, a spherical unphysical soft boundary (with a radius of 10Å and an exterior potential barrier of 1 kcal mol^{-1} Å^{-2}) was imposed on each drug molecule around certain DNA residues [10]. We divided the 160 bps into seven sections (since there are seven drug molecules) as evenly as possible, i.e. six groups with 23 bps and one with 22 bps, and the centroids of the spherical constraints were then set to be the centroid of each group. For example, for the first drug molecule, its constraint included the first 23 bps (i.e. residues 1 to 23 and 298 to 320). The reason behind such a choice is that, since the diameter of a DNA "cylinder" is about 20Å [49, 245], choosing a 10Å diameter boundary can more-or-less ensure the constrained molecule to interact with the DNA 100% of the simulation time. Moreover, due to the aromaticity of the chromophores of the drug molecules, it has been seen in unconstrained test simulations that two drug molecules stack upon each other and form a duplex. Not only do such duplex systems hinder intercalating actions of individual molecules because of the increased overall thickness, the mode of normal interactions with DNA can be drastically changed as well, primarily because of the doubling of the total mass and the change in electrostatics of the molecule (or complex). The imposition of such a constraint as described above can effectively prevent multiplexing from occurring, at least in the relatively early stage of simulations.

6.3.4 Data analysis

Analyses in this chapter can be broadly divided into two categories, *viz.* qualitative and quantitative. Qualitative analysis includes the evaluation of the shape of plasmids during the course of simulations. This was mostly done via inspection of simulation snapshots using the visualisation software VMD. The main features we are most interested in include the evolution of the shapes of plasmids and the positions of the drug molecules during the simulations.

On the other hand, quantitative analysis entails the evaluation of the time evolution of ΔWr, i.e. the writhe number. A Python wrapper code was written for the calculation of this quantity using the library `pyknotid`. Each full 100 ns simulation (in 20,000 snapshots) was divided into 200 chunks of 0.5 ns (100 snapshots each frame). The instantaneous coordinates

[10] A residue in a biological macromolecule is a basic building block of it. For example, in proteins a residue can be an amino acid, whereas in DNA a residue is typically a whole nucleotide.

of the atoms at the final snapshot of each chunk were recorded. Then a virtual “ribbon" was formed from the centroids of the bps in the DNA, which are defined as the average of the coordinates of the six super-atoms representing the nucleobases within base pairs. The “ribbon" was then smoothed by taking samples once every eight base pairs, so that the smoothed “ribbon" consisted of 20, rather than all 160 points. This is because for a non-“canonical B-form" DNA, since the nucleobases are tilted, the centroids would form a helix by themselves which in turn has its own Tw and Wr values, which would interfere with the global values. Whilst using a smoothed “ribbon" may risk losing some of the features, it gives a better representation of the system on the tertiary structure level. With this, the writhe value of the smoothed ribbon $\Delta Wr(t')$ at a particular time t' was then calculated using the aforementioned wrapper code. The whole $\Delta Wr(t)$ profile was formed by joining all the ΔWr values at the 200 time frames, plus the post-minimisation configuration as the first frame at $t = 0$.

The essence of the quantitative analyses lies in the next part, where the writhe profiles obtained in the previous step were fitted to the form

$$Wr(t) = A \exp\left(-\frac{t}{\tau}\right) + B \tag{6.18}$$

where A and B are fitting parameters and τ is the time constant. Here, A represents the total span of the time-dependence writhe value and B is the steady-state value for the writhe (i.e. at $t \to \infty$). To probe the *rate of change* of Wr, we took the time derivative of Eq. 6.18 to get

$$\frac{d}{dt} Wr(t) = -\frac{A}{\tau} \exp\left(-\frac{t}{\tau}\right). \tag{6.19}$$

In particular, we obtained the *initial* rate by taking $t = 0$, resulting in

$$R_0 = \left.\frac{d}{dt} Wr(t)\right|_{t=0} = -\frac{A}{\tau}. \tag{6.20}$$

In terms of error analysis, since curve-fittings were performed using the `curve_fit` function in the python library `scipy.optimize`, the standard deviations of the A, B and τ parameters were obtained by directly taking the square roots of the diagonal elements of the covariance matrix from the outputs. As for the error (standard deviation) of the initial rate R_0, we have exploited the relation

$$\begin{aligned} \sigma_{R_0} &= \left|\frac{\partial R_0}{\partial A}\right| \sigma_A + \left|\frac{\partial R_0}{\partial \tau}\right| \sigma_\tau \\ &= \frac{1}{\tau}\sigma_A + \frac{A}{\tau^2}\sigma_\tau \end{aligned} \tag{6.21}$$

where σ_A and σ_τ are the standard deviations of A and τ obtained using the method described above. The absolute values of the partial derivatives are taken because errors accumulate.

6.4 Results and discussion

In this section we present the results from the simulations of the cccDNA-drug systems. The flow of presentation in each of the system will follow roughly the list below:

1. Discussion on the shape of the plasmid
2. Discussion on the interaction mode between the drug molecules and the DNA
3. Discussion on the writhe number profile
4. Comments on the rate of supercoiling

In the context of physiology, human blood serum can be divided into three groups according to the sodium concentration, *viz.* isonatremic, hyponatremic and hypernatremic. The meanings of these terms can be seen from their respective Greek roots: isonatremic environments are those having similar sodium concentration to physiological systems (~ 153.8mM), whereas hypo- and hypernatremic environments correspond to those with significantly lower and higher sodium contents than the number quoted above. In the discussion below, we will use this categorisation for our systems of different sodium concentrations.

6.4.1 Bare DNA — control

Isonatremic environment We first investigate the effect of salinity on the supercoiling behaviours of a bare plasmid. We consider the system in isonatremic environment first, as this is the most natural form of the system possible, and should serve well as a control experiment.

Figure 6.4 shows the snapshots taken from the simulation of bare DNA in isonatremic environment. It can be clearly seen (especially from the lower subfigure) that during the process of supercoiling, the plasmid formed kinks on opposite sides of the circle first, then the crossing of the strands slid up the circle until it reached roughly halfway up the circle. This is clearly counter-intuitive, as one would logically deduce that with the natural preference of symmetry, the simplest way of creating an “8-shaped" knot is to twist the two opposite ends at the same rate but in opposite directions.

Figure 6.5 shows the time evolution of the writhe number of the system. The blue curve (i.e. the data obtained directly from the simulation) shows that the ΔWr value started from a value very close to zero and gradually decreased to rather near -1. The rate of decrease of ΔWr is noted to decrease with time, and the value tended to a saturation point near the end of the simulation. The profile was fitted to the exponential form explained above (Eq. 6.18) and we obtained for the parameters $A = 1.034 \pm 0.028, B = -0.970 \pm 0.007$ and $\tau = (12.788 \pm 0.606)$ns. Hence the relative errors (percent errors), i.e. $\frac{\sigma_A}{A}$ etc., of the parameters are 2.7%, 0.7% and 4.7% (rounded to 1 decimal place) respectively, in turn showing that exponential regression is suitable for use in the analysis of the time-dependent writhe value. However, although the aforementioned regression model is proven to be appropriate, the *instantaneous* deviation of the data points from the fitted curve is worth mentioning. From Figure 6.5, we can see that the ΔWr value fluctuates rather wildly at times (cf. near

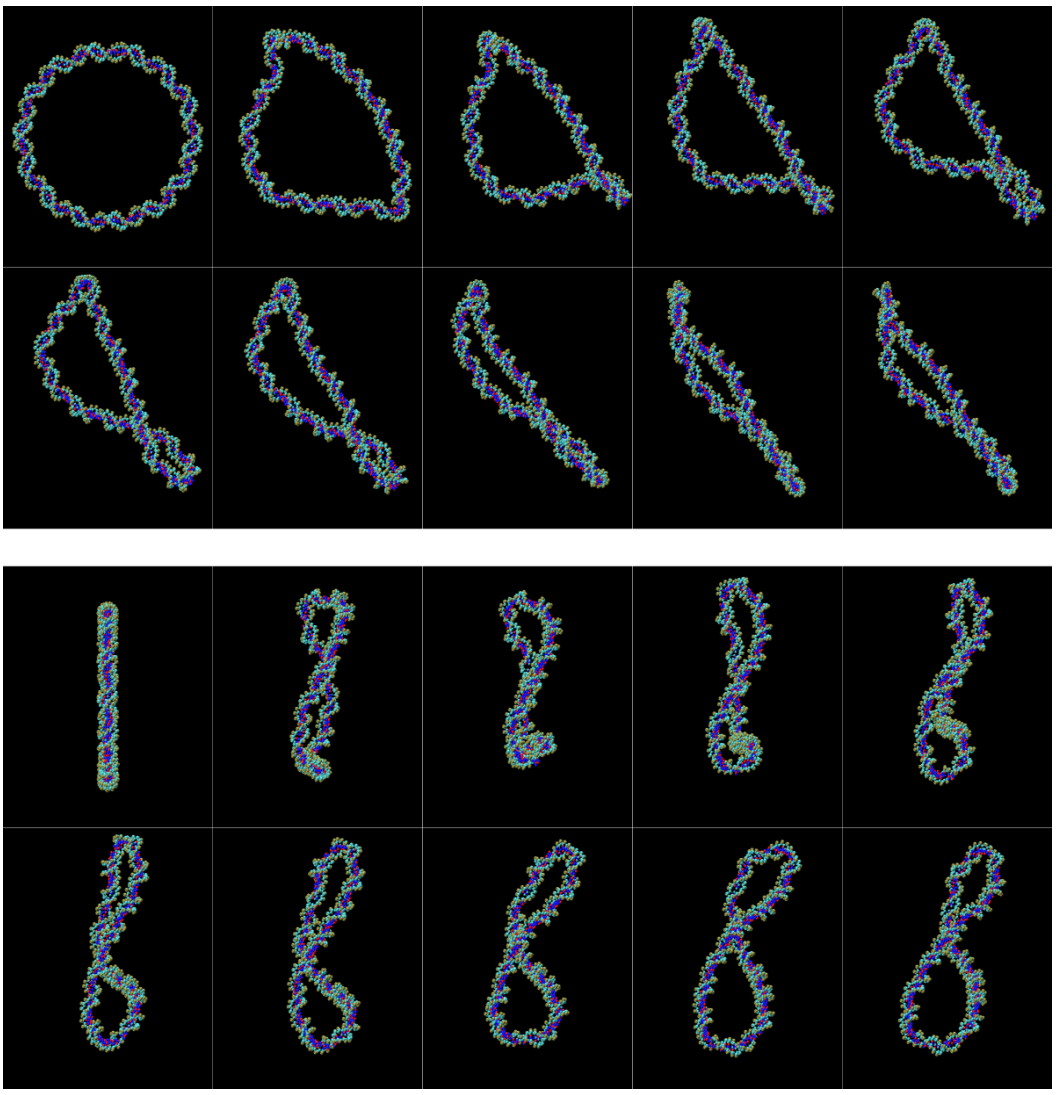

Figure 6.4: Snapshots from simulation of bare DNA in isonatremic environment, taken at 10ns intervals. Lower subfigure shows the same snapshots as the upper but rotated about the $+z$ axis by 90° anti-clockwise. Direction: Left-to-right, then top-to-bottom.

the start). This is because ΔWr, whilst telling how tightly a loop is supercoiled by *counting the number of strand crossings*, is highly sensitive towards the structural topology of the knot, which includes features such as local twists and bends and the non-planarity of the knot. This then makes the non-integer values easily justifiable: if these are not taken into account, then regardless of how twisted or bent the knot is, so long as there is one crossing between the strands $|\Delta Wr|$ must be exactly unity [11], which is clearly not the case. To consider the same argument from another perspective, there exists infinitely many conformations of the system which could give rise to the same ΔWr as the positions of the atoms, i.e. points on the knot, are continuous quantities (cf. Eq. 6.8). Therefore, the absolute value of ΔWr at a particular time does not necessarily give adequate information about the system conformation, whereas the trend which it takes over time conveys a more important message, which is how the plasmid topology transforms macroscopically.

[11] The sign of ΔWr only tells the directionality of the supercoil.

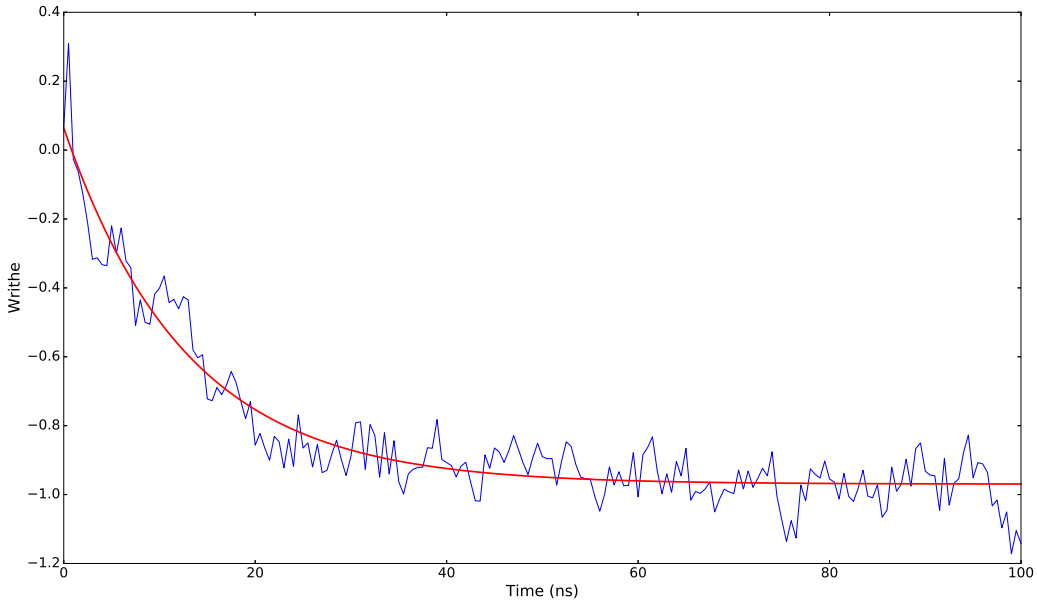

Figure 6.5: Writhe number profile for the bare DNA in isonatremic environment. Blue: data from simulation; Red: exponential fit of data.

$[Na^+]$	A	σ_A	B	σ_B	τ	σ_τ	R_0	σ_{R_0}
0.1	1.024	0.019	-0.843	0.010	23.163	1.030	-44.217	2.785
0.2	0.681	0.025	-0.662	0.007	15.412	0.981	-44.214	4.301
0.5	1.094	0.017	-0.970	0.008	20.383	0.715	-53.692	2.740
1.0	0.997	0.019	-1.019	0.011	23.513	1.100	-42.393	2.799
2.0	1.099	0.023	-0.985	0.007	14.801	0.570	-74.277	4.399
5.0	1.249	0.025	-1.229	0.006	11.725	0.405	-106.233	5.792
10.0	0.766	0.036	-0.611	0.006	6.661	0.498	-114.972	13.986

Table 6.2: Exponential fitting parameters for bare DNA system in different salt concentrations (in multiples of $c = 0.1538$M). Units: $[\tau]$ = ns, $[R_0] = \text{ps}^{-1} = 1000\text{ns}^{-1}$.

Non-isonatremic environments Table 6.2 shows the parameters used in the exponential fitting of data. We can see that for systems in hyponatremic solutions, the fluctuation in the initial supercoiling rates of the DNA is relatively small. Moreover, the values are all comparable to the isonatremic case. However, on the other hand, we see that for the hypernatremic cases, the value of R_0 increases rather drastically with the concentration of ions, whereas the time constant τ, which tells how quickly the system attains stable state, decreases quickly with the increase in solvent concentration. This may imply that whilst the interatomic electric force extends to infinity, the dominance of the short-ranged and long-ranged effects can vary a lot. Bearing in mind that supercoiling of cccDNA is mainly facilitated by the long-ranged component, we can easily see why the systems exhibit such behaviours as described above. Firstly, the drop in the dielectric constant (cf. Table 6.1) when the salinity increases from $0.1c$ to $1.0c$ ($74.042 \rightarrow 72.595$) is relatively small compared to that when the salinity further increases from $1.0c$ to $10.0c$ ($72.595 \rightarrow 59.510$). Since the dielectric constant measures how strongly the electric field (or force) is weakened, this suggests that the long-ranged interaction is more heavily screened in an environment with a higher dielectric constant than in one with a lower dielectric constant, due to the inverse-square law of the force. This in turn means that long-ranged interactions are much less prevalent in low salinity systems

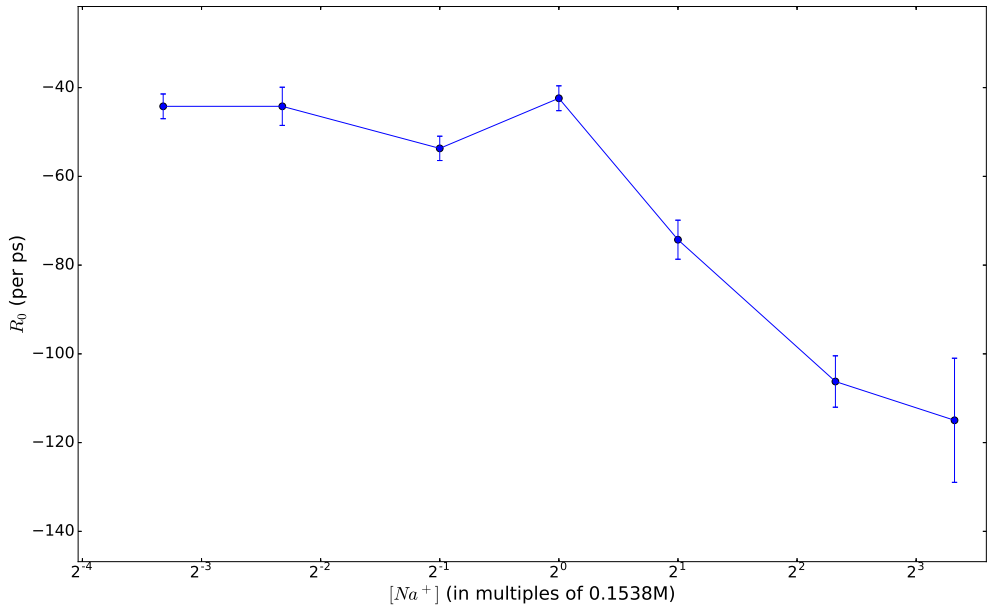

Figure 6.6: Graph showing initial rate R_0 as a function of salinity for bare DNA system.

than in high salinity systems. This is also the reason why we see, in Table 6.2, the nonlinear increase of the rate of supercoiling with the solvent salinity.

6.4.2 Daunomycin and doxorubicin

[Na+]	A	σ_A	B	σ_B	τ	σ_τ	R_0	σ_{R_0}
0.1	0.420	0.014	-0.640	0.013	32.344	3.299	-12.982	1.761
0.2	0.870	0.018	-0.880	0.021	41.064	2.673	-21.189	1.817
0.5	0.846	0.024	-0.866	0.007	14.109	0.747	-59.960	4.911
1.0	1.002	0.022	-0.956	0.007	16.541	0.696	-60.604	3.856
2.0	0.971	0.017	-1.027	0.007	19.372	0.721	-50.110	2.739
5.0	0.899	0.015	-0.956	0.006	18.576	0.663	-48.398	2.560
10.0	0.912	0.022	-0.876	0.005	10.615	0.435	-85.898	5.607

Table 6.3: Exponential fitting parameters for DNA-DAU system in different salt concentrations (in multiples of c = 0.1538M). Units: $[\tau]$ = ns, $[R_0]$ = ps^{-1} = 1000ns^{-1}.

Table 6.3 shows the parameters determined from the exponential fitting of data points obtained from the DNA-DAU system. We can see that the trend of the initial rate R_0 does not follow strictly that of the control experiment without drug molecules. Firstly, in the cases where the system was under extreme hyponatremia, we observe that the R_0 values are extremely low, when compared with values from systems with higher sodium concentrations and those from their counterparts in the control experiments. This is contributed both from the very low values of A (the extent of supercoiling) and the abnormally high values of τ. Secondly, we see quite clearly that the variation in R_0 with respect to the salinity from [Na+] = 0.5 onward is not as high as that in the control experiments. Moreover, for the vast majority of R_0 values in the DNA-DAU systems, they are lower than their respective counterparts

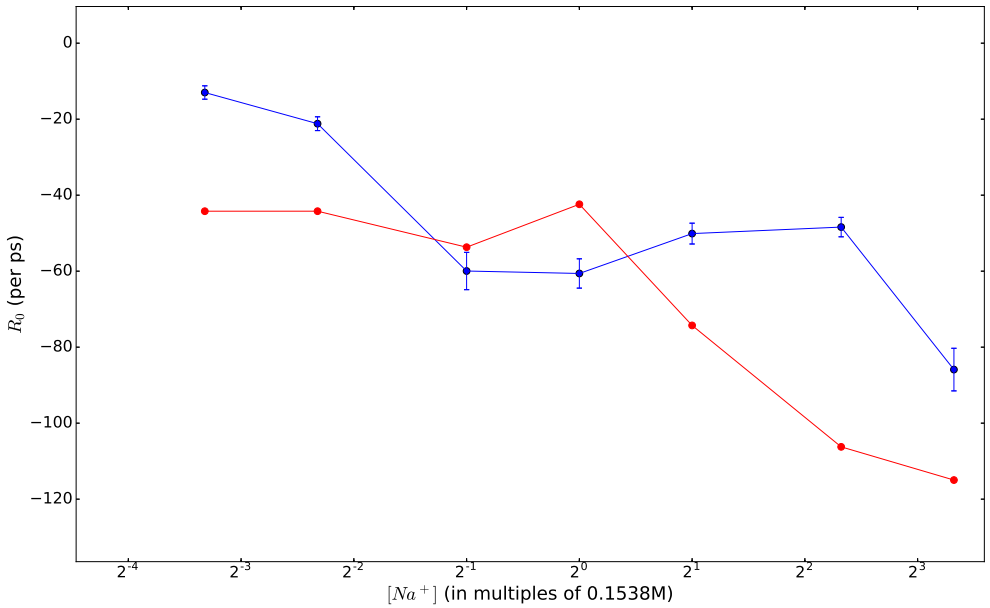

Figure 6.7: Graph showing initial rate R_0 as a function of salinity for DNA-DAU system. R_0 for bare DNA plotted in red for comparison.

in previous discussions. This can be attributed to the interactions between the drug and the DNA.

[Na+]	A	σ_A	B	σ_B	τ	σ_τ	R_0	σ_{R_0}
0.1	0.704	0.009	-0.718	0.005	23.866	0.712	-29.506	1.239
0.2	0.729	0.013	-0.755	0.006	21.345	0.847	-34.169	1.953
0.5	1.039	0.018	-0.993	0.006	15.008	0.480	-69.202	3.394
1.0	0.863	0.025	-0.830	0.006	10.564	0.513	-81.678	6.316
2.0	0.807	0.020	-0.800	0.004	10.492	0.443	-76.887	5.172
5.0	1.128	0.019	-1.178	0.006	15.965	0.504	-70.674	3.395
10.0	1.107	0.061	-1.082	0.006	17.772	0.528	-62.272	2.764

Table 6.4: Exponential fitting parameters for DNA-DOX system in different salt concentrations (in multiples of $c = 0.1538$M). Units: $[\tau]$ = ns, $[R_0]$ = ps^{-1} = 1000ns^{-1}.

Table 6.4 shows the parameters determined from the exponential fitting of data points obtained from the DNA-DOX system. Since the basic method of analysis for doxorubicin systems is the same as that for daunomycin systems, we will not repeat the procedures here. However, in terms of the results, we do see an interesting phenomenon in the DOX case which does not occur in DAU, which is the decrease in R_0 when salinity goes higher than the physiological value of $1c = 153.8$mM. We assert that this is related to the structural difference between the two molecules. From Fig. 1.8 we see that whilst the structures of DAU and DOX are nearly identical to each other, the relatively inert acetyl side chain in DAU is swapped into a more reactive carboxyl group. This means that while DAU interacts weakly with the DNA via induced dipole moments, the interaction between DOX and DNA is much stronger as it is a direct electrostatic interaction. This also implies that DOX when interacting with DNA, actively changes the electrostatic behaviour within the DNA

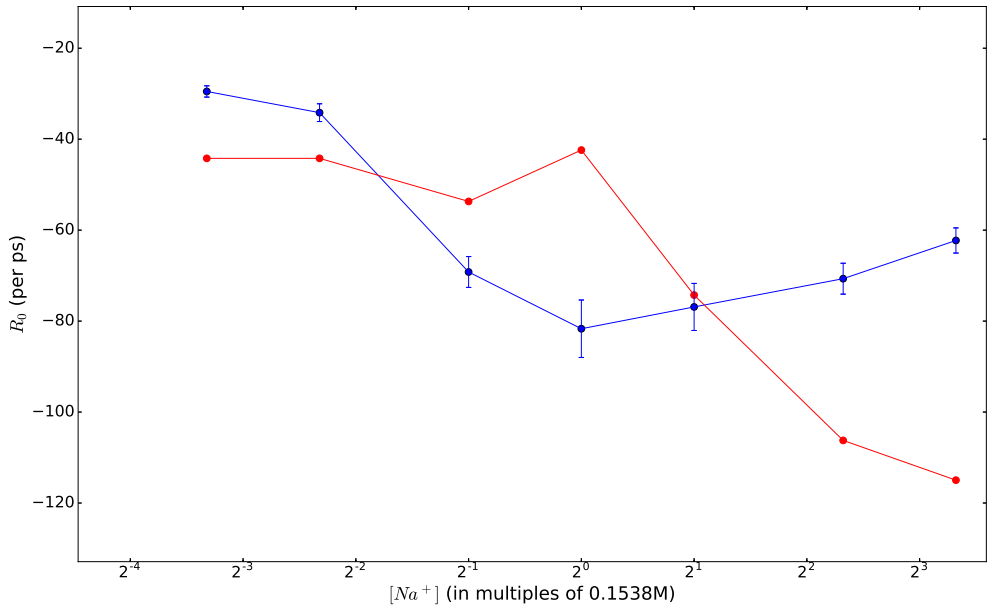

Figure 6.8: Graph showing initial rate R_0 as a function of salinity for DNA-DOX system. R_0 for bare DNA plotted in red for comparison.

substrate, with effect especially prevalent in the long-ranged forces.

6.4.3 Idarubicin

[Na+]	A	σ_A	B	σ_B	τ	σ_τ	R_0	σ_{R_0}
0.1	0.792	0.012	-0.787	0.010	29.833	1.305	-26.999	1.622
0.2	0.477	0.013	-0.526	0.004	13.632	0.688	-34.982	2.743
0.5	0.921	0.015	-0.925	0.006	19.425	0.683	-47.396	2.446
1.0	1.112	0.023	-0.977	0.026	41.233	2.630	-26.695	2.266
2.0	0.998	0.028	-1.072	0.028	34.607	3.012	-28.824	3.314
5.0	1.147	0.031	-0.939	0.006	9.547	0.429	-120.116	8.654
10.0	1.414	0.025	-1.333	0.014	23.338	0.987	-60.569	3.623

Table 6.5: Exponential fitting parameters for DNA-IDA system in different salt concentrations (in multiples of $c = 0.1538$M). Units: $[\tau]$ = ns, $[R_0]$ = ps^{-1} = 1000ns^{-1}.

A similar phenomenon occurs with systems involving idarubicin as in those involving the two other drugs, where we see that apart from one particular case, the supercoiling of the cccDNA had been slowed down by the interaction with idarubicin. However, what is interesting here is the case where $[Na^+] = 5c$, where a sudden spike in R_0 can be seen.

The reason for this can be understood from Fig. 6.10, the snapshot from the simulation at $t \sim 35$ns. We can see rather clearly that two of the seven IDA molecules have formed a duplex, which we have discussed before, and are bridging across two segments of the DNA. We assert that this bridging action is facilitated by the high charge density of the backbone of the DNA and the polarisability of the drug molecules. Due to the interaction, the two sides of the arc are held relatively close together, producing a short region within

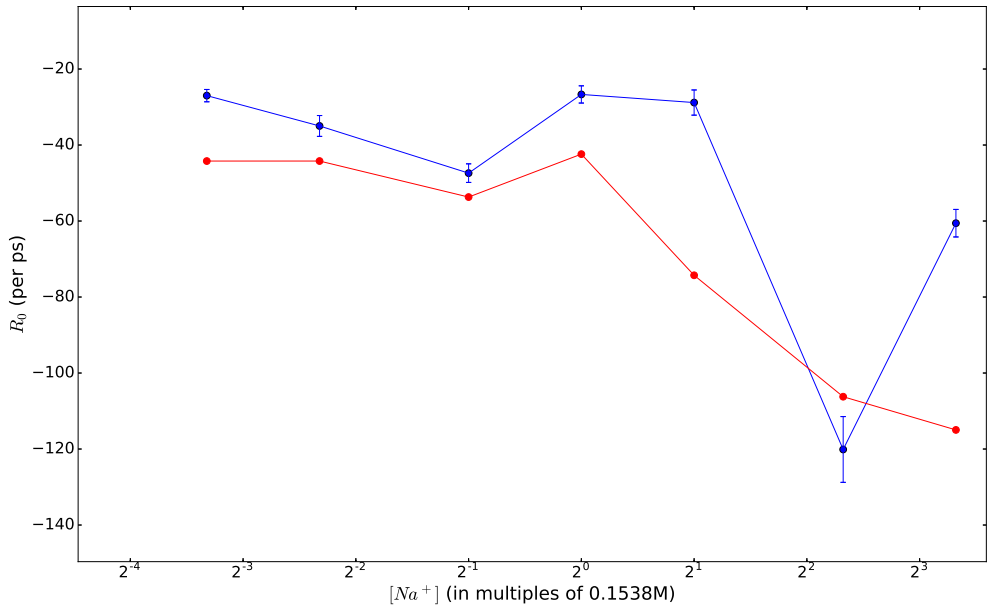

Figure 6.9: Graph showing initial rate R_0 as a function of salinity for DNA-IDA system. R_0 for bare DNA plotted in red for comparison.

the DNA with a very high curvature. This also explains why the value of τ is so abnormally low, which in turn implies an early attainment of steady-state.

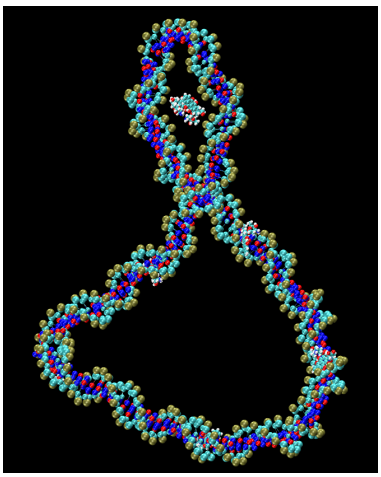

Figure 6.10: An IDA duplex bridging across an arc of the DNA.

It is natural to suspect whether there are correlations between such bridging effects and the spherical positional constraints we imposed on the idarubicin molecules. The answer to this is rather complicated, as it is both affirmative and negative, depending on the stage of simulation. If there were no constraints imposed, whilst the DNA would still undergo supercoiling, since the drug molecules are free to move, the formation of bridges would be purely probabilistic. However, as we have seen in the case with constraints, the occurence of bridging is certain. This can be attributed mostly to the regular placement of the drug molecules, which implies that regardless of how the DNA carries out supercoiling, at least one drug molecule could be found around where the strands cross each other, which then would be eligible for exerting the bridging action. Moreover, for the same reason, duplexes are ensured to form as the DNA supercoils and this further strengthens the bridging between the strands. Therefore, we assert that whilst the constraints are not the direct cause for the bridging action, they *ensure* such interaction happens and *enhance* its strength.

We stress here, that the aforementioned phenomenon is likely to be due to the concentration effect of the drug. To illustrate this, we take the typical dosage of idarubicin. As an injectable

solution, idarubicin is introduced to the (adult) patient in 1 mg/mL concentration (which converts to 2.010 mM) and 12 mg/m^2 amount per day [290]. With the dilution effect of the blood serum, cell cytoplasm and the nucleoplasm, the concentration of the drug actually interacting with the DNA should be much lower. This in turn suggests that the probability of two IDAs meeting each other and forming a duplex which bridges across a short section of a DNA is, in reality, rather unlikely. Moreover, as elucidated earlier, DNA segments in cells are typically thousands to millions of base pairs long — a 160b plasmid is extremely short.

6.5 Summary

In this Chapter, we have studied the effect of salinity and anthracycline drugs on the supercoiling behaviours of cccDNA. In particular, we have derived a one-equation description for the estimation of the dielectric constant of a sodium chloride solution, as a function of both salinity and temperature (suitable up to physiological temperatures). We have also devised an exponential fit for the time-dependent changes in the writhe number of the cccDNA in action.

In terms of results, firstly, we have shown that without any external molecules, the rate of supercoiling of the DNA increases with the concentration of the solvent. However, we discovered that the increase in the rate is not linear with solvent concentration, due to the non-linearity of the dielectric constant with respect to salinity. Nevertheless, we assert that the overall trend is expected, as the driving force of DNA supercoiling is long-ranged electrostatic interactions which are inversely proportional to the dielectric constant of the solvent.

Secondly, we have shown that with the introduction of external drug molecules, the supercoiling behaviour has changed. For example, in both the cases of daunomycin and doxorubicin, supercoiling at low and high salinities are suppressed, whereas at near-physiological salinities, the rates of supercoiling are more-or-less maintained at the value without the drugs.

Thirdly, in nearly all cases with idarubicin, we saw that the rates of supercoiling were being suppressed. However, in the case where the salinity was five times the physiological value, supercoiling was speeded up. We attributed it to the duplexing of two adjacent idarubicins and the bridging action of the duplex between two short segments of the DNA which forces the formation of a sharp nick in the DNA.

Finally, as a consequence of the general retardation of supercoiling due to interactions between anthracycline molecules and plasmid, we assert that such overall action of the drugs and DNA, and the topological variations thereof, may also cause the mode of interactions among DNA, anthracyclines and topoisomerases (the enzymes whose functions are inhibited by the above-said drugs) to change, hence suggesting another pathway of pharmacological actions of the drugs.

Chapter 7

Energetics studies of DNA intercalation

7.1 Motivation of studies

In the previous chapters, we have studied the effects of different binding modes of anthracycline drugs on DNA. In Chapters 4 and 5, we have studied the structural perturbation induced by intercalative binding. In Chapter 6, we have studied how groove-binding (i.e. *non*-intercalative) actions of the drugs change the supercoiling behaviour of closed circular DNAs.

In this Chapter, we will revisit intercalative interactions of anthracyclines, but from the perspective of the energetics. In particular, we calculate the free energy change associated with the intercalation of the drugs. The free energy *change* ΔG of a reaction, as explained in Chapter 2, is directly linked with the equilibrium constant through the equation

$$K_{\mathrm{eq}} = \exp\left(-\beta \Delta G\right)$$

which implies that, for a reversible reaction, the more negative the ΔG value, the further the equilibrium shifts towards the products side. In the case of intercalative interactions, the equation means the more negative the ΔG is, the more likely the intercalation would happen.

This has a pivotal importance in this work, since we would like to learn which drugs have higher intercalation probability. The probability of action is one of the most vital measures of the effectiveness of a drug. However, the effectiveness of a drug does not solely depend on the overall reactivity on the substrate. Another important feature of DNA- or protein-binding drugs could come from the sequence-specificity of their action. This is because these drugs, especially anti-cancer agents, should target specific site(s) of the DNA or proteins which contribute most towards the uncontrollable growth of the cancerous cells or the devastating effects of those cells on the body. If an intercalating drug does not have a clear site-specificity, meaning that the tendency or probability of intercalation into other irrelevant sites are significant as well, then it might also exhibit other pharmacokinetic side-effects, such as cardiotoxicity [283], on the patient's body, which could be potentially fatal.

7.2 Methodology

7.2.1 System hierarchy

In this Chapter, we calculate the ΔG for the drugs' intercalation using the eABF algorithm described in Sec. 2.7.2. Moreover, the procedures which have been adopted for the calculations follow strictly the geometrical transformation, or the "stacking the matryoshki", method, as explained in Sec. 2.7.1. Therefore, before going into the technicalities regarding how the simulations or calculations are done, it is necessary to first explain the nomenclature of our systems, and to explore what contributes to every layer of "matryoshka".

Fig. 7.1 shows the hierarchical structure of the systems used in this Chapter, with the CVs whose associated PMFs are calculated enclosed in purple hexagonal nodes.

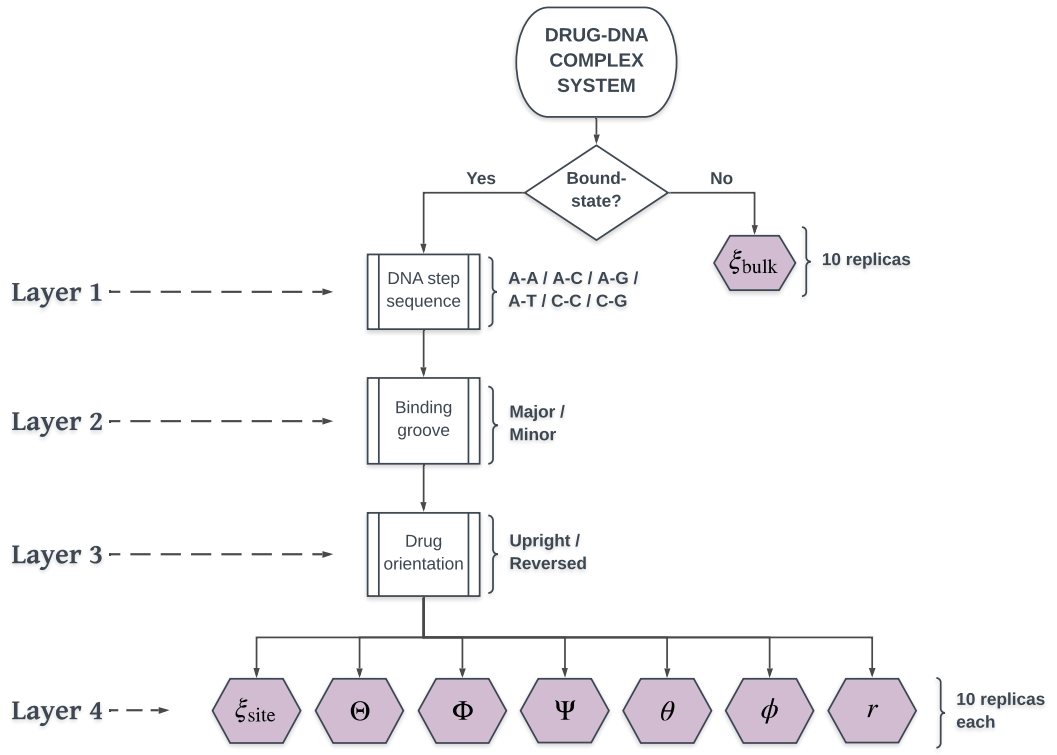

Figure 7.1: Flowchart showing the hierarchical structure of systems in Chapter 7.

The first layer constitutes the binary step sequence of the DNA, that is, the base sequence of the two nucleotides which make up the intercalation site. Naïvely speaking, because there are four types of nucleotides, the total number of possible combinations of a two-step sequence is $4^2 = 16$. However, some of these dinucleotide steps are not unique and can be re-created by means of swapping and inverting of other dinucleotide steps. As such, it is widely accepted in the biochemical community to define ten such steps which are mutually independent of each other, *viz.* A-A, A-G, G-A, G-G, A-C, A-T, G-C, C-A, C-G and

T-A [195] [1].

Svozil *et al.* [265] performed fully quantum mechanical calculations on the strengths of 5′-to-3′ directional stacking interactions of the 10 unique steps in B-DNA. It was discovered through their study that although the stacking energy (determined using the CBS(T) level theory) range from −15.27 kcal mol^{-1} (for the G-C step) to −9.93 kcal mol^{-1} (for the C-C step), the 5′-to-3′ stacking potentials differ from the 3′-to-5′ by only a few percents. For instance, the C-G step (-15.22 kcal mol^{-1}) is only about 0.4% higher than the G-C step. In view of this, we assert that the whilst omitting this small contribution may lead to a slightly larger error in our calculations it can effectively reduce the computational cost. As such, in this Chapter, we have further reduced the unique dinucleotide steps to six, *viz.* A-A, A-T, A-C, A-G, C-C and C-G.

The second layer only has two components, and it accounts for the binding mode of the drug, *viz. major groove* binding and *minor groove* binding.

The third layer, also having two components, accounts for the orientation of the drug as it intercalates into the DNA. From the discussion in Chapter 1, we know that all anthracycline drugs have a planar chromophore which serves as the intercalating component, and one or more side chains ("tails") which stick out into the groove after the chromophore has intercalated. Owing to the rigid stereochemistry of the ring structure of the chromophore, these tails cannot freely rotate, and hence could be used to determine the orientation of the molecule. Fig. 7.2 shows the "upright" and the "reversed" orientations of a daunomycin molecule. All the three anthracycline drugs we use throughout this thesis have a six-membered sugar ring [2] which is by far the largest side chain at their "tails"; we then call these sugar rings the major chain. The "upright" orientation is, hence, one which has the major chain parallel to the direction of the helical axis, whereas the "reversed" orientation is one which has the major chain point at the *reversed* direction of the helical axis.

Lastly, the fourth layer has seven components, which accounts for the seven *bound-state* collective variables in the ABF and eABF formalisms (cf. Secs. 2.7.1 and 2.7.2), namely the bound-state RMSD (ξ_{site}), the three orientational (Euler) angles (Θ, Φ, Ψ), the two conformational angles (θ, φ) and the radial contribution r.

The above hierarchical system applies to all the *bound-state* drug-DNA complex systems. For the *unbound* state, we have an alternative CV, *viz.* ξ_{bulk}, which accounts for the PMF contribution from conformational changes of the ligand in the *bulk*, i.e. free in the solution.

System nomenclature Because of the complexity of the hierarchy of the systems, we have devised a nomenclature which can be used for all systems in this Chapter for easier communication. The general form reads "**dddXXyZz**", where the first three letters "ddd" denote the drug species, using the acronym for the drug name — "dau" for daunomycin, "dox" for doxorubicin and "ida" for idarubicin.

The next two letters "XX" denote *both Layers 1 and 2* in the discussion above. The letters

[1] The exact combinations of the unique dinucleotide steps are *not* unique *per se*. However, regardless of how the combination is defined, the set should always contain 10 such steps.

[2] In the IUPAC nomenclature the formal names of daunomycin, doxorubicin and idarubicin all end with "hexopyranoside", meaning "compounds derived from the hexopyranose sugar".

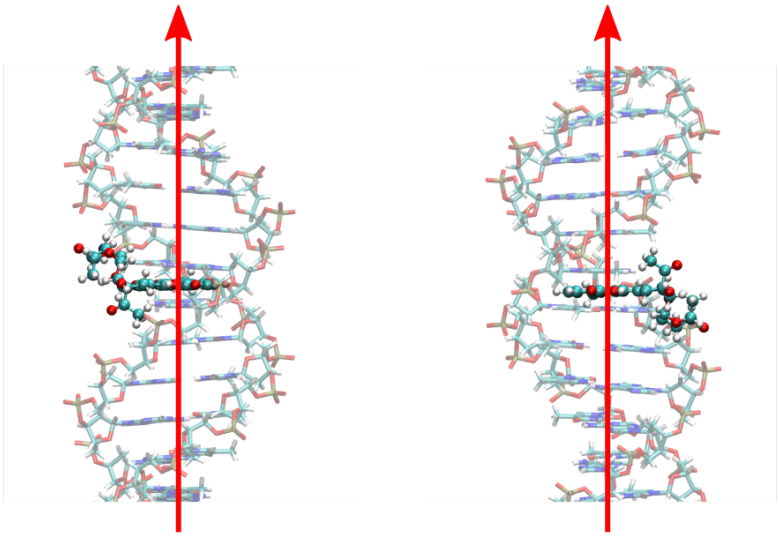

Figure 7.2: Schematic diagram showing the two possible orientations of the intercalator. Left — "Upright"; Right — "Reversed". Red arrows denote directionality of helical axis.

themselves tell the binary step sequence (i.e. Layer 1) whereas the cases of the letters tell the binding mode (i.e. Layer 2) — uppercase means major groove binding and lowercase means minor groove binding. For instance, "CG" would mean a major groove binding in a C-G step, whereas "ac" would mean a minor groove binding in an A-C step.

The single (lowercase) letter "y" denotes the third layer in the hierarchy which accounts for the orientation of the bound drug. A "u" means the drug is intercalated in the *upright* orientation, whereas an "r" means it is intercalated in the *reversed* orientation.

Finally, the last two letters "Zz" depict the fourth layer in the hierarchy which tells which CV whose PMF is being calculated. The codes for the angular CVs use the first two letters from their respective *anglicised* spellings, with the case of the letters preserved. For instance, "Ph" (for *Phi*) is used in place of Φ, whereas "ph" (for *phi*) is substituted for φ, et cetera. Moreover, for the bound-state RMSD (ξ_{site}) the code is "Bd", whereas for the radial component (r) it is "Rr".

As an example of a real case studied in this Chapter, the bound-state system "**idaCGrPs**" would be, if spelt out in full, "an **ida**rubicin intercalated into a **C-G** step from the **major** groove, in the **reversed** orientation; the CV concerned is the Euler angle **Ψ**".

For the unbound state, since the system does not have the four-layer hierarchy (cf. right hand branch in Fig. 7.1), the code is much simpler and only has the form "**dddUb**" where "ddd" is still the drug's code and "Ub" says that the drug is unbound. Hence, the system "**dauUb**" is "a **dau**nomycin in the **u**n**b**ound state".

7.2.2 System specifications

In this subsection, we will discuss the specifications of the systems used for the study in this Chapter. As discussed in the previous subsection, the systems can be broadly divided into two categories, *viz.* unbound-state and bound-state, and so should also be the discussion about their respective specifications.

Unbound state For systems involving a lone intercalator in the unbound state, the drug molecules were solvated in a truncated octahedral box of WT4 coarse-grained water molecules. The box size was set to obey the rules such that the closest distance between any atom from a WT4 molecule and any atom from the solute (i.e. the drug molecule) is at minimum 10Å. For the drugs used in this Chapter, counterions were not added to the system as the solutes are all charge neutral.

Bound state From the discussion in the previous subsection, we see that there are a total of $3 \times 6 \times 2 \times 2 = 72$ bound-state drug-DNA complex systems being studied in this Chapter. Therefore in order to simplify the communication here, we only use one of them, say "doxACr", as example, and the rest of the systems follow suit.

For the bound-state DNA, we used a short 10 base pair oligomer of the sequence $d(AC)_{10}$ [3]. The DNA was created using the all-atom/coarse-grain (AA/CG) method, where the three base pairs on either end were coarse-grained and the central four base pairs remained in all-atom scale. Then the intercalator, which in this case was a doxorubicin, was inserted manually between the fifth and the sixth base pairs from the major groove, in the reversed orientation.

The drug-DNA complex system was then charge-neutralised by introducing 18 sodium counterions, since the 10b DNA has an overall charge of $-18e$ [4]. Lastly, as in the unbound case, the whole system was solvated using a truncated octahedral box of WT4 water, with a shell width of 10Å.

7.2.3 Definitions of atom groups and CVs

Before we proceed onto the discussion of how calculations were actually performed, it is necessary to first explain how the CVs concerned in this Chapter are defined. Fig. 7.3 provides a schematic depiction of how atoms in the DNA (the substrate) and in the intercalating drug (the ligand) are defined.

For the groups in the DNA, the orange stars labelled P_i denote the backbone phosphate groups of the i-th nucleotide, whereas the "ladder rungs" labelled Ns_i are the i-th nucleoside (i.e. the nucleobase plus the sugar). From these we define three groups of atoms for use in calculation of CVs, *viz.* $\mathcal{S}_1, \mathcal{S}_2$ and $\mathcal{S}_3$ (cf. Fig. 7.3, left). $\mathcal{S}_1$ is defined to be the phosphate

[3] Here the subscripted number denote the total number of base pairs in the oligomer, *not* the number of repetition of the motif as in the polymer chemistry sense.

[4] For an N-bp double-stranded DNA there are $2(N-1)$ backbone phosphate groups, hence the overall charge is $-2(N-1)e$, where $e \approx 1.602 \times 10^{-19} C$ is the standard electron charge.

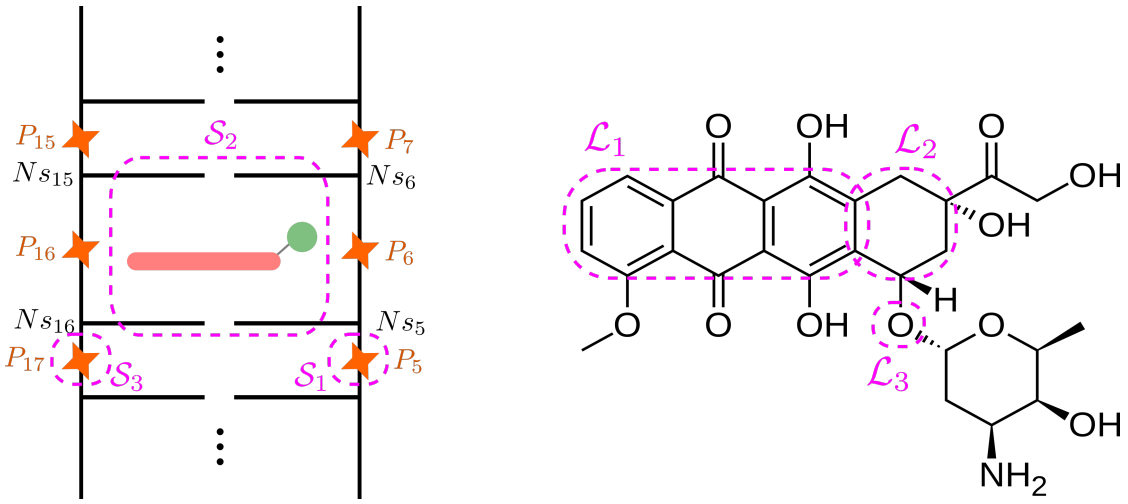

Figure 7.3: Schematic diagrams showing the definitions of the atom groups for CV calculations in Chapter 7, with daunomycin being used as a representative of the ligand.

group P_5 on the Watson strand, $\mathcal{S}_2$ to be *all* atoms from the base pairs making up the intercalation site, and lastly, $\mathcal{S}_3$ consists of the atoms in the phosphate group P_{17} on the Crick strand.

Similarly, three groups of atoms can be defined within the ligand, which we call $\mathcal{L}_1$, $\mathcal{L}_2$ and $\mathcal{L}_3$ (cf. Fig. 7.3, right). $\mathcal{L}_1$ contains all carbon atoms in the three front rings (i.e. the aromatic region of the intercalating part), $\mathcal{L}_2$ consists of the carbon atoms in the non-aromatic ring closest to the tail of the molecule, and $\mathcal{L}_3$ contains only the oxygen linkage between $\mathcal{L}_2$ and the major tail chain of the molecule.

With the atom groups defined, we can also define the CVs in terms of these groups. Following the method described in Woo *et al.* [307], we define the CVs as below:

- Θ = $\mathcal{L}_2$—$\mathcal{L}_1$—$\mathcal{S}_2$ angle
- Φ = $\mathcal{L}_3$—$\mathcal{L}_2$—$\mathcal{L}_1$—$\mathcal{S}_2$ dihedral angle
- Ψ = $\mathcal{L}_2$—$\mathcal{L}_1$—$\mathcal{S}_2$—$\mathcal{S}_3$ dihedral angle
- θ = $\mathcal{L}_2$—$\mathcal{S}_2$—$\mathcal{S}_3$ angle
- φ = $\mathcal{L}_2$—$\mathcal{S}_2$—$\mathcal{S}_3$—$\mathcal{S}_1$ dihedral angle
- r = $\mathcal{L}_1$—$\mathcal{S}_2$ distance

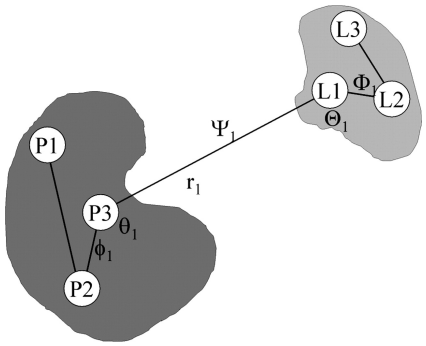

Figure 7.4: Schematic diagram showing the definitions of CVs in [307]. Image taken from [307].

It is noted in Gumbart *et al.* [122] that the choice of the $\mathcal{L}$ and $\mathcal{S}$ triplets, i.e. the definitions of the atom groups, and their combinations in forming the CVs do *not* matter. This is because, as can be seen from Eq. 2.46, the six bound-state CVs here, alongside the RMSD, contribute to the entirety of the overall bound-state PMF.

Thus, although individual PMFs of the CVs will change upon the alteration of definitions of the constituent groups, the overall PMF should be an invariant, provided that the constraints on PMFs are applied in a correct order.

7.2.4 Simulation protocols

7.2.4.1 Energy minimisation

For all the systems simulated in this Chapter, the energy minimisation process consisted of two steps, *viz.* linear (static) minimisation and simulated annealing (dynamic).

In the linear minimisation procedure, the systems were energy minimised using 1,000 steps of conjugate gradient method with velocity quenching to ensure fast convergence to nearest *local* minimum.

The procedure was followed by a 15-cycle simulated annealing. In each of the annealing cycles, the system was rapidly heated up from 0K to 300K in 3,000 steps (3 ps). This was done by coupling the system to a Langevin heat bath which had a stepwise increment in temperature of 10K per 100 steps. The system was then allowed to stay at 300K for 5,000 steps (5 ps) before cooling down. In the cooling stage, the system, being coupled to the heat bath, was slowly cooled back down to 0K in 15,000 steps (15 ps); the heat bath here had a stepwise decrement in temperature of 2K per 100 steps. After that, the system was allowed another 5,000 steps for atoms to stop moving. At the end of all 15 cycles, a further 50,000 steps was run with the heat bath being kept at 0K. Hence, the entire annealing procedure consisted of 470,000 steps (470 ps).

We assert that the annealing procedure is crucial in the work which will be presented in this Chapter. This is because not only do we need the systems to be in the state with the lowest possible energy [5] but we also need the values of the CVs at that particular state (cf. Sec. 2.7.1) as they represent the reference value in the eABF constraints which take the Hookean form.

7.2.4.2 MD simulation protocol

The MD simulation procedure was one where the PMFs of the CVs in a particular drug-DNA compound system were calculated. Basically, they all follow the same heating protocol, where systems were heated from 0K to the physiological temperature of 310K in 31,000 steps, by coupling to a Langevin heat bath which had a stepwise increment in temperature of 1K per 100 steps. The heating stage was followed by an NPT simulation where the PMFs were calculated on-the-fly; the pressure was maintained by means of a Langevin piston at 1 atm and the temperature was kept at 310K. The simulation time of this stage was determined by the CV being calculated — the full duration for r (the radial contribution) calculations was 1,000,000 steps (1 ns) whereas those for the calculations of other CVs were 500,000 steps (500 ps).

[5] With such complex a system, it is nearly impossible to tell whether the *ground state* has been attained without the use of techniques such as Monte Carlo simulations, which would involve huge computational cost.

Ten repetitions were performed for each of the CVs, so there were altogether $72 \times 7 \times 10 = 5040$ bound-state calculations as there are 72 systems which has seven bound-state CVs each. Moreover, for each of the drugs there is a CV accounting for the unbound state, hence adding another $3 \times 10 = 30$ simulation runs.

Lastly, in terms of the technicalities regarding the simulations performed for this Chapter, all simulations were run on the VIKING supercomputer at University of York, using two nodes (80 CPUs) per simulation run. Typical run-time for a bound-state minimisation run was about 15 minutes each, whereas that for a bound-state simulation run was about 35 minutes each. All simulations were performed with explicit coarse-grained solvent and periodic boundary conditions as described in the previous section.

7.2.4.3 Order of calculations and application of constraints

As elucidated in previous sections, the ABF method is applied by performing calculations in an arbitrary order of CVs one after another, adding harmonic constraints to the previous CVs.

In the work presented in this Chapter, we have followed the sequence $\xi_{\text{site}} \rightarrow \Theta \rightarrow \Phi \rightarrow \Psi \rightarrow \theta \rightarrow \varphi \rightarrow r$, i.e. the left-to-right direction in Layer 4 of the system hierarchy in Fig. 7.1 for the bound-states CVs. In terms of the example used above, the codes for the system's CVs in this order would be "doxACrBd" → "doxACrTh" → "doxACrPh" → ⋯ → "doxACrRr".

In order to "fix" the shape of the drugs whilst in the intercalated state, we have used a strong spring constant of $k_\xi = 25$ kcal mol^{-1} Å^{-2} for the CV ξ_{site}. On the other hand, since the values of the other CVs fluctuate rather wildly and should be allowed to do so, the spring constants for them were set to be only 1 kcal mol^{-1} Å^{-2}. For the unbound-state RMSD, in order for the calculations to be consistent with the bound-state RMSD case, the spring constant was set to be $k_\xi = 25$ kcal mol^{-1} Å^{-2} for ξ_{bulk}, i.e. the same as that for ξ_{site}.

7.2.5 Data manipulation and analysis

Since the data (PMFs) obtained from the simulations are expressed in general coordinates (the CVs' internal coordinates), they cannot be used directly and have to be transformed according to the methods described in Sec. 2.7.1 in order to be mapped back to the familiar cartesian coordinates.

As suggested by Gumbart *et al.* [122], these transformation integrals can be most easily done graphically, using softwares such as XMGRACE. However, whilst it is feasible for tiny systems with merely a few CVs to consider, as presented as the example case in the aforementioned tutorial, it is most definitely hopeless for systems in this work which have 24 subsystems and 10 replicas each, contributing to nearly 1,700 CVs per system on aggregate. Moreover, with such vast data set, manual evaluation of errors would also be extremely difficult.

Therefore, a suite of Python programs have been written to automate the calculation (transformation) of the ΔG for each of the CVs and their associated errors. Another program in the suite calculates the overall ΔG value for an entire drug-DNA complex system once all

the components from the subsystems' CVs have been calculated and consolidated. This program also performs statistical calculations for the systems, ranking the 24 subsystems according to their ΔG values and outputting the results graphically to allow easy visualisation of the binding mode preference of the drug.

7.3 Results and discussion

In this section, we present the results we have obtained from the work regarding the free energy change in DNA intercalation by antineoplastic drugs. But before diving into the discussion of the three drugs separately, we should first take a look at some general features which occur in the calculations.

Population distribution of CVs The distribution of CVs in a calculation is a good parameter in the determination of whether a calculation has converged. This is because the convergence of a (semi-)stochastic process is signified by the attainment of the stationary state of the probability density function, i.e. when the partial differential equation $\partial_t P = 0$ is fulfilled.

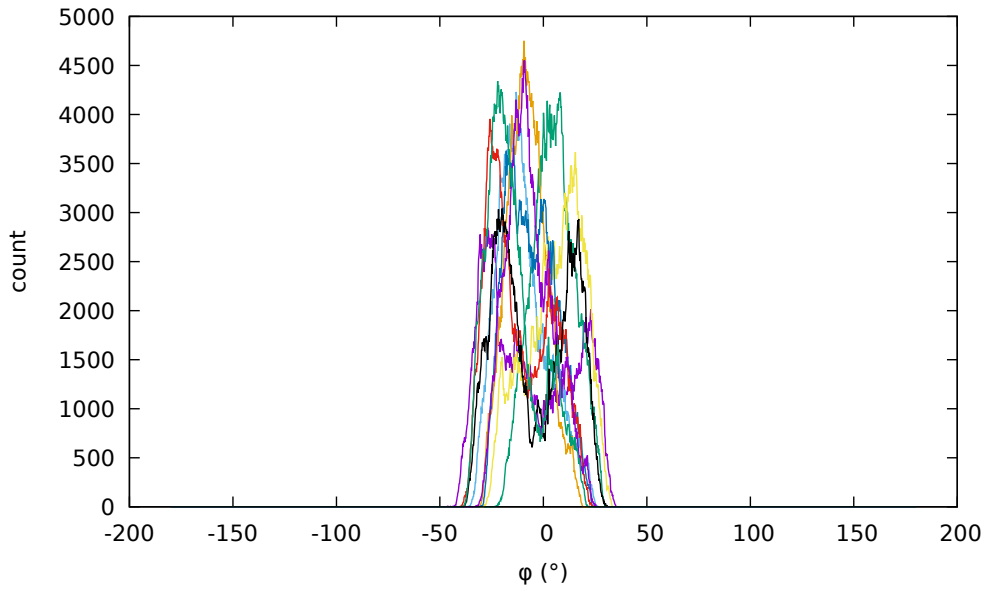

Figure 7.5: Sample distribution of 10 replicas of free energy calculation of CV (φ) under harmonic biasing force. Each line represents data from one replica.

In the case of a simulation with a harmonic biasing potential acting on a particular CV, we have proved in Sec. B.5, that the steady-state population distribution should be Gaussian. In Fig. 7.5 which shows the distribution of a CV from 10 replicas of a particular subsystem. We can see that whilst the overall distribution follows the above-said Gaussian shape, some individual replicas have either slightly skewed or even bimodal distributions.

The calculation of the distribution function in Sec. B.5 takes into account only two degrees of freedom for the system, *viz.* the CV of interest in the simulation and its spatial gradient,

and hence is an ideal solution. For biological macromolecular systems like those we are studying, there are far more degrees of freedom than in the ideal case and thus deviations must occur. Moreover, the ξ_0 values, i.e. the "ground-state" values for CVs only correspond to when the temperature tends to 0K. Once a non-zero temperature is introduced, the internal energy in the system may be raised to a level enough to cross the nearest free energy barrier, hence causing the polymodal distributions seen in some of the replicas in Fig. 7.5. This is also the primary reason for which we performed multiple replicas on each subsystem, each starting with a different random seed for the initial stochastic force term in the Langevin formalism, in order to better sample the phase space and produce a good overall approximation to the ideal distribution.

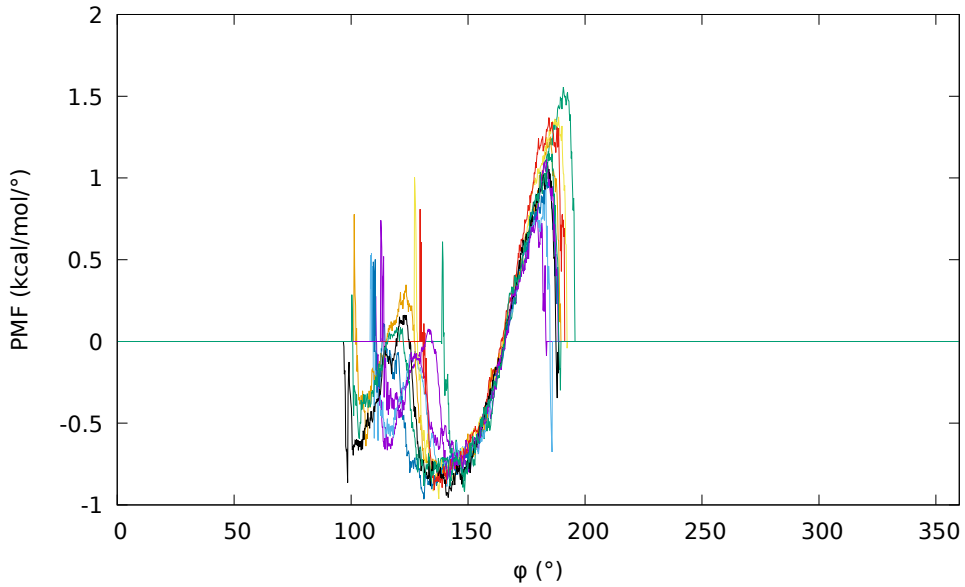

Figure 7.6: Sample gradients of PMF from 10 replicas of free energy calculation of CV under harmonic biasing force.

Gradients of PMFs The gradient of the calculated PMFs, as mentioned in Sec. 2.7.1, is the negative of the average value of the force experienced by the system along the CV being calculated. As a result, since we expect from the external harmonic potential a PMF curve of the quadratic form, centred at the mean value, ξ_0, of the CV, the gradient of the PMF should be linear and should be zero at ξ_0. Fig. 7.6 shows the gradient profiles from the simulation of the CV "idaCCrph", which has a mean value at $\varphi_0 \approx 165°$. Here we can see that gradient profiles around φ_0 are linear, which signifies that the PMF profiles are quadratic.

Ideally speaking, the quadraticity of the PMFs, and hence the linearity of their gradients, should extend to the entire domain of calculation, which in this example is the entirety of $\varphi \in [0°, 360°]$. However, as can be seen in Fig. 7.6, this is clearly not the case. Instead of a long linear line we observe a bend around $\varphi \approx 135°$. This is because with the 1 kcal mol^{-1} Å^{-2} force constant on the CV, the fictitious potential prevented the anomalous regions from being sufficiently sampled. Nevertheless, since the anomalies start to occur sufficiently far away from the region around φ_0, the integrands in the transforming integrals (Eq. 2.44)

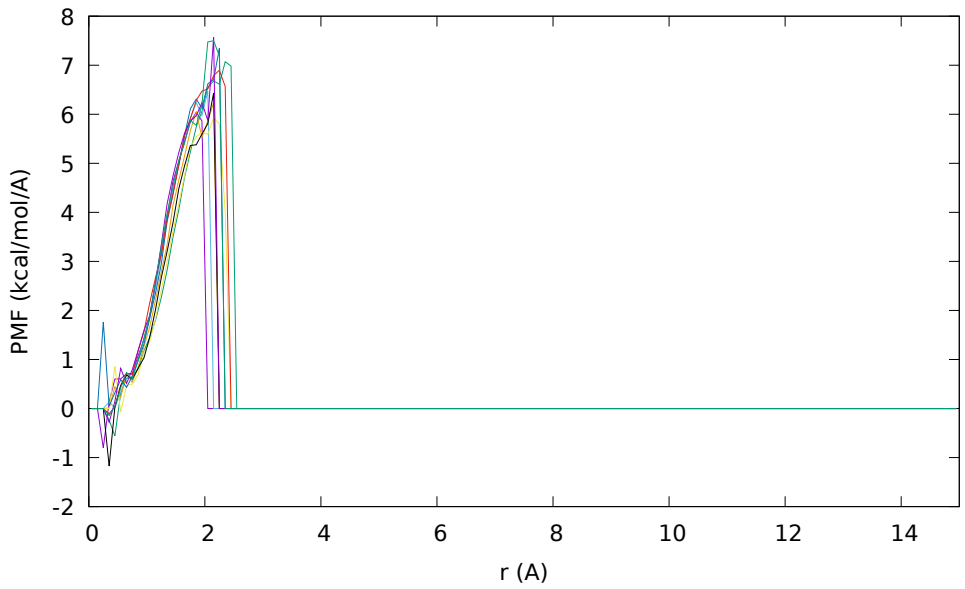

Figure 7.7: Sample gradients of PMF from 10 replicas of free energy calculation of radial CV.

become vanishingly small and so should not affect the results very much.

The gradients of PMFs from r, the radial CV, provides another piece of important information. Fig. 7.7 shows the gradient of the PMF of the CV "idaCCrRr". From Eq. 2.50, we see that the PMF $w(r)$ only appears in the I^* integral but not in the S^* integral. This means that the PMF calculated is one corresponding to the energy involved in the (un)binding action of the ligand by means of the removal of it from its binding site in the direction *perpendicularly* away from the helical axis. The peak value for the gradient is about 7 kcal mol^{-1} $Å^{-1}$ around $r = 2Å$. This means that as one attempts to pull the intercalator out perpendicularly, the intercalator would experience a force of this magnitude pushing it back into the site. With a conversion factor of about 69.5 pN for 1 kcal mol^{-1} $Å^{-1}$, the molecule experiences a force of 486.5 pN, which is very large for a molecule of about 0.5 kg mol^{-1}.

7.3.1 Doxorubicin

In this subsection, we present the results obtained for the doxorubicin-DNA complex systems. As mentioned in the methodology section above, we have performed nearly 2,000 simulations just for this system. The vast number of data sets precludes dealing with them one by one, but we can conveniently study them by grouping them in two different ways, namely vertically and horizontally.

Vertical grouping involves the linear combination of the different contributions from the respective CVs within the same system (e.g. AAu, cgr, etc.), which is justifiable as shown in Eq. 2.53. The associated errors propagate linearly as well, as the original quantities combine linearly. The resultant of such a grouping method is the total Gibbs free energy of interaction in a particular mode.

On the other hand, horizontal grouping involves the comparison among different systems and hence producing the ensemble average of the quantity of interest. Note that the ensemble average is no longer a simple summation over terms, so the error does not propagate linearly (cf. Sec. B.5).

	major	**minor**	
AA	-5.91 ± 0.42	-6.90 ± 0.44	**upright**
	-5.72 ± 0.40	-6.06 ± 0.18	**reverse**
AC	-7.00 ± 0.61	-8.93 ± 0.44	**upright**
	-7.49 ± 0.67	-4.56 ± 0.24	**reverse**
AG	-8.34 ± 0.24	-5.44 ± 0.44	**upright**
	-5.19 ± 0.43	-5.67 ± 0.27	**reverse**
AT	-5.70 ± 0.25	-4.19 ± 0.22	**upright**
	-6.28 ± 0.33	-5.25 ± 0.27	**reverse**
CC	-8.66 ± 0.28	-8.99 ± 0.30	**upright**
	-7.68 ± 0.33	-5.68 ± 0.25	**reverse**
CG	-6.01 ± 0.22	-5.49 ± 0.24	**upright**
	-7.04 ± 0.32	-5.20 ± 0.25	**reverse**

Table 7.1: Average free energy (in kcal mol^{-1}) across 10 replicas of all 24 subsystems of DOX:DNA complex. Data are divided into 3 separate dimensions, namely base step, groove of approach and orientation of ligand. For example, for the system "agr", look up the cell in **AG–minor–reverse**, which gives $\Delta G_{\text{agr}} = -5.67 \pm 0.27$ kcal mol^{-1}.

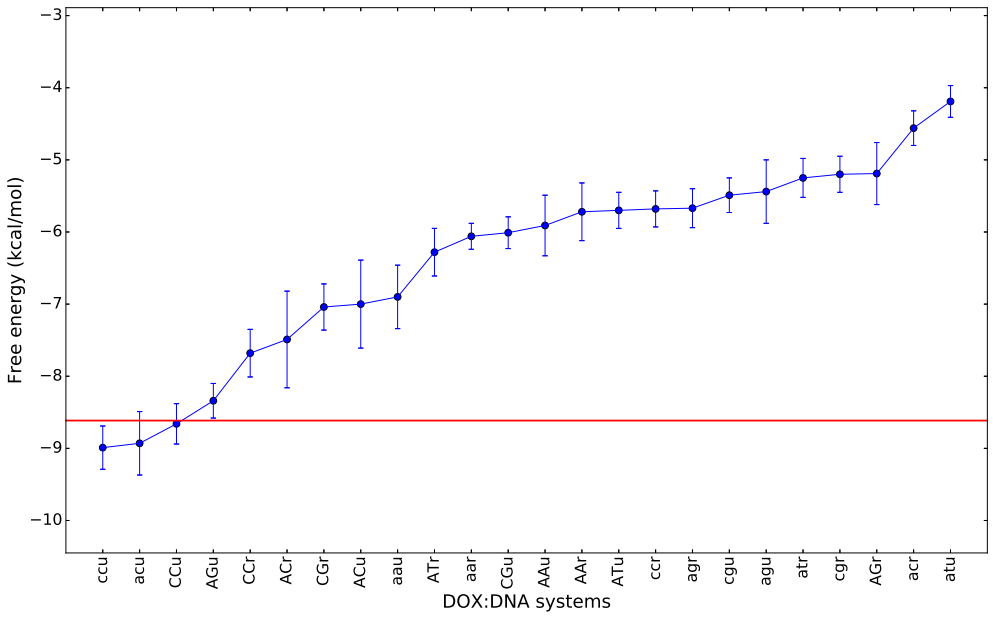

Figure 7.8: ΔG for the 24 subsystems in the DOX:DNA complex system. Red horizontal line denotes average value.

Table 7.1 shows the average intercalation free energy of each of the 24 systems (at a confidence level of 90%), and the sorted (in increasing order of energy) sequence is shown in Fig. 7.8. It is evident, that there is a preference of interaction sites, from the fact that some systems have more negative energy than others. The range of energies spans from −8.99 (ccu) to −4.19 (atu) kcal mol^{-1}, which converts to the equilibrium constants of $2.16 \times 10^6\,\text{M}^{-1}$

and $9.04 \times 10^2\,\mathrm{M}^{-1}$, at 310K, respectively, using Eq. 2.43. This means at equilibrium state the concentration of DOX:DNA complexes which intercalate through the ccu mode is more than 2300 times higher than that through the atu mode. Furthermore, from the same table, the ensemble average free energy (Eq. B.18) is calculated to be -8.61 ± 0.33 kcal mol^{-1} (shown as the thick red horizontal line in Fig. 7.8), which is in close proximity to previous results of -8.9 ± 0.3 kcal mol^{-1}, obtained by experiments involving the intercalation of DOX into calf thymus DNA [34, 52].

AA	9	11	13	14	$\Delta G_{AA} = -6.54 \pm 0.43$	$K_{\mathrm{AA}} = (4.04 \pm 2.82) \times 10^4$
AC	2	6	8	23	$\Delta G_{AC} = -8.73 \pm 0.41$	$K_{\mathrm{AC}} = (1.43 \pm 0.96) \times 10^6$
AG	4	17	19	22	$\Delta G_{AG} = -8.27 \pm 0.23$	$K_{\mathrm{AG}} = (6.68 \pm 2.51) \times 10^5$
AT	10	15	20	24	$\Delta G_{AT} = -5.98 \pm 0.33$	$K_{\mathrm{AT}} = (1.63 \pm 0.88) \times 10^4$
CC	1	3	5	16	$\Delta G_{CC} = -8.77 \pm 0.29$	$K_{\mathrm{CC}} = (1.54 \pm 0.72) \times 10^6$
CG	7	12	18	21	$\Delta G_{CG} = -6.73 \pm 0.33$	$K_{\mathrm{CG}} = (5.54 \pm 2.98) \times 10^4$

Table 7.2: Energy ranking of base steps (in increasing order of the rank disregarding the actual subsystem), followed by respective ΔG (in kcal mol^{-1}) and equilibrium constants (in M^{-1}).

Another approach one can take to investigate the energy values is, rather than to look at the system one by one, to read them in blocks. One of the blocking method can be by base steps. For example, for the A-A step, we have the 4 systems of AAu, AAr, aau and aar. Since from Fig. 7.8, the distribution of the energies across the systems is fairly uniform, which implies the distribution of the equilibrium constants should be roughly exponential, then it should be safe to assert that if a block has many members in the first half of the energy rank (i.e. having more negative energy), it should have a much higher dominance in steady state than a block with members mainly in the second half.

For instance, for steps like A-C (**2**-**6**-**8**-23) and C-C (**1**-**3**-**5**-16) which both have 3 subsystems in the top half of the energy ranking table (Table 7.2), their equilibrium constants (K_{AC} and K_{CC}) should be much higher than K_{AT}, whose subsystems are mostly in the lower half of the table; and this is exactly the case as shown in Table 7.2. Actually, K_{CC} is nearly 95 times higher than K_{AT}, showing that there is a strong sequence preference towards C-C steps.

In fact, because of the exponential dependence of the equilibrium constant on the interaction energy, only one or two subsystems which have very negative interaction energies are needed to drastically pull the overall equilibrium constant for the group up. For example, if we compare A-G and A-T, we may intuitively guess that A-G should have a rather similar equilibrium constant to, if not lower than, that of A-T. However, just because of the AGu subsystem which ranks as high as fourth energetically, which also dominates the reactions within the group, K_{AG} is more than 40 times higher than K_{AT}, making the A-G step the third most favourable step.

7.3.2 Daunomycin

Daunomycin is the second drug we studied in this Chapter. Since we follow the same procedure when analysing the data in this part, only the data will be presented with descriptions.

We start the analysis by considering Table 7.3 and Fig. 7.9. Apart from the fact that the red

	major	minor	
AA	-7.81 ± 0.27	-7.60 ± 0.26	**upright**
	-6.03 ± 0.26	-6.57 ± 0.41	**reverse**
AC	-6.99 ± 0.38	-6.02 ± 0.62	**upright**
	-6.61 ± 0.21	-5.60 ± 0.26	**reverse**
AG	-6.90 ± 0.20	-7.59 ± 0.27	**upright**
	-5.62 ± 0.46	-7.19 ± 0.26	**reverse**
AT	-6.82 ± 0.24	-7.61 ± 0.16	**upright**
	-6.60 ± 0.33	-6.56 ± 0.27	**reverse**
CC	-7.56 ± 0.19	-6.88 ± 0.32	**upright**
	-6.18 ± 0.32	-5.82 ± 0.27	**reverse**
CG	-6.12 ± 0.48	-6.79 ± 0.21	**upright**
	-5.80 ± 0.34	-6.27 ± 0.37	**reverse**

Table 7.3: Average free energy (in kcal mol^{-1}) across 10 replicas of all 24 subsystems of DAU:DNA complex.

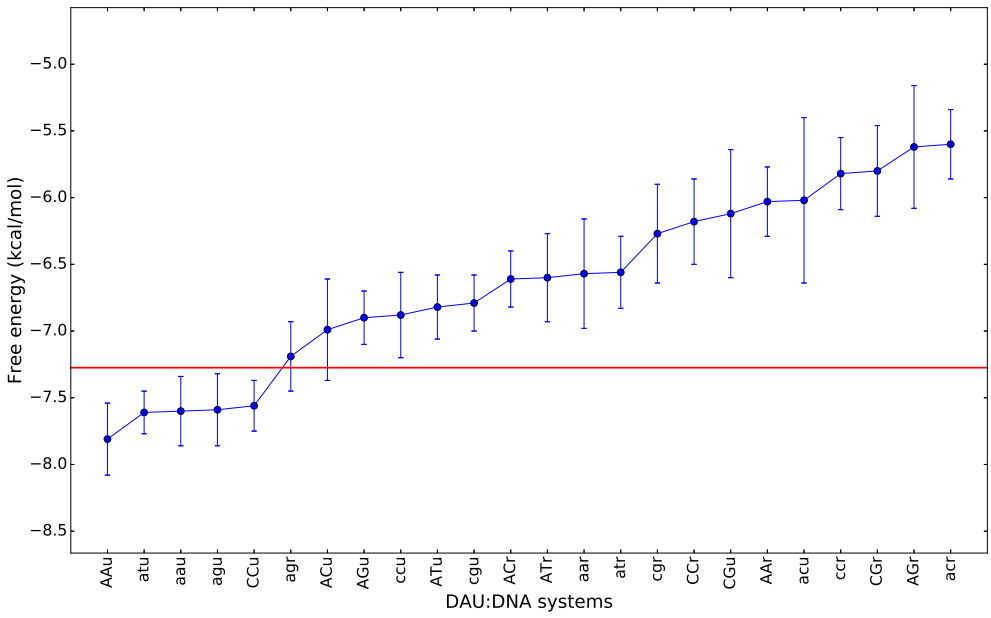

Figure 7.9: ΔG for the 24 subsystems in the DAU:DNA complex system. Red horizontal line denotes average value.

AA	1	3	14	19	$\Delta G_{AA} = -7.59 \pm 0.26$	$K_{\mathrm{AA}} = (2.25 \pm 0.94) \times 10^5$
AC	7	12	20	24	$\Delta G_{AC} = -6.69 \pm 0.34$	$K_{\mathrm{AC}} = (5.24 \pm 2.92) \times 10^4$
AG	4	6	8	23	$\Delta G_{AG} = -7.32 \pm 0.26$	$K_{\mathrm{AG}} = (1.45 \pm 0.61) \times 10^5$
AT	2	10	13	15	$\Delta G_{AT} = -7.25 \pm 0.16$	$K_{\mathrm{AT}} = (1.28 \pm 0.34) \times 10^5$
CC	5	9	17	21	$\Delta G_{CC} = -7.24 \pm 0.19$	$K_{\mathrm{CC}} = (1.27 \pm 0.40) \times 10^5$
CG	11	16	18	22	$\Delta G_{CG} = -6.46 \pm 0.25$	$K_{\mathrm{CG}} = (3.59 \pm 1.46) \times 10^4$

Table 7.4: Energy ranking of base steps (in increasing order of the rank disregarding the actual subsystem), followed by respective ΔG (in kcal mol^{-1}) and equilibrium constants (in M^{-1}).

line from Fig. 7.9 shows that the ensemble average of ΔG for the DAU:DNA intercalation is (-7.27 ± 0.23) kcal mol^{-1} , which is more than 1 kcal mol^{-1} higher than that for DOX:DNA,

we also notice the range of the energies across the 24 subsystems. Previously, we saw that the averages of ΔG for DOX:DNA subsystems covered a wide range of values, with a span of nearly 5 kcal mol^{-1} (cf. Fig. 7.8). In the case of DAU:DNA, the span of ΔG values dropped to just 2.21 kcal mol^{-1}. Moreover, whilst there are only three DOX:DNA subsystems which have their ΔG values below the red line, there are five in the case of DAU:DNA (or seven, taking into account the error bar at 90% confidence level). These observations are good evidence that, albeit having extremely similar chemical structures, daunomycin and doxorubicin exhibit very different physical interactions with DNA.

Furthermore, this evidence can be confirmed using Table 7.4, which consolidates all the individual subsystems into groups by their base steps. We can see rather clearly that the ranks are quite evenly distributed across the six base steps; none of them have three or more subsystems in the lower quartile (cf. the C-C step in DOX:DNA). This is why the equilibrium constants are also very close together. In fact, the most preferred step of A-A has a K_{eq} value about only six times higher than that of the least preferred step C-G. This further shows the non-specificity of daunomycin's intercalation.

Last but not least, we note that the overall ΔG value, (-7.27 ± 0.23)kcal mol^{-1}, which we have obtained through this work is, again, in close proximity to previously reported experimental results of (-7.9 ± 0.3)kcal mol^{-1} [35].

7.3.3 Idarubicin

Having performed computational calculations for the two drugs, *viz.* daunomycin and doxorubicin, which have been widely studied in the past decades experimentally, and produced theoretical results in close proximity to these experiments, we claim that the method we have used is suitable for use in the determination of the free energy changes, and thus the associated equilibrium constants, of intercalation-type binding actions. With this, we have further carried out calculations of these quantities for the intercalation of idarubicin for which there are no experimental results so far.

	major	**minor**	
AA	-6.24 ± 0.26	-6.89 ± 0.38	**upright**
	-4.81 ± 0.44	-7.46 ± 0.32	**reverse**
AC	-7.13 ± 0.32	-6.28 ± 0.41	**upright**
	-6.33 ± 0.35	-6.99 ± 0.36	**reverse**
AG	-6.83 ± 0.31	-8.31 ± 0.18	**upright**
	-5.39 ± 0.32	-5.17 ± 0.47	**reverse**
AT	-6.66 ± 0.28	-7.09 ± 0.32	**upright**
	-6.14 ± 0.50	-6.14 ± 0.24	**reverse**
CC	-8.33 ± 0.20	-5.98 ± 0.33	**upright**
	-7.30 ± 0.27	-6.10 ± 0.34	**reverse**
CG	-6.33 ± 0.46	-5.04 ± 0.55	**upright**
	-7.68 ± 0.27	-4.82 ± 0.61	**reverse**

Table 7.5: Average free energy (in kcal mol^{-1}) across 10 replicas of all 24 subsystems of IDA:DNA complex.

As before, Table 7.5 shows the free energy change in all 24 subsystems of IDA:DNA, Fig. 7.10

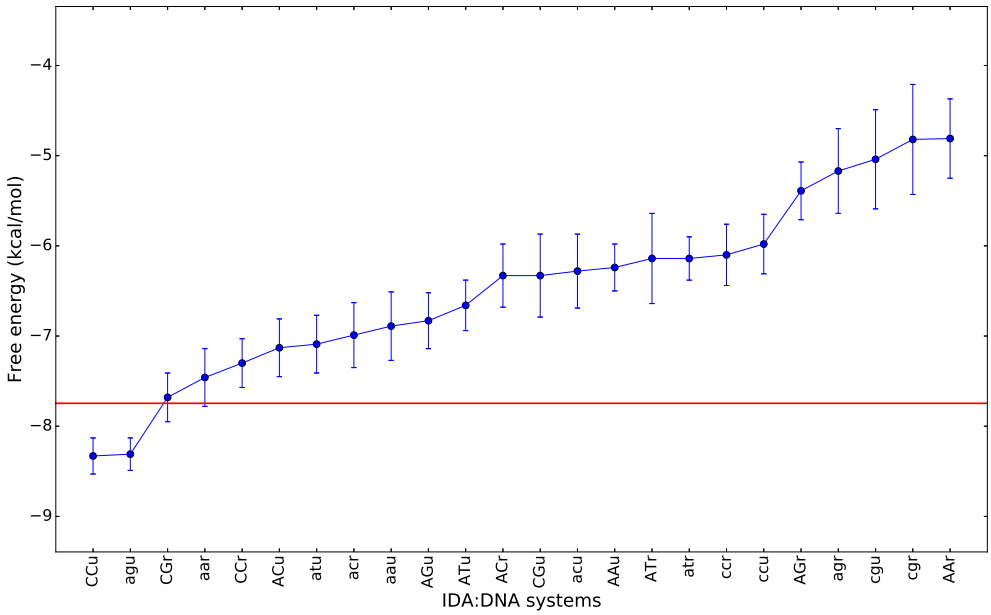

Figure 7.10: ΔG for the 24 subsystems in the IDA:DNA complex system. Red horizontal line denotes average value.

AA	4	9	15	24	$\Delta G_{AA} = -7.18 \pm 0.33$	$K_{\mathrm{AA}} = (1.15 \pm 0.62) \times 10^5$
AC	6	8	12	14	$\Delta G_{AC} = -6.90 \pm 0.34$	$K_{\mathrm{AC}} = (7.29 \pm 3.97) \times 10^4$
AG	2	10	20	21	$\Delta G_{AG} = -8.15 \pm 0.16$	$K_{\mathrm{AG}} = (5.55 \pm 1.44) \times 10^5$
AT	7	11	16	17	$\Delta G_{AT} = -6.77 \pm 0.31$	$K_{\mathrm{AT}} = (5.90 \pm 2.97) \times 10^4$
CC	1	5	18	19	$\Delta G_{CC} = -8.08 \pm 0.18$	$K_{\mathrm{CC}} = (5.00 \pm 1.48) \times 10^5$
CG	3	13	22	23	$\Delta G_{CG} = -7.49 \pm 0.23$	$K_{\mathrm{CG}} = (1.90 \pm 0.71) \times 10^5$

Table 7.6: Energy ranking of base steps (in increasing order of the rank disregarding the actual subsystem), followed by respective ΔG (in kcal mol^{-1}) and equilibrium constants (in M^{-1}).

shows the graphical representation of the above-said data, and Table 7.6 consolidates them in to groups by their original base steps.

The free energy profile appearing in Fig. 7.10 bears resemblance to those of daunomycin and doxorubicin in different aspects. Firstly, the span of the ΔG values is about 3.5 kcal mol^{-1} which is rather narrow — this is much like the case of daunomycin. Secondly, however, there are only three (arguably four) subsystems with energies below the average value — this resembles the case of doxorubicin instead.

Moreover, if we inspect Table 7.6 closely, we find that the distribution of the energies of the subsystems is the most even among the three drug-DNA complex systems — nearly all groups (base steps) have two of their four subsystems in the lower half. From the same table, we observe that the highest equilibrium constant (A-G step) is only about 9.4 times the value of the lowest one (A-T step). Using the same logic as in the case of daunomycin, we assert that the selectivity of base sequence by idarubicin is low, relative to doxorubicin. Furthermore, because of the similarity in the ΔG values of the two systems with the most negative ΔG (*viz.* A-G and C-C), their K_{eq} values are very close to each other as well.

Lastly, we report that through our calculations, the ensemble average of the interaction free energy for the IDA:DNA systems is determined to be (-7.75 ± 0.17) kcal mol^{-1}, which corresponds to the mean equilibrium constant of $(2.89 \pm 0.81) \times 10^5$ M^{-1}.

7.4 Summary

In this Chapter, we studied the energetics of the intercalation of three drugs, *viz.* daunomycin, doxorubicin and idarubicin, into DNA. In particular, we used the ABF and eABF methods to calculate the free energy changes regarding their intercalations computationally.

Firstly, concerning the free energies, we found that the ΔG values for the three drugs are (-8.61 ± 0.33) kcal mol^{-1}, (-7.27 ± 0.23) kcal mol^{-1} and (-7.75 ± 0.17) kcal mol^{-1} for the DOX:DNA, DAU:DNA and IDA:DNA systems respectively. These numbers are directly related to the equilibrium constants of the drugs' intercalative actions through Eq. 2.43, and are thus translated as $(1.18 \pm 0.64) \times 10^6$M^{-1}, $(1.34 \pm 0.50) \times 10^5$M^{-1} and $(2.89 \pm 0.81) \times 10^5$M^{-1} for the three systems respectively.

Secondly, we noted the difference in the sequence-specificity or selectivity of the drugs. For instance, doxorubicin showed very high preference towards intercalation in a C-C or A-C step with K_{eq} values tens of times higher than those of the other steps, whereas daunomycin and idarubicin did not show such trend, and their most preferred step sequences have K_{eq} values less than one order of magnitude those of the least preferred sites. Moreover, the so-called "more preferred" sites seem to vary a lot depending on the drug and such behaviour does not appear to have an obvious trend to be traced. For instance, doxorubicin prefers C-C or A-C steps, daunomycin has a slightly higher preference towards A-A step, whereas idarubicin prefers A-G or C-C step. We assert that, this lack of predictability of the sequence-specificity in the drugs' prefered intercalation site may be attributed to a variety of reasons, including the geometries, their side chains, or even the electronic structures, as in the case of the some other topoisomerase II poisons [103]. These factors may appear to be relatively minor especially in the case of the anthracyclines studied in this work, as the differences between any pair of the drugs come only from their stereochemistry but not their chemical compositions. The real cause of the amplification of effects on the interaction modes is yet to be discovered.

Thirdly, in terms of analysis, we deduced the formula for calculating the ensemble average of the interaction free energy. We noted that since the function is exponential, the averaging would be much weighted towards systems with more negative ΔG values. This in turns suggests that, the likelihood of having a clear preference of intercalation site increases with the spread of the free energies across the "microstates". As a result, as the "microstates" are grouped by their original base steps, it is not necessary that all microstates have very low ΔG for a base step to be preferred. Instead, it only takes one or two very dominant microstates in a base step group to make the step much preferred over other steps.

Lastly, and perhaps most importantly, through the computational study of doxorubicin and daunomycin on which experiments have been conducted extensively in the past four

decades, we found out that our theoretical values calculated in this work are both in close proximity to the experimental values reported in the past. This in turns has proved that the extended-system ABF algorithm works for these two particular systems and hence may also be useful in the determination of the free energy changes and equilibrium constants associated with intercalation-type interactions. Hence, this work may shed light on future researches in pharmacology as the calculation of free energy changes, which is the pièce-de-resistance in evaluating the effectiveness of a drug, could possibly be sped up.

Chapter 8

Conclusions

8.1 Summary of research

Molecular dynamics has been successfully used, as the backbone of this project, to investigate the different aspects of DNA-anthracycline complex systems, including the structural perturbation caused by intercalation, the effect on supercoiling behaviours of closed-circular DNA by non-intercalative binding of drugs, and the energetics associated with intercalative interactions. To aid the quantification of these effects, we have also employed different techniques and theories alongside classical MD methods, with exmaples including simulated annealing, X-ray diffraction, ribbon theory and extended-system adaptive biasing force method.

8.1.1 Structural studies of DNA-anthracycline complexes

In the first part of this research, we investigated the changes induced by intercalation of anthracyclines. Three difference sequences, *viz.* dA_{72}, $d(AC)_{72}$ and dC_{72} (representing 0%, 50% and 100% GC contents respectively) were simulated, starting from canonical A- and B-forms.

It was discovered, from the X-ray diffraction simulation, that just by the intercalation of a single molecule, the diffractions pattern would look vastly different from those of the respective canonical forms — the characteristic dumbbell- and X-shapes for the A- and B-forms. In fact, all of them had turned rather fuzzy with extra longitudinal fringes which are much less visible in the canonical cases which shows the breakage of general molecular symmetries. Moreover, all of the simulated XRD patterns with intercalations show that they all bear some resemblance to *both* canonical forms to different extents; some have a more prevalent dumbbell shape while the others have more visible X-shapes. However, using the modified Van Hove R-factor we devised, we showed that the B-form seems to have higher resilience to structural changes than A-form, which is probably the reason why the B-like conformation is more dominant in biological systems in normal environments.

The second part of this research has been about the effects of non-intercalative binding interactions between DNA and anthracycline drugs. Seven anthracycline drugs were place near the grooves of a pre-twisted ($Tw = -2$) 160b closed-circular DNA. The systems were

simulated using an all-atom/coarse-grained hybrid model for 100ns. An analysis program was written for the calculation of the time-evolution of the writhing number (ΔWr) of the DNA.

First of all, we found out that by altering the salinity (and hence the dielectric constant) of the simulation environment, the rate of supercoiling of the DNA changes accordingly. To be precise, the rate of supercoiling increases monotonically with the salinity. However, this increment is nonlinear due to the nonlinear nature of the dielectric constant as a function of the salinity — the linearity breaks as salinity gets higher than about 1.5M. On the other hand, we have shown that this monotonicity of the supercoiling rate with respect to the solvent salinity only holds for systems with bare DNA (hence, without external molecules). In the case with anthracyclines, we could interpolate the data and find that for the vast majority of them, the initial rates of supercoiling appear as a cusp with a peak near the physiological salinity of 153.8mM.

Secondly, we discovered that the aromatic components (or the chromophores, for they are responsible for the fluorescence properties of the molecules) of the drugs were capable of forming multiplexes among themselves by forming a $\pi - \pi^*$ stacking between the chromophores. Moreover, once the multiplexes are formed, they can bridge strongly across two adjacent segments of the cccDNA, thus increasing the rate of supercoiling and enhancing the structural stability of the supercoiled DNA.

8.1.2 Free energy calculation of DNA-anthracycline complexes

The third part of this project includes the evaluation of free energy changes associated with the intercalation processes of anthracycline drugs. In particular, we have used the method of extended-system adaptive biasing force to perform the calculations.

Firstly, for all the three drugs being studied, we noticed that while all the angular components of the free energies have similar values to each other, the components responsible for the root-mean-square structural deviations of the intercalators are in general about twice the values of the angular components. This implies that once intercalated, the intercalator has more freedom to change its position inside the intercalation site than to bend or twist itself. Moreover, the radial components of the free energies have even larger magnitudes than the RMSD components: they can get up to six times the values for the angular components in some subsystems, making the radial component the most dominant contribution to the overall free energy change of the interaction.

Secondly, we noted the magnitude of the gradients of the potentials of mean force as the force being experienced by the intercalator (and obviously also the DNA due to Newton's third law of motion). We paid particular attention to those of the radial components, which could be as high as 7 kcal mol^{-1} $Å^{-1}$ or more, implying an extremely strong force of about half a nano-newton being acted on the relatively light intercalator molecule to draw it back to the intercalation site. We assert that such stability of the intercalator inside the interstitial site is because it forms *two* sets of $\pi - \pi^*$ stacking interactions with the base pairs directly above and below it. In order for the intercalator to successfully de-intercalate from the site, *both* sets of orbital stacking must be completely broken.

Thirdly, we computed the ensemble averages of the free energies of the intercalations of the anthracycline drugs. Our theoretical study showed that the overall free energy changes of intercalations of daunomycin and doxorubicin to be (-7.27 ± 0.23) and (-8.61 ± 0.33) kcal mol^{-1} respectively. We note that these numbers are in close proximity to previous experimental results; our error bars have overlapping regions with those of the experimental outcomes. With this, we performed similar calculations on idarubicin which is far less studied than the two drugs above, and predicted that the free energy change of intercalation of this drug is (-7.75 ± 0.17) kcal mol^{-1}.

Last but not least, we devised a method to decouple the base sequence-specific contribution of the free energies from the overall energies. Through this analysis we found that sequence-specificity of drug intercalation may be individual to the drug, and could be very different for two drugs even if they have high structural resemblance with each other. For instance, the three anthracyclines used in this work only differ slightly with each other in the compositions of their side chains, but they exhibit vastly different sequence-specificity and sequence-selectivity. We found out that, of the three drugs, only doxorubicin has a clear preference towards an A-A or A-C step with the equilibrium constants for these two modes nearly 100 times those of the other modes. On the other hand, the sequence-selectivity of daunomycin or idarubicin is much lower than that of doxorubicin: the most favourable combinations of base steps only have K_{eq} values a few times higher than those of the least favourable combinations.

8.2 Future work

Though this work paved the way for a quick method for the determination of the binding free energy changes, and hence binding likelihoods, of drugs on DNA, by using three anthracyclines as examples, it leaves a great potential for further usage of this method.

For instance, the drug prototypes we used here represent only the tip of an iceberg in the family of anthracyclines; similar treatments can be taken in the theoretical studies of the rest of the drugs in the same family. Moreover, since anthracyclines are topoisomerase inhibitors (mostly of Topo I or Topo IIa), these enzymes can be used instead of DNA to investigate the bonding potential of the drugs on different topoisomerases. Ultimately, because one of the proposed pathways of interactions between topoisomerases, inhibitors and DNA is the formation of a cleavable three-component complex, such a system could be simulated for a more holistic view of these interactions.

In terms of the study around anthracyclines with cccDNA, since one of the aims of this project was to investigate the effects of these drugs on the supercoiling behaviours of cccDNA and we have shown that they generally slow down the supercoiling process, a natural follow-up question would then be: How does that affect the interactions between the DNA and other biological molecules surrounding it, especially those which would bind to it (especially topoisomerases)? To answer this question, one may, again, perform similar simulations to the ones presented in this work but also with the protein present.

Another aspect of interactions one may study is the causality. From earlier discussions, we

know that anthracyclines can bind to topoisomerases whilst being a potent DNA intercalator itself. Then, a big question to be asked regarding this would be: Which of these come first — is it that the drug intercalates into the DNA first then forms a cleavable complex with the incoming protein, or does the drug bind to the enzyme first and the complex intercalates into the DNA using the chromophore of the drug as the intercalating component? To this end, we assert that accelerated MD could be an excellent method to use for the simulation, as it has been proven through this work to massively boost the likelihood of occurrence of binding-type rare events.

Appendix A

Time variation of DNA structural parameters

In this appendix, we present the time variation of a selection of DNA parameters for *all* simulations performed in Chapter 3. The data will be presented as contour plots, with the horizontal axis being the time axis, and the vertical axis the base pair number.

The parameters presented can be broadly divided into two groups, *viz.* groove parameters and base pair parameters. The groove parameters are the widths and depths of the major and minor grooves, and the base parameters are: helical rise, helical twist, buckle, roll, shift and slide.

For each system, the graphs will be arranged in the following order:-

Major Width	Major Depth
Minor Width	Minor Depth
Helical Rise	Helical Twist
Buckle	Roll
Shift	Slide

The base pair parameters can be verbally described as below:

- Helical rise and twist: translation and rotation of successive base pairs along and around the helical axis [165].
- Buckle: the angle between the bases within a pair which “buckles" with respect to the helical axis.
- Roll: The extent of non-parallelarity between two successive base pairs in the groove-ward directions
- Shift: The relative translation between two successive base pairs in the groove-ward directions
- Slide: The relative translation between two successive base pairs in the backbone-ward directions

Graphical representation of the above parameters can be found in [82] and rigid mathematical treatments adopted in CURVES+ and CANAL are explained and derived in [166,167,262].

A.1 d(A)$_{72}$, A-start (AA-A) series

A.1.1 Bare DNA (AA-A-bare)

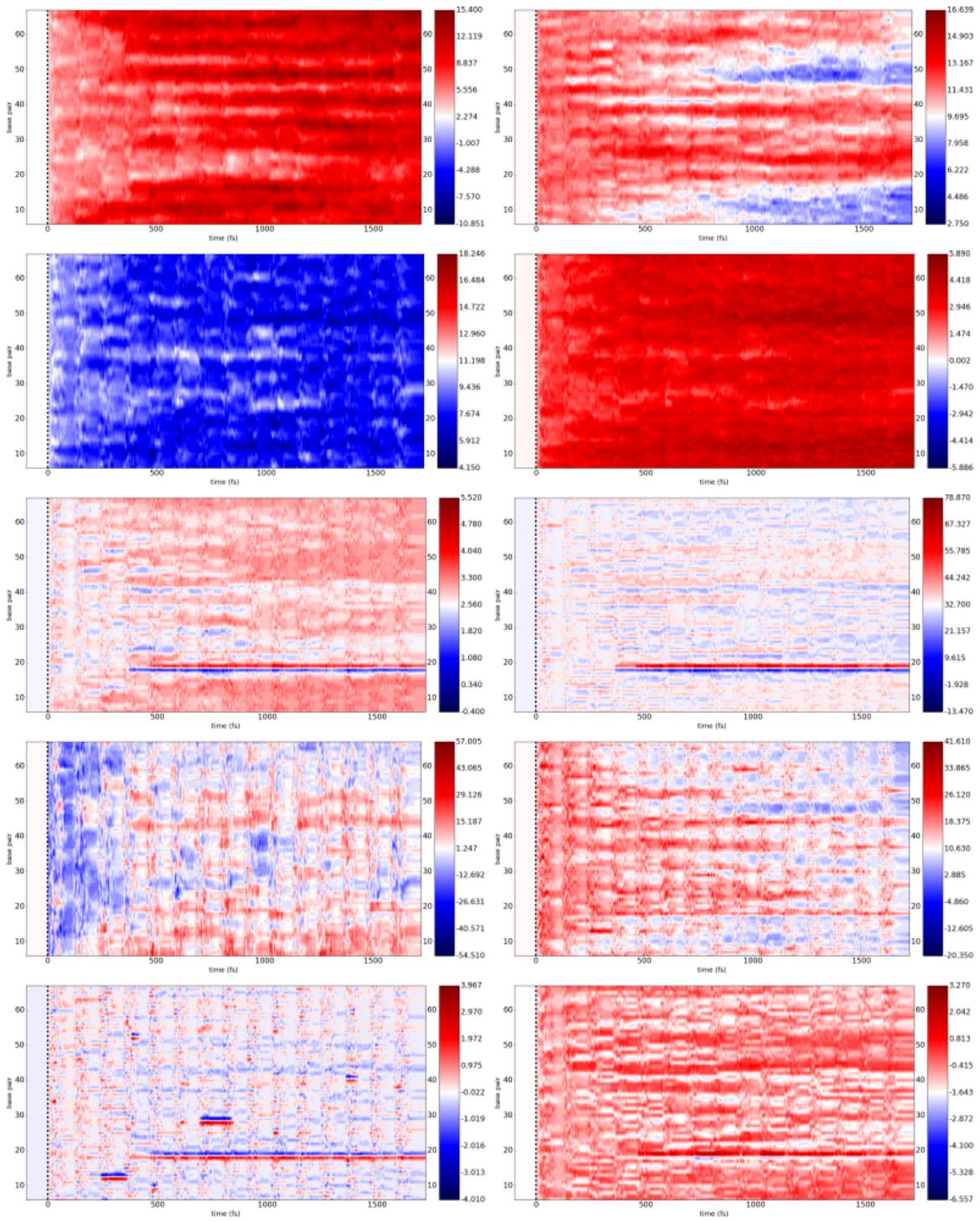

Figure A.1: Groove parameters for the AA-A-bare system. See first page of Appendix A for ordering of graphs.

A.1.2 Daunomycin (AA-A-dau)

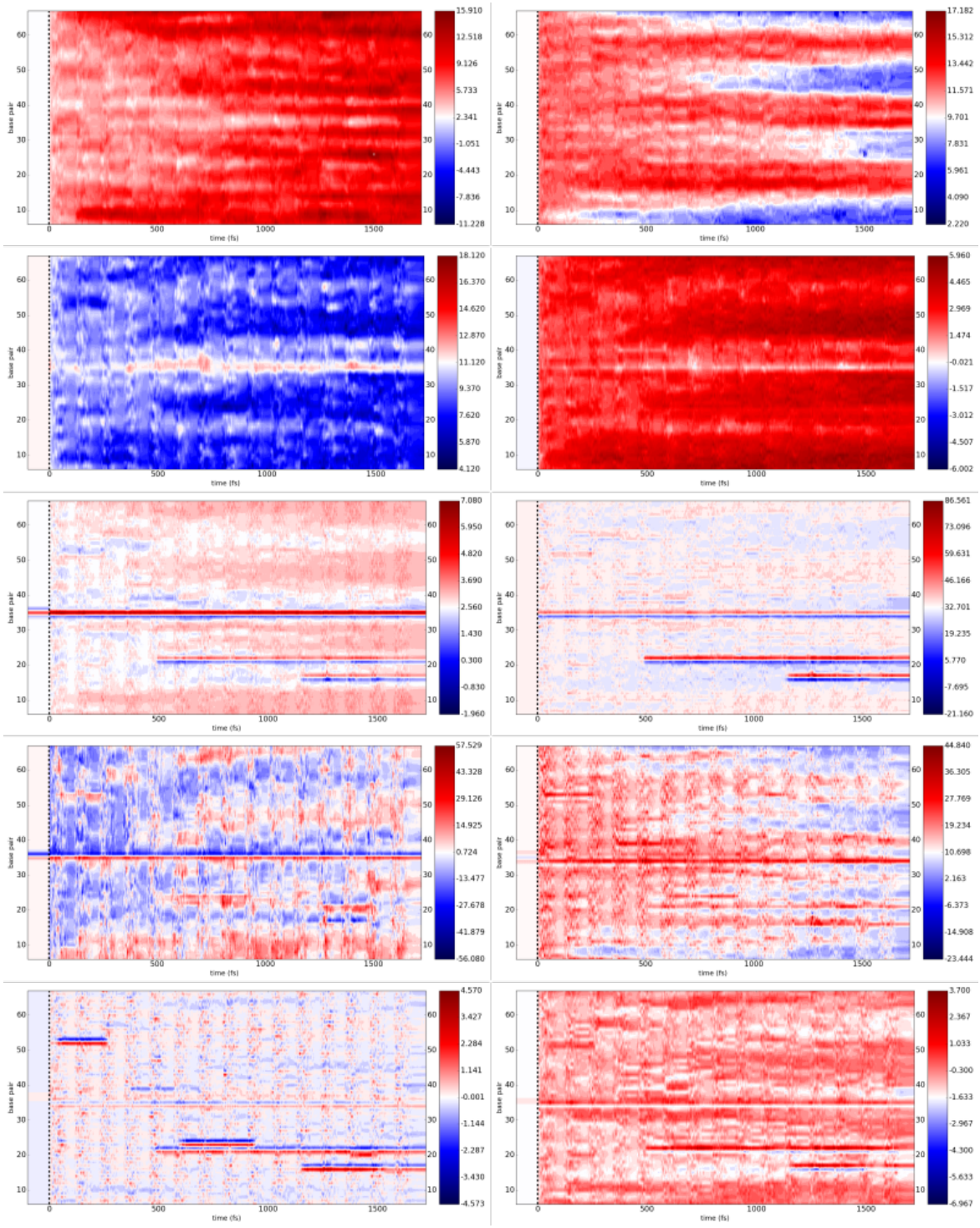

Figure A.2: Groove parameters for the AA-A-dau system. See first page of Appendix A for ordering of graphs.

A.1.3 Doxorubicin (AA-A-dox)

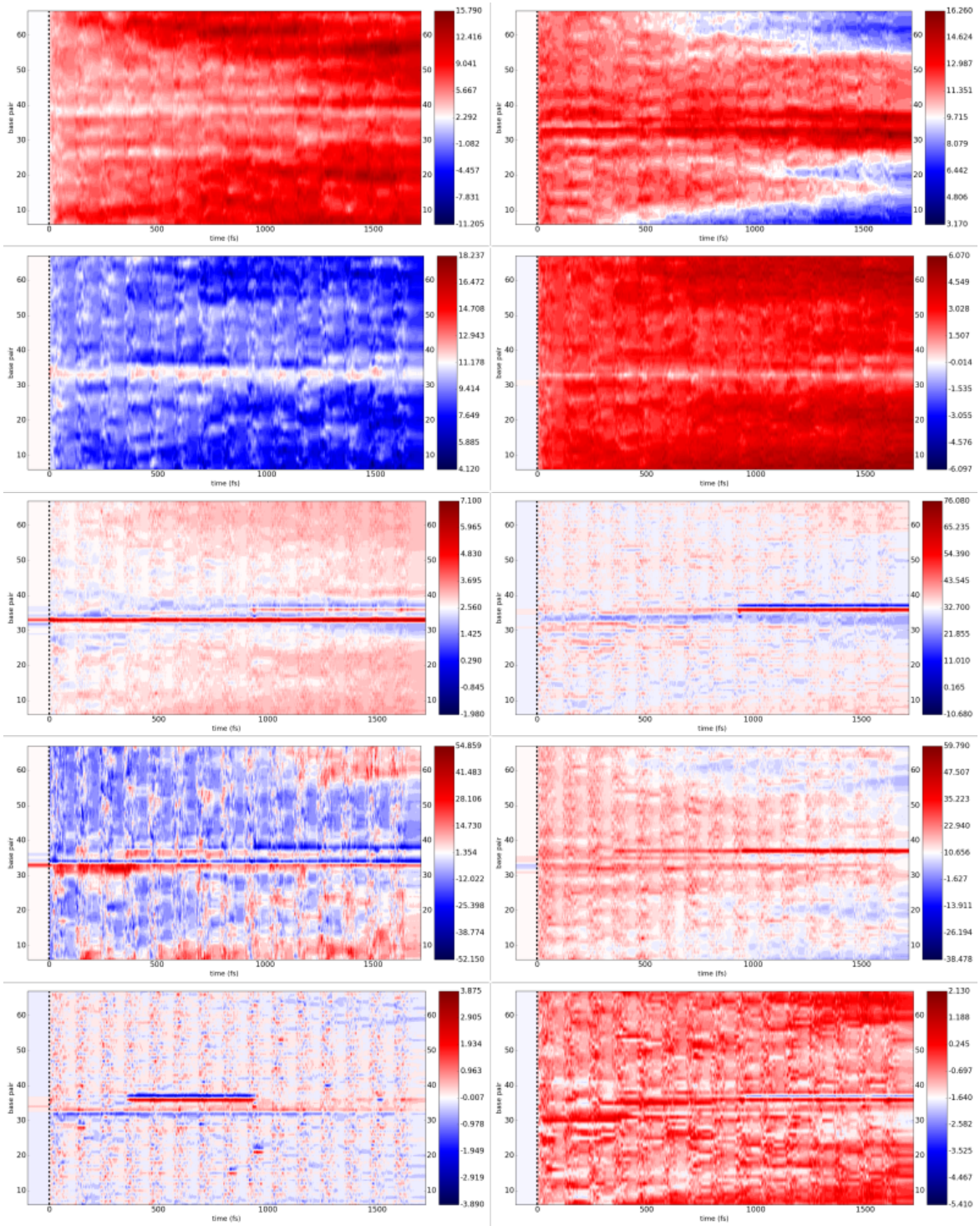

Figure A.3: Groove parameters for the AA-A-dox system. See first page of Appendix A for ordering of graphs.

A.1.4 Idarubicin (AA-A-ida)

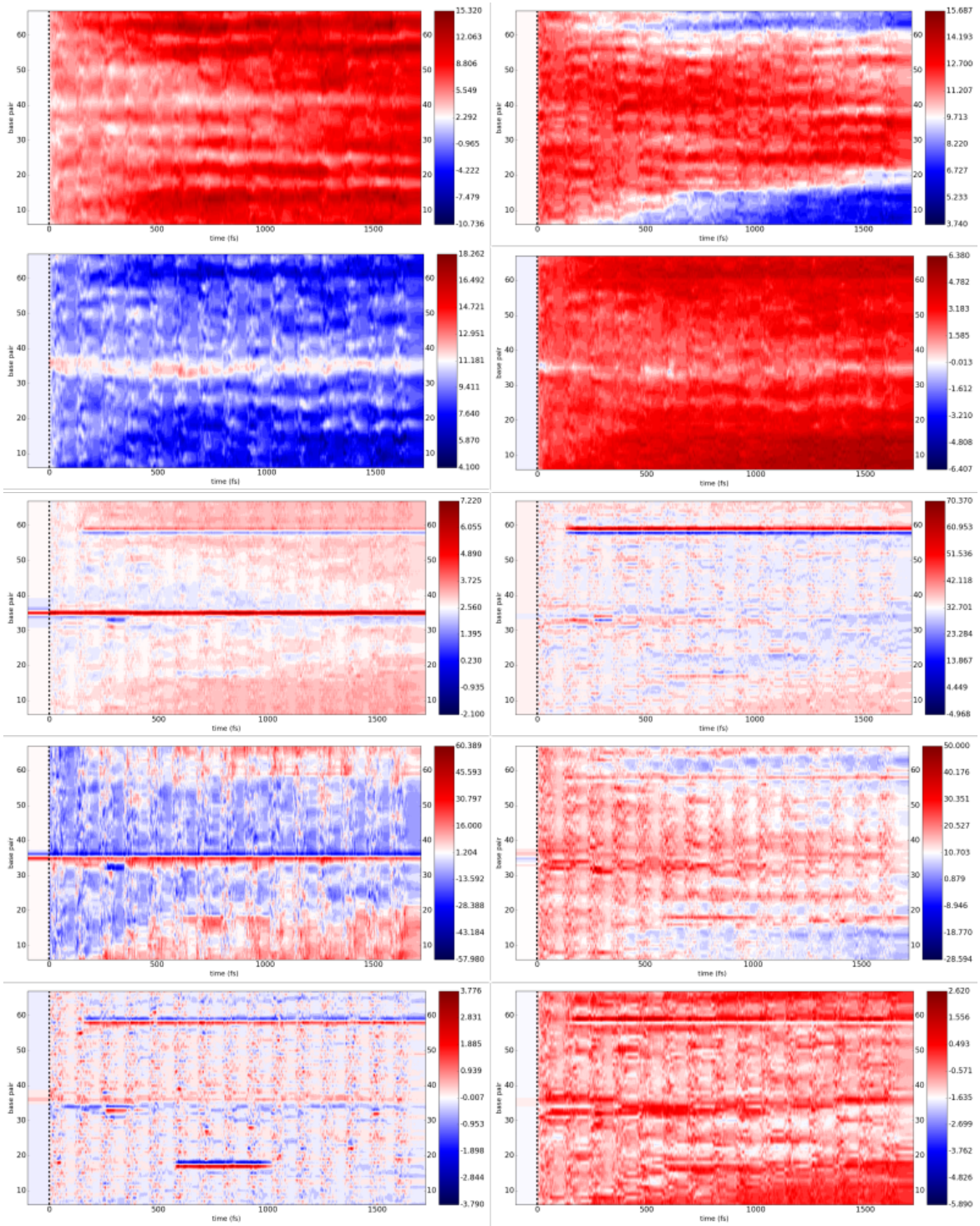

Figure A.4: Groove parameters for the AA-A-ida system. See first page of Appendix A for ordering of graphs.

A.2 $d(A)_{72}$, B-start (AA-B) series

A.2.1 Bare DNA (AA-B-bare)

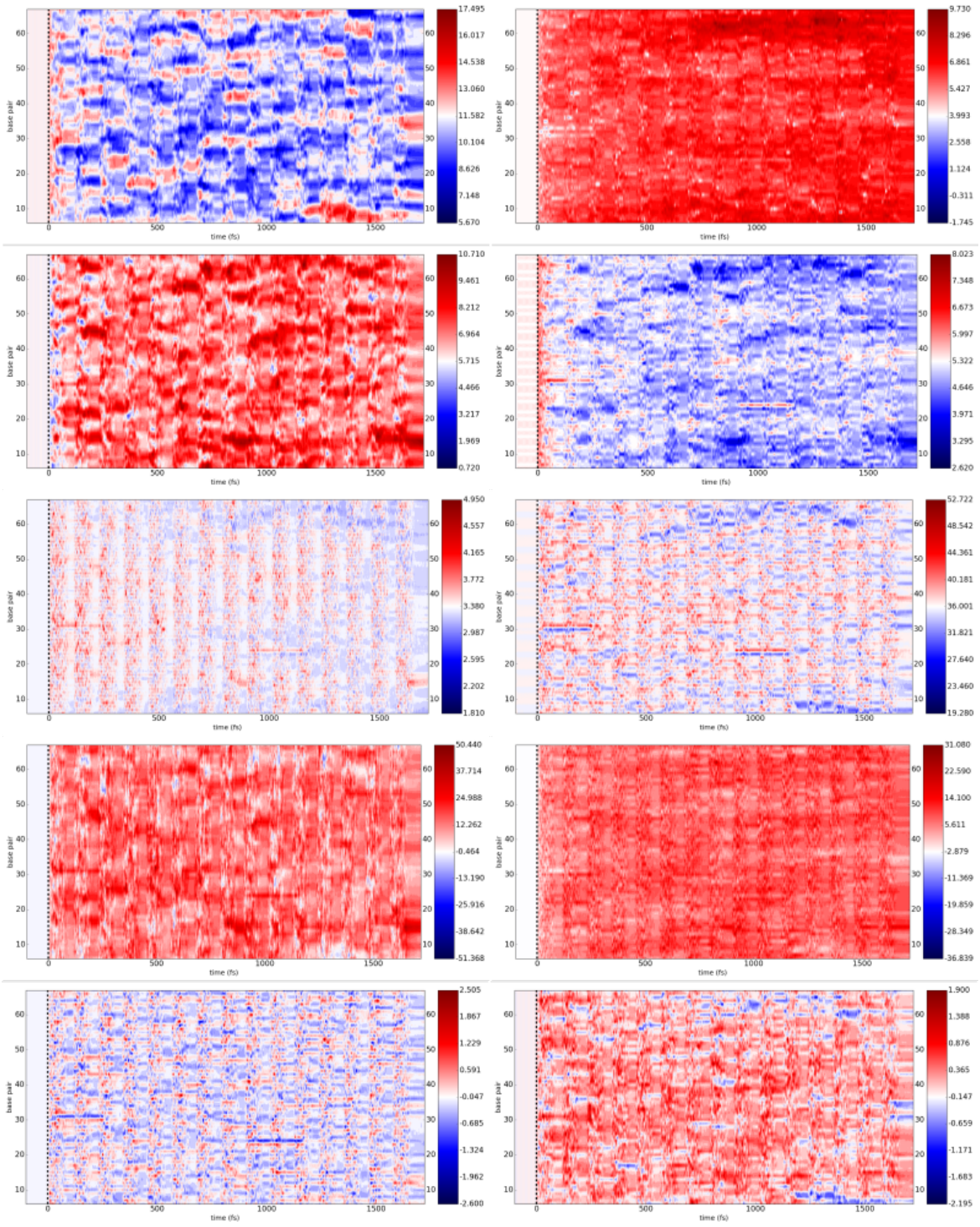

Figure A.5: Groove parameters for the AA-B-bare system. See first page of Appendix A for ordering of graphs.

A.2.2 Daunomycin (AA-B-dau)

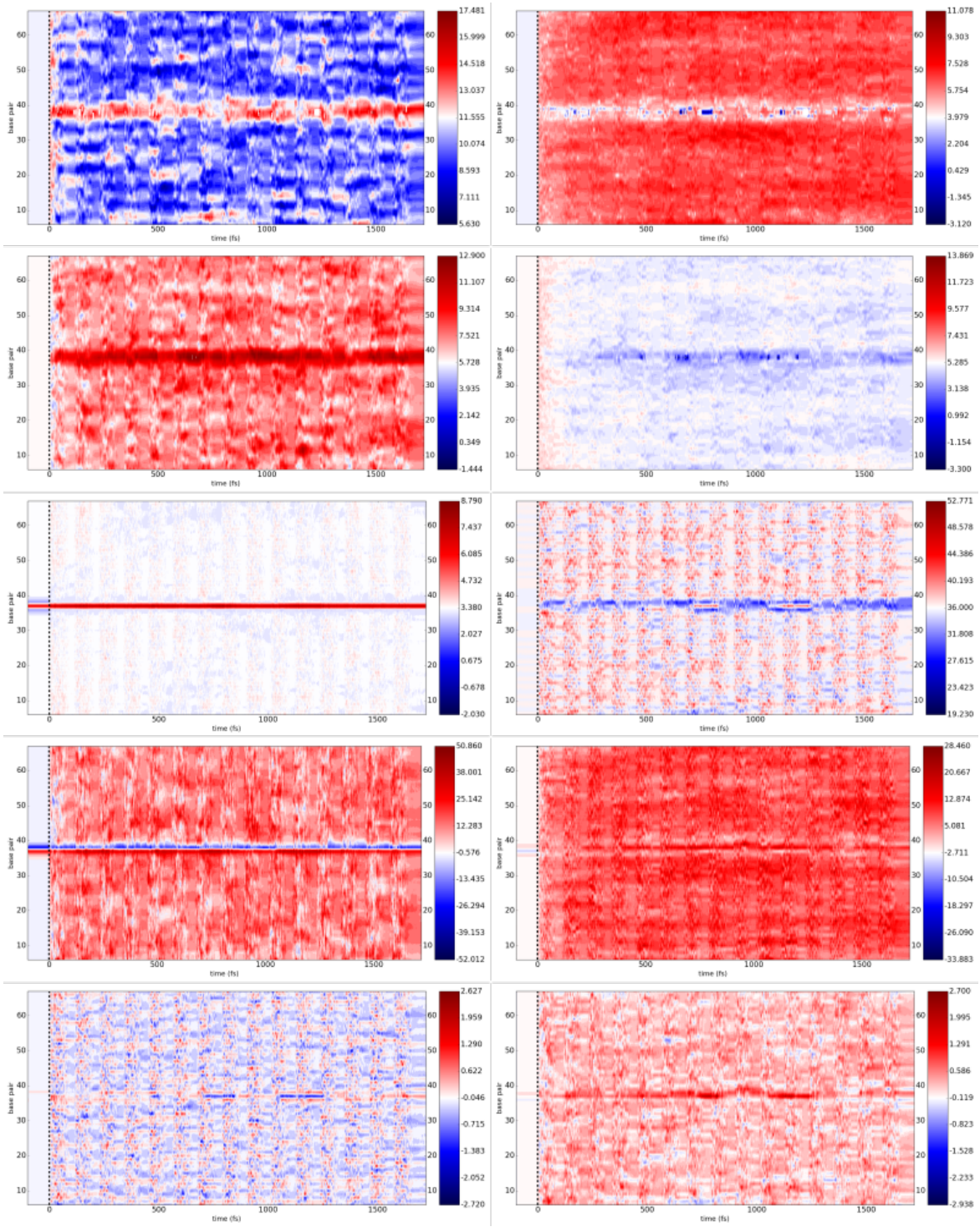

Figure A.6: Groove parameters for the AA-B-dau system. See first page of Appendix A for ordering of graphs.

A.2.3 Doxorubicin (AA-B-dox)

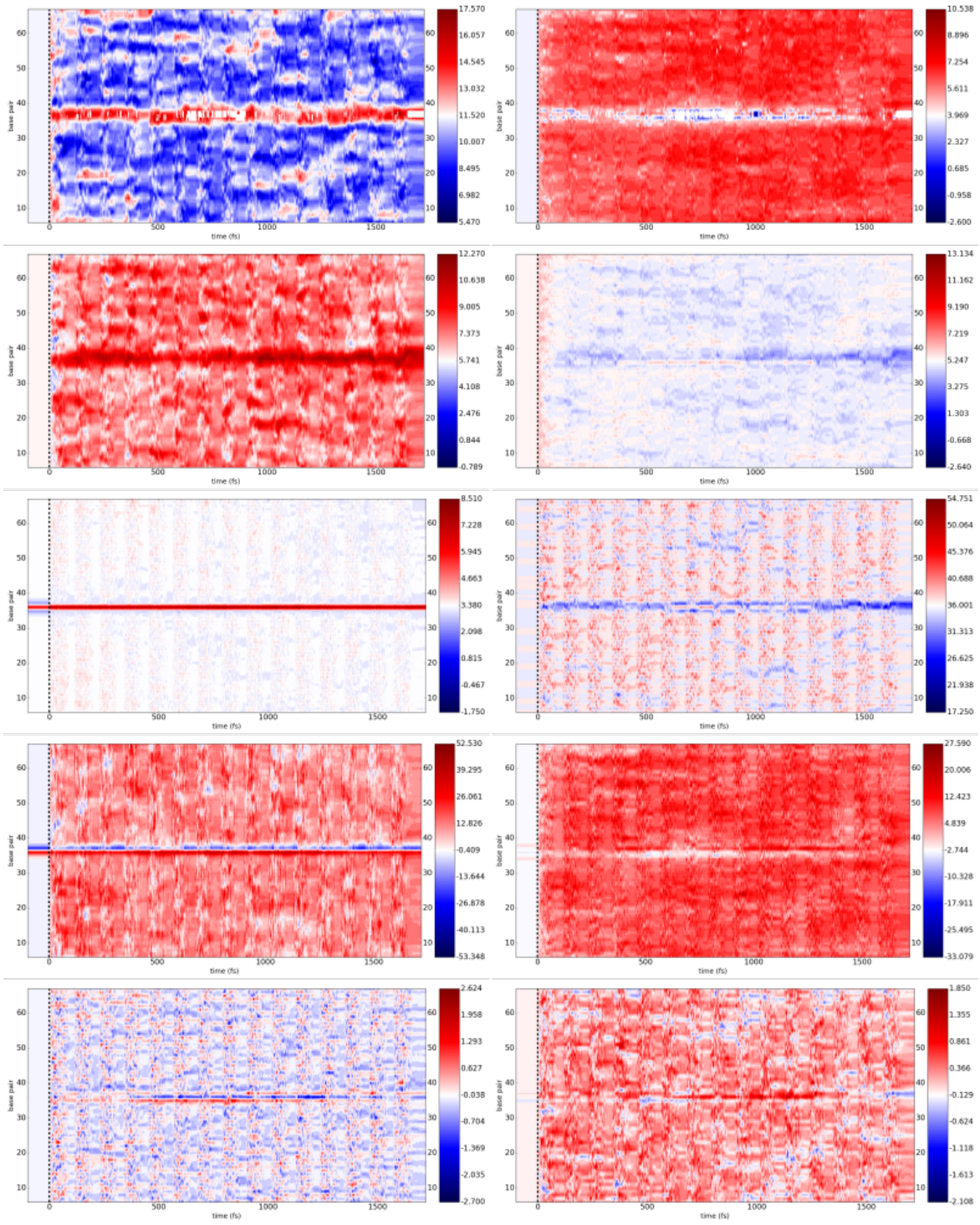

Figure A.7: Groove parameters for the AA-B-dox system. See first page of Appendix A for ordering of graphs.

A.2.4 Idarubicin (AA-B-ida)

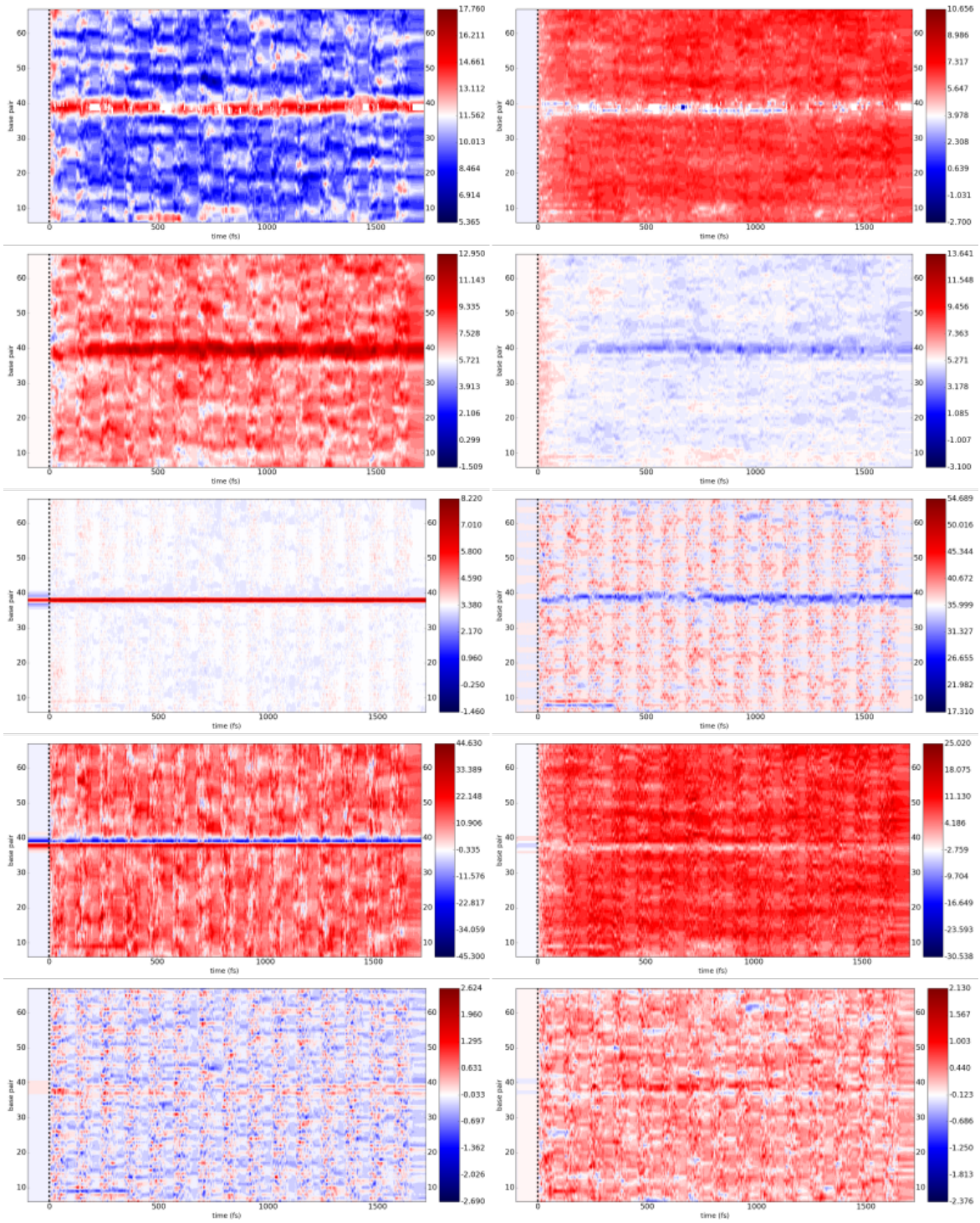

Figure A.8: Groove parameters for the AA-B-ida system. See first page of Appendix A for ordering of graphs.

A.3 d(AC)$_{72}$, A-start (AC-A) series

A.3.1 Bare DNA (AC-A-bare)

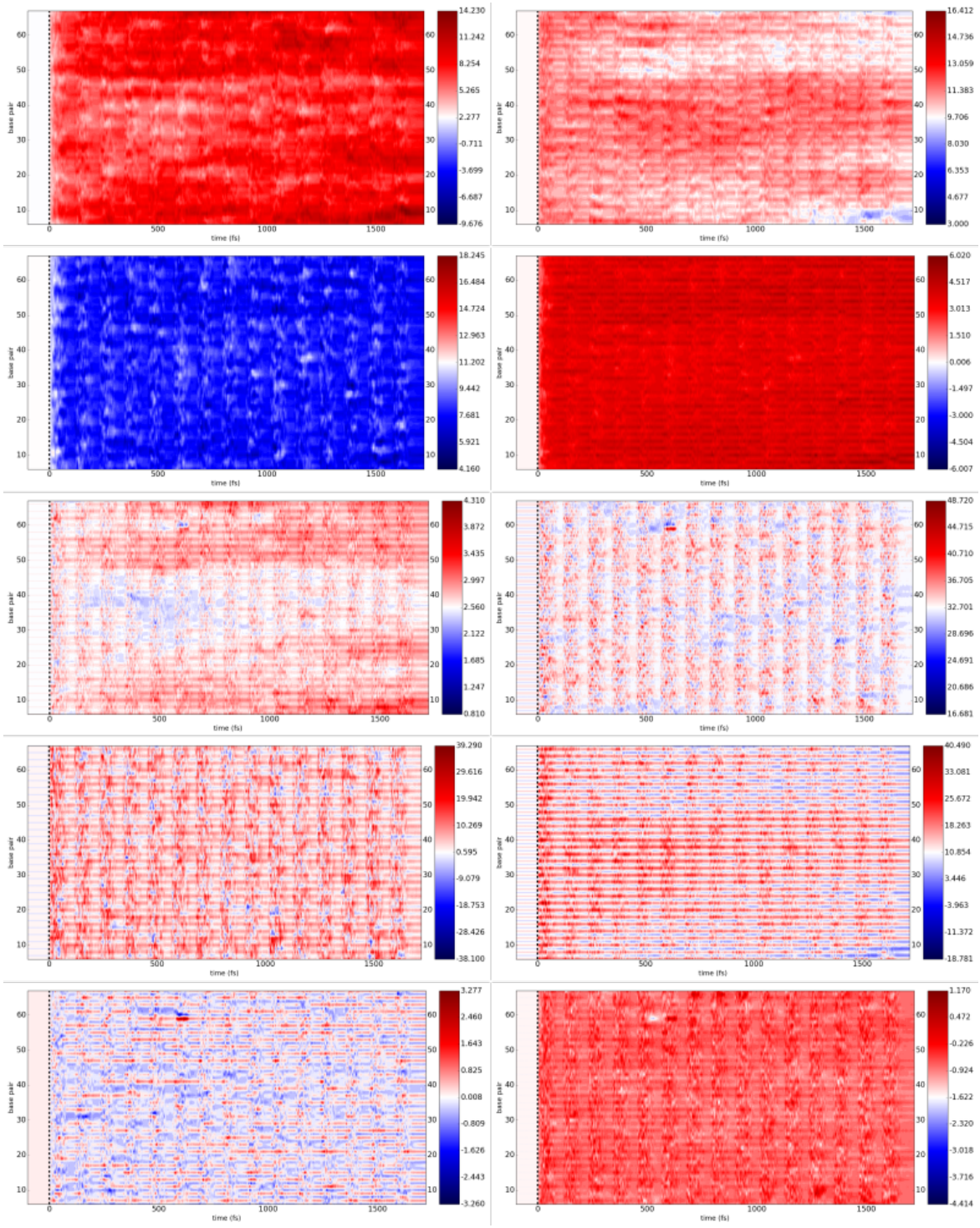

Figure A.9: Groove parameters for the AC-A-bare system. See first page of Appendix A for ordering of graphs.

A.3.2 Daunomycin (AC-A-dau)

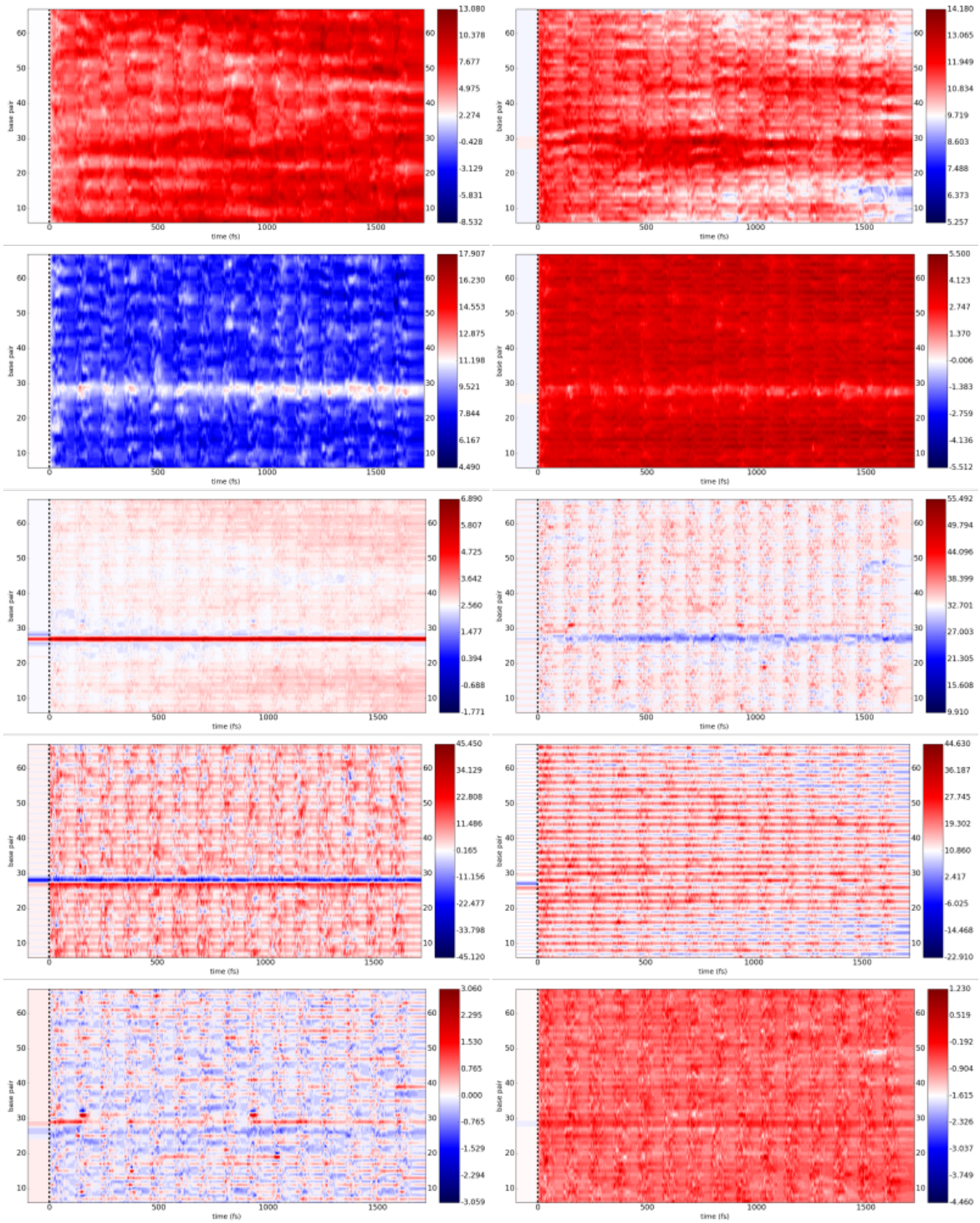

Figure A.10: Groove parameters for the AC-A-dau system. See first page of Appendix A for ordering of graphs.

A.3.3 Doxorubicin (AC-A-dox)

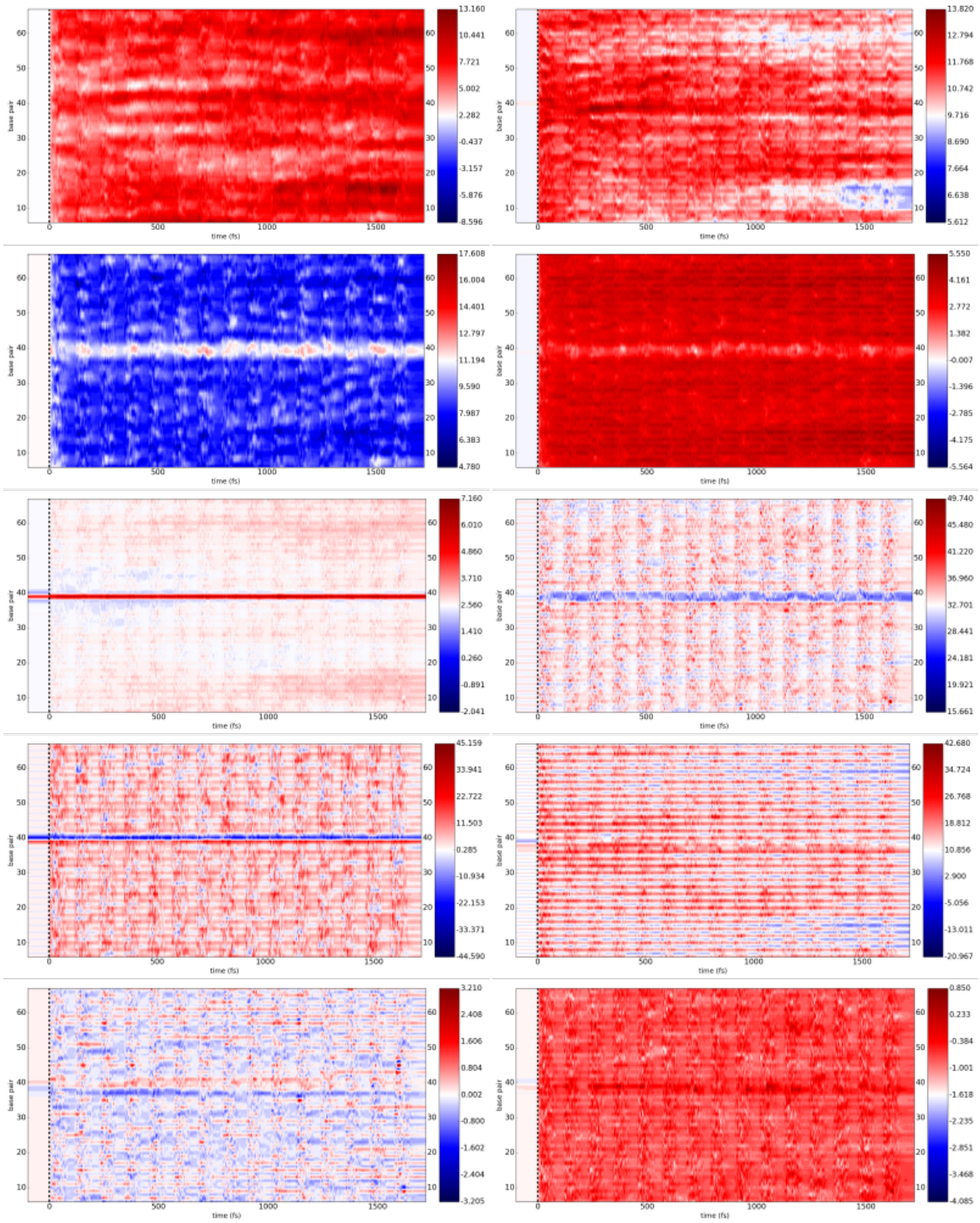

Figure A.11: Groove parameters for the AC-A-dox system. See first page of Appendix A for ordering of graphs.

A.3.4 Idarubicin (AC-A-ida)

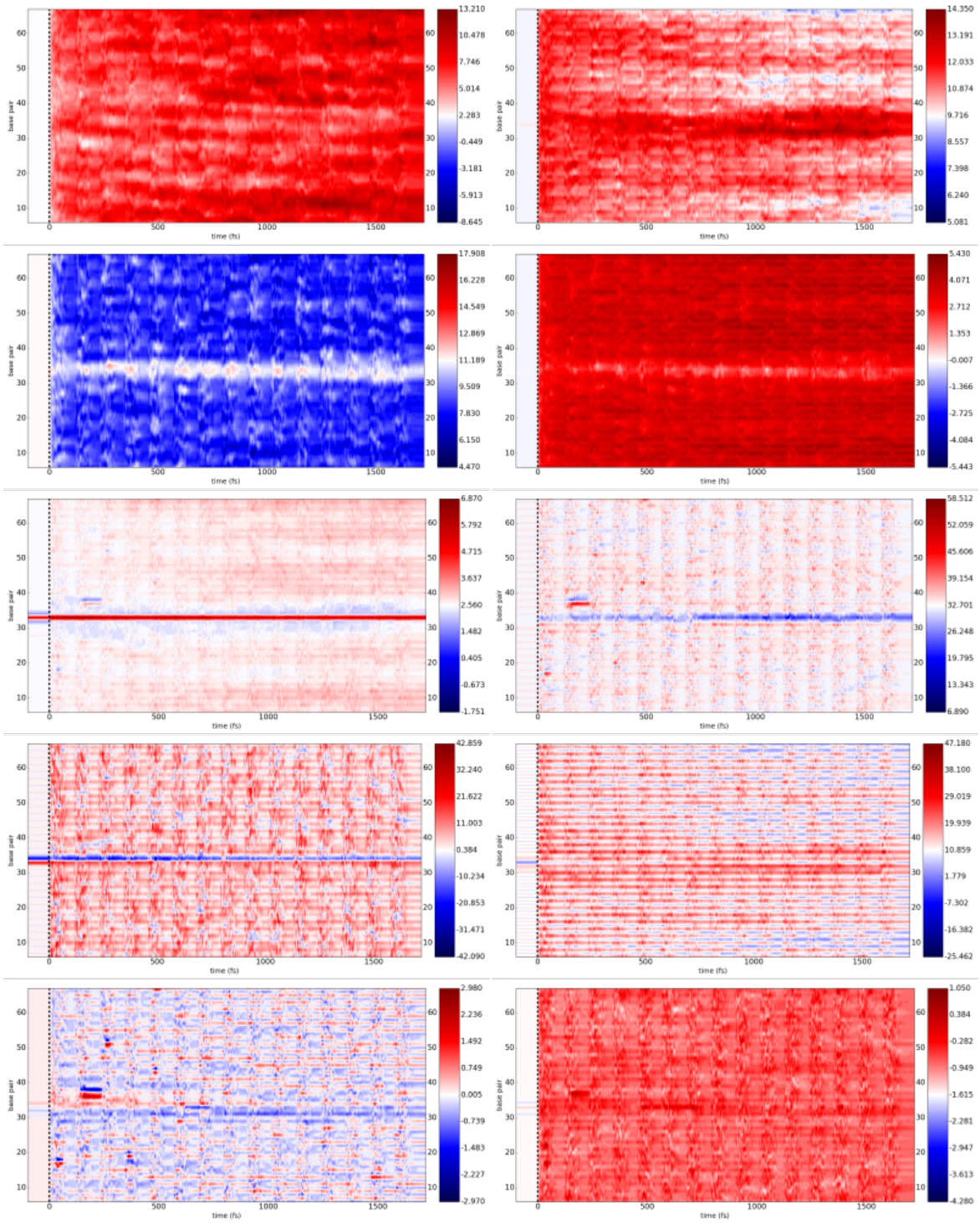

Figure A.12: Groove parameters for the AC-A-ida system. See first page of Appendix A for ordering of graphs.

A.4 d(AC)$_{72}$, B-start (AC-B) series

A.4.1 Bare DNA (AC-B-bare)

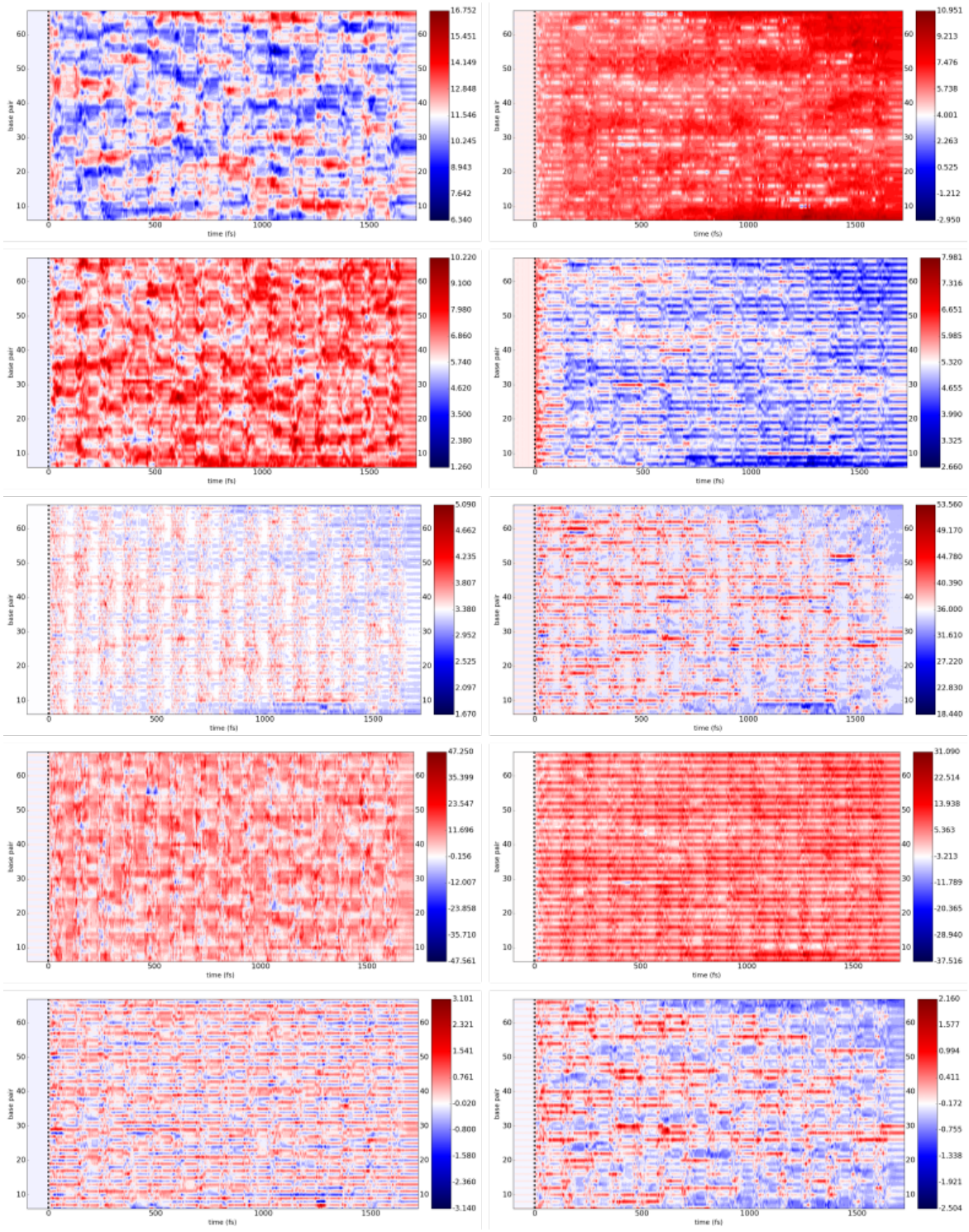

Figure A.13: Groove parameters for the AC-B-bare system. See first page of Appendix A for ordering of graphs.

A.4.2 Daunomycin (AC-B-dau)

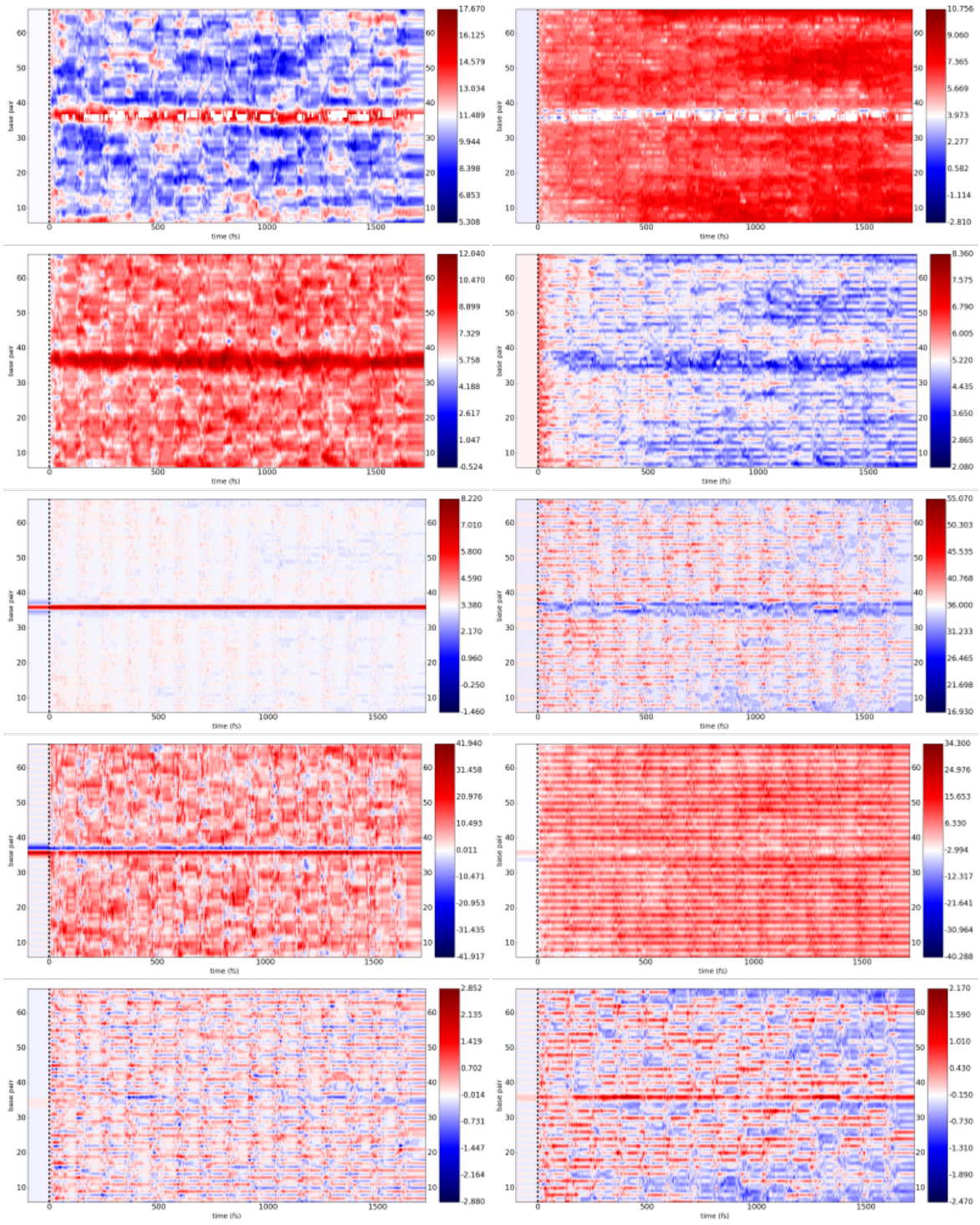

Figure A.14: Groove parameters for the AC-B-dau system. See first page of Appendix A for ordering of graphs.

A.4.3 Doxorubicin (AC-B-dox)

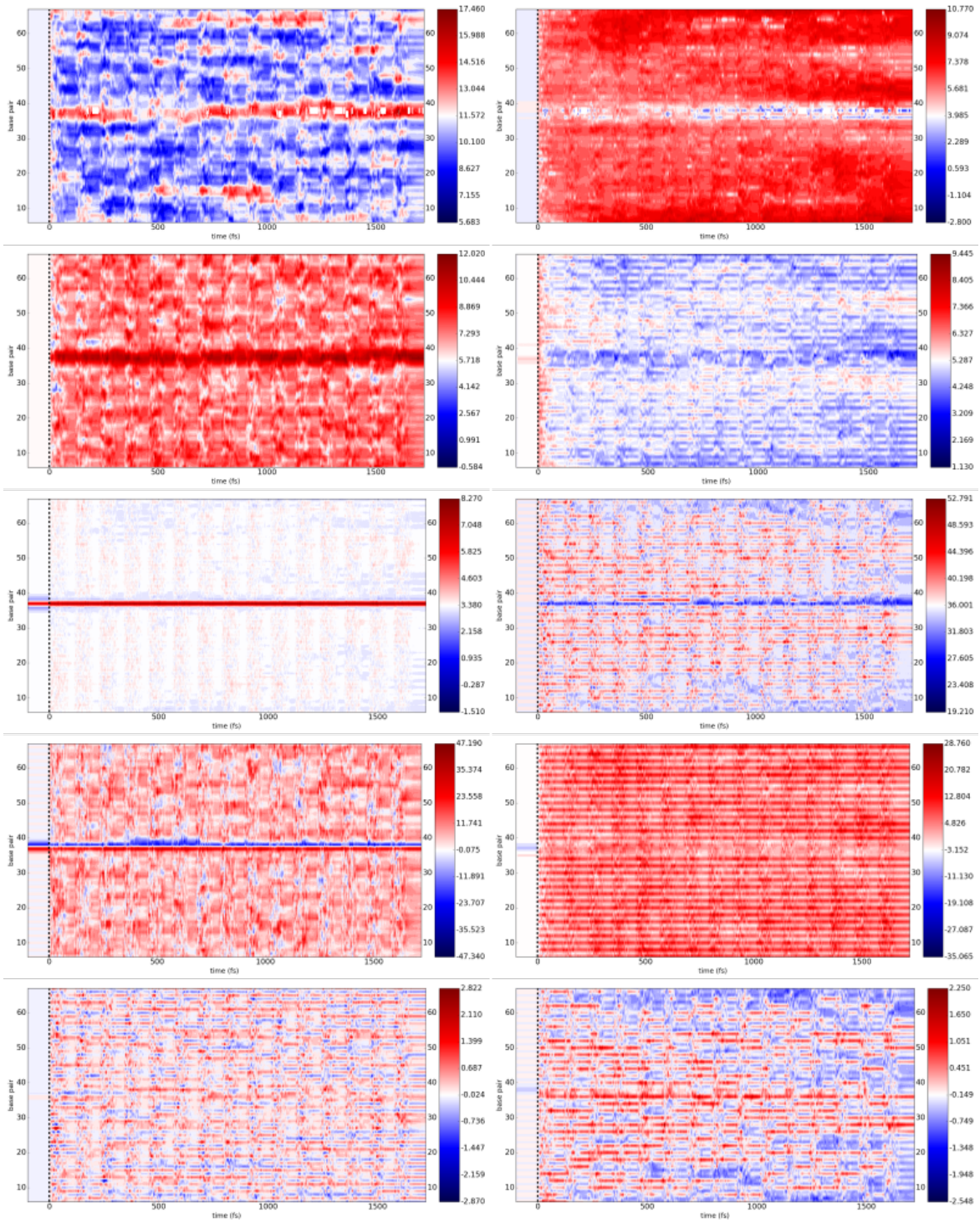

Figure A.15: Groove parameters for the AC-B-dox system. See first page of Appendix A for ordering of graphs.

A.4.4 Idarubicin (AC-B-ida)

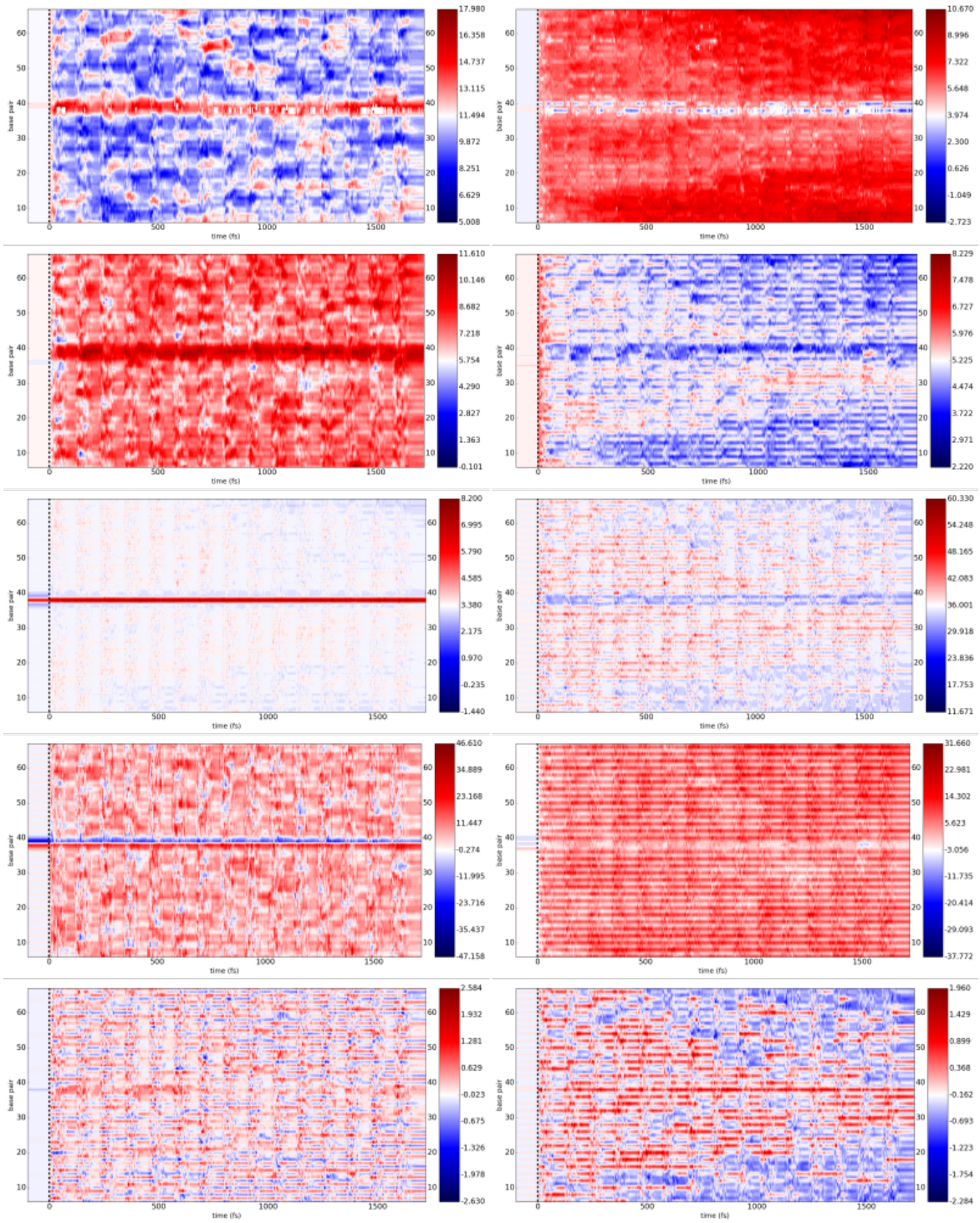

Figure A.16: Groove parameters for the AC-B-ida system. See first page of Appendix A for ordering of graphs.

A.5 d(C)$_{72}$, A-start (CC-A) series

A.5.1 Bare DNA (CC-A-bare)

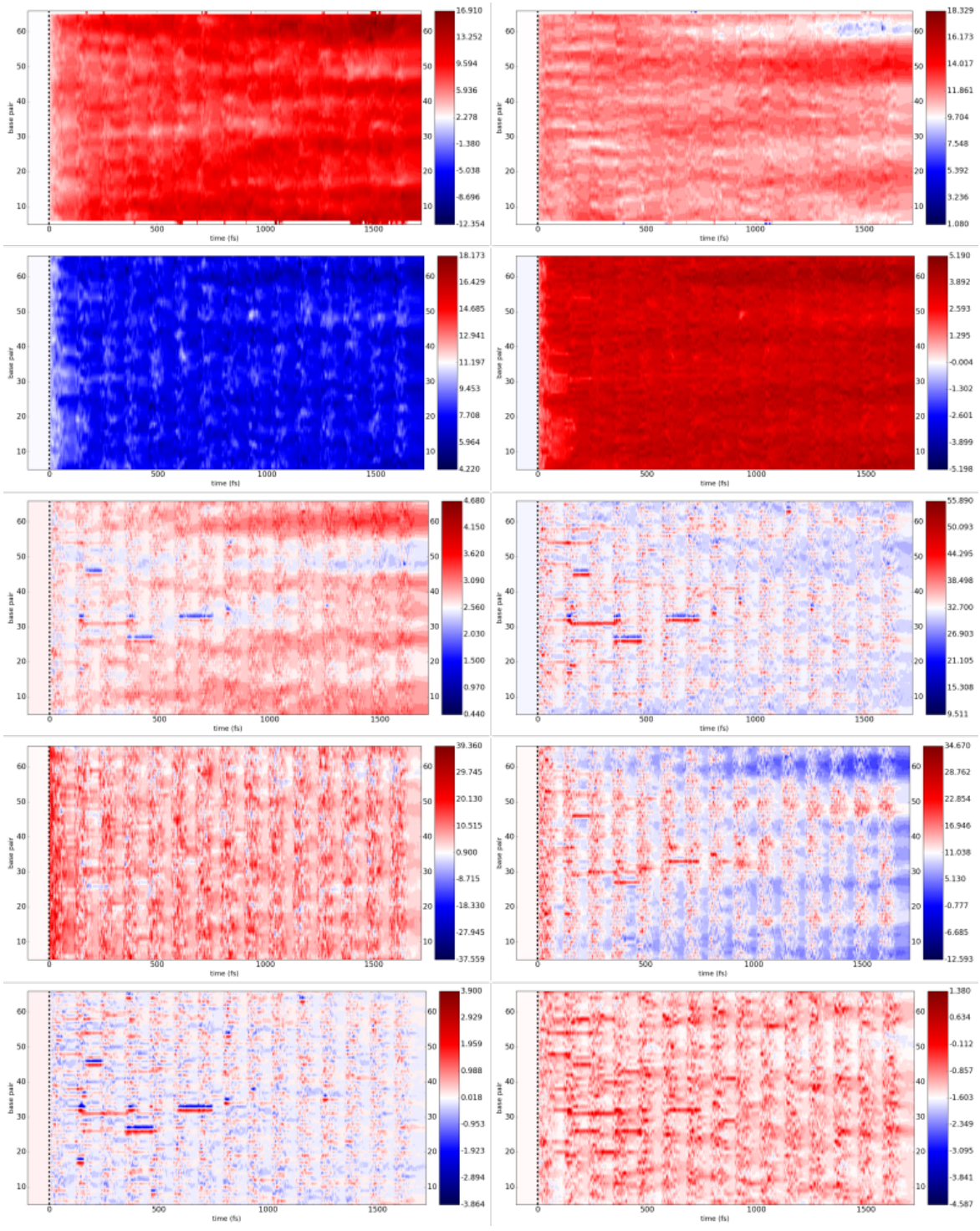

Figure A.17: Groove parameters for the CC-A-bare system. See first page of Appendix A for ordering of graphs.

A.5.2 Daunomycin (CC-A-dau)

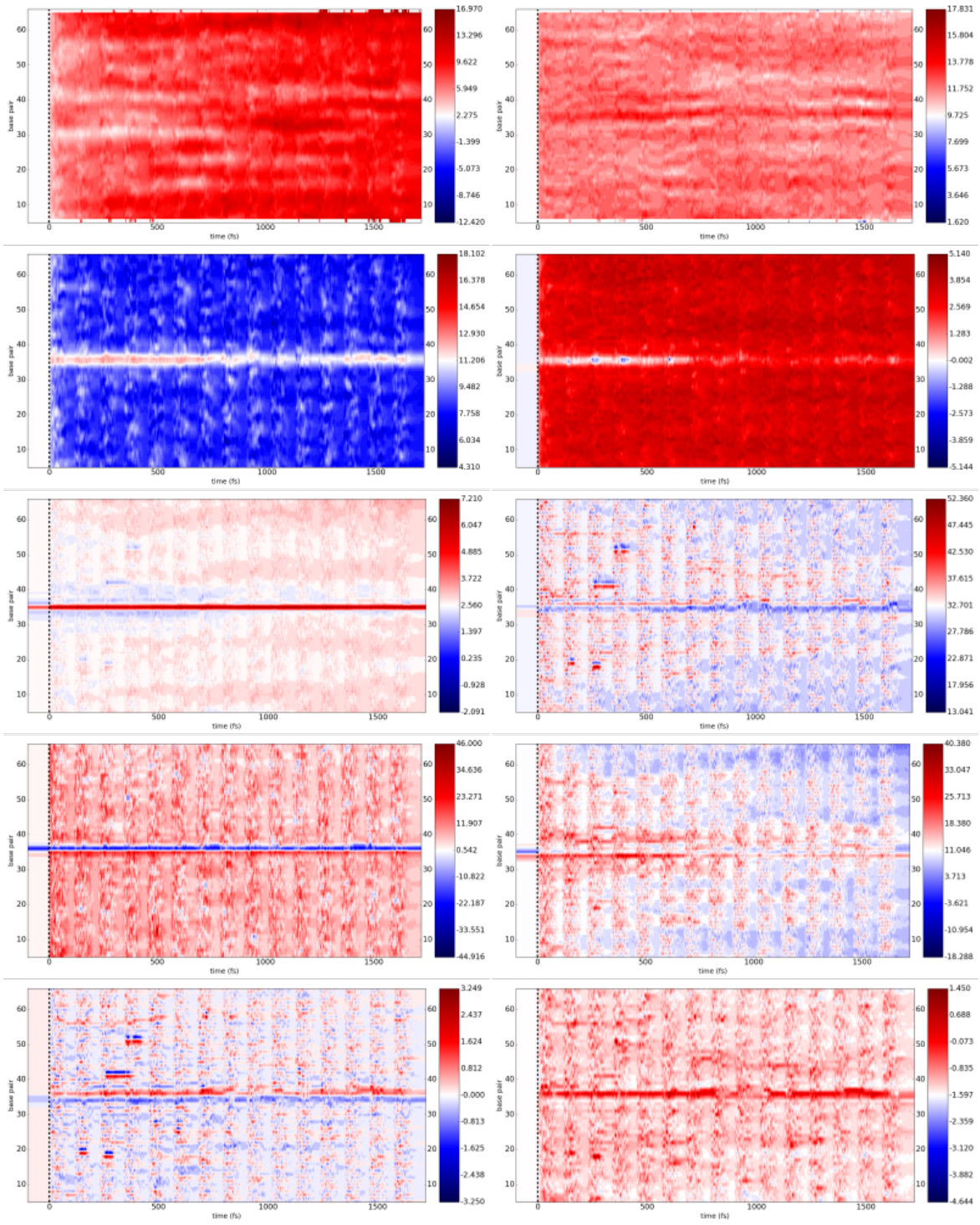

Figure A.18: Groove parameters for the CC-A-dau system. See first page of Appendix A for ordering of graphs.

A.5.3 Doxorubicin (CC-A-dox)

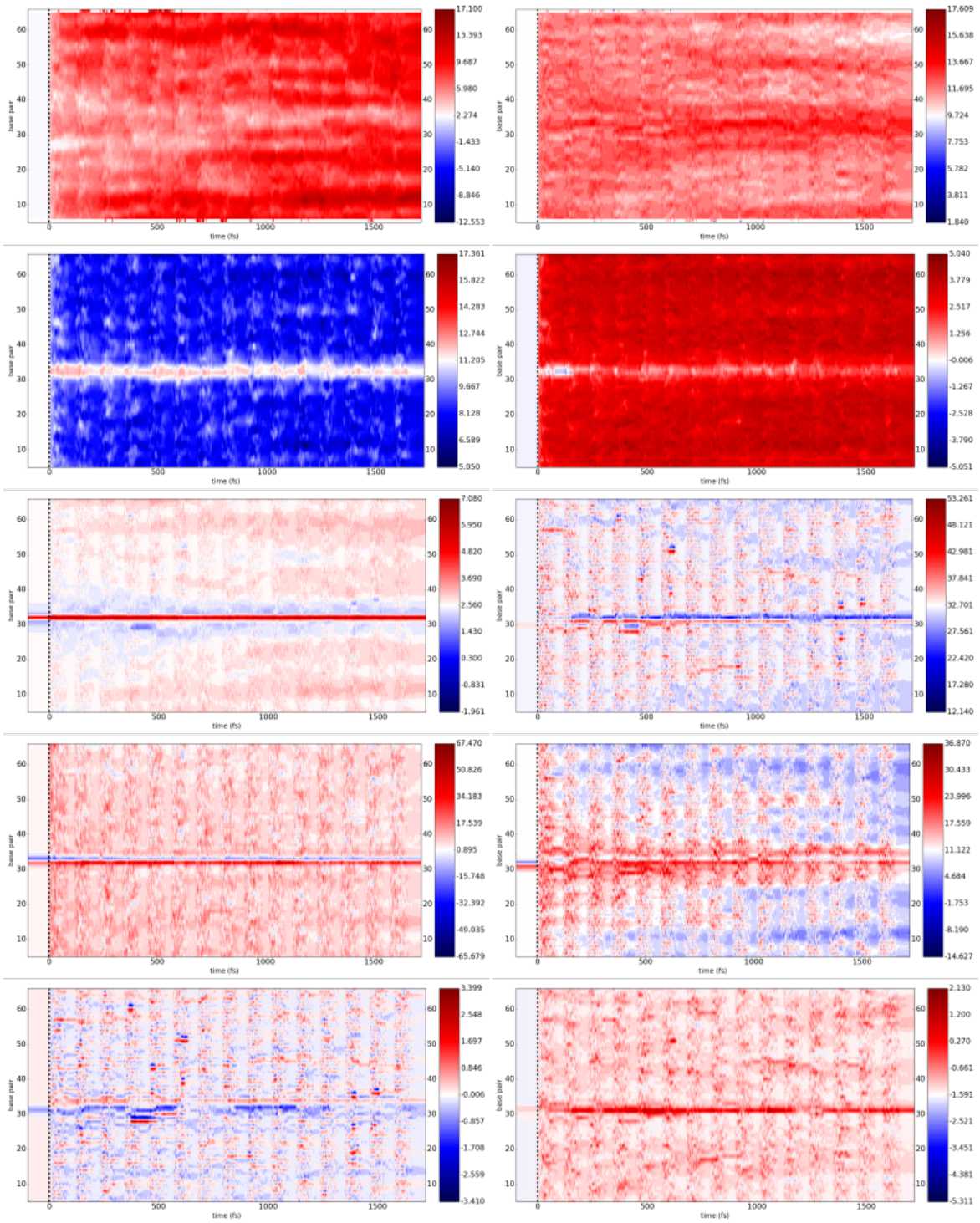

Figure A.19: Groove parameters for the CC-A-dox system. See first page of Appendix A for ordering of graphs.

A.5.4 Idarubicin (CC-A-ida)

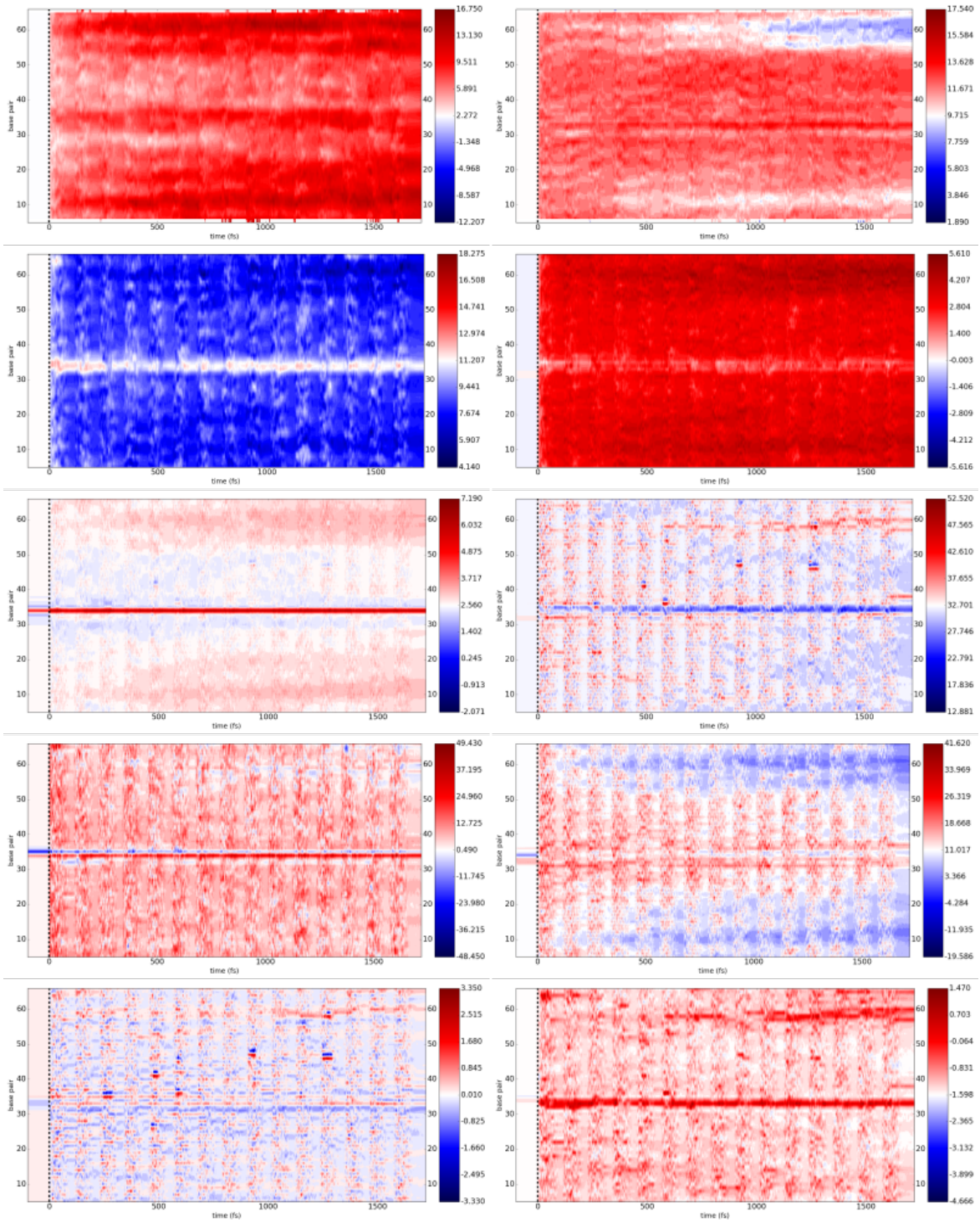

Figure A.20: Groove parameters for the CC-A-ida system. See first page of Appendix A for ordering of graphs.

A.6 d(C)$_{72}$, B-start (CC-B) series

A.6.1 Bare DNA (CC-B-bare)

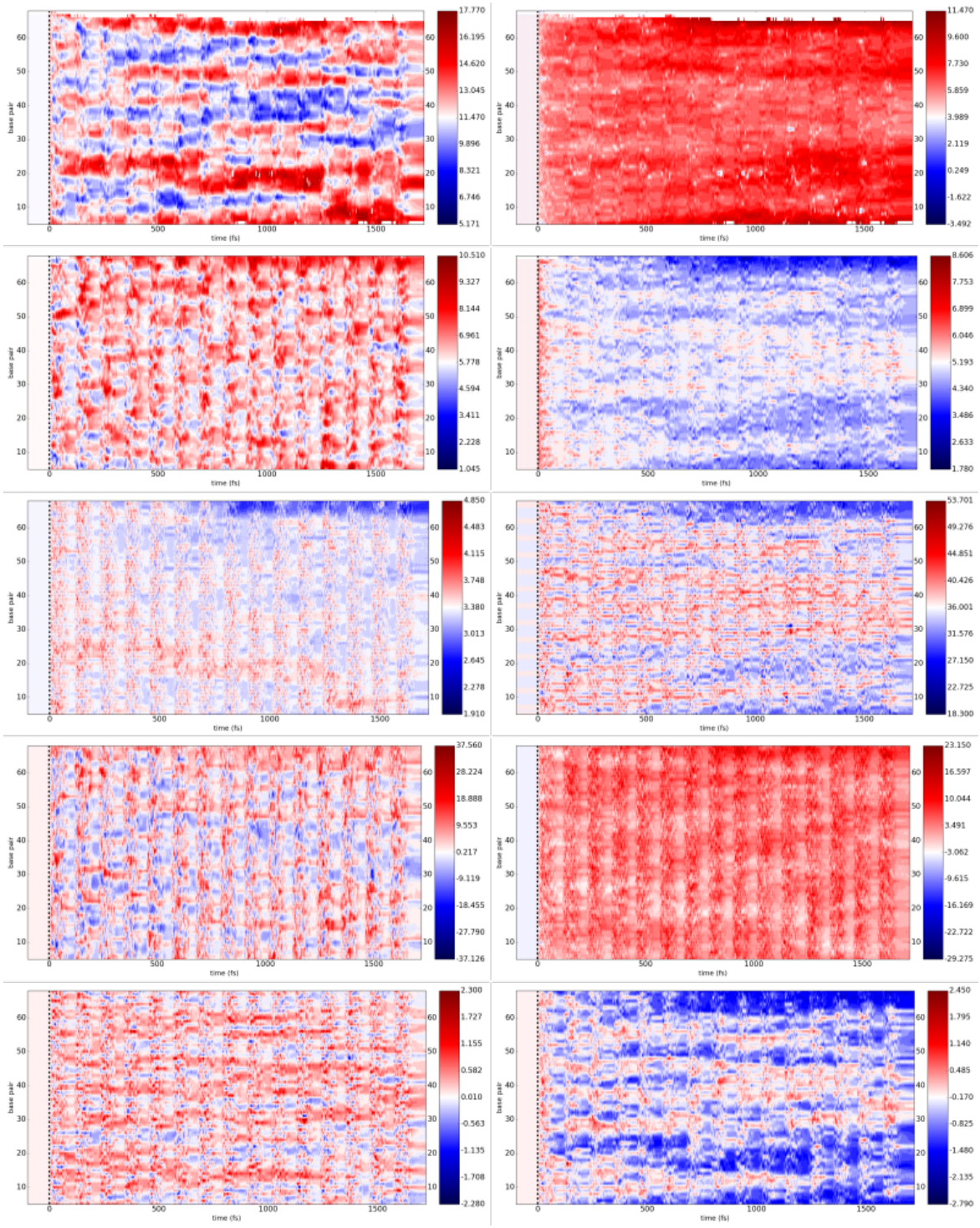

Figure A.21: Groove parameters for the CC-B-bare system. See first page of Appendix A for ordering of graphs.

A.6.2 Daunomycin (CC-B-dau)

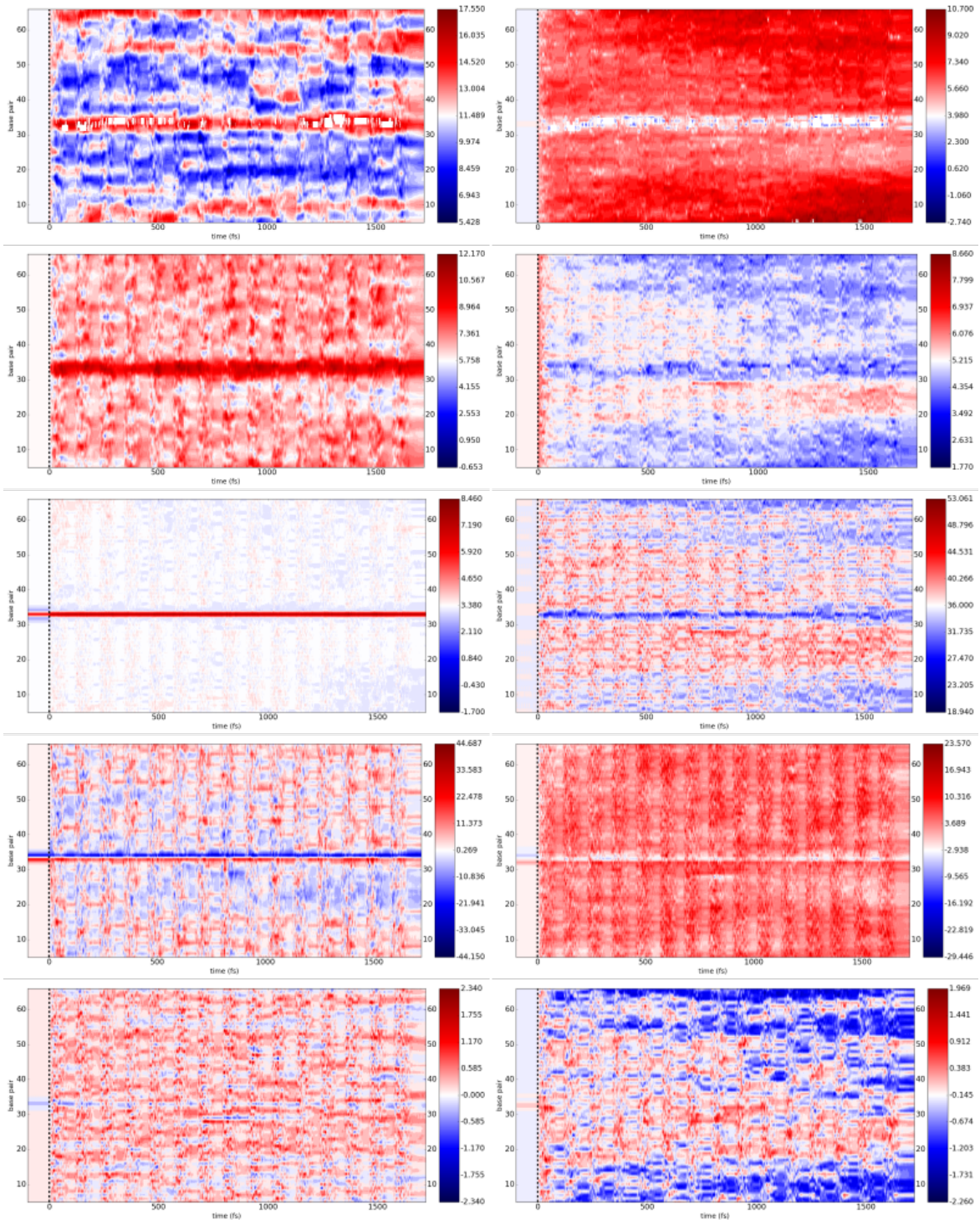

Figure A.22: Groove parameters for the CC-B-dau system. See first page of Appendix A for ordering of graphs.

A.6.3 Doxorubicin (CC-B-dox)

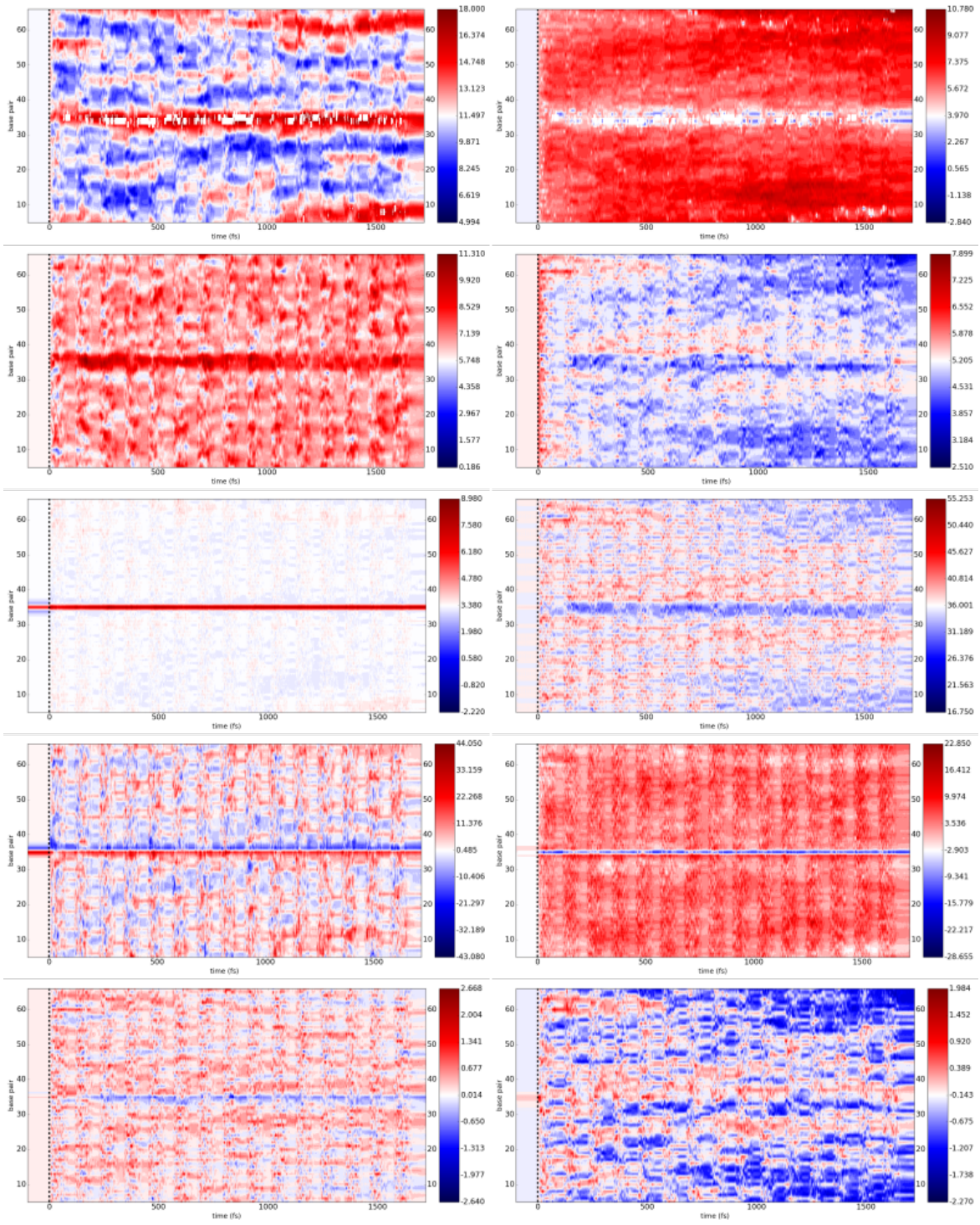

Figure A.23: Groove parameters for the CC-B-dox system. See first page of Appendix A for ordering of graphs.

A.6.4 Idarubicin (CC-B-ida)

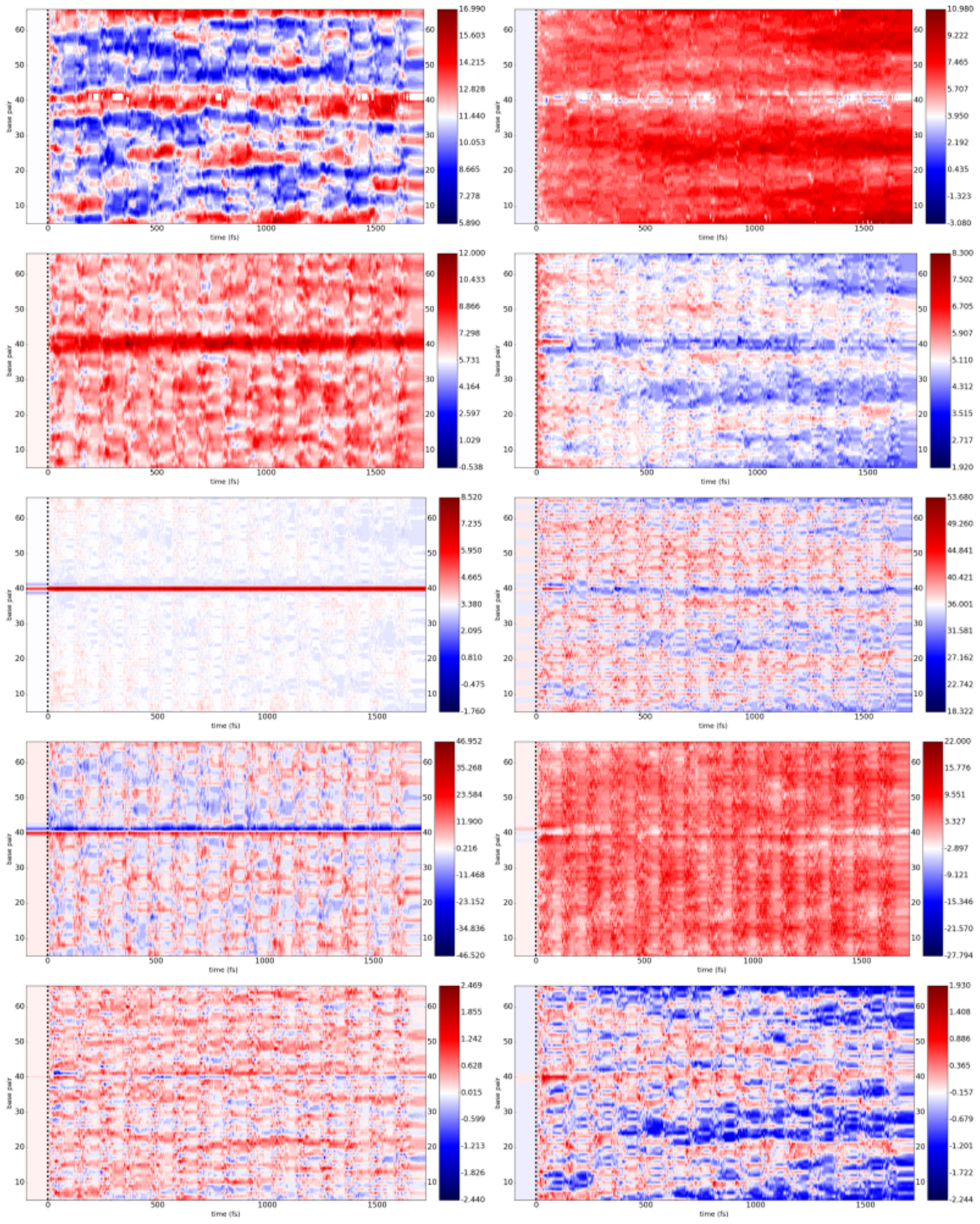

Figure A.24: Groove parameters for the CC-B-ida system. See first page of Appendix A for ordering of graphs.

Appendix B

Derivations of important results

B.1 Mapping between equilibrium constant and probabilities of states

In P. 62 when we discussed about the equilibrium constant, we asserted that the equilibrium constant K_c can be mapped onto the probability ratio. There are two different routes of deriving the same result and they are listed below [1]:-

Straightforward derivation Consider a binding interaction $L + S \leftrightharpoons L:S$ where S is the substrate, L is the ligand and $L:S$ is the bound state. Then the equilibrium constant (known specifically as the binding constant K_a in this case) of the interaction is given by

$$K_a = \frac{\gamma_{LS}[L:S]}{\gamma_L[L]\gamma_S[S]}C \tag{B.1}$$

where $[\dots]$ is the equilibrium concentration of the enclosed species (the subscripted "eq" is omitted for simplicity), $\gamma_{\dots}$ is the activity coefficient of a species, and $C = \frac{N}{V}$ is the number density of system at equilibrium which appears such that K_a dimensionless. A special case occurs in standard conditions where $N = 1\,\text{mol}$ and $V = 1\,\text{dm}^3$, in which case $K_a = K_a{}^\circ$ is known as the *standard* binding constant. Moreover, it is customary to assume that the activities γ are close to unity so they are usually left out from Eq. B.1 [203].

Now, K_a can be expressed readily in terms of the number of molecules of each species,

$$\begin{aligned} K_a &\approx \frac{[L:S]}{[L][S]}C \\ &= \frac{N_{LS}/V}{(N_L/V)\,(N_S/V)} = \frac{N_{LS}}{N_L N_S}VC \\ &= \frac{p_{LS}N}{p_L N\, p_S N}VC = \frac{p_{LS}}{p_L p_S}\frac{V}{N}C \\ \implies K_a &\approx \frac{p_{LS}}{p_L p_S} \end{aligned} \tag{B.2}$$

[1] In this derivation we only deal with binding by a single ligand per receptor and multiple binding is out of scope.

where $p_{...}$ is the probability of getting a specific species in the equilibrium mixture. The denominator $p_L p_S$ is justified as the probability of finding unreacted residues should be the same as that of getting one each of the ligand and the substrate in successive random draws, which is the product of the two respective probabilities.

Rigorous derivation We now use the simplified common notation of the binding constant as a basis and note that

$$K_a = \frac{[L:S]}{[L][S]}. \tag{B.3}$$

Now set π_{LS} and π_S to be the proportion of substrates bound with *exactly one* ligand, and that of *unbound* substrates. Clearly, since there are only two outcomes, *viz.* bound and unbound, $\pi_{LS}+\pi_S$ must equal unity. Moreover, if we assume the concentration of substrates to be sufficiently lower than that of ligands, $[L]$ can be assumed to be constant throughout the reaction. Then it follows that $[L:S] = \pi_{LS}[S]_0$ and $[S] = \pi_S[S]_0$ where $[S]_0$ is the initial concentration of substrates. Then Eq. B.3 becomes

$$K_a = \frac{1}{[L]}\frac{\pi_{LS}}{\pi_S} \tag{B.4}$$

B.2 Derivation of the atomic form factor in X-ray diffraction

It is stated without proof in Cantor and Schimmel [32] (Eq. 13-23) that, for a spherically symmetrical charge distribution $\rho(r) = zNe^{-kr^2}$, the form factor of the X-ray diffraction of such distribution is given by

$$f(S) = z\exp\left(-\frac{\pi^2 S^2}{k}\right) \tag{B.5}$$

where $S = \|\boldsymbol{S}\|$ is the norm of the scattering vector $\boldsymbol{S}$. The following section is dedicated to the derivation of this relation.

We first start by realising that the X-ray scattering pattern on a plane is the spatial Fourier transform of the sample charge distribution [32], hence

$$\begin{aligned}
f(\boldsymbol{S}) &= \iiint_{\text{all space}} d^3\boldsymbol{r}\,\rho(\boldsymbol{r})e^{i2\pi\boldsymbol{S}\cdot\boldsymbol{r}} \\
&= \int_0^{2\pi} d\phi \int_0^{\pi} d\theta \sin\theta \int_0^{\infty} dr\, r^2\rho(\boldsymbol{r})e^{i2\pi\boldsymbol{S}\cdot\boldsymbol{r}} \\
&= 2\pi\int_0^{\infty} dr\, r^2\rho(r)\int_0^{\pi} d\theta \sin\theta e^{i2\pi Sr\cos\theta} \\
&= -2\pi\int_0^{\infty} dr\, r^2\rho(r)\int_1^{-1} dx\, e^{i2\pi Srx} \\
&= 4\pi\int_0^{\infty} dr\,\rho(r)r^2\frac{\sin(2\pi Sr)}{2\pi Sr} \\
&\equiv f(S)
\end{aligned} \tag{B.6}$$

which is Eq. 13-21 in [32].

Now for $\rho(r) = zNe^{-kr^2}$,

$$\begin{aligned} f(S) &= 4\pi \int_0^\infty dr\, \rho(r) r^2 \frac{\sin(2\pi Sr)}{2\pi Sr} \\ &= \frac{2}{S} \int_0^\infty dr\, rzNe^{-kr^2} \sin(2\pi Sr) \\ &= \frac{2zN}{S} \underbrace{\int_0^\infty dr\, re^{-kr^2} \sin(2\pi Sr)}_{\mathcal{I}} \end{aligned} \tag{B.7}$$

Consider the substitution

$$\begin{cases} u = \sin(2\pi Sr) \\ dv = re^{-kr^2} dr \end{cases} \implies \begin{cases} du = 2\pi S \cos(2\pi Sr)\, dr \\ v = -\dfrac{1}{2k} e^{-kr^2} \end{cases},$$

then the integral $\mathcal{I}$ becomes, using integration by parts

$$\begin{aligned} \mathcal{I} &= \cancel{-\frac{1}{2k} e^{-kr^2} \sin(2\pi Sr) \Big|_0^\infty}^{\to 0} + \frac{\pi S}{k} \int_0^\infty dr\, e^{-kr^2} \cos(2\pi Sr) \\ &= \frac{\pi S}{2k} \int_{-\infty}^\infty dr\, e^{-kr^2} \cos(2\pi Sr) \\ &= \frac{\pi S}{2k} \sqrt{\frac{\pi}{k}} \exp\left(-\frac{\pi^2 S^2}{k}\right) \end{aligned} \tag{B.8}$$

where the result is obtained from Abramowitz and Stegun [2] and by making use of the even nature of the integrand. Now substituting Eq. B.8 back into Eq. B.7,

$$\begin{aligned} f(S) = \frac{2zN}{S} \mathcal{I} &= \frac{2zN}{S} \frac{\pi S}{2k} \sqrt{\frac{\pi}{k}} \exp\left(-\frac{\pi^2 S^2}{k}\right) \\ &= zN \left(\frac{\pi}{k}\right)^{3/2} \exp\left(-\frac{\pi^2 S^2}{k}\right) \end{aligned} \tag{B.9}$$

Now, normalisation of the charge distribution enforces that

$$\begin{aligned} \iiint_{\text{all space}} d^3\boldsymbol{r} \rho(\boldsymbol{r}) &= z \\ \int_0^{2\pi} d\phi \int_0^\pi d\theta \sin\theta \int_0^\infty dr\, r^2 \rho(r) &= z \\ 4\pi N \int_0^\infty dr\, r^2 e^{-kr^2} &= 1 \\ 4\pi N \frac{1}{2} \frac{\sqrt{\pi}}{2k^{3/2}} &= 1 \\ \therefore N &= \left(\frac{\pi}{k}\right)^{-3/2} \end{aligned} \tag{B.10}$$

Hence,

$$f(S) = z \left(\frac{\pi}{k}\right)^{-3/2} \left(\frac{\pi}{k}\right)^{3/2} \exp\left(-\frac{\pi^2 S^2}{k}\right)$$

$$= z \exp\left(-\frac{\pi^2 S^2}{k}\right)$$

Thus proves Eq. B.5.

B.3 Derivation of relationship between intensity and number of photons

The formal definition of the intensity of an electromagnetic wave is the time-averaged flux of the wave [89]. Hence, mathematically,

$$I(t) = \frac{E}{At} \tag{B.11}$$

for a fixed-energy and fixed-area system. Now for a single photon, its energy ε is given by the Einstein equation $\varepsilon = h\nu = \dfrac{hc}{\lambda}$. Then for a system with n electrons, the total energy is $E = n\varepsilon = \dfrac{nhc}{\lambda}$. Hence,

$$\begin{aligned} I &= \frac{nhc}{\lambda A t} \\ &\propto \frac{n}{t} \end{aligned} \tag{B.12}$$

which implies the intensity detected is proportional to the number of photons hitting the detector per unit time, assuming the monochromaticity of the light source.

B.4 Calculation of ensemble average of interaction free energy

In Chapter 7, we have calculated the free energy associated with the intercalation of three anticancer drugs into DNA. We asserted that, for each of the drug-DNA complex system, a total of $6 \times 2 \times 2 = 24$ subsystems are needed for the computation of the overall free energy using the ABF or eABF algorithm. However, since these 24 subsystems only contribute to the final energy change, it is necessary to derive an analytical formalism in order to combine these numbers together, and this section will be dedicated for this task.

In this section, we will first derive the ensemble average for a simple two-state binding reaction, then we will analyse how errors propagate in our formalism. Finally, this formalism will be applied to the systems studied in this work.

Simple two-state binding system Consider a binding reaction $A + B \leftrightharpoons A : B$ where the colon means a physical or chemical bond formed between the chemicals A and B. Also assume that there are N modes of binding. Then, to obtain the probability p_i of binding through the i-th mode, we need the ratio between the equilibrium concentration of the

complex in the particular mode and that of the entirety of the mixture of the complex, i.e.

$$p_i = \frac{[A:B]_i}{\sum_{j=1}^{N}[A:B]_j} \tag{B.13}$$

Now if we divide both the numerator and the denominator by $[A][B]$, i.e. the equilibrium concentrations of the unbound reactants, we obtain

$$p_i = \frac{[A:B]_i}{[A][B]} \Bigg/ \frac{\sum_j [A:B]_j}{[A][B]} \tag{B.14}$$

If we compare the numerator and the denominator in Eq. B.14 and the equilibrium constant, which is given by

$$K_i = \frac{[A:B]_i}{[A][B]}, \tag{B.15}$$

we can see that the probability can be expressed in terms of the equilibrium constant, i.e.

$$p_i = \frac{K_i}{\sum_j K_j}. \tag{B.16}$$

Now using Eq. 2.43, Eq. B.16 becomes

$$p_i = \frac{\exp(-\beta\Delta G_i)}{\sum_j \exp(-\beta\Delta G_j)}. \tag{B.17}$$

To obtain the *average* free energy of the reaction,

$$\begin{aligned} \langle \Delta G \rangle &= \sum_i p_i \Delta G_i \\ &= \sum_i \Delta G_i \frac{\exp(-\beta\Delta G_i)}{\sum_j \exp(-\beta\Delta G_j)} \end{aligned} \tag{B.18}$$

Error propagation Since ΔG is determined either experimentally or computationally, which means what one obtains as ΔG is actually $\langle \Delta G \rangle \pm \delta\Delta G$, i.e. a mean value with an error. However, since the ensemble average (Eq. B.18) is rather complicated, its error propagated from those of its components is not intuitive. In this section we derive an expression for the error term.

We start off by simplifying Eq. B.18, by noting that both numerator and denominator are linearly independent of each other. We can then write

$$\langle \Delta G \rangle = \sum_i \Delta G_i \frac{\exp(-\beta\Delta G_i)}{\sum_j \exp(-\beta\Delta G_j)} = \frac{\mathcal{N}(\Delta G_i)}{\mathcal{D}(\Delta G_j)} \tag{B.19}$$

where $\mathcal{N}(\Delta G_i) = \sum_i \Delta G_i \exp(-\beta \Delta G_i)$ and $\mathcal{D}(\Delta G_j) = \sum_j \exp(-\beta \Delta G_j)$.

Then from the definitions of derivatives, we have

$$\delta\langle \Delta G \rangle = \left| \frac{\partial \langle \Delta G \rangle}{\partial \mathcal{N}} \right| \delta \mathcal{N} + \left| \frac{\partial \langle \Delta G \rangle}{\partial \mathcal{D}} \right| \delta \mathcal{D} \tag{B.20}$$

where the two absolute-valued partial derivatives can be trivially evaluated as

$$\begin{cases} \dfrac{\partial \langle \Delta G \rangle}{\partial \mathcal{N}} = \dfrac{1}{\mathcal{D}(\Delta G_j)} \\ \dfrac{\partial \langle \Delta G \rangle}{\partial \mathcal{D}} = \dfrac{\mathcal{N}(\Delta G_j)}{(\mathcal{D}(\Delta G_j))^2} = \dfrac{1}{\mathcal{D}(\Delta G_j)} \langle \Delta G \rangle \end{cases} \tag{B.21}$$

Similarly, the errors in $\mathcal{N}$ and $\mathcal{D}$ can be calculated as

$$\begin{cases} \delta \mathcal{N} = \sum_i |1 - \beta \Delta G_i| \exp(-\beta \Delta G_i)\, \delta(\Delta G_i) \\ \delta \mathcal{D} = \beta \sum_j \exp(-\beta \Delta G_j)\, \delta(\Delta G_j) \end{cases} . \tag{B.22}$$

Here, $\delta(\Delta G_i) = \delta(\Delta G_j)$ are the errors from the raw data calculated for each binding mode. Thus, Eq. B.20 reads, in fully expanded form,

$$\delta\langle \Delta G \rangle = \frac{\sum_i |1 - \beta \Delta G_i| \exp(-\beta \Delta G_i)\, \delta(\Delta G_i) + \beta \langle \Delta G \rangle \sum_i \exp(-\beta \Delta G_i)\, \delta(\Delta G_i)}{\sum_j \exp(-\beta \Delta G_j)} \tag{B.23}$$

Now, since the equilibrium constant of a reaction is related with the free energy change via Eq. 2.43, the ensemble average should also take the same form, hence

$$\langle K \rangle = \exp(-\beta \langle \Delta G \rangle). \tag{B.24}$$

Then naturally, the associated error of $\langle K \rangle$ can be obtained using the same logic as above,

$$\delta \langle K \rangle = \left| \frac{d \langle K \rangle}{d \langle \Delta G \rangle} \right| \delta \langle \Delta G \rangle = \beta \langle K \rangle \delta \langle \Delta G \rangle \tag{B.25}$$

with the ordinary differential operator d being used instead of the partial differential operator ∂ since $\langle K \rangle$ only explicitly depends on $\langle \Delta G \rangle$.

Similar arguments can be made for the probability of specific binding modes p_i (cf. Eq. B.16), once the binding constant K_i and its associated error δK_i are found. As before, we define $\mathcal{N}(K_i) = K_i$ and $\mathcal{D}(K_j) = \sum_j K_j$. The only difference between here and when we

calculated $\delta\langle\Delta G\rangle$ is the subsequent errors of $\mathcal{N}$ and $\mathcal{D}$, which are now much more trivially

$$\begin{cases} \delta\mathcal{N} = \delta K_i \\ \delta\mathcal{D} = \sum_j \delta K_j \end{cases} . \tag{B.26}$$

Then following the steps before, we can easily arrive at the conclusion that

$$\delta p_i = \frac{1}{\sum_j K_j}\left(\delta K_i + p_i \sum_j \delta K_j\right). \tag{B.27}$$

However, whilst this is the correct form of the error in p_i, there is a very important caveat in using it. We note that in this formalism, p_i is dependent on all the modal equilibrium constants, which in turn are exponential functions of the corresponding free energies. Using this formalism, we have attempted the calculation of the probabilities of binding modes and their associated statistical errors. We discovered that such errors would, more likely than not, exceed the value of 100%, which does not make any mathematical sense at all, as the probability has a natural domain of $[0,1]$.

Application in DNA intercalation In Chapter 7 and in earlier discussion, we have already mentioned a few times that the intercalative interactions between drugs and DNA are extremely complex and can be considered in four layers, namely base sequence b = {AA, AC, AG, AT, CC, CG}, groove g = {major, minor}, orientation o = {upright, reversed} and their respective CVs. Therefore the summations in the previous equations are in fact multiple sums over all these layers. Due to the complexity of Eqs. B.18 and B.23, the expanded forms of these equations into the layers are not listed here, but the generalisation is straightforward.

However, there is one term in Eq. B.23 which is worth special care, which is the $\delta(\Delta G_i)$ term, which becomes $\delta(\Delta G_{b,g,o})$ in the layered form. This term accounts for the error in each of the DNA-drug complex system. Note that we did not add in the layer of the CVs. This is because of the linearly additive nature of the energies of the CVs, it implies that the associated errors of these energies are also linearly additive. As elucidated before, this term comes directly from the raw data obtained from the simulations, thus the error should reflect the distribution of the data. In this work, we have used the central limit theorem (CLT) [27] which states that assume a finite set of data x_i follows normal distribution, its error is given by

$$\delta x_i \sim \frac{\sigma_{x_i}}{\sqrt{N}} z_{\alpha/2} \tag{B.28}$$

where σ_{x_i} is the standard deviation of the data set, N is the number of samples (data points) in the data sets, and $z_{\alpha/2}$ is the z-score associated with a user-defined confidence level p. $z_{\alpha/2}$, basically the same as the quantile function of the zero-mean and unity-variance normal distribution $\mathcal{N}(0,1)$, is defined as

$$z_{\alpha/2} = \sqrt{2}\,\mathrm{erf}^{-1}(2p-1) \tag{B.29}$$

where erf^{-1} is the inverse error function.

Two messages can be obtained from Eqs. B.28 and B.29. Firstly, given the injectivity (i.e. one-to-one mapping) of the inverse error function, the z-score is a constant given a fixed value of p. This implies that δx_i decreases as the number of data points increases, given a constant p. Secondly, and as an indirect consequence of the first point, if one wants to keep δx_i whilst shrinking the data size, his confidence in data points lying within this error range must decrease accordingly.

B.5 Calculation of projected population density in simulations

In Chapter 7, when we performed the simulations, the eABF algorithm imposes a harmonic constraint on the virtual extended degree of freedom which is coupled to the real CV. In order to justify whether a calculation is correct, one of the ways is to compare the population from the binned data with analytical solution for the population distribution. In this section, we derive the population distribution for a harmonic oscillator under Langevin dynamics, which involves first the reformulation of the Langevin equations into the generalised Fokker-Planck equation.

We first define a function $P(\boldsymbol{x},\boldsymbol{v},t)$ such that $P(\boldsymbol{x},\boldsymbol{v},t)d\boldsymbol{x}d\boldsymbol{v}$ is the probability of finding the particle between $\boldsymbol{x}$ and $(\boldsymbol{x}+\boldsymbol{dx})$, with its velocity within the range of $(\boldsymbol{v},\boldsymbol{v}+\boldsymbol{dv})$. Then from probability theory, we know that

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} P(\boldsymbol{a},t)d\boldsymbol{x}d\boldsymbol{v} \equiv 1. \tag{B.30}$$

where $\boldsymbol{a} = (\boldsymbol{x},\boldsymbol{v})$ is the phase-space coordinate of the particle. Then, for a confined phase-space $V \in \mathbb{R}^6$ with a simple closed manifold boundary ∂V, the continuity equation must hold and reads

$$\frac{d}{dt}\iiint_V P(\boldsymbol{a},t)d\boldsymbol{x}d\boldsymbol{v} = -\oiint_{\partial V} P(\boldsymbol{a},t)\,\dot{\boldsymbol{a}}\cdot\boldsymbol{dS} \tag{B.31}$$

where $\boldsymbol{S}$ is the vector pointing outwards from V. Then, the generalised Gauss theorem transforms the RHS into also a volume integral, i.e.

$$\frac{d}{dt}\iiint_V P(\boldsymbol{a},t)\,d\boldsymbol{x}d\boldsymbol{v} = -\iiint_V \nabla\cdot(P(\boldsymbol{a},t)\dot{\boldsymbol{a}})\,d\boldsymbol{x}d\boldsymbol{v} \tag{B.32}$$

Hence,

$$\begin{aligned}\frac{\partial}{\partial t}P(\boldsymbol{a},t) &= -\nabla\cdot(P(\boldsymbol{a},t)\dot{\boldsymbol{a}}) \\ &= -\nabla_{\boldsymbol{x}}(P(\boldsymbol{a},t)\dot{\boldsymbol{x}}) - \nabla_{\boldsymbol{v}}(P(\boldsymbol{a},t)\dot{\boldsymbol{v}})\end{aligned} \tag{B.33}$$

which in the case of the Langevin equation reads

$$\begin{aligned}\frac{\partial P}{\partial t} &= -\nabla\cdot(P\dot{\boldsymbol{a}}) \\ &= -\nabla_{\boldsymbol{x}}\cdot(P\dot{\boldsymbol{x}}) - \nabla_{\boldsymbol{v}}\cdot\left(P(-\frac{\gamma}{m}\boldsymbol{v}+\frac{\boldsymbol{F}}{m}(\boldsymbol{x})+\frac{1}{m}\boldsymbol{\xi}(t))\right)\end{aligned}$$

$$
\begin{aligned}
&= -\boldsymbol{v}\cdot\nabla_{\boldsymbol{x}}P+\frac{\gamma}{m}\nabla_{\boldsymbol{v}}\cdot(P\boldsymbol{v})-\frac{1}{m}\boldsymbol{F}\cdot\nabla_{\boldsymbol{v}}P-\frac{1}{m}\boldsymbol{\xi}\cdot\nabla_{\boldsymbol{v}}P \\
&= -\boldsymbol{v}\cdot\nabla_{\boldsymbol{x}}P+\frac{\gamma}{m}P+\frac{\gamma}{m}\boldsymbol{v}\cdot\nabla_{\boldsymbol{v}}P-\frac{1}{m}\boldsymbol{F}\cdot\nabla_{\boldsymbol{v}}P-\frac{1}{m}\boldsymbol{\xi}\cdot\nabla_{\boldsymbol{v}}P
\end{aligned}
\tag{B.34}
$$

which can be readily written using partial differential operators into

$$
\frac{\partial P}{\partial t}=-(\hat{L_S}+\hat{L_D})P \tag{B.35}
$$

where

$$
\begin{cases}
\hat{L_S}=\boldsymbol{v}\cdot\nabla_{\boldsymbol{x}}-\dfrac{\gamma}{m}-\dfrac{\gamma}{m}\boldsymbol{v}\cdot\nabla_{\boldsymbol{v}}+\dfrac{1}{m}\boldsymbol{F}\cdot\nabla_{\boldsymbol{v}} \\
\hat{L_D}=\dfrac{1}{m}\boldsymbol{\xi}\cdot\nabla_{\boldsymbol{v}}
\end{cases}
\tag{B.36}
$$

Here the subscripts S and D denotes "streaming" and "diffusion" respectively, as they resemble the similar terms in fluid dynamics. Following some sophisticated maths (see [259]) one finds that

$$
\frac{\partial P}{\partial t}=-\boldsymbol{v}\cdot\nabla_{\boldsymbol{a}}P-\frac{1}{m}\nabla_{\boldsymbol{a}}\cdot((\gamma\boldsymbol{v}-\boldsymbol{F})\,P)+\frac{1}{2m}(\nabla_{\boldsymbol{a}}\cdot\underline{\underline{\mathbf{D}}}\cdot\nabla_{\boldsymbol{a}})P \tag{B.37}
$$

where $\underline{\underline{\mathbf{D}}}$ is a diagonal diffusion tensor. Eq. B.37 is called the Fokker-Planck equation. Note that in this equation the stochastic term involving $\boldsymbol{\xi}$ does not exist. This is because the effect of $\boldsymbol{\xi}$ is recorded intrinsically in P in each realisation of it and because of its stochastic nature, $P(t)$ is different in each realisation. As a result, the probability P in Eq. B.37 is, in fact, the average effect of the random force on the particle. Hence, $P=\langle\boldsymbol{\xi}\rangle$.

Moreover, we can simplify Eq. B.37 by noticing that it can be written back into the differential form of the equation of continuity, hence,

$$
\frac{\partial P}{\partial t}=-\left(\nabla_{\boldsymbol{a}}\cdot\boldsymbol{w}(\boldsymbol{a})-\frac{1}{2}\nabla_{\boldsymbol{a}}\cdot\underline{\underline{\mathbf{D}}}\cdot\nabla_{\boldsymbol{a}}\right)P. \tag{B.38}
$$

Now, analogous to the operators in Eq. B.36, the two terms here correspond to the streaming and diffusive behaviour to the probability flow.

For the Langevin equation (Eq. 2.23), which can be decoupled into two equations:

$$
\begin{cases}
\dfrac{d\boldsymbol{x}}{dt}=\boldsymbol{v} \\
\dfrac{d\boldsymbol{v}}{dt}=\dfrac{1}{m}\left(-\gamma\boldsymbol{v}+\boldsymbol{F}(\boldsymbol{x})+\boldsymbol{\xi}(t)\right)
\end{cases}
\quad , \tag{B.39}
$$

we can rewrite it in the matrix form, where

$$
\boldsymbol{w}(\boldsymbol{a})=\left(\boldsymbol{v},-\frac{\gamma}{m}\boldsymbol{v}+\frac{1}{m}\boldsymbol{F}(\boldsymbol{x})\right)^{T};\ \underline{\underline{\mathbf{D}}}=\begin{pmatrix}0 & 0\\ 0 & \dfrac{2\gamma k_BT}{m^2}\end{pmatrix} \tag{B.40}
$$

with which the Fokker-Planck equation (Eq. B.38) is fully expanded as

$$
\frac{\partial P}{\partial t}=-\nabla_{\boldsymbol{x}}(\boldsymbol{v}P)-\nabla_{\boldsymbol{v}}\left(\left(-\frac{\gamma}{m}\boldsymbol{v}+\frac{1}{m}\boldsymbol{F}\right)P\right)+\frac{2\gamma k_BT}{m^2}{\nabla_{\boldsymbol{v}}}^2P \tag{B.41}
$$

where ${\nabla_{\boldsymbol{v}}}^2\equiv\nabla_{\boldsymbol{v}}\cdot\nabla_{\boldsymbol{v}}$ which is the laplacian with respect to $\boldsymbol{v}$. Since at equilibrium, $\partial_tP=0$,

then the equilibrium probability must satisfy the homogeneous differential equation

$$-\nabla_{\boldsymbol{x}}\left(\boldsymbol{v}P\right)-\nabla_{\boldsymbol{v}}\left(\left(-\frac{\gamma}{m}\boldsymbol{v}+\frac{1}{m}\boldsymbol{F}\right)P\right)+\frac{2\gamma k_B T}{m^2}\nabla_{\boldsymbol{v}}{}^2 P=0 \tag{B.42}$$

For a holonomic system, the Hamiltonian of the system is the sum of the kinetic and potential energies, so

$$H=\frac{1}{2}m\boldsymbol{v}^2+V(\boldsymbol{x}) \tag{B.43}$$

with which Eq. B.42 can be written as

$$\begin{aligned}-\nabla_{\boldsymbol{x}}\left(P\nabla_{\boldsymbol{v}}H\right)+\nabla_{\boldsymbol{v}}\left(P\nabla_{\boldsymbol{x}}H\right)+\gamma\nabla_{\boldsymbol{v}}\left(\frac{1}{m}P\nabla_{\boldsymbol{v}}H+\frac{k_BT}{m}\nabla_{\boldsymbol{v}}P\right)=0\\ \Longleftrightarrow\gamma\nabla_{\boldsymbol{v}}\left(\frac{1}{m}P\nabla_{\boldsymbol{v}}H+\frac{k_BT}{m}\frac{dP}{dH}\nabla_{\boldsymbol{v}}H\right)=0\end{aligned} \tag{B.44}$$

assuming P is solely dependent on H. In order for this equation to hold, it must reduce to the ordinary differential equation

$$P+k_BT\frac{dP}{dH}=0 \tag{B.45}$$

which has the general solution

$$P(H)=Ce^{-\beta H}\Longleftrightarrow P(\boldsymbol{x},\boldsymbol{v})=Ce^{-\frac{1}{2}\beta m\boldsymbol{v}^2-\beta V(\boldsymbol{x})} \tag{B.46}$$

where C is a system-specific constant.

The probability density function in Eq B.46 is a joint distribution on both displacement and velocity. In order to obtain the equivalent distribution along a given parameter, we have to perform a "partial integration" on the other parameter. Hence,

$$\begin{cases}P(\boldsymbol{x})=\iiint d^3\boldsymbol{v}\,P(\boldsymbol{x},\boldsymbol{v})\sim e^{-\beta V(\boldsymbol{x})}\\ P(\boldsymbol{v})=\iiint d^3\boldsymbol{x}\,P(\boldsymbol{x},\boldsymbol{v})\sim e^{-\frac{1}{2}\beta m\boldsymbol{v}^2}\end{cases} \tag{B.47}$$

For instance, if the potential profile is harmonic, as in the case for the fictitious degree of freedom in the eABF formalism (see Sec. 2.7.2), the distribution takes the form

$$P(\xi)\sim e^{-\frac{1}{2}\beta k(\xi-\xi_0)^2} \tag{B.48}$$

which is a Gaussian distribution with the mean at $\xi=\xi_0$.

Bibliography

[1] *"Free Energy" — Oxford Dictionary of Physics*. Oxford University Press, 6th edition, 2009.

[2] M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. National Bureau of Standards, 1972.

[3] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular Biology of the Cell*. Garland Science, 4th edition, 2002.

[4] A. Ambrogelly, S. Palioura, and D. Söll. Natural expansion of the genetic code. *Nature Chemical Biology*, 3(1):29–35, 2007.

[5] C. Andersen, H. Molecular dynamics at constant pressure and/or temperature. *J. Chem. Phys.*, 72:2384–2393, 1980.

[6] F. Arcamone, G. Cassinelli, G. Fantini, A. Grein, P. Orezzi, C. Pol, and C. Spalla. Adriamycin, 14-hydroxydaimomycin, a new antitumor antibiotic from *S. Peucetius* var. *caesius*. *Biotechnology and Bioengineering*, 11(6):1101–1110, 1969.

[7] S. A. Arrhenius. Über die Dissociationswärme und den Einfluß der Temperatur auf den Dissociationgrad der Elektrolyte. *Z. Phys. Chem.*, 4:96–116, 1889.

[8] S. A. Arrhenius. Über die Reaktionsgeschwindigkeit bei der Invasion von Rohrzucker durch Säuren. *Z. Phys. Chem.*, 4:226–248, 1889.

[9] O. T. Avery, C. M. Macleod, and M. McCarty. Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types : Induction of Transformation By a Desoxyribonucleic Acid Fraction Isolated From Pneumococcus Type III. *J. Exp. Med.*, 79(2):137–58, 1944.

[10] J. Awrejcewicz. *Numerical Analysis: Theory and Application*. InTech, 2011.

[11] L. V. Azaroff. *Elements of X-Ray Crystallography*. McGraw-Hill, New York, 1968.

[12] N. R. Bachur, S. L. Gordon, and M. V. Gee. Anthracycline antibiotic augmentation of microsomal electron transport and free radical formation. *Mol. Pharmacol.*, 13:901–910, 1977.

[13] N. R. Bachur, S. L. Gordon, and M. V. Gee. A general mechanism for microsomal activation of quinone anticancer agents. *Cancer Res.*, 38:1745–1750, 1978.

[14] A. Barducci, G. Bussi, and M. Parrinello. Well-tempered metadynamics: A smoothly converging and tunable free-energy method. *Phys. Rev. Lett.*, 100(2):1–4, 2008.

[15] C. I. Bayly, P. Cieplak, W. D. Cornell, and P. A. Kollman. A Well-Behaved Electrostatic Potential Based Method Using Charge Restraints for Deriving Atomic Charges: The RESP Model. *J. Phys. Chem.*, 97:10269–10280, 1993.

[16] D. Beeman. Some multistep methods for use in molecular dynamics calculations. *J. Comp. Phys.*, 20(2):130–139, 1976.

[17] N. B. Benchekroun, B. K. Sinha, and J. Robert. Doxorubicin-induced oxygen free radical formation in sensitive and doxorubicin resistant variants of rat glioblastoma cells. *FEBS Lett.*, 322:295–298, 1993.

[18] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak. Molecular Dynamics with Coupling to an External Bath. *J. Chem. Phys.*, 81(8):3684–3690, 1984.

[19] J. M. Berg, J. L. Tymoczko, L. Stryer, and N. D. Clarke. *Biochemistry*. W. H. Freeman and Company, 2002.

[20] R. Bernardi *et al*. *NAMD User's Guide (Version 2.12)*. Theoretical Biophysics Group, University of Illinois and Beckman Institute, 405 N. Mathews, Urbana, IL 61801, 2016.

[21] P. Beroza and D. A. Case. calculation of proton binding thermodynamics in proteins. *Methods Enzymol.*, 295:170–189, 1998.

[22] A. S. Biebricher, I. Heller, R. F. H. Roijmans, T. P. Hoekstra, E. J. G. Peterman, and G. J. L. Wuite. The impact of DNA intercalators on DNA and DNA-processing enzymes elucidated through force-dependent binding kinetics. *Nature Comms.*, 6:7304, 2015.

[23] M. Binaschi, M. Bigioni, A. Cipollone, C. Rossi, C. Goso, C. A. Maggi, G. Capranico, and F. Animati. Anthracyclines: selected new developments. *Curr. Med. Chem. Anticancer Agents*, 1(2):113–130, 2001.

[24] C. K. Birdsall and A. B. Langdon. *Plasma Physics via Computer Simulations*. McGraw-Hill Book Company, 1985.

[25] C. Bissantz, B. Kuhn, and M. Stahl. A Medicinal Chemist's Guide to Molecular Interactions. *J. Med. Chem.*, 53, 2010.

[26] W. Blaschke. *Einführung in die Differentialgeometrie*. Springer-Verlag, 1950.

[27] M. L. Boas. *Mathematical Methods in the Physical Sciences*. John Wiley & Sons, Inc., 3rd edition, 2006.

[28] A. P. Breen and J. A. Murphy. Reactions of oxyl radicals with DNA. *Free Radic. Biol. Med.*, 18:1033–1077, 1995.

[29] R. Buchner, G. Hefter, and P. May. Dielectric Relaxation of Aqueous NaCl Solutions. *J. Phys. Chem. A*, 103(1):1, 1999.

[30] J. Bustamente, M. Galleano, E. E. Medrano, and A. Boveris. Adriamycin effects on hydroperoxide metabolism and growth of human breast tumor cells. *Breast Cancer Res. Treat.*, 17:145–153, 1990.

[31] E. Calendi, A. Di Marco, M. Reggiani, B. Scarpinato, and L. Valentini. On physicochemical interactions between daunomycin and nucleic acids. *Biochim. Biophys. Acta*, 103:25–49, 1965.

[32] C. R. Cantor and P. R. Schimmel. *Biophysical Chemistry. Part II: Techniques for the study of biological structure and function*. W. H. Freeman and Company, 1980.

[33] R. Car and M. Parrinello. Unified approach for molecular dynamics and density-functional theory. *Phys. Rev. Lett.*, 55:2471–2474, 1985.

[34] J. B. Chaires, N. Dattagupta, and D. M. Crothers. Studies on interaction of anthracycline antibiotics and deoxyribonucleic acid: equilibrium binding studies on interaction of daunomycin with deoxyribonucleic acid. *Biochemistry*, 21:3933–3940, 1982.

[35] Jonathan B Chaires. *Small Molecule DNA and RNA Binders: From Synthesis to Nucleic Acid Complexes. Volume 2.* WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim, 2003.

[36] U. Chandra Singh and P. A. Kollman. An approach to computing electrostatic charges for molecules. *J. Comp. Chem.*, 5(2):129–145, 1984.

[37] S. Chandrasekhar. Stochastic problems in physics and astronomy. *Rev. Mod. Phys.*, 15(1):1–89, 1943.

[38] Chaplin, M. Water structure and science. [Online, URL: `http://www1.lsbu.ac.uk/water/water_structure_science.html`; accessed 9-May-2019].

[39] E. Chargaff, R. Lipshitz, and C. Green. Composition of the desoxypentose nucleic acids of four genera of sea-urchin. *J. Chem. Biol.*, 195(1):155–160, 1952.

[40] A. Cheong, S. McGrath, and S. Cutts. Anthracyclines. *WikiJournal of Medicine*, 5(1):1, 2018.

[41] C. J. Chetsanga, M. Lozon, C. Makaroff, and L. Savage. Purification and characterization of Escherichia coli formamidopyrimidine-DNA glycosylase that excises damaged 7-methylguanine from deoxyribonucleic acid. *Biochemistry*, 20(18):5201–5207, 1981.

[42] C. Chipot. Frontiers in Free-Energy Calculations of Biological Systems. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 4:71–89, 2014.

[43] C. Chipot and A. Pohorille. *Free energy calculations. Theory and applications in chemistry and biology*. Springer-Verlag, 2007.

[44] J. Choe and B. Kim. Determination of Proper Time Step for Molecular Dynamics Simulation . *Bull. Korean Chem. Soc.*, 21(4):419–424, 2000.

[45] G. Ciccotti, R. Kapral, and E. Vanden-Eijnden. Blue moon sampling, vectorial reaction coordinates, and unbiased constrained dynamics. *Chem. Phys. Chem.*, 6:1809–1814, 2005.

[46] S. J. Clark, M. D. Segall, C. J. Pickard, P. J. Hasnip, M. I. J. Probert, K. Refson, and M. C. Payne. First principles methods using CASTEP. *Zeitschrift für Kristallographie*, 220:567–570, 2005.

[47] P. Cluzel, A. Lebrun, C. Heller, R. Lavery, J. Viovy, D. Chatenay, and F. Caron. DNA: An Extensible Molecule. *Science*, 271(5250), 1996.

[48] J. Comer, J. C. Gumbart, J. Hénin, T. Lelièvre, A. Pohorille, and C. Chipot. The adaptive biasing force method: Everything you always wanted to know but were afraid to ask. *J. Phys. Chem. B*, 119(3):1129–1151, 2015.

[49] B.N. Conner, K. Takano, S. Tanaka, K. Itakura, and R.E. Dickerson. The molecular structure of d(^{1}CpCpGpG), a fragment of right-handed double helical A-DNA. *Nature*, 295:294–299, 1982.

[50] L. Cooper and D. Steinberg. *Introduction to Methods of Optimization*. W. B. Saunders Company, 1970.

[51] W.D. Cornell, P. Cieplak, C.I. Bayly, I.R. Gould, K.M. Merz, D.M. Ferguson, D.C. Spellmeyer, T. Fox, J.W. Caldwell, and P.A. Kollman. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society*, 117(19):5179–5197, 1995.

[52] J. J. Correia and J. B. Chaires. Analysis of drug-DNA binding isotherms: a monte carlo approach. *Methods Enzymol.*, 240:593–614, 1994.

[53] R. Courant. *Differential and Integral Calculus, Vol. 2.* Blackie, London., 1936.

[54] N.R. Cozzarelli, T.C. Boles, and J.H. White. *Primer on the topology and geometry of DNA supercoiling. In DNA topology and its biological effects.* Cold Spring Harbor Laboratory Press, 1990.

[55] C. J. Cramer and D. G. Truhlar. Implicit solvation models: equilibria, structure, spetra, and dynamics. *Chem. Rev.*, 99:2161–2200, 1999.

[56] F. H. C. Crick. Linking numbers and nucleosomes. *Proc. Natl. Acad. Sci.*, 73(8):2639–2643, 1971.

[57] F. H. C. Crick, L. Barnett, S. Brenner, and R. J. Watts-Tobin. General nature of the genetic code for proteins. *Nature*, 192(4809):1227–1232, 1961.

[58] G. Călugăreanu. Sur les classes d'isotopie des nœuds tridimensionnels et leurs invariants. *Czechoslovak Math. J,*, 11(86):588–625, 1961.

[59] R. Dahm. Friedrich Mieschner and the discovery of DNA. *Developmental Biology*, 278:274–288, 2005.

[60] R. Dahm. Discovering DNA: Friedrich Miescher and the early years of nucleic acid research. *Human Genetics*, 122(6):565–581, 2008.

[61] Daintith, J. Intercalation compound — Oxford Dictionary of Chemistry, 2008. [Online, URL: `http://www.oxfordreference.com/view/10.1093/acref/9780199204632.001.0001/acref-9780199204632-e-2226?rskey=xmADap&result=2479`; accessed 8-May-2019].

[62] P. D. Dans, A. Zeida, M. R. Machado, and S. Pantano. A Coarse Grained Model for Atomic-Detailed DNA Simulations with Explicit Electrostatics. *J. Chem. Theory Comput.*, 6:1711–1725, 2010.

[63] L. Darré, M. R. Machado, P. D. Dans, F. E. Herrera, and S. Pantano. Another Coarse Grain Model for Aqueous Solvation: WAT FOUR? *J. Chem. Theory Comput.*, 6(12):3793–3807, 2010.

[64] L. Darré, R. Machado, A. F. Brandner, H. C. González, S. Ferreira, and S. Pantano. SIRAH: A Structurally Unbiased Coarse-Grained Force Field for Proteins with Aqueous Solvation and Long-Range Electrostatics. *J. Chem. Theory Comput.*, 11:723–739, 2015.

[65] L. Darré, A. Tek, M. Baaden, and S. Pantano. Mixing Atomistic and Coarse Grain Solvation Models for MD Simulations: Let WT4 Handle the Bulk. *J. Chem. Theory Comput.*, 8(10):3880–3894, 2012.

[66] E. Darve and A. Pohorille. Calculating free energies using average force Adaptive biasing force method for scalar and vector free energy calculations Calculating free energies using average force. *J. Chem. Phys.*, 115:9169–9183, 2001.

[67] K. J. Davies. The broad spectrum of responses to oxidants in proliferating cells: a new paradigm for oxidative stress. *IUPMB Life*, 48:41–47, 1999.

[68] R. Day, D. Paschek, and A. E. Garcia. Microsecond simulations of the folding/unfolding thermodynamics of the Trp-cage miniprotein. *Proteins*, 78(8), 2010.

[69] P. Debye and E. Hückel. Zur Theorie der Electrolyte. I. Gefrierpunktserniedrigung und verwandte Erscheinungen. *Physikalische Zeitschrift*, 24(9):185–206, 1923.

[70] Advanced Chemistry Development. ACD/ChemSketch (Freeware). Version 2018.2.5.

[71] M. J. S. Dewar, E. G. Zoebisch, E. F. Healy, and J. J. P. Stewart. Development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model. *J. Am. Chem. Soc.*, 107(13):3902–3909, 1985.

[72] A. Di Marco, Silvestrini R., S. Di Marco, and Dasdia T. Inhibiting effect of the new cytotoxic antibiotic daunomycin on nucleic acids and mitotic activity of HeLa cells. *J. Cell. Biol.*, 27:545–550, 1965.

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

[74] R. E. Dickerson and T. K. Chiu. Helix bending as a factor in protein/DNA recognition. *Biopolymers*, 44(4):361–403, 1997.

[75] R. Ditchfield, W. J. Hehre, and J. A. Pople. Self-Consistent Molecular-Orbital Methods. IX. An Extended Gaussian-Type Basis for Molecular-Orbital Studies of Organic Molecules. *J. Chem. Phys.*, 54(2):724–728, 1971.

[76] J. Donohue. Selected topics in hydrogen bonding. In *Structural Chemistry and Molecular Biology*, pages 443–465. Freeman, San Francisco, 1968.

[77] H. R. Drew, R. M. Wing, T. Takano, C. Broka, S. Tanaka, K. Itakura, and R. E. Dickerson. Structure of a B-DNA dodecamer: conformation and dynamics. *Proc. Natl. Acad. Sci. U.S.A.*, 78:2179–2183, 1981.

[78] M. Eckert. Max von Laue and the discovery of X-ray diffraction in 1912. *Annalen der Physik*, 524(5):A83–A85, 2012.

[79] A. Einstein. Über einen die Erzeugung und Verwandlung des Lichtes betreffenden heuristischen Gesichtspunkt. *Annalen der Physik*, 17(6):132–148, 1905.

[80] M. Engels, E. Jacoby, P. Krüger, J. Schlitter, and A. Wollmer. The T ↔ R structural transition of insulin; pathways suggested by targeted energy minimization. *Protein Engineering, Design and Selection*, 5(7):669–677, 1992.

[81] L. Euler. *Institutionum calculi integralis*. 1768.

[82] European Molecular Biology Organization. Definitions and nomenclature of nucleic acid structure parameters. *The EMBO Journal*, 8(1):1–4, 1989.

[83] P. P. Ewald. *Fifty Years of X-Ray Diffraction*. Oosthoek, Utrecht, The Netherlands, 1962.

[84] A. R. Faheem, T. H. Bohkari, S. Roohi, A. Mushtaq, and M. Sohaib. 99mTc-Daunorubicin a potential brain imaging and theranostic agent: synthesis, quality control, characterization, biodistribution and scintigraphy. *Nucl. Med. Bio.*, 40:148–152, 2013.

[85] S. E. Feller, Y. Zhang, R. W. Pastor, and B. R. Brooks. Constant pressure molecular dynamics simulation: The Langevin piston method. *J. Chem. Phys.*, 103:4613–4621, 1995.

[86] L. R. Ferguson and B. C. Baguley. Mutagenicity of anti-cancer drugs that inhibit topoisomerase enzymes. *Mutat. Res.*, 355:91–101, 1996.

[87] D. Fincham. Choice of timestep in molecular dynamics simulation. *Comp. Phys. Comms.*, 40(2–3):263–269, 1986.

[88] G. Fiorin, M. L. Klein, and J Hénin. Using collective variables to drive molecular dynamics simulations. *Mol. Phys.*, 111(22-33):3345–3362, 2013.

[89] P. M. Fishbane, S. G. Gasiorowicz, and S. T. Thornton. *Physics for Scientists and Engineers with Modern Physics: International Edition*. Pearson Prentice Hall, 3rd edition, 2005.

[90] J. M. Fogg, D. J. Catanese Jr., G. L. Randall, M. C. Swick, and L. Zechiedrich. *Differences between positively and negatively supercoiled DNA that topoisomerases may distinguish. In Mathematics of DNA Structure, Function and Interactions.* Springer, 2009.

[91] H. Föll. Science of Lattices and Crystals. [Online, URL: `https://www.tf.uni-kiel.de/matwis/amat/iss/kap_4/illustr/s4_2_1.html`; accessed 5-September-2019].

[92] H. Föll. X-Ray Diffraction. [Online, URL: `https://www.tf.uni-kiel.de/matwis/amat/iss/kap_4/illustr/s4_2_1.html`; accessed 10-September-2019].

[93] M. Frankenberg-Schwager. Induction, repair and biological relevance of radiation-induced DNA lesions in eukaryotic cells. *Radiat. Environ. Biophys.*, 29:273–292, 1990.

[94] R. Franklin and R. Gosling. The structure of sodium thymonucleate fibres. I. The influence of water content. *Acta. Cryst.*, 6:673–677, 1953.

[95] R. E. Franklin and R. G. Gosling. Molecular Configuration in Sodium Thymonucleate. *Nature*, 171:740–741, 1953.

[96] P. L. Freddolino, C. B. Harrison, Y. X. Liu, and K. Schulten. Challenges in protein folding simulations: Timescale, representation, and analysis. *Nat. Phys.*, 6, 2010.

[97] P. L. Freddolino, S. Park, B. Roux, and K. Schulten. Force field bias in protein folding simulations. *Biophys. J.*, 96, 2009.

[98] D. Frenkel and B. Smit. *Understanding Molecular Simulation: From Algorithms to Applications*. Academic Press, 1996.

[99] E. C. Friedberg, G. C. Walker, W. Siede, R. D. Wood, R. A. Schultz, and T. Ellenberger. *DNA Repair and Mutagenesis*. ASM Press, 2nd edition, 2006.

[100] S. J. Froelich-Ammon and N. Osheroff. Topoisomerase poisons: harnessing the dark side of enzyme mechanism. *J. Biol. Chem.*, 270:21429–21432, 1995.

[101] A. Fujiwara, T. Hoshino, and J. Westley. Anthracycline Antibiotics. *Crit. Revs. Biotech.*, 3(2):133–157, 1985.

[102] F. B. Fuller. The writhing number of a space curve. *Proc. Natl. Acad. Sci.*, 68(4):815–819, 1971.

[103] H. Gao, E. F. Yamasaki, K. K. Chan, L. L. Shen, and R. M. Snapka. DNA Sequence Specificity for Topoisomerase II Poisoning by the Quinoxaline Anticancer Drugs XK469 and CQS. *Mol. Pharmacol.*, 63(6), 2003.

[104] V. Gapsys, S. Michielssens, J. H. Peters, B. L. de Groot, and H. Leonov. Calculation of Binding Free Energies. In *Molecular Modeling of Proteins (in Methods in Molecular Biology), Vol. 1215*, pages 173–209. Springer Science+Business Media New York, 2015.

[105] N. Gavish and K. Promislow. Dependence of the dielectric constant of electrolyte solutions on ionic concentration: a microfield approach. *Phys. Rev. E*, 94:012611, 2016.

[106] Genome Reference Consortium. Human Genome Assembly GRCh38. [Online, URL: https://www.ncbi.nlm.nih.gov/grc/human/data?asm=GRCh38; accessed 25-Oct-2017].

[107] D. A. Gewirtz. A critical evaluation of the mechanisms of action proposed for the antitumor effects of the anthracycline antibiotics adriamycin and daunorubicin. *Biochem. Pharmacol.*, 57(7):727–741, 1999.

[108] M. K. Gilson. Theory of electrostatic interactions in macromolecules. *Curr. Opin. Struct. Biol.*, 5:216–223, 1995.

[109] F. Giustino. *Materials Modelling unsing Density Functional Theory: Properties and Predictions*. Oxford University Press, first edition, 2014.

[110] R. I. Glazer, K. D. Hartmann, and C. L. Richardson. Cytokinetic and biochemical effects of 5-iminodaunorubicin in hum colon carcinoma in culture. *Cancer Res.*, 42:117–121, 1982.

[111] H. Goldstein. *Classical Mechanics*. Addison-Wesley Publishing Company, inc., 10th edition, 1973.

[112] D. T. Goodhead. The initial damage produced by ionizing radiations. *Int. J. Radiat. Biol.*, 56:623–634, 1989.

[113] M. F. Goodman, M. J. Bessman, and N. R. Bachur. Adriamycin and daunorubicin inhibition of mutant T4 DNA polymerase. *Proc. Natl. Acad. Sci. USA*, 71:1193–1196, 1974.

[114] M. F. Goodman and G. M. Lee. Adriamycin interactions with T4 DNA polymerase. *J. Biol. Chem.*, 252:2670–2674, 1977.

[115] D. L. Goodstein. *States of Matter*. Dover Publications, 1985 (republished 2002).

[116] R. J. Greenall. Side-by-side Models of DNA. In *Topics in Nucleic Acid Structure: Part 3*, pages 133–162. The Macmillan Press, 1987.

[117] W. Greiner. *Classical mechanics: systems of particles and Hamiltonian dynamics*. Springer-Verlag New York, inc., 2003.

[118] S. Greitzer. Many Cheerful Facts. *Arbelos*, 4(5):14–17, 1986.

[119] F. Griffith. The significance of pneumococcal types. *J. Hyg. (Lond).*, 27(2):113–159, 1928.

[120] P. Gross. Quantifying how DNA stretches, melts and changes twist under tension. *Nature Physics*, 7(9):731–736, 2011.

[121] B. Guillot. A reappraisal of what we have learnt during three decades of computer simulations on water. *Journal of Molecular Liquids*, 101(1-3):219–260, 2002.

[122] J. Gumbart, B. Roux, and C. Chipot. Protein:ligand standard binding free energies: A tutorial for alchemical and geometrical transformations, 2017.

[123] J. C. Gumbart, B. Roux, and C. Chipot. Standard binding free energies from computer simulations: What is the best strategy? *J. Chem. Theor. Comput.*, 9:794–802, 2013.

[124] B. Q. Guo, A. Tam, S. A. Santi, and A. M. Parissenti. Role of autophagy and lysosomal drug sequestration in acquired resistance to doxorubicin in MCF-7 cells. *BMC Cancer*, 16(762):1–18, 2016.

[125] D. Hamelberg, J. Mongan, and J. A. McCammon. Accelerated molecular dynamics: A promising and efficient simulation method for biomolecules. *J. Chem. Phys.*, 120(24):11919–11929, 2004.

[126] M. D. Hanwell, D. E. Curtis, D. C. Lonie, T. Vandermeersch, E. Zurek, and G. R. Hutchison. Avogadro: An advanced semantic chemical editor, visualization, and analysis platform. *J. Cheminformatics*, 4:17, 2012.

[127] D. R. Hartree. The Wave Mechanics of an Atom with a Non-Coulomb Central Field. Part I. Theory and Methods. *Proc. Camb. Phil. Soc.*, 24:89–110, 1928.

[128] D. R. Hartree. The Wave Mechanics of an Atom with a Non-Coulomb Central Field. Part II. Some Results and Discussion. *Proc. Camb. Phil. Soc.*, 24:111–132, 1928.

[129] D. R. Hartree. Self-consistent field, with exchange, for beryllium. *Proc. R. Soc. Lond. A*, 150(869):9–33, 1935.

[130] S. Hassani. *Mathematical Methods for Students of Physics and Related Fields*. Springer, 2nd edition, 2009.

[131] J. B. Hasted, D. M. Ritson, and C. H. Collie. Dielectric Properties of Aqueous Ionic Solutions. Parts I and II. *J. Chem. Phys.*, 16(1):1, 1948.

[132] J. E. Hearsst, S. T. Isaacs, D. Kanne, H. Rapoport, and K. Straub. The reaction of the psoralens with deoxyribonucleic acid. *Q. Rev. Biophys.*, 17:1–44, 1984.

[133] J. Hénin, G. Fiorin, C. Chipot, and M. L. Klein. Overcoming free energy barriers using unconstrained molecular dynamics simulations. *J. Chem. Phys.*, 121:2904–2914, 2004.

[134] J. Hénin, G. Fiorin, C. Chipot, and M. L. Klein. Exploring multidimensional free energy landscapes using time-dependent biases on collective variables. *J. Chem. Theory Comput.*, 6(1):35–47, 2010.

[135] A. Hershey and M. Chase. Independent functions of viral protein and nucleic acid in growth of bacteriophage. *J. Gen. Physiol.*, 36(1):39–56, 1952.

[136] R. W. Hockney. The potential calculation and some applications. *Methods Comp. Phys.*, 9:136–211, 1970.

[137] P. Hohenberg and W. Kohn. Inhomogeneous electron gas. *Phys. Rev.*, 136:B864–B871, 1964.

[138] S. R. Holbrook, A. H. Wang, A. Rich, and S. H. Kim. Local mobility of nucleic acids as determined from crystallographic data. II. Z-form DNA. *J. Mol. Biol.*, 187(3):429–440, 1986.

[139] B. Honig and A. Nicholls. Classical electrostatics in biology and chemistry. *Science*, 268:1144–1149, 1995.

[140] K. Hoogsteen. The crystal and molecular structure of a hydrogen-bonded complex between 1-methylthymine and 9-methyladenine. *Acta Crystallogr.*, 16(9):907–916, 1963.

[141] W. G. Hoover. Canonical dynamics: Equilibrium phase-space distributions. *Phys. Rev. A*, 31:1695–1697, 1985.

[142] W. G. Hoover. Constant pressure equations of motion. *Phys. Rev. A*, 34:2499–2500, 1986.

[143] A.J. Hopfinger and R.A. Pearlstein. Molecular mechanics force-field parameterization procedures. *Journal of Computational Chemistry*, 5(5):486–499, 1984.

[144] M. Hosier. Probing the Structure and Dynamics of DNA-cation Systems. BSc thesis, University of York, 2016.

[145] W. Humphrey, A. Dalke, and K. Schulten. VMD - Visual Molecular Dynamics. *J. Mol. Graphics*, 14:33–38, 1996.

[146] L. H. Hurley. DNA and its associated processes as targets for cancer therapy. *Nat. Rev. Cancer*, 2:188–200, 2002.

[147] F. Hutchinson. Chemical changes induced in DNA by ionizing radiation. *Prog. Nucleic Acid Res.*, 32:115–154, 1985.

[148] M. Iannuzzi, A. Liao, and M. Parrinello. Efficient exploration of reactive potential energy surfaces using Car-Parrinello molecular dynamics. *Phys. Rev. Lett.*, 90:238302, 2003.

[149] J. D. Jackson. *Classical Electrodynamics*. John Wiley & Sons (Asia) Pte. Ltd., 3rd edition, 1999.

[150] A. Jakalian. *Fast, efficient generation of high-quality atomic charges*. PhD thesis, Concordia University, 2000.

[151] A. Jakalian, B. L. Bush, D. B. Jack, and C. I. Bayly. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: I. Method. *J. Comp. Chem.*, 21(2):132–146, 2000.

[152] A. Jakalian, D. B. Jack, and C. I. Bayly. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and Validation. *J. Comp. Chem.*, 23(16):1623–1641, 2002.

[153] C Jarzynski. A nonequilibrium equality for free energy differences. *Phys. Rev. Lett.*, 78(14):2690–2693, 1997.

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

[155] R. K. Kaul. Topological Quantum Field Theories – A Meeting Ground for Physicists and Mathematicians, 1999. [Online; accessed 14-August-2019].

[156] G. Khandelwal and J. Bhyaravabhotla. A Phenomenological Model for Predicting Melting Temperatures of DNA Sequences. *PLoS ONE*, 5(8).

[157] M. P. Knapp. Sines and Cosines of Angles in Arithmetic Progression. *Mathematics Magazine*, 82(5):371–372, 2009.

[158] W. Kohn and L. J. Sham. Self-consistent equations including exchange and correlation effects. *Phys. Rev.*, 140:A1133, 1965.

[159] M. Korth. Third-generation hydrogen-bonding corrections for semiempirical qm methods and force fields. *J. Chem. Theory Comput.*, 6, 2010.

[160] D. A. Koster, K. Palle, E. S. M. Bot, M.-A. Bjonrsti, and N. H. Dekker. Antitumour drugs impede DNA uncoiling by topoisomerase I. *Nature*, 448:213–217, 2007.

[161] P.R. Kramer and R.R. Sinden. Measurement of unrestrained negative supercoiling and topological domain size in living human cells. *Biochemistry*, 36:3151–3158, 1997.

[162] P. J. M. Laarhoven and E. H. L. Aarts. *Simulated annealing: theory and applications*. Kluwer Academic Publishers Norwell, MA, USA, 1987.

[163] A. Laio and M. Parrinello. Escaping Free-Energy Minima. *Proc. Natl. Acad. Sci. U. S. A.*, 99:12562, 2002.

[164] P. Langevin. Sur la théorie du mouvement brownien. *C. R. Acad. Sci. Paris*, 146:530–533, 1908.

[165] R. Lavery and M. Pasi. Curves+ User Guide V3.0, 2016. [Online, URL: `http://curvesplus.bsc.es/static/doc/Curves+new.pdf`; accessed 5-October-2019].

[166] R. Lavery and H. Sklenar. The Definition of Generalized Helicoidal Parameters and of Axis Curvature for Irregular Nucleic Acids. *J. Biomol. Struct. Dyn.*, 6(1):63–91, 1988.

[167] R. Lavery and H. Sklenar. Defining the Structure of Irregular Nucleic Acids: Conventions and Principles. *J. Biomol. Struct. Dyn.*, 6(4):655–667, 1989.

[168] A. R. Leach. *Molecular Modelling: Principles and Applications*. Addison Wesley Longman Ltd., 1st edition, 1996.

[169] P. Leder and M. W. Nirenberg. RNA codewords and protein synthesis, III. On the nucleotide sequence of a cysteine and a leucine RNA codeword. *Proc. Natl. Acad. Sci. U. S. A.*, 52(6):1521, 1964.

[170] P. M. Lee. *Bayesian Statistics: An Introduction*. Wiley, 4th edition, 2012.

[171] A. L. Lehninger. *Biochemistry*. Worth, 1975.

[172] T. Lelièvre, M. Rousset, and G. Stoltz. Computation of free energy profiles with parallel adaptive dynamics. *J. Chem. Phys.*, 126:134111, 2007.

[173] T. Lelièvre, M. Rousset, and G. Stoltz. *Free Energy Computations: A Mathematical Perspective*. Imperial College Press, London, 2010.

[174] L.S. Lerman. Structural considerations in the interaction of DNA and acridines. *J. Mol. Biol.*, 3(1):18–30, 1961.

[175] A. Lesage, T. Lelièvre, G. Stoltz, and J. Hénin. Smoothed Biasing Forces Yield Unbiased Free Energies with the Extended-System Adaptive Biasing Force Method. *J. Phys. Chem. B*, 121(15):3676–3685, 2017.

[176] A. G. W. Leslie, S. Arnott, R. Chandrasekaran, and R. L. Ratliff. Polymorphism of DNA double helices. *Journal of Molecular Biology*, 143(1):49–72, 1980.

[177] J. T. Lett. Damage to DNA and chromatin structure. *Prog. Nucleic Acid Res. Mol. Biol.*, 39:305–352, 1990.

[178] T. Lindahl. Instability and decay of the primary structure of DNA. *Nature*, 22(362):709–715, 1993.

[179] M. B. Lion. Search for a mechanism for the increased sensitivity of 5-bromouracil-substituted DNA to ultraviolet light. *Biochim. Biophys. Acta*, 155:505–520, 1968.

[180] H. Lodish, C. A. Kaiser, A. Bretscher, A. Amon, A. Berk, M. Krieger, H. Ploegh, and M. P. Scott. *Molecular Cell Biology*. W. H. Freeman and Company, 7th edition, 2013.

[181] A. A. Lucas. Rosetta Stone of the genetic language. *Int. J. Quant. Chem.*, 90:1491–1504, 2002.

[182] A. A. Lucas. A-DNA and B-DNA: Comparing Their Historical X-ray Fiber Diffaction Images. *J. Chem. Educ.*, 85(5):737–743, 2008.

[183] A. A. Lucas and P. Lambin. Diffraction by DNA, carbon nanotubes and other helical nanostructures. *Rep. Prog. Phys.*, 68:1181–1249, 2005.

[184] A. A. Lucas, P. Lambin, R. Mairesse, and M. Mathor. Revealing the Backbone Structure of B-DNA from Laser Optical Simulations of Its X-ray Diffraction Diagram. *J. Chem. Educ.*, 76(3):378–383, 1999.

[185] R. Luo, L. David, and M. K. Gilson. Accelerated Poisson-Boltzmann calculations for static and dynamic systems. *J. Comp. Chem.*, 23:1244–1253, 2002.

[186] M. R. Machado, E. E. Barrera, F. Klein, M. Sóñora, S. Silva, and S. Pantano. The SIRAH force field 2.0: Altius, Fortius, Citius. *J. Chem. Theory. Comput.*, 15(4):2719–2733, 2019.

[187] M. R. Machado, P. D. Dans, and S. Pantano. A hybrid all-atom/coarse grain model for multiscale simulations of DNA. *Phys. Chem. Chem. Phys.*, 13:18134–18144, 2011.

[188] M. R. Machado, H. C. González, and S. Pantano. MD Simulations of Viruslike Particles with Supra CG Solvation Affordable to Desktop Computers. *J. Chem. Theory Comput.*, 13(10):5106–5116, 2017.

[189] M. R. Machado and S. Pantano. Exploring LacI-DNA Dynamics by Multiscale Simulations Using the SIRAH Force Field. *J. Chem. Theory Comput.*, 11(10):5012–5023, 2015.

[190] M. R. Machado, A. Zeida, L. Darré, and S. Pantano. From Quantum to Subcellular Scales: Multiscale Simulation Approaches and the SIRAH Force Field. *Interface Focus*, page 20180085, 2019.

[191] B. Maddox. *Rosalind Franklin: The Dark Lady of DNA.* HarperCollins, 2002.

[192] J. D. Madura, M. E. Davis, M. K. Gilson, R. C. Wade, B. A. Luty, and J. A. McCammon. Biological applications of electrostatic calculations and Brownian dynamics simulations. *Ref. Comp. Chem.*, 5:229–267, 1994.

[193] M. W. Mahoney and W. L. Jorgensen. A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions. *J. Chem. Phys.*, 112(20):8910–8922, 2000.

[194] F. Mandl. *Statistical Physics*. John Wiley & Sons, 2nd edition, 1988.

[195] A. Marathe and M. Bansal. An ensemble of B-DNA dinucleotide geometries lead to characteristic nucleosomal DNA structure and provide plasticity required for gene expression. *BMS Struct. Biol.*, 11(1).

[196] S. J. Marrink, H. J. Risselada, S. Yefimov, D. P. Tieleman, and A. H. de Vries. The MARTINI Force Field: Coarse Grained Model for Biomolecular Simulations. *J. Phys. Chem. B*, 111, 2007.

[197] Wolfram MathWorld. Frame. [Online; accessed 14-August-2019].

[198] T. Meissner and F. J. Wentz. The Complex Dielectric Constant of Pure and Sea Water from Microwave Satellite Observations. *IEEE Transactions on Geoscience and Remote Sensing*, 42(9):1836, 2004.

[199] E. Merzbacher. *Quantum Mechanics*. John Wiley & Sons, Inc., 3rd edition, 2004.

[200] Y. Miao, V. A. Feher, and J. A. McCammon. Gaussian Accelerated Molecular Dynamics: Unconstrained Enhanced Sampling and Free Energy Calculation. *J. Chem. Theory Comput.*, 11(8):3584–3595, 2015.

[201] Y. Miao, W. Sinko, L. Pierce, D. Bucher, R. C. Walker, and J. A. McCammon. Improved Reweighting of Accelerated Molecular Dynamics Simulations for Free Energy Calculation. *J. Chem. Theory Comput.*, 10:2677–2689, 2014.

[202] F. Miescher. His W et al (eds): Die histochemischen und physiologischen arbeiten von Friedrich Miescher - aus dem wissenschaftlichen briefwechsel von F. Miescher, 1869.

[203] M. Mihailescu and M. K. Gilson. On the Theory of Noncovalent Binding. *Biophys J.*, 87:23–36, 2004.

[204] B. Milholland, X. Dong, L. Zhang, X. Hao, Y. Suh, and G. Vijg. Differences between germline and somatic mutation rates in humans and mice. *Nature Comms.*, 8(85183):1–8.

[205] G. Minotti, P. Menna, E. Salvatorelli, G. Cairo, and L. Gianni. Anthracyclines: molecular advances and pharmacologic developments in antitumor activity and cardiotoxicity. *Pharmacol Rev.*, 56(2):185–229, 2004.

[206] A. V. Morozov, T. Kortemme, K. Tsemekhman, and D. Baker. Close agreement between the orientation dependence of hydrogen bonds observed in protein structures and quantum mechanical calculations. *Proc. Natl. Acad. Sci. USA*, 101(18), 2004.

[207] A. Motte. *Newton's Principia : the mathematical principles of natural philosophy*. 1729.

[208] A. Mukherjee, R. Lavery, B. Bagchi, and J. T. Hynes. On the molecular mechanism of drug intercalation into DNA: A simulation study of the intercalation pathway, free energy, and DNA structural changes. *J. Am. Chem. Soc.*, 130(30):9747–9755, 2008.

[209] R. S. Mulliken. Electronic Population Analysis on LCAO-MO Molecular Wave Functions. I. *J. Chem. Phys.*, 23(10):1833–1840, 1955.

[210] S. Neidle. *Principles of Nucleic Acid Structure*. Academic Press, 2008.

[211] P. C. Nelson. DNA elasticity. [Online, URL: `https://www.sas.upenn.edu/~pcn/html-physics/mcgraw2/mcglatex.html`; accessed 19-August-2019].

[212] I. Newton. *Philosophiæ Naturalis Principia Mathematica*. Londini, Jussu Societatis Regiæ ac Typis Josephi Streater. Prostat apud plures Bibliopolas., 1687.

[213] S. Nosé. A molecular dynamics method for simulation in the canonical ensemble. *Mol. Phys.*, 52:255–268, 1984.

[214] S. Nosé. A unified formulation of the constant temperature molecular dynamics method. *J. Chem. Phys.*, 81:511–519, 1984.

[215] Royal Society of Chemistry. ChemSpider — Daunorubicin. [Online, URL: `https://www.chemspider.com/Chemical-Structure.28163.html?rid=f2e92027-69b1-4eaa-abe0-9eb9ed8ad905`; accessed 23-September-2019].

[216] Royal Society of Chemistry. ChemSpider — Doxorubicin. [Online, URL: `https://www.chemspider.com/Chemical-Structure.29400.html?rid=2671ec59-ae87-4714-a7bc-5b2a52d6a18b`; accessed 23-September-2019].

[217] Royal Society of Chemistry. ChemSpider — Epirubicin. [Online, URL: `https://www.chemspider.com/Chemical-Structure.38201.html?rid=e6b57261-76cc-4bda-9f47-fa27cf273d01`; accessed 23-September-2019].

[218] Royal Society of Chemistry. ChemSpider — Idarubicin. [Online, URL: `https://www.chemspider.com/Chemical-Structure.39117.html?rid=780ae5d3-f829-4441-a622-a3d1b1b28f6e`; accessed 23-September-2019].

[219] R. Okazaki, T. Okazaki, K. Sakabe, K. Sugimoto, and A. Sugino. Mechanism of DNA chain growth. I. Possible discontinuity and unusual secondary structure of newly synthesized chains. *PNAS*, 59(2):598–605, 1968.

[220] A. Onufriev. The generalised Born model: its foundation, applications, and limitations, 2010. [Available online, URL: `http://people.cs.vt.edu/~onufriev/PUBLICATIONS/gbreview.pdf`; accessed 9-May-2019].

[221] A. Onufriev, D. Bashford, and D. Case. Exploring protein native states and large-scale conformational changes with a modified generalized Born model. *Proteins*, 55:383–394, 2004.

[222] A. Onufriev, D. A. Case, and D. Bashford. Effective Born radii in the generalized Born approximation: The importance of being perfect. *Journal of Computational Chemistry*, 23(14):1297–1304, 2002.

[223] T. E. Ouldridge. *Coarse-grained modelling of DNA and DNA self-assembly*. Springer, Berlin, 2012.

[224] T. E. Ouldridge, A. A. Louis, and J. P. K. Doye. Structural, mechanical, and thermodynamic properties of a coarse-grained DNA model. *J. Chem. Phys.*, 134, 2011.

[225] R. Owczarzy, Y. You, B. G. Moreira, J. A. Manthey, L. Huang, M. A. Behlke, and J. A. Walder. Effects of sodium ions on DNA duplex oligomers: improved predictions of melting temperatures. *Biochemistry*, 43(12).

[226] T. Paramanathan, I. Vladescu, M. J. McCauley, I. Rouzina, and M. C. Williams. Force spectroscopy reveals the DNA structural dynamics that govern the slow binding of Actinomycin D. *Nucleic Acid Res.*, 40:4925–4932, 2012.

[227] L. Pauling. The Nature of the Chemical Bond. IV. The Energy of Single Bonds and the Relative Electronegativity of Atoms. *J. Amer. Chem. Soc.*, 54(9), 1932.

[228] M. C. Payne, M. P. Teter, D. C. Allan, T. A. Arias, and J. D. Joannopoulos. Iterative minimization techniques for ab initio total-energy calculations – molecular-dynamics and conjugate gradients. *Rev. Mod. Phys.*, 64:1045–1097, 1992.

[229] J. B. Pendry. Reliability factors for LEED calculations. *J. Phys. C: Solid St. Phys.*, 13:937–944, 1980.

[230] E. Pennisi. ENCODE Project Writes Eulogy for Junk DNA. *Science*, 337(6099):1159–1161, 2012.

[231] B. G. Pfrommer, M. Cote, S. G. Louie, and M. L. Cohen. Relaxation of crystals with the quasi-Newton method. *J. Comput. Phys.*, 131:233–240, 1997.

[232] D. R. Phillips, A. Di Marco, and F. Zunino. The interaction of daunomycin with polydeoxynucleotides. *Eur. J. Biochem.*, 85:487–492, 1978.

[233] R. I. Porter. *Further elementary analysis*. G. Bell & Sons Ltd., 4th edition, 1970.

[234] V. V. Prasolov and A. B. Sossinsky. *Knots, links, braids and 3-manifolds*, volume 154 of *Translations of Mathematical Monographs*. American Mathematical Society, Providence, RI, 1997. An introduction to the new invariants in low-dimensional topology, Translated from the Russian manuscript by Sossinsky [Sosinskiĭ].

[235] E. Prince, editor. *International Table for Crystallography, Volume C: Mathematical, Physical and Chemical Tables*. Kluwer Academic Publishers, 3rd edition, 2004.

[236] M. I. J. Probert. Improved algorithm for geometry optimisation using damped molecular dynamics. *J. Comp. Phys.*, 191:130–146, 2003.

[237] D. C. Rapaport. *The Art of Molecular Dynamics Simulation*. Cambridge University Press, 2nd edition, 2004.

[238] A. N. Real. *Molecular dynamics simulations of AT-rich DNA and DNA-spermine complexes*. PhD thesis, University of York, 2001.

[239] M. Reuter and D. T. F. Dryden. The kinetics of YOYO-1 intercalation into single molecules of double-stranded DNA. *Biochemical and Biophysical Research Communications*, 403(2):225–229, 2010.

[240] D. T. Richens. *The Chemistry of Aqua Ions: Synthesis, Structure and Reactivity: a Tour Through the Periodic Table of the Elements*. Wiley & Sons, 1997.

[241] P. C. Riedi. *Thermal Physics*. Oxford Science Publications, 2nd edition, 1987.

[242] P. A. Riley. Free radicals in biology: oxidative stress and the effects of ionizing radiation. *Int. J. Radiat. Biol.*, 65:27–33, 1994.

[243] S. F. A. Rizvi, S. Tariq, and M. Mehdi. Anthracyclines: Mechanism of Action, Classification, Pharmacokinetics and Future – A Mini Review. *Int. J. Biotech & Bioeng.*, 4(4):81–85, 2018.

[244] O. Rodrigues. Des lois géometriques qui regissent les déplacements d'un système solide dans l'espace, et de la variation des coordonnées provenant de ces déplacement considérées indépendant des causes qui peuvent les produire. *J. Math. Pures Appl.*, 5:380–440, 1840.

[245] W. Saenger. *Principles of Nucleic Acid Structure*. Springer-Verlag, 1984.

[246] W. Saenger and D. Suck. The relationship between hydrogen bonding and base stacking in crystalline 4-thiouridine derivatives. *Eur. J. Biochem.*, 32:473–478, 1973.

[247] M. Sayar, B. Avşaroğlu, and A. Kabakçıoğlu. Twist-writhe partitioning in a coarse-grained DNA minicircle model. *Phys. Rev. E*, 81:041916, 2010.

[248] Aylwyn Scally. The mutation rate in human evolution and demographic inference. *Current Opinion in Genetics & Development*, 41:36–43, 2016.

[249] M. Scarsi, J. Apostolakis, and A. Caflisch. Continuum electrostatic energies of macromolecules in aqueous solutions. *J. Phys. Chem. A*, 101:8098–8106, 1997.

[250] J. Schlitter, M. Engels, and P. Kruger. Targeted molecular dynamics: A new approach for searching pathways of conformational transitions. *Journal of Molecular Graphics*, 12(2):84–89, 1994.

[251] J. Schlitter, M. Engels, P. Krüger, E. Jacoby, and A. Wollmer. Targeted Molecular Dynamics Simulation of Conformational Change-Application to the T ↔ R Transition in Insulin. *Molecular Simulation*, 10(2-6):291–308, 1993.

[252] E. Schrödinger. *What is Life?* Cambridge University Press, 1944.

[253] J.B. Schvartzman and A. Stasiak. A topological view of the replicon. *EMBO Rep.*, 5:256–261, 2004.

[254] J. W. Shepherd. Classical and Quantum Simulations of DNA/spermine systems. Master's thesis, University of York, 2015.

[255] J. W. Shepherd. *Manipulating DNA with Magneto-Optical Tweezers and Multiscale Simulations*. PhD thesis, University of York, 2018.

[256] M. J. Shoura, I. Gabdank, L. Hansen, J. Merker, J. Gotlib, S. D. Levene, and A. Z. Fire. Intricate and Cell Type-Specific Populations of Endogenous Circular DNA (eccDNA) in *Caenorhabditis elegans* and *Homo sapiens*. *G3: GENES, GENOMES, GENETICS*, 7(10):3295–3303, 2017.

[257] G. Sigalov, P. Scheffel, and A. Onufriev. Incorporating variable dielectric environments into the generalized Born model. *J. Chem. Phys.*, 122(9), 2005.

[258] B. K. Sinha. Free radicals in anticancer drug pharmacology. *Chem. Biol. Interact.*, 69:293–317, 1989.

[259] L. Sjögren. *Lecture notes on stochastic processes*. University of Gothenburg, Sweden.

[260] R. E. Smallman. *Modern Physical Metallurgy*. Butterworth & Co (Publishers) Ltd., 3rd edition, 1970.

[261] P. J. Stephens, C. D. Greenman, B. Fu, F. Yang, G. R. Bignell, L. J. Mudie, E. D. Pleasance, K. W. Lau, D. Beare, L. A. Stebbings, S. McLaren, M. Lin, D. J. McBride, I. Varela, S. Nik-Zainal, C. Leroy, M. Jia, A. Menzies, A. P. Butler, J. W. Teague, M. A. Quail, J. Burton, H. Swerdlow, N. P. Carter, L. A. Mosberger, C. Iacobuzio-Donahue, G. A. Follows, A. R. Green, A. M. Flanagan, M. R. Stratton, P. A. Futreal, and P. J. Campbell. Massive Genomic Rearrangement Acquired in a Single Catastrophic Event during Cancer Development. *Cell*, 144(1), 2011.

[262] E. Stofer and R Lavery. Measuring the Geometry of DNA Grooves. *Biopolymers*, 34:337–346, 1994.

[263] G. Stokes. The Effect of Polyamines on the Structure of DNA. BSc thesis, University of York, 2016.

[264] K. Sugimoto, T. Okazaki, and R. Okazaki. Mechanism of DNA chain growth, II. Accumulation of newly synthesized short chains in E. coli infected with ligase-defective T4 phages. 60(4):1356–1362, 1968.

[265] D. Svozil, P. Hobza, and J. Šponer. Comparison of Intrinsic Stacking Energies of Ten Unique Dinucleotide Steps in A-RNA and B-DNA Duplexes. Can We Determine Correct Order of Stability by Quantum-Chemical Calculations? *J. Phys. Chem. B*, 114(2).

[266] D. Swigon. The Mathematics of DNA. In *Mathematics of DNA Structure, Function and Interactions*, pages 293–320. Springer, 2009.

[267] W. C. Swope, H. C. Anderson, P. H. Berens, and K. R. Wilson. A Computer Simulation Method for the Calculation of Equilibrium Constants for the Formation of Physical Clusters of Molecules: Application to Small Water Clusters. *J. Chem. Phys.*, 76:637–649, 1982.

[268] M. Tanaka and S. Yoshida. Mechanism of the inhibition of calf thymus DNA polymerases α and β by daunomycin and adriamycin. *J. Biochem. (Tokyo)*, 87:911–918, 1980.

[269] R. Téoule. Radiation-induced DNA damage and its repair. *Int. J. Radiat. Biol.*, 51:573–589, 1987.

[270] The Nobel Foundation. Nobel Prizes and Laureates: The Nobel Prize in Physiology or Medicine 1962. Online, URL: `https://www.nobelprize.org/nobel_prizes/medicine/laureates/1962/`; accessed 25-Oct-2017.

[271] C. M. Thomas and D. Summers. Bacterial plasmids. In *Encyclopedia of Life Sciences*. Wiley, 2008.

[272] N. W. Timoféeff-Ressovsky, K. G. Zimmer, and M. Delbrück. Über die Natur der Genmutation und der Genstruktur. *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Fachgruppe VI*, 1(13):189–245, 1935.

[273] G. M. Torrie and J. P. Valleau. Nonphysical Sampling Distributions in Monte Carlo Free-Energy Estimation: Umbrella Sampling. *J. Chem. Phys.*, 23:187–199, 1977.

[274] A Travers and G. Muskhelishvili. DNA supercoiling - a global transcriptional regulator for enterobacterial growth? *Nat. Rev. Microbiol.*, 3:157–169, 2005.

[275] M. E. Tuckerman. *Statistical Mechanics: Theory and Molecular Simulation*. Oxford University Press, 2010.

[276] P. Šulc, F. Romano, T. E. Ouldridge, L. Rovigatti, J. P. K. Doye, and A. A. Louis. Sequence-dependent thermodynamics of a coarse-grained DNA model. *J. Chem. Phys.*, 137, 2012.

[277] W. F. van Gunsteren and H. J. C. Berendsen. Computer Simulation of Molecular Dynamics: Methodology, Applications, and Perspectives in Chemistry. *Angewandte Chemie International Edition in English*, 29(9):992–1023, 1990.

[278] M. A. Van Hove, S. Y. Tong, and M. H. Elconin. Surface structure refinements of 2H-MoS_2, 2H-$NbSe_2$ and W(100)p(2×1)-O via new reliability factors for surface crystallography. *Surface Science*, 64:85–95, 1977.

[279] Y. S. Vassetzky, G. C. Alghisi, and S. M. Gasser. DNA topoisomerase-II mutations and resistance to antitumor drugs. *Bioessays*, 17:767–774, 1995.

[280] L. Verlet. Computer “Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Phys. Rev.*, 159:98–103, 1967.

[281] S. N. Vinogradov and R. H. Linnell. *Hydrogen Bonding*. Van Nostrand Reinhold, New York, 1971.

[282] D. Voet and A. Rich. The crystal structures of purines, pyrimidines and their intermolecular complexes. *Prog. Nucleic Acid Res. Mol. Biol.*, 10:186–265, 1970.

[283] M. Volkova and III. Raymond Russell. Anthracycline Cardiotoxicity: Prevalence, Pathogenesis and Treatment. *Curr. Cardiol. Rev.*, 7(4):214–220, 2011.

[284] C. von Sonntag. *The Chemical Basis of Radiation Biology*. Taylor & Francis, London, UK, 1987.

[285] J. F. Ward. DNA damage produced by ionizing radiation in mammalian cells: identities, mechanisms of formation and reparability. *Prog. Nucleic Acid Res. Mol. Biol.*, 35:98–125, 1988.

[286] J. F. Ward. The yield of DNA double-strand breaks produced intracellularly by ionizing radiation: a review. *Int. J. Radiat. Biol.*, 57:1141–1150, 1990.

[287] M. Waring. Variation of the supercoils in closed circular DNA by binding of antibiotics and drugs: Evidence for molecular models involving intercalation. *Journal of Molecular Biology*, 54(2):247–279, 1970.

[288] Water Science School, U.S. Geological Survey. The Water in You: Water and the Human Body, 2019. [Online, URL: `https://www.usgs.gov/special-topic/water-science-school/science/water-you-water-and-human-body?qt-science_center_objects=0#qt-science_center_objects`; accessed 8-May-2019].

[289] J. D. Watson and F. H. C. Crick. Molecular structure of nucleic acids. *Nature*, 171:737–738, 1953.

[290] WebMD. Idarubicin: dosing & uses, 2019. [Online; accessed 9-December-2019].

[291] S. Weinberg. *Lectures on Quantum Mechanics*. Cambridge University Press, 1st edition, 2013.

[292] S.J. Weiner, P.A. Kollman, D.T. Nguyen, and D.A. Case. An all atom force field for simulations of proteins and nucleic acids. *Journal of Computational Chemistry*, 7(2):230–252, 1986.

[293] J. H. White. Self-linking and the gauss integral in higher dimensions. *Am. J. Math.*, 91(3):693–728, 1969.

[294] Wikibooks. Structural biochemistry/chemical bonding/hydrogen bonds — wikibooks, the free textbook project, 2019. [Online, URL: `https://en.wikibooks.org/w/index.php?title=Structural_Biochemistry/Chemical_Bonding/Hydrogen_bonds&oldid=3528197`; accessed 2-April-2019].

[295] Wikipedia contributors. Ribbon (mathematics) — Wikipedia, the free encyclopedia, 2017. [Online; accessed 14-August-2019].

[296] Wikipedia contributors. Mulliken population analysis — Wikipedia, the free encyclopedia, 2018. [Online, URL: https://en.wikipedia.org/w/index.php?title=Mulliken_population_analysis&oldid=867231356; accessed 1-April-2019].

[297] Wikipedia contributors. Annealing (metallurgy) — Wikipedia, the free encyclopedia, 2019. [Online, URL: https://en.wikipedia.org/w/index.php?title=Annealing_(metallurgy)&oldid=890463949; accessed 2-May-2019].

[298] Wikipedia contributors. Hopf link — Wikipedia, the free encyclopedia, 2019. [Online; accessed 14-August-2019].

[299] Wikipedia contributors. Ionic strength — Wikipedia, the free encyclopedia, 2019. [Online; accessed 21-December-2019].

[300] Wikipedia contributors. Linking number — Wikipedia, the free encyclopedia, 2019. [Online; accessed 14-August-2019].

[301] Wikipedia contributors. Okazaki fragments — Wikipedia, the free encyclopedia, 2019. [Online; accessed 22-November-2019].

[302] Wikipedia contributors. Persistence length — Wikipedia, the free encyclopedia, 2019. [Online; accessed 19-August-2019].

[303] Wikipedia contributors. Transcription (biology) — Wikipedia, the free encyclopedia, 2019. [Online; accessed 22-November-2019].

[304] Wikipedia contributors. Verlet integration — Wikipedia, the free encyclopedia, 2019. [Online, URL: https://en.wikipedia.org/w/index.php?title=Verlet_integration&oldid=895861512; accessed 14-May-2019].

[305] M. H. F. Wilkins, A. R. Stokes, and H. R. Wilson. Molecular structure of deoxypentose nucleic acids. *Nature*, 171:738–740, 1953.

[306] R. K. Williams. *Molecular conformational studies of deoxyribonucleic acid by potential energy minimisation with normal mode analysis*. PhD thesis, University of Keele, 1990.

[307] H. J. Woo and B. Roux. Calculation of absolute protein-ligand binding free energy from computer simulations. *Proc. Natl. Acad. Sci. USA*, 102:6825–6830, 2005.

[308] S. Wu, K. M. Turner, N. Nguyen, R. Raviram, M. Erb, J. Santini, J. Luebeck, U. Rajkumar, Y. Diao, B. Li, W. Zhang, N. Jameson, M. R. Corces, J. M. Granja, X. Chen, C. Coruh, A. Abnousi, J. Houston, Z. Ye, R. Hu, M. Yu, H. Kim, J. A. Law, R. G. W. Verhaak, M. Hu, F. B. Furnari, H. Y. Chang, B. Ren, V. Bafna, and P. S. Mischel. Circular ecDNA promotes accessible chromatin and high oncogene expression. *Nature*, 575:699–703, 2019.

[309] N. B.-y. Yee. The Structure and Dynamics of DNA-cation Systems. Master's thesis, University of York, 2015.

[310] F. Zunino, R. Gambetta, and A. Di Marco. The inhibition *in vitro* of DNA polymerase and RNA polymerases by daunomycin and adriamycin. *Biochem. Pharmacol.*, 24:309–311, 1975.