COMPUTER SIMULATIONS OF PEPTIDE NUCLEIC ACID UNDER EXTERNAL FORCE

JACK GOODMAN

A thesis submitted in partial fulfilment of the requirements of the University of the West of England, Bristol for the degree of Doctor of Philosophy

This research was carried out in collaboration with GKN Aerospace

Faculty of Health and Applied Science, University of the West of England, Bristol

August 2023

Contents

Abstract

The research question, “Can molecular dynamics be used to assess and screen the single-molecular binding properties of a candidate bioadhesive?" is answered in this thesis using molecular dynamics simulations. A ‘nearest-neighbour’ model was produced that related the candidate bioadhesive peptide nucleic acid’s (PNA’s) primary sequences with equilibrium binding enthalpies and could predict experimental binding enthalpies with an accuracy of 8.7%. In addition, the relationship between PNA rupture forces and loading rates at high loading rates was established for two distinct loading axes, and internal cohesive energies between two bound strands under external force were expressed as a function of displacements along unbinding coordinates. In addition, a novel coarse-grained model for ds-PNA that is natively integrable into other related coarse-grained models was produced and found capable of replicating both experimental structures and rupture forces as determined by all-atom models.

This thesis presents the first time that the relationship between primary sequence and PNA binding energies have been derived. In addition, it presents the first time that the relationship between rupture force and loading rate has been established for PNA and the first time this relationship has been expressed in terms of inter-strand energies for PNA-containing nucleic acids. This thesis is of general interest for the development of a PNA bioadhesive by providing the single-molecular framework against which macroscopic observables can be interpreted and compared. In addition, the methods presented are broadly available and do not require specialist equipment, making them of interest to developing single-molecular interpretations of bioadhesive properties without significant financial or experimental investment.

List of figures

List of tables

List of equations

1-1: $\mathrm{G} = \mathrm{U} + \mathrm{PV} - \mathrm{TS}$

1-2: $F = \frac{k_B T}{x^{\ddagger}} \ln \frac{r x^{\ddagger}}{k_B T k}$

1-3: $-\Delta G^{\ddagger} = k_B T \ln \frac{kh}{k_B T}$

1-4: $U = \sum_i E_i p_i$

1-5: $p_i \propto e^{\frac{-E_i}{k_B T}}$

1-6: $p_i = \frac{e^{\frac{-E_i}{k_B T}}}{Z} = \frac{e^{\frac{-E_i}{k_B T}}}{\sum_i e^{\frac{-E_i}{k_B T}}} = \frac{e^{\frac{-E_i}{k_B T}}}{\int e^{\frac{-E_i}{k_B T}} d\vec{q}}$

1-7: $G = -k_B T \ln Z$

1-8: $\Delta G_{ij} = -k_B T \ln \frac{Z_i}{Z_j} = -k_B T \ln \frac{\left(\int e^{\frac{-E_i}{k_B T}} d\vec{q} \right)_i}{\left(\int e^{\frac{-E_i}{k_B T}} d\vec{q} \right)_j}$

1-9: $K = \frac{\alpha}{\left(\frac{C}{n}\right)^{n-1} (1-\alpha)^n}$

1-10: $ln\, K = -\frac{\Delta H}{RT} + \frac{\Delta S}{R}$

1-11: $\Delta\chi = a\Delta x_{init.GC} + b\Delta x_{init.AT} + \sum_i j_i \Delta x_i + c\Delta x_{sym.}$

1-12: $(\sum j_i \Delta x_i + \Delta x_{init})_E = a(\sum j_i \Delta x_i + \Delta x_{init})_C + b$

2-1: $V = \sum_{bonds} k(b - b_0)^2 + \sum_{angles} k(\theta - \theta_0)^2 + \sum_{torsions} k[1 + \cos(n\phi - \delta)] + \varepsilon_{nonbonded}$

2-2a: $\boldsymbol{v}\left(t + \frac{1}{2}\Delta t\right) = \boldsymbol{v}\left(t - \frac{1}{2}\Delta t\right) + \frac{\Delta t}{m}\boldsymbol{F}(t)$

2-2b: $\boldsymbol{r}(t + \Delta t) = \boldsymbol{r}(t) + \Delta t \boldsymbol{v}\left(t + \frac{1}{2}\Delta \mathrm{t}\right)$

2-3: $\frac{1}{2}nkT = E_{\text{kinetic}} = \frac{1}{2}\sum_i m_i v_i^2$

2-4: $\frac{dT}{dt} = \frac{T_0 - T}{t}$

2-5: $\frac{dP}{dt} = \frac{P_0 - P}{t}$

2-6: $V_{LJ} = 4\epsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right]$

2-7: $V_{el} = \frac{q_i q_j}{4\pi\varepsilon_0 r}$

2-8: $p_{\xi\,\text{unbiased}} = \frac{\sum_i g_i^{-1} h_i}{\sum_j n_j g_j^{-1} \exp\left(\frac{1}{k_B T} w_j - f_j\right)}$

2-9: $p_\xi = \frac{e^{-G_\xi + g_\xi}}{Z}$

2-10: $\Delta G = -k_B T \ln \langle e^{-\frac{E_B - E_A}{k_B T}} \rangle_A$

2-11: $\Delta G = \int_{\lambda=0}^{\lambda=1} \frac{\partial G}{\partial \lambda} d\lambda$

2-12: $\sum_{i=1}^{n_i} \frac{1}{1+e^{\left(\ln\left(\frac{n_i}{n_j}\right) + \frac{\Delta E_{ij} - \Delta G}{k_B T}\right)}} - \sum_{j=1}^{n_j} \frac{1}{1+e^{\left(\ln\left(\frac{n_j}{n_i}\right) + \frac{\Delta E_{ji} - \Delta G}{k_B T}\right)}} = 0$

2-13: $\Delta G_{ij} = -k_B T \ln \frac{\langle M e^{\frac{-E_i}{k_B T}} \rangle_j}{\langle M e^{\frac{-E_j}{k_B T}} \rangle_i}$

2-14: $\Delta G = \Delta E_{MM} + \Delta G_{solv} - T\Delta S$

2-15: $\Delta E_{MM} = \Delta E_{bonded} + \Delta E_{non-bonded}$

2-16: $\Delta G_{solv} = \Delta G_{polar} + \Delta G_{non-polar}$

2-17: $\Delta G_{non-polar} = mSA + b$

2-18: $\Delta G_{polar} = \left(1 - \frac{1}{\epsilon}\right)\frac{1}{2}\sum_{ij} \frac{q_i q_j}{f_{GB}}$

4-1: $K = \frac{\alpha}{2C(1-\alpha)^2}$

4-2: $\boldsymbol{C} = \left(\frac{1}{k_B T}\boldsymbol{H}\right)^{-1}$

4-3: $\Delta G^{\circ} = \Delta G - RT \ln\left(\frac{V_{\mathrm{com}}}{V^{\circ}}\right)$

4-4: $\underbrace{\begin{pmatrix} \Delta\chi_1 \\ \vdots \\ \vdots \\ \vdots \\ \Delta\chi_N \end{pmatrix}}_{\boldsymbol{\chi}} = \underbrace{\begin{pmatrix} j_1^1 & \cdots & j_v^1 \\ \vdots & & \vdots \\ \vdots & \ddots & \vdots \\ \vdots & & \vdots \\ j_1^N & \cdots & j_v^N \end{pmatrix}}_{\boldsymbol{S}} \cdot \underbrace{\begin{pmatrix} \Delta\Delta\chi_1 \\ \vdots \\ \Delta\Delta\chi_{v-2} \\ \Delta\chi_{init.GC} \\ \Delta\chi_{init.AT} \end{pmatrix}}_{\boldsymbol{\varepsilon}}$

4-5: $\boldsymbol{\sigma^{-1} \cdot \chi = \sigma^{-1} \cdot S \cdot \varepsilon}$

6-1: $V_{RB} = \frac{1}{2} k \frac{(\cos\theta - \cos\theta_0)^2}{\sin^2\theta}$

6-2: $\Delta G_{s_1 \to s_2} = \Delta G_{s_1 \to \emptyset} - \Delta G_{s_2 \to \emptyset}$

6-3: $V_{ij} = D\left(1 - e^{-\beta(r - r_0)}\right)^2$

Unit Conversion: 1 kcal mol^{-1} = 4.184 kJ mol^{-1} = 1.689 $\mathrm{k_B T}$

Abbreviations

AFM: Atomic force microscope.

SMD: Steered molecular dynamics.

PNA: Peptide nucleic acid.

RNA: Ribonucleic acid.

DNA: Deoxyribonucleic acid.

PMO: Phosphorodiamidate morpholino.

LNA: Locked nucleic acid.

AWH: Accelerated weight histogram.

US: Umbrella sampling.

BAR: Bennett acceptance.

FEP: Free energy perturbation.

TI: Thermodynamic integration.

RMSD: Root mean square atomic deviation.

MM-PB(GB)SA: Molecular mechanics Poisson-Boltzmann (generalised Born) surface area.

STA: Single trajectory approach.

MTA: Multiple trajectory approach.

QH: Quasi-harmonic.

CHARMM: Chemistry at Harvard Macromolecular Mechanics.

LINCS: LINear Constraint Solver.

ds-*x*NA: Double-stranded x-nucleic acid.

ss-*x*NA: Single-stranded x-nucleic acid.

For clarity, ds-nucleic acids are often written with their complementary strand implied. For example, ds-PNA GTAGATCACT is shorthand for ds-PNA GTAGATCACT:CATCTAGTGA. A colon : is used to separate nucleobases of opposing strands whereas nucleobases written in sequence are not separated, i.e., G:C refers to a ds-xNA guanine-cytosine base pair, whereas GC refers to an ss-xNA sequence consisting of 2 base pairs, guanine and then cytosine, in sequence.

Chapter 1: Introduction

1.1 Bioadhesion

1.1.1 Methods and examples of bioadhesion

Adhesion is an association between structures such as two surfaces typically mediated through microscopic interactions. Such interactions can arise from both mechanical and chemical interactions.[1] Mechanical adhesion can arise from, for example, the interlocking of physical structures with one another such as in a hook-and-loop fastener wherein a surface functionalised with hooks mechanically interlocks itself with a surface functionalised with loops. In doing so, the two surfaces are affixed such that their separation typically requires an external macroscopic force.

Chemical adhesion, by contrast, is a result of either bonded or non-bonded chemical interactions between two surfaces at the nanometre or sub-nanometre resolution. In this thesis, non-bonded refers to all intermolecular interactions including hydrogen bonding. Bonded refers primarily to covalent bonding. Associations arising from bonded interactions are typically stronger than those arising from non-bonded interactions.[1] This can be demonstrated by epoxy resins, which are typically liquids containing chains of epoxy polymers that interact with one another through non-bonded chemical interactions. Before curing, an epoxy resin administered between two surfaces provides only a limited adhesive force. During curing, which can be achieved by mixing with a hardener, for example, the strength of adhesion increases significantly because of the formation of covalent crosslinks between the polymer strands.[2]

Both non-bonded and bonded methods of adhesion are also observed in nature. This bioadhesion can either refer to the adhesion between surfaces wherein at least one of the surfaces is biological in nature[3] or can refer to the adhesive substance itself being of biological origin. This definition has been extended to include synthetic

substances that are biologically inspired. Bioadhesion has been observed in both microscopic organisms, for example in the colonisation of surfaces in a host organism by pathogenic bacteria, and in macroscopic organisms, for example in the strong associations between marine organisms like mussels and limpets with surfaces in the tidal environment.[4] Like epoxy adhesion, mussel adhesion also involves both bonded and non-bonded interactions.[5]

The non-bonded element of mussel adhesion is achieved using hydroxyl functional groups in the 'adhesive pad' of the mussel (Figure 1-1). These hydroxyl groups are present in a novel amino acid, 3,4-dihydroxy-L-phenylalanine (dopa) and enable the adhesive pad to strongly interact with inorganic surfaces. At the same time, oxidation of the dopa molecules results in the formation of even stronger covalent bonds with organic surfaces, though at the expense of decreasing the bonding strength with inorganic surfaces. Mussels consequently use a combination of both bonded and non-bonded interactions to fine-tune their associations with surfaces in the marine environment.[5]

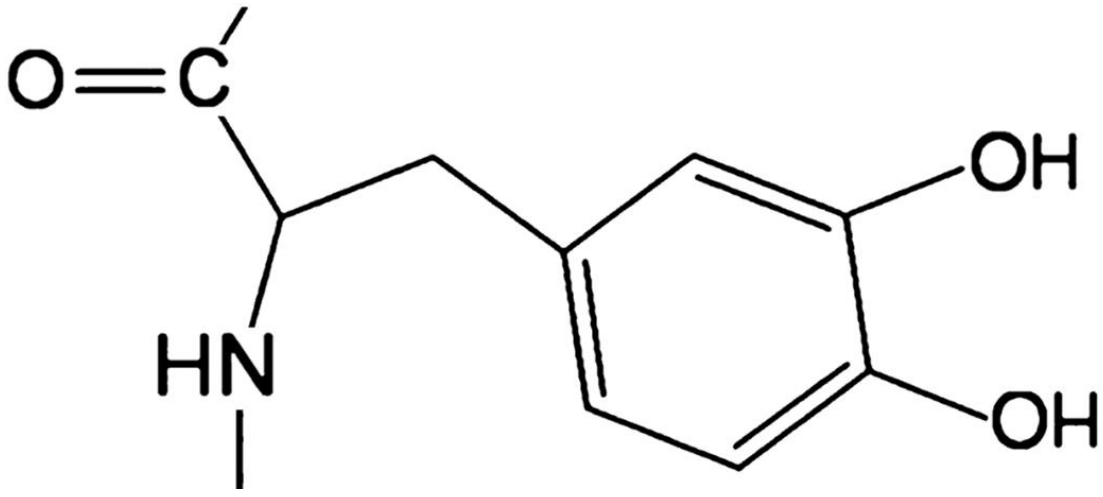


Figure 1-1: Skeletal formula of a dopa amino acid used in non-bonded and bonded mussel adhesion.

The forces required to rupture the interactions between a single dopa molecule and its surface have been measured to be approximately 0.8 and 2.2 nN for non-bonded and bonded (oxidised dopa) interactions respectively. A dissociation energy of 22.2 kcal/mol for the non-bonded case has also been determined. In epoxy adhesion, a continuous chain of covalent bonds is formed, and observed single-molecule rupture behaviours would differ depending on which covalent bond failed. These may also differ depending on how the epoxy was cured. For example, for an epoxy resin hardened with ethylenediamine, new C-N and C-O bonds are created, and the bond dissociation energies of these have been reported as 72.90 and 85.56 kcal/mol respectively. This suggests them to be several times stronger than the nonbonded dopa interaction.[6] Studies on the macroscale suggest that, depending on how mussel adhesives are cured, their maximum tensile strengths can range from approximately 0.3 to 1.0 MPa.[7] By contrast, maximum tensile strengths for epoxies have been reported between 30 and 40 MPa, suggesting them to be at least an order of magnitude stronger than mussel adhesives.[8]

Though mussel bioadhesion is only one example of a bioadhesive its properties are similar to other available bioadhesives. For example, its maximum tensile strength is comparable to that of a commercially-available fibrin bioadhesive,[7] and its non-bonded single-molecule rupture force is in the same order of magnitude as a carbohydrate-carbohydrate interaction involved in cell-to-cell adhesion (190 – 310 pN vs. 800 pN).[9] Despite their lower tensile strengths relative to epoxy adhesives, however, bioadhesives such as fibrin have found use in medicine, indicating that there are applications for bioadhesives wherein tensile strengths comparable to traditional alternatives like epoxies are not critical.[10]

Though bioadhesion in nature is often facilitated through proteins and carbohydrates, a nucleic acid bioadhesive based on the formation of deoxyribonucleic acid (DNA) complexes from individual strands of DNA has been demonstrated.[11] In a DNA bioadhesive interactions arise from the non-bonded association between the individual strands, though despite the lack of bonded

interactions macroscopic adhesive gels have been demonstrated. A unique property of DNA bioadhesives is the formation of ‘sequence-specific’ associations wherein sequence refers to the order and position of individual DNA ‘nucleotide’ monomers in the complex. These monomers preferentially associate with ‘complementary bases’ on the opposing strand, meaning that a single strand of DNA interacts most strongly with an opposing strand with the exact ‘complementary sequence.’ Due to this, it has been demonstrated that DNA can be used to form bioadhesive colloids only when beads functionalised with one strand are introduced in solution to beads functionalised with the exact complement.[11] This contrasts with other bioadhesion, like mussel adhesion, which can occur between the organism and a variety of organic or inorganic surfaces.

1.1.2 Advantages of bioadhesion over traditional adhesion

Bioadhesives typically have two primary advantages relative to traditional adhesives, like epoxies. These are biocompatibility, meaning the bioadhesive produces minimum toxicity to the host organism, and their lower environmental impact, accounting for aspects such as biodegradability and effect on human health. As a result of the biocompatibility of many bioadhesives they have been applied in clinical applications as alternative methods of wound closure relative to invasive methods such as suturing. Various surgical bioadhesives have been demonstrated to have low immunogenicity, to have anti-inflammatory or antibacterial properties, to undergo reversible but tough adhesion, and to act as scaffolds for the delivery and localisation of pharmaceutical molecules into cells.[12,13]

The lower environmental impact is the second major advantage of bioadhesion relative to traditional adhesion. Life-cycle analysis (LCA) is a method that characterises the various avenues of environmental harm caused by a material throughout its use and expresses these values numerically using normalisation and weighting. LCAs conducted on traditional petrochemical adhesives, such as epoxies, and bioadhesives suggests that, over each material’s lifetime,

petrochemical adhesives have a more negative impact on the environment. This assessment was in part due to the significantly lower fossil fuel use of bioadhesives, but also accounted for their reduced negative impacts on human health relative to petrochemical adhesives.[14]

1.1.3 Screening molecules using modelling

Since bioadhesives have human health and environmental benefits relative to traditional adhesives, methods for the inexpensive selection of candidate bioadhesives could be advantageous for long-term human and environmental health by assisting in the replacement of petrochemical adhesives with biological ones when possible. Regardless of the bioadhesive or its use case, the interaction strength of the bond is a relevant selection criterion since an adhesive is, by definition, a substance used to hold together two structures, particles, or surfaces. Consequently, the interaction strength is relevant to screening candidate bioadhesives and an inexpensive computational approach to this screening is therefore explored in this present work.

Given that bioadhesives are a broad category of molecules including proteins, carbohydrates, and nucleic acids, then a general modelling technique that does not need to be drastically adapted or redeveloped for new molecule types is appropriate. All-atom molecular dynamics is a suitable method which meets this requirement since it represents molecules using their atomic structures and models the system according to the behaviours between these atoms.[15] Consequently, the variable molecular structures of different candidate bioadhesives are intrinsically accounted for by their differing atomic structures and the method is thereby extensible regardless of molecular structure. In addition to this, by basing the model on the atomic, single-molecular scale, the ‘baseline’ stability of any bioadhesive, separated from additional effects such as density, macromolecular composition, and similar, can be established. This baseline stability is a necessary starting point for a rigorous study on the efficacy of a bioadhesive since in the absence of a single-molecular

understanding any observations on the effect of molecular structure, which is what is changed when a new bioadhesive is screened, on the bond stability can not necessarily be separated from other macroscopic contributions.

1.1.4 Research question

Bioadhesives have favourable properties and can theoretically be studied using computational modelling. The intention of this present work is to perform a method for studying the baseline, single-molecular stabilities of a candidate bioadhesive such that observations can be separated from macroscopic effects. Consequently, an all-atom molecular dynamics approach is used. The central research question is therefore:

“Can molecular dynamics be used to assess and screen the single-molecular binding properties of a candidate bioadhesive?"

This central research question will be discussed with respect to three technical research questions relating to Chapters 4, 5 and 6 respectively. These are, in order:

(1) “Can molecular dynamics be used to generate a model that allows binding properties to be predicted from structure alone, and what quantities contribute to the binding stability of the candidate bioadhesive?”

(2) “Can molecular dynamics be used to assess the single-molecular properties of a candidate bioadhesive under external force, and what quantities contribute to this?”

(3) “Can coarse graining be used to replicate properties observed in atomic molecular dynamics simulations?”

Technical research question (1) is proposed since a model that relates structure with binding is a screening tool that enables a user to select or eliminate candidate

molecules based on how likely they are to have the desired binding properties. Technical research question (2) is proposed since the purpose of an adhesive is to couple structures such that the relative displacement of the structures by an external force is resisted. It is therefore important that molecular dynamics can be used to assess the properties of a bioadhesive under external force. Technical research question (3) is proposed since there are computational resource limitations to all-atom simulations. Coarse graining can be used to overcome these limitations by representing the bioadhesive in a simplified form.

These research questions are discussed in this thesis in relation to the example candidate bioadhesive, peptide nucleic acid (PNA). This is since the ability of its naturally occurring counterpart, deoxyribonucleic acid (DNA), to form adhesive gels has already been demonstrated.[11] PNA was selected over DNA for the present work since it has several advantageous properties with regards to the stability of its bound complex relative to DNA. The nature of these molecules, and the current state of the literature on their single molecular binding properties, is presented in subchapters 1.4, 1.5 and 1.6. Subchapters 1.2 and 1.3 introduce the theoretical and methodological background relevant to interpreting the available literature.

1.2 Binding energies and external forces

1.2.1 Binding energy

A molecule has kinetic energy, from the movement of atoms, and potential energy from the relative position of its atoms. These kinetic and potential energies together equal the total energy of the molecule, but this energy fluctuates with time since, at the molecular level, particles are constantly experiencing thermal motion. This total energy is consequently distributed around a mean value called the internal energy U of the system.[16]

When a molecule is coupled thermally and mechanically to its surroundings its energy is affected by the external pressure P and temperature T. The total energy of the system fluctuates around the Gibbs free energy G (Equation 1-1), the change in which, ΔG, approaches zero as the system approaches equilibrium. When two molecules bind, ΔG between the initial (unbound) and final (bound) states is called the binding free energy. The binding free energy is routinely used as a measure of the interaction strength of bound complexes.

$$G = U + PV - TS \qquad [\text{Eq. } 1\text{–}1]$$

Where V is the volume and S the entropy. $U + PV$ is equivalent to the enthalpy H.

1.2.2 Binding energies under external forces

According to classical transition state theory, the bound and unbound states are separated from one another by a transition barrier with an energy $\Delta G^{\ddagger}$ at some distance $x^{\ddagger}$ (Figure 1-2).[17] This barrier must be surmounted for bond rupture to occur. In 1978, Bell observed that external force affected the rate of escape k over this transition barrier in biomolecules, demonstrating that the stability of biomolecular interactions was a force-dependent phenomenon.[18] This relationship between the rupture force, this being the force needed to rupture a biomolecular interaction, and loading rate, this being the rate of application of external force in Newtons per second, is described by a force-loading curve (Figure 1-3).

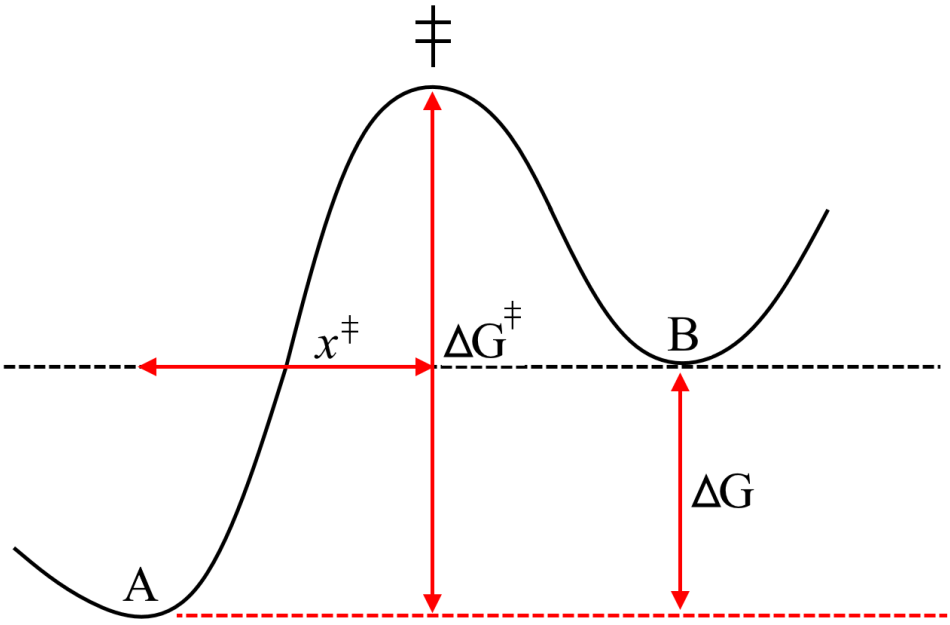


Figure 1-2: A bound state A is separated from an unbound state B by a barrier with height $\Delta G^{‡}$ and distance from the bound state $x^{‡}$. Particles move over the barrier ‡ with an escape rate k.

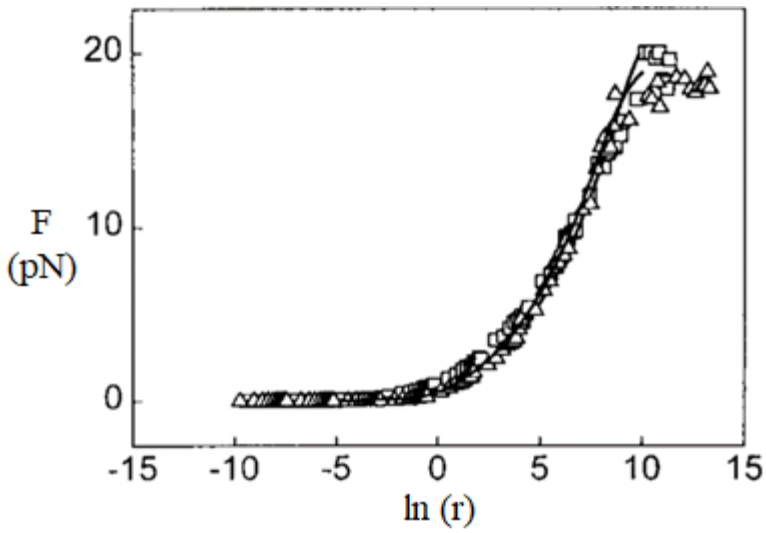


Figure 1-3: Loading rate r dependence on most probable rupture force F showing the transition between two regimes as loading rate increases. Reprinted from Evans and Ritchie, Copyright (1997), with permission from Elsevier (License No.: 5559281509342).[19]

The existence of the force-loading curve has implications for the comparability of stabilities under external force. Namely, it means that the equilibrium binding free

energy ΔG is not sufficient for comparing the stabilities of complexes at varying loading rates. A different approach based on fitting phenomenological models which relate the force-loading curve to the underlying free energy landscape can be used instead. Evans and Ritchie were the first, in 1997, to use the force-loading curve to derive free energy landscape parameters (Equation 1-2).[19]

$$[\text{Eq. 1–2}] \qquad F = \frac{k_B T}{x^{\ddagger}} \ln \frac{r x^{\ddagger}}{k_B T k}$$

Where r is the loading rate and F the most probable rupture force at that loading rate. The barrier height $\Delta G^{\ddagger}$ can be calculated from the rate of escape k in the absence of external force (Equation 1-3). k_B is the Boltzmann constant.

$$[\text{Eq. 1–3}] \qquad -\Delta G^{\ddagger} = k_B T \ln \frac{kh}{k_B T}$$

Where h is Planck's constant. Other solutions to this problem have been presented to address some of the limitations, such as its assumption of linearity which does not hold for higher loading rates, [20,21] of the above 'Bell-Evans' model. For example, Dudko, Hummer and Szabo proposed a generalisation of the Bell-Evans model,[22,23] though their model is itself limited in that, in the limit of infinitely high rupture forces (and hence loading rates), the rate of escape approaches zero, which is explicitly non-physical.[24]

Models like the Bell-Evans and Dudko-Hummer-Szabo model are called 'force spectroscopy' models. They enable the comparison of the stabilities of different binding complexes under conditions of external force to be evaluated in a loading-rate independent manner. However, the quantities associated with these models are only apparent, relying on assumptions such as a single transition barrier, or a barrier

with a particular curvature, and have therefore been demonstrated to produce impossible quantities such as picometre barrier widths for large biomolecular complexes.[21,22]

1.2.3 Thermodynamic energies have statistical meaning

Microscopic potential and kinetic energies are associated with atomic positions and momenta, and a particular configuration of atomic positions and momenta is called a microstate. Each microstate has an energy and a probability. Therefore, just as internal energy can be thought of as the mean value of the sum of microscopic kinetic and potential energies, it can equally be calculated by summing the energies of each microstate weighted by their probability (Equation 1-4).[16]

[Eq. 1–4] $$U = \sum_i E_i p_i$$

The probability of each microstate at equilibrium is dependent on its energy (Equation 1-5). [16]

[Eq. 1–5] $$p_i \propto e^{\frac{-E_i}{k_B T}}$$

The probability distribution of a system's energy at equilibrium is called the Boltzmann distribution. It is obtained by dividing the right side of Equation 1-5 by the sum of the probabilities of all possible microstates (Equation 1-6). This normalisation term, Z, is otherwise known as the partition function.[16]

[Eq. 1–6] $$p_i = \frac{e^{\frac{-E_i}{k_B T}}}{Z} = \frac{e^{\frac{-E_i}{k_B T}}}{\sum_i e^{\frac{-E_i}{k_B T}}} = \frac{e^{\frac{-E_i}{k_B T}}}{\int e^{\frac{-E_i}{k_B T}} d\vec{q}}$$

The discrete sum form of the partition function can also be represented as the integral form on the right-hand side. This integral is taken over a volume in ‘phase space’ which refers to a coordinate system with axes for both the positions and momenta of the particles, represented by the vector $\vec{q}$ in that phase space.

The free energy of the system can be calculated from the partition function (Equation 1-7).[16]

[Eq. 1–7] $$G = -k_B T \ln Z$$

And the free energy difference between two states i and j can be calculated from the ratio of partition functions (Equation 1-8).

[Eq. 1–8] $$\Delta G_{ij} = -k_B T \ln \frac{Z_i}{Z_j} = -k_B T \ln \frac{\left(\int e^{\frac{-E_i}{k_B T}} d\vec{q} \right)_i}{\left(\int e^{\frac{-E_i}{k_B T}} d\vec{q} \right)_j}$$

All of this is to say, if the energy of all microstates are known, then these energies can be used to calculate the probabilities of all microstates at equilibrium and the free energy. This means that, if a chemical system can be atomically rendered and its microscopic energy distributions accurately simulated, then it is theoretically possible to derive useful quantities related to the stability of biomolecular interaction.

1.3 Methods to quantify binding stabilities

1.3.1 Molecular dynamics

Computer simulations can directly compute microscopic potential and kinetic energies from atomic positions, enabling the derivation of quantities associated with binding stability, such as the binding free energy.[25] Using simulations can avoid experimental difficulties like reagent acquisition and financing since chemical structures are essentially just text files with a specific format. Online repositories like the Protein Data Bank[26] further trivialise this by enabling the direct download of biomolecular crystal structures. In combination with accelerations in computing power, simulations have become fast and accurate methods for the study of biomolecular interactions.

Simulations, in general, make compromises between accuracy and resource consumption. Highly accurate quantum mechanical *ab initio* methods simulate chemical systems entirely from first principles but are expensive to compute, often scaling with N^4 or higher where N can be thought of as a system's size or its degrees of freedom.[27,28] Molecular dynamics is a simulation process that approximates atoms as points in space with discrete charges and solves the evolution of their positions over time using Newtonian, as opposed to quantum, mechanics.[25] Files containing force constants, distances and other parameters associated with interatomic interactions, called molecular mechanics force fields, allow the forces on all the atoms to be calculated.[29,30]

Any molecular dynamics simulation follows the same general workflow:[25]

(1) Firstly, input conditions are initialised from structure and topology files. Structure files contain atomic positions and topology files contain information about the molecular mechanics forcefield. The combination of atomic positions and the forcefield allow the potential energy to be computed as a function of atomic positions since the forcefield encodes force constants for bonded and non-bonded interactions.

(2) Secondly, the force on an atom is computed from the gradient in the potential energy surface $F_i = -\partial V / \partial r_i$ where F_i is the force on the i^{th} atom and r_i its position vector, with V being the potential energy.

(3) Thirdly, the force on any atom results in an acceleration vector according to Newton's second law of motion $F = ma$. This results in a position update since atoms have moved, and steps 2 and 3 repeat until…

(4) The output step where atomic positions, energies and so forth are written into an analysable data file.

Though the general case for molecular dynamics is simplistic, there is significant nuance in how each step is calculated. For example, the choice of the molecular mechanics forcefield directly affects the potential energy. Not only that, but the experienced forces also depend on parameters such as the set cut-off distance (how far atoms need to be before their forces on one another are ignored) and environmental factors such as how the coupling of the simulation's temperature and pressure to an external bath is defined. Another quirk of molecular dynamics is that time is discrete instead of continuous, and consequently the position update step requires Newton's equations of motion to be solved over discrete timesteps on the order of femtoseconds. These numerical solutions to Newton's equations of motion over discrete time are achieved by algorithms called integrators, and the choice of integrator affects how the system evolves over time.[25]

A completed molecular dynamics simulation containing its positional and velocity data is called a trajectory. Trajectories can be analysed to determine quantities such as free energies along reaction coordinates of interest, rupture forces of biomolecular associations at varying loading rates, and enthalpies and entropies. The analysis of molecular dynamics trajectories is therefore an appropriate method

by which to calculate quantities associated with binding stabilities. The methods of trajectory analysis vary and are discussed at length in Chapter 2: Theory and Methods.

1.3.2 Thermal melting experiments

Beyond trajectory analysis, the binding free energy can also be experimentally measured using thermal melting experiments. By measuring the heat absorbed by a bound complex, calorimetry can determine thermodynamic quantities like binding enthalpy, entropy and free energy, for example. For nucleic acids it is also possible to obtain these quantities using optical melting experiments. During optical melting experiments, the amount of light absorbed by a solution containing nucleic acids is measured as the temperature is raised. Since single-stranded nucleic acids, which are the individual binding partners in a double-stranded nucleic acid complex, absorb more light at a 260nm wavelength than the double-strand, then the melting of double-stranded into single-stranded nucleic acids as temperature increases can be measured (Figure 1-4).[31,32]

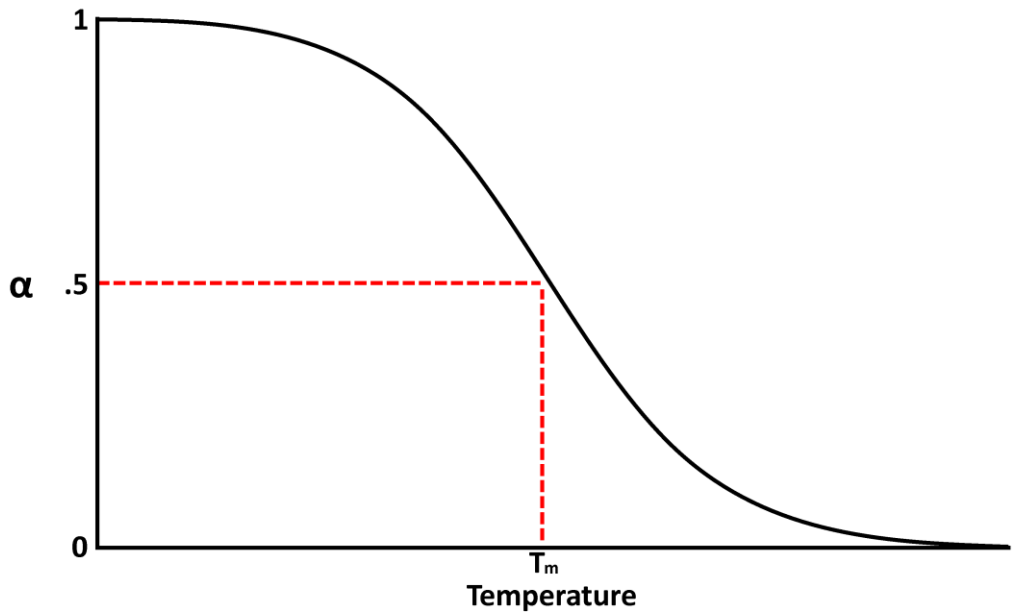


Figure 1-4: Normalised absorbance curve for an optical melting experiment. α is the normalised absorbance. The melting temperature T_m is defined as the point at which $\alpha = 0.5$.

From this absorbance curve, the melting temperature of the complex can be obtained from the midpoint when the normalised absorbance $\alpha = 0.5$. The melting temperature of a double-stranded nucleic acid is the temperature at which half of the double-stranded complexes in a system have dissociated into single strands. Likewise, an equilibrium constant, which is the ratio of products to reactants (melted to un-melted complexes) can be determined (Equation 1-9).[31]

[Eq. 1–9] $$K = \frac{\alpha}{\left(\frac{C}{n}\right)^{n-1}(1-\alpha)^{n}}$$

n is the number of binding partners which is two for a bimolecular reaction, such as the formation of ds-DNA from ss-DNA. C is the total concentration of the binding partners so long as each binding partner is present in equal concentration.

This equilibrium constant is related to the binding entropy and enthalpy through a linear form of the van't Hoff equation (Equation 1-10),[32,33] allowing binding enthalpies, entropies and therefore free energies to be determined from optical melting experiments.

[Eq. 1–10] $$ln\,K = -\frac{\Delta H}{RT} + \frac{\Delta S}{R}$$

1.3.3 Atomic force microscopy

Rupture forces can be measured at different loading rates using an atomic force microscope (AFM).[34,35,36,37] The AFM is able to measure forces on a surface using a probe with a well-defined spring constant. When the probe encounters a surface, the two exert a force on one another which deflects the probe. By shining a laser onto

the back of the probe, and reflecting this laser into a detector, the force on the probe can be measured by the movement of the laser since the spring constant of the probe is known (Figure 1-5).[38]

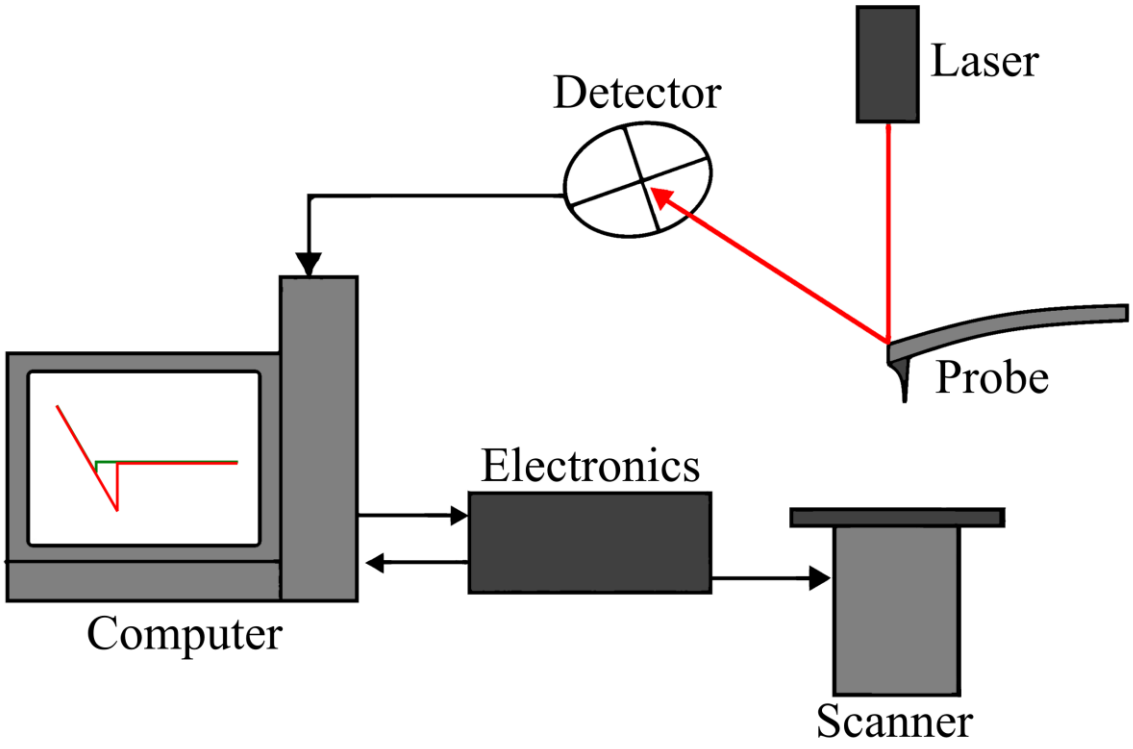


Figure 1-5: Schematic of an AFM in force spectroscopy mode.

In force spectroscopy mode the force on the probe is measured over a cycle consisting of an approach to and a retraction from the surface (Figure 1-6).[38] Typically, the probe is functionalised with one binding partner, and the surface with another. When the probe and surface come into contact a binding event theoretically occurs between the two binding partners, and this bond must then be broken during retraction. Hence, during retraction, a pull-out force is measured corresponding to the breakage of the specific interaction, and this corresponds to the rupture force of the bound complex.

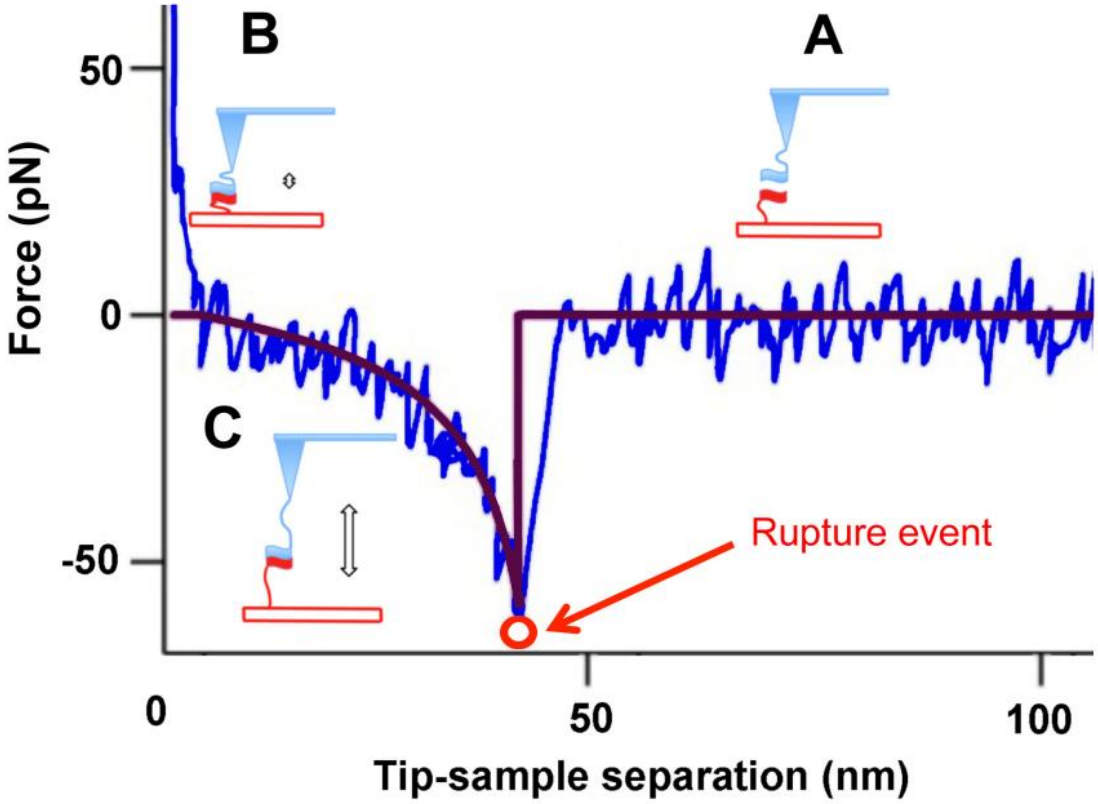


Figure 1-6: Example force spectroscopy experiment reproduced from Lyubchenko under a Creative Commons Attribution License.[39] From points A to B to C the probe approaches, contacts, and finally ruptures from, the surface.

The results of force spectroscopy experiments can be fitted to force spectroscopy models such as the Bell-Evans model, enabling the approximation of parameters associated with the underlying free energy surface and thereby the rate-independent comparison of different biomolecular complexes. Methods such as force spectroscopy, thermal melting experiments and molecular dynamics can and have been applied to the study of nucleic acid complexes. Such complexes are discussed in the following chapter and are used in the present work as example candidate molecules for the study of single-molecular adhesion via molecular dynamics simulations.

1.4 Peptide nucleic acids

1.4.1 Structure of deoxyribonucleic acid

The discovery of the structure of a deoxyribonucleic acid (DNA) salt was famously presented by F. Crick and J. Watson in their 1953 paper ‘Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid’. The double-helical structure they demonstrated, with binding between two backbones facilitated by the sequence-specific pairing of chemical bases, is a foundation of modern genetics (Figure 1-7). In their own words, “the specific pairing that we have postulated immediately suggests a possible copying mechanism for the genetic material.”[40]

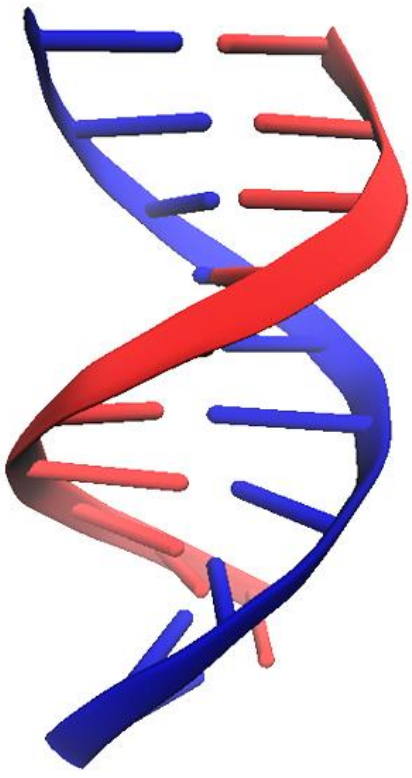


Figure 1-7: Visual Molecular Dynamics (VMD)[41] cartoon of a short DNA double-helix.[42] Each chain is highlighted in a different colour and nucleobases are represented by rods.

The specific pairing that Crick and Watson refer to is the pairing between the nucleobases adenine (A) with thymine (T), and guanine (G) with cytosine (C), in

the DNA double-helix (Figure 1-9).[43] Each pair consists of a double-ringed purine structure (A, G) and a single-ringed ‘pyrimidine’ (T, C). These base pairs are connected in sequence to one another through a negatively charged deoxyribose phosphate backbone. This backbone has a directionality and, by convention, is read from its 5’ to 3’ terminus (Figure 1-8).

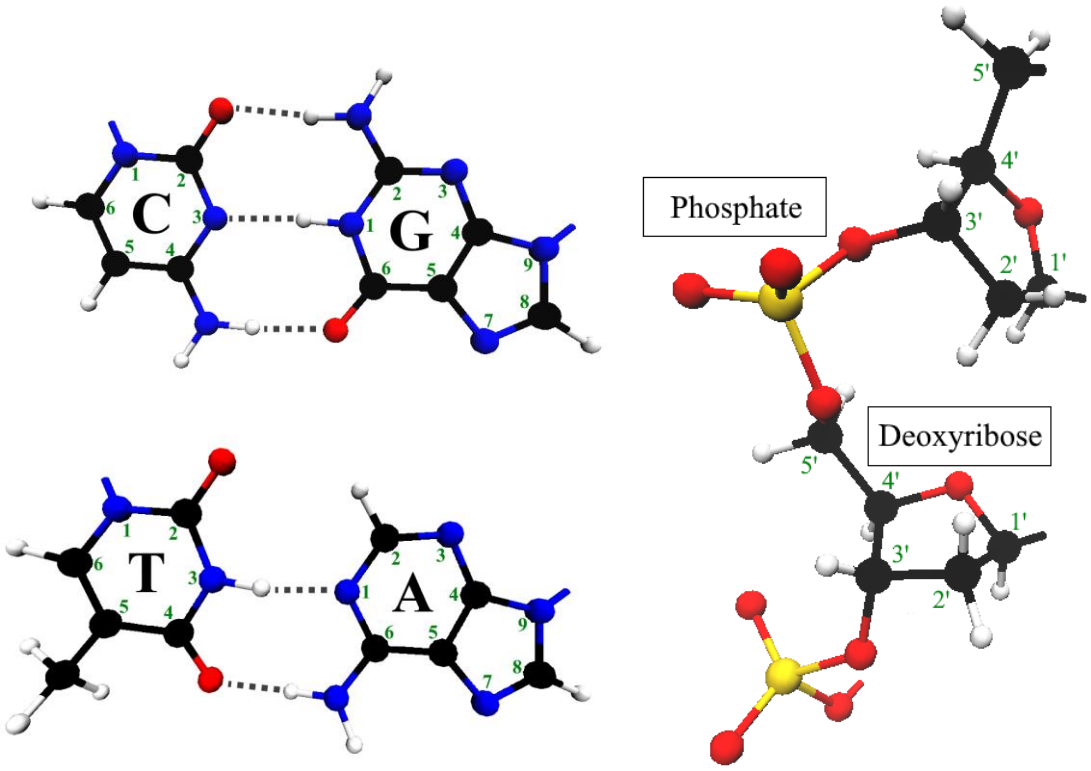


Figure 1-8: Left: VMD image of the A:T and G:C nucleobase pairs with hydrogen bonds indicated by dashed lines.[41] Right: VMD image of two repeating units of the deoxyribose phosphate backbone running from the upstream 5’ to downstream 3’ terminus.[41] H, C, N, O, P elements coloured white, black, blue, red, and yellow respectively.

Pairing between complementary bases is favoured relative to pairing between non-complementary bases, for example A with C. Resultantly, a strand of DNA with a particular sequence of nucleobases, called its primary sequence, encodes its opposing strand. The copying mechanism that Crick and Watson suggest, now understood as the process of DNA replication, is implied from this since, if one

strand encodes another, then strands act as templates to produce their own opposing strands.

The DNA double-helix formed by the pairing between base sequences is described by a set of helical parameters. A particular set of these parameters constitutes a form of DNA, for example the B-form (Figure 1-7, Table 1-1), which is the form typically taken by DNA in aqueous solutions.

Table 1-1: Approximate B-form values of select helical parameters.[44]

Helical Parameter	Approximate B-form value
Helix rotation per base pair	34.3°
Base pairs per turn (360° rotation of helix)	10 – 11
Rise (distance along axis between base pairs)	0.332 nm
Pitch (distance along axis per turn)	3.320 – 3.652 nm
Helical diameter	2.000 nm
Handedness (direction of rotation)	Right

1.4.2 Stability of deoxyribonucleic acid

Structural factors such as the length of the helix, its base composition, and the position of the bases in sequence all affect the double-helix's melting temperature. The dependence of stability on length can be understood by longer sequences having a greater surface area over which to form stabilising interactions. Similarly, the effect of composition can be understood in that G:C pairs have more stabilising hydrogen bonds than A:T pairs, and hence, per unit of interacting area, G:C pairs contribute more to stability.[45]

The effect of position is explained by the existence of another stabilising interaction in the double-helix. When bases are oriented parallel and adjacent to one another, as in the double-helix, a hydrophobic stacking interaction arises between their ring

systems (Figure 1-9).[44] Consequently, the stability provided by a base in sequence also depends on the bases adjacent to it, resulting in a positional effect on stability. These stacking interactions occur in both single-stranded (ss-) and double-stranded (ds-) DNA between bases (ss-stacks) and base pairs (ds-stacks) respectively.

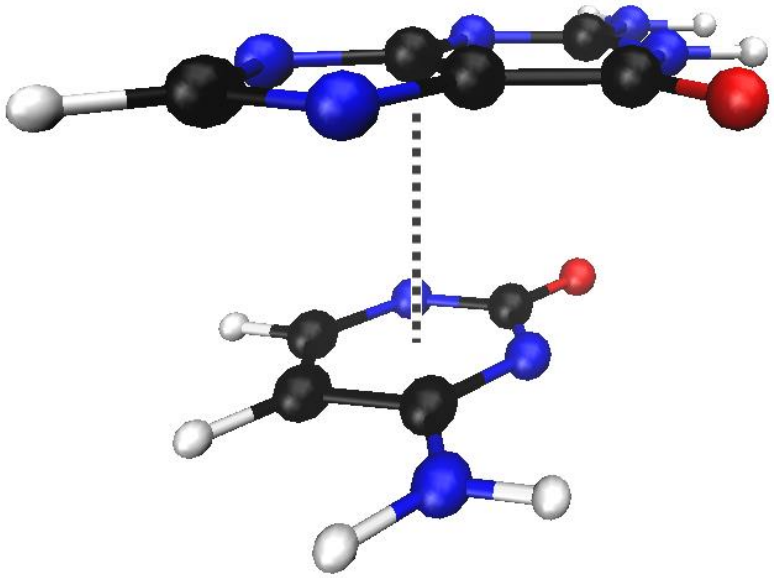


Figure 1-9: VMD image of an ss-stack with a dashed grey line indicating the stacking interaction.[41]

The stability of the DNA double-helix is highest near physiological conditions of 0.2M ionic strength and pH 7 and decreases towards the extremes of these variables.[46,47] The dependence on ionic strength is a consequence of the negatively charged backbone of DNA, since in the absence of counterions these negative backbones repel one another in the double-helix, leading to destabilisation in pure water. Similar reductions in stability occur due to the presence of nucleic acid degrading enzymes called nucleases, which break down DNA in organisms, reducing bioavailability and resulting in low serum half-lives.[48]

1.4.3 Structure of peptide nucleic acid

The function of DNA as a hereditary molecule requires that it performs sequence-specific recognition. Physical laws do not exclude other chemical structures from potentially performing the same function. Nucleic acid analogue molecules that perform sequence-specific recognition but also have additional desirable properties have therefore been produced. These include the locked nucleic acids (LNAs) and phosphorodiamidate morpholinos (PMOs). For both LNA and PMO, backbone monomers consist of carbon ring systems adjoined to a phosphorus centre in a manner similar to biologically occurring nucleic acids.[49] Relative to DNA, both have a higher thermal stability[50] though PMO additionally forms stable double-helices in pure water at pH 7 since its backbone is neutrally charged under these conditions.[51]

Peptide nucleic acid (PNA) is another nucleic acid analogue, except that unlike LNA or PMO its backbone is not composed of ring systems and phosphorus centres. Instead, its backbone has been entirely substituted with a neutrally charged pseudopeptide called N-(2-aminoethyl)glycine (Figure 1-10). Whereas the directionality of DNA is by convention read from the 5’ to 3’ carbon of the pentose ring, the directionality of PNA proceeds from an N-terminal amine to the C-terminal carboxyl like a polypeptide.[52]

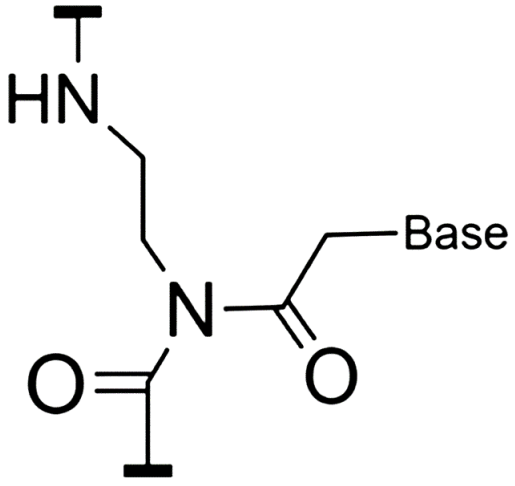


Figure 1-10: Skeletal formula of an N-(2-aminoethyl)glycine monomer.

Whereas double-stranded DNA (ds-DNA) forms a B-form helix, double-stranded PNA (ds-PNA) forms a P-form helix (Figure 1-11). Unlike the B-form, which has a preferred helix handedness, the P-form exists as a mixture of right- and left-handed double helices since the PNA backbone is itself achiral,[53] meaning mirror images of the backbone monomers are superimposable. The P-form is also wider (2.8nm) than the B-form helix (2.0nm) and has a greater pitch of 18 base pairs against the 10 – 11 of the B-form.[54] Like the B-form, the bases are oriented normal to the longitudinal axis of the P-form helix, though the rise of the helix per base pair is slightly lower in the P-form than the B-form (Table 1-2).[55,56,57] Single-stranded PNA (ss-PNA) can form heteroduplexes by binding to single-stranded DNA (ss-DNA), producing a helix with a structure intermediate of the B- and P-forms.[54]

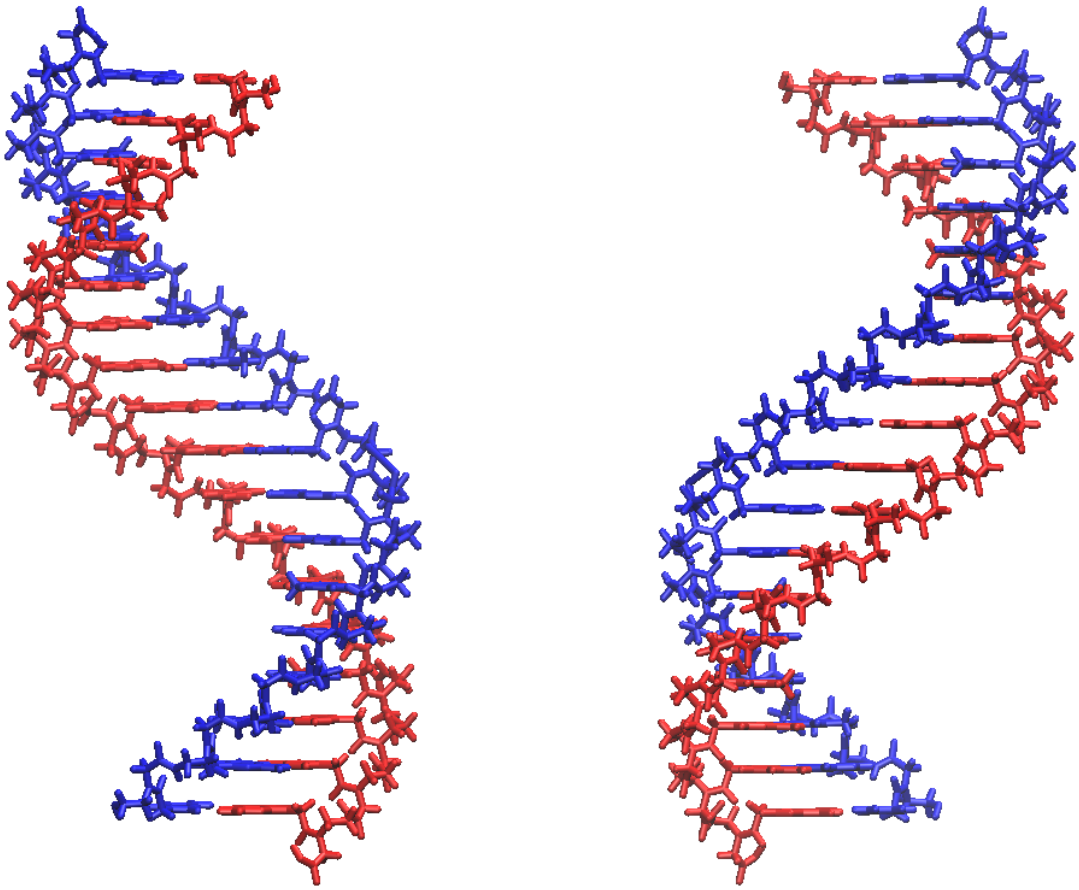


Figure 1-11: VMD licorice image[41] of the P-form helix from the perspective of the major groove (left) and the minor groove (right).

Table 1-2: Approximate P-form values of select helical parameters.[54-57]

Helical Parameter	Approximate P-form value
Helix rotation per base pair	20°
Base pairs per turn (360° rotation of helix)	18
Rise (distance along axis between base pairs)	0.29 nm
Pitch (distance along axis per turn)	5.22 nm
Helical diameter	2.80 nm
Handedness (direction of rotation)	Indiscriminate

1.4.4 Stability and solubility of peptide nucleic acid

The stability of ds-PNA across a range of solvent conditions is well-reported.[58,59,60,61,62] Firstly, its stability is independent of the concentration of ions in the solvent due to its neutral charge,[58,59] resulting in no net like-charge repulsion between its backbones, similarly to PMO.[60] Secondly, in human serum, PNA was found to resist proteolytic enzymes and experienced no significant degradation in conditions that significantly degraded peptides.[61] Like both PMO and LNA, PNA also resists hydrolysis by nucleolytic enzymes,[59,60] making it biostable against both typical protein-degrading and nucleic acid-degrading enzymes despite having structural properties of both proteins (amide bonds) and nucleic acids (bases). Thirdly, PNA resists degradation at low pH, unlike DNA which loses purine bases under acidic conditions.[59] Finally the melting temperature of ds-PNA only decreases slightly for an increase in organic co-solvent concentrations up to 70%, raising the possibility of stable PNA binding in organic solvent conditions.[62] Compared with DNA or RNA, however, and as indicated by its tolerance to organic conditions, the solubility of PNA in aqueous solutions is lower, though a charged lysine tag can be incorporated to mitigate this.[63] Given these properties, it is plausible that a PNA bioadhesive could have improved performance relative to the DNA-based bioadhesives[11] demonstrated in prior research.

The stability of PNA across a broad range of solvents contributes to its versatility. Not only is ds-PNA stable across these conditions, but it is also more stable than both ds-DNA and another naturally occurring nucleic acid, double-stranded ribonucleic acid (ds-RNA), under standard conditions. The melting temperature of a ds-nucleic acid 10-mer was determined to be more than 30K higher in ds-PNA (347.0K) than in ds-RNA (314.6K).[64] Another 10-mer has reported ds-PNA melting temperatures of 369.1,[65] 362.3[58] and 368.2K[62] against ds-DNA melting temperatures of 334.2, 341.0, 333.8K respectively. The melting temperatures of hybrid PNA:DNA heteroduplexes are intermediate of these.[58,62,65] This represents a significant increase in the stability of PNA-containing relative to solely DNA- or RNA-containing duplexes.

The above properties of ds-PNA drove its selection over ds-DNA for the present work since broader environmental stabilities and higher melting temperatures indicated that it could have improved properties relative to ds-DNA in relation to adhesion. This is since, for adhesion, the stability of the bound complex is related to the efficacy of the adhesive because a more stable bound complex can theoretically resist more external force.

In addition to the theoretical applications of ds-PNA in adhesion, PNA-containing heteroduplexes have been studied for *in vivo* applications. For example, PNA can perform strand invasion, meaning it can bind to ds-DNA to form a triple helix, or triplex.[66] This has led to their design as gene therapeutics in which an ss-PNA is encoded such that it invades a particular gene, binding with the ds-DNA to hinder the transcription of that gene. This strand invasion into ds-DNA makes ss-PNA an antigene drug, meaning a drug that works by targeting nuclear DNA to downregulate the function of a particular gene.[67,68] By modifying PNA chemistry, for example, by using synthetic nucleobases, the ability of PNA to downregulate genes can be further enhanced such as by introducing cross-linking between itself and the target gene.[69]

Since PNA also stably binds RNA, it has found applications as an antisense therapeutic as well.[68] Antisense technologies work by targeting cellular messenger RNA (mRNA) to regulate its function. Since mRNA, which is transcribed from DNA, is transported from the nucleus to be translated by ribosomal assemblies into proteins, antisense PNAs act upstream of antigene PNAs. By binding to the translational start sites of mRNAs, PNAs can sterically block recognition of the mRNA by the ribosomal assembly, or by binding further along the sequence they can arrest the process of translation before completion.[70] As with antigene therapies, antisense therapies using PNA are advantageous due to a combination of their binding strength and sequence selectivity, limiting off-site effects whilst resulting in strong targeted inhibition.[71]

Beyond antisense and antigene applications, PNA has potential in diagnostic assays by sensing target sequences of nucleic acids in real-time.[72] This has been demonstrated as potentially advantageous in cancer diagnoses, for example,[73] since they have a high signal-to-noise ratio. By printing PNAs onto surfaces, they have also been used for the development of microarrays that can quantify a range of target sequences from crude extracts. Like other biosensors, PNAs can be modified with fluorescent labels to enable ease of visualisation and as such have even been applied *in vivo* to detect target sites in lung cancer cells and for the detection of specific mutations in melanoma samples.[71] Lastly, their stable binding with DNA has been further utilised for the development of highly sensitive PNA-based polymerase chain reaction (PCR) kits for the sequence-specific amplification of minute quantities of DNA with no cross reactivity.[74]

1.5 The nearest-neighbour binding model

1.5.1 The binding energies of peptide nucleic acids

Tomac *et al*.,[58] Ratilainen *et al*.,[75] Nielsen and Sen,[62,65] and Schwarz, Robinson and Butler[76] have all reported binding energies for PNA heteroduplexes across a variety of aqueous, ionic and organic conditions. In aqueous conditions, it was typically

observed that PNA:DNA duplexes had lower enthalpies of binding than ds-DNA, though this was sequence dependent. In some cases, binding enthalpy was higher, though this was typically compensated by a higher binding entropy such that the heteroduplex was more stable than ds-DNA despite its higher binding enthalpy. This enthalpy-entropy compensation[75] has been attributed to the release of ions upon binding of the two strands, since the negatively charged ss-DNA is associated with counterions that are released upon binding with ss-PNA.[58]

Another important parameter of PNA hybridisation with DNA or RNA is how well it discriminates mismatched pairs. Hence, the effects of mismatching on PNA heteroduplexes have been quantified. Jensen *et al*. measured the melting temperatures of 12 mismatched sequences, each having a single different mismatch relative to a control.[77] For PNA:DNA pairs they found that any mismatch, at a minimum, decreased the melting temperature by 8K and resulted in rates of dissociation that were, in most cases, an order of magnitude higher than for the control sequence. For PNA:RNA a similar high sensitivity to mismatching was demonstrated. This sensitivity to mismatching is crucial to PNA's function as an antigene and antisense technology since limiting off-site effects reduces the likelihood of unintended patient outcomes.[66,68,70]

Alongside quantifying PNA heteroduplex binding energies, many of the studies also quantified the melting temperatures and binding energies of ds-PNA. Regardless of sequence length, composition or the nature of the solvent, ds-PNA always had a higher melting temperature than ds-DNA, ds-RNA and PNA heteroduplexes. Extremely high melting temperatures of 382K (84°C)[77] and 386K (88°C) were reported for 15- and 17-mer ds-PNAs respectively.[78] An 8-mer ds-PNA had a reported melting temperature of 353K (55°C)[78] approximately equal to that of a 15-mer ds-DNA which had a lower melting temperature of 351.5K (53.5°C) despite being almost half its sequence length.[77] These melting temperatures are reflected in the across-the-board lower binding free energies of ds-PNA (Table 1-3).

Table 1-3: Aqueous ds-PNA binding free energies against ds-DNA binding free energies determined by a predictive model.[79] Values and standard errors are either averages from the literature when multiply-sourced or are directly obtained.

Sequence	$-\Delta G_{PNA}$ ($\mathbf{kcal\ mol^{-1}}$)	$-\Delta G_{DNA}$ ($\mathbf{kcal\ mol^{-1}}$)
TGATCTAC	13.10[78]	6.97
GTAGATCACT	18.99 ± 0.87[58,62,65,75,78,80]	9.71
TGTTACGACT	21.08 ± 1.00[64]	10.89
AGGTAACCAG	18.76 ± 0.60[65]	10.85
AGTGAAGCAG	19.02 ± 0.82[65]	11.52
GTAGATCACTGT	21.40[78]	12.64
GTAGATCACTGTCAC	26.40[78]	17.04
GTAGATCACTGTCACAG	28.70[78]	19.80
GTAGATCACTGTCACAGAT	36.30[78]	21.83

The absolute value of the binding free energy of ds-PNA from the literature is anywhere from approximately 50% greater, for longer sequences, to twice as large for shorter sequences, representing a significant improvement in the stabilisation of these double helices under aqueous conditions. In addition, whereas PNA:DNA has been reported to have higher binding entropies in aqueous conditions than ds-DNA due to the release of counterions from the DNA backbone,[58] binding entropies for ds-PNA are typically equal to or lower than ds-DNA binding entropies.[58,75,65] This indicates that, whereas PNA:DNA heteroduplexes may partly be stabilised relative to ds-DNA due to entropic contributions, ds-PNA may instead be stabilised more on account of its lower binding enthalpy. In this sense, ds-PNA may experience enthalpy-entropy compensation in the opposite way to PNA:DNA heteroduplexes with a lower binding entropy being compensated for by a lower binding enthalpy, though with only limited sequences tested this conclusion is tenuous.

1.5.2 Historical development of a predictive binding model

Directly comparing the binding energies of different nucleic acid sequences, though useful for demonstrating the increased stability of PNA-containing complexes relative to ds-DNA, does not allow any predictions of stability based on primary sequence to be made.

An early step in predicting binding energy from sequence was made by Privalov *et al*. in 1965 when they measured the heats of absorption of ds-DNA in solutions at different pH and ionic concentrations.[46] In doing so they measured the melting enthalpy of the ds-DNA, which they translated into an enthalpy per base pair by dividing by sequence length. In 1969, they extended their research by predicting the per base pair binding free energy, this being the negative of the per base pair melting free energy since it is the opposite process, of ds-DNA to be $-1.22\ \mathrm{kcal/mol}$ under physiological conditions.[81]

In 1959, however, Marmur and Doty had already noted that there was a linear relationship between the melting temperature of ds-DNA and its base composition with a 1K rise in the melting temperature of ds-DNA correlating with a 2.5 mole-% increase in the G:C content of a sequence.[82] This relationship, expanded in 1962 using 40 samples,[83] demonstrates that the per base pair melting enthalpy of Privalov *et al*. is insufficient since it fails to account for the effect of base composition (% G:C or inversely % A:T). A similar study by Bunville *et al.* in 1965, this time tabulating the mole-% G:C values against the enthalpies of denaturation of ds-DNA sequences, found that enthalpy was similarly correlated with the G:C content.[84]

A scheme that decomposes binding energies by length and composition, however, fails to account for the impact of base stacking, since stacking emerges as a result of a base pair's position in the sequence relative to other pairs. Bunville *et al*. were already aware of this when they themselves stated that, "There have been many suggestions… that hydrophobic interactions are important in stabilizing the DNA

helical structure".[84] In 1970, Wells *et al*. stated that "DNA's which have the same base composition but different nucleotide sequences (sequence isomers) do not show identical helix-coil transitions",[85] referring to the transition during DNA melting. Their results demonstrated that both base composition and position were important in determining the melting temperatures of ds-DNA, with the positional effect being due to stacking.

The dual positional and compositional effects of sequence on binding energy results in the nearest-neighbour model of ds-DNA binding energy. In the nearest-neighbour model the binding energy of a ds-DNA is calculated from a linear superposition of ds-stacks, where a ds-stack consists of two base *pairs* in sequence. The prefix ds- is used to emphasise that such a stack accounts for both pairing and stacking energies by highlighting its double-stranded nature, as opposed to an ss-stack which would account only for the stacking energy between nucleobases on the same strand. Using the nearest-neighbour model, the binding energy of the sequence GATC:CTAG would be calculated from the sum of the 'incremental energies' of the ds-stacks GA:CT, AT:TA and TC:AG, for example.[79] The ds-stacks are superimposed since the first pair of the following stack is the latter pair of the preceding stack, for example A:T from AT:TA is the same pair as A:T from GA:CT. Bases separated by a colon refer to the paired bases of the opposite strand whereas bases not separated by a colon are in sequence on the same strand.

In 1973, Tinoco *et al*. demonstrated that incremental free energies could be used to predict the most stable secondary structure taken by that sequence.[86] In that sense, the primary sequence of a nucleic acid directly predicted its secondary structure, opening the way for the design of RNA sequences with desired secondary structures. This was the first case of the nearest-neighbour model being used to understand structure-function relations of nucleic acid molecules. In 1978, Filmonov and Privalov established the incremental enthalpies of some DNA ss-stacks,[87] followed by Gotoh and Tagashira in 1981[88] and Marky and Breslauer in

1982[89] both determining the incremental enthalpies of all DNA ds-stacks in fully complementary duplexes.

Marky and Breslauer went on to demonstrate that incremental enthalpies could predict the binding enthalpy of ds-DNA from its primary sequence in 1986.[90] They incorporated a helix initiation energy, this being a base-independent energy associated with the initiation of helix formation. In 1996 SantaLucia Jr. *et al.* additionally differentiated this initiation energy depending on whether the helix had a terminal A:T pair or not, hence accounting for the end-stabilisation of the helix by G:C-only termini.[91] The binding energy of a double-helix can then be expressed as a sum of its occurrence-weighted incremental energies and termini-dependent helix initiation energy. The various values assigned to these energies from different authors were eventually compiled into a unified model by SantaLucia Jr. in 1998 (Table 1-4).[79]

Table 1-4: Unified nearest-neighbour binding energies and entropies at 310K reproduced from SantaLucia. Jr, 1998.[79] The symmetry term accounts for an additional entropic correction needed for self-complementary ds-DNA.

Parameter	$-\Delta \boldsymbol{H}$ ($\mathbf{kcal\ mol^{-1}}$)	$-\Delta \boldsymbol{S}$ ($\mathbf{cal\ K^{-1} mol^{-1}}$)	$-\Delta \boldsymbol{G}$ ($\mathbf{kcal\ mol^{-1}}$)
AA:TT	7.9	22.2	1.00
AT:TA	7.2	20.4	0.88
TA:AT	7.2	21.3	0.58
CA:GT	8.5	22.7	1.45
GT:CA	8.4	22.4	1.44
CT:GA	7.8	21.0	1.28
GA:CT	8.2	22.2	1.30
CG:GC	8.6	27.2	2.17
GC:CG	9.8	24.4	2.24
GG:CC	8.0	19.9	1.84

Initiation (GC)	-0.1	2.8	-0.98
Initiation (AT)	-2.3	-4.1	-1.03
Symmetry	0	1.4	-0.43

Equation 1-11 below is a master equation for the nearest-neighbour model, describing how a binding free energy, entropy or enthalpy, generalised as $\Delta\chi$, is calculated from incremental and initiation energies or entropies. To distinguish them from the total $\Delta\chi$, incremental values are denoted Δx.

$$[\text{Eq. 1–11}] \qquad \Delta\chi = a\Delta x_{init.GC} + b\Delta x_{init.AT} + \sum_i j_i \Delta x_i + c\Delta x_{sym.}$$

Where $\Delta x_{init.GC}$ is the helix initiation energy if there are no terminal A:T and $\Delta x_{init.AT}$ is the helix initiation energy if there is an A:T terminus. $\Delta x_{sym.}$ is a symmetry correction term for self-complementary ds-DNA. $\sum_i j_i \Delta x_i$ is the sum of the occurrence-weighted j_i incremental energies or entropies Δx_i for each ds-stack i where the occurrence is the number of times a ds-stack appears in a sequence. a and b evaluate to 1 if there are no terminal A:Ts or there is at least one terminal A:T respectively, else they evaluate to 0. c evaluates to 1 for self-complementary sequences and otherwise 0.[33,79]

1.5.3 Molecular dynamics of pairing and stacking energies

Given that molecular dynamics can estimate thermodynamic quantities, it is theoretically possible that it can calculate the nearest-neighbour parameters for nucleic acid hybridisations. The first steps to showing the application of molecular dynamics in calculating nearest-neighbour parameters were taken in 2015 when Sakuraba, Asai and Kameda measured the difference in free energy between RNA complexes.[92] By numerically converting one RNA primary sequence (state A) to

another (state B), achieved by smoothly interpolating structure and topology information, they could measure the difference in binding free energy of the two strands using a thermodynamic cycle (Figure 1-12). This process of interpolating topologies is called an alchemical transformation, and the calculation of free energy using it an alchemical free energy calculation (Chapter 2.4), though this method is limited in that it is computationally expensive.[93]

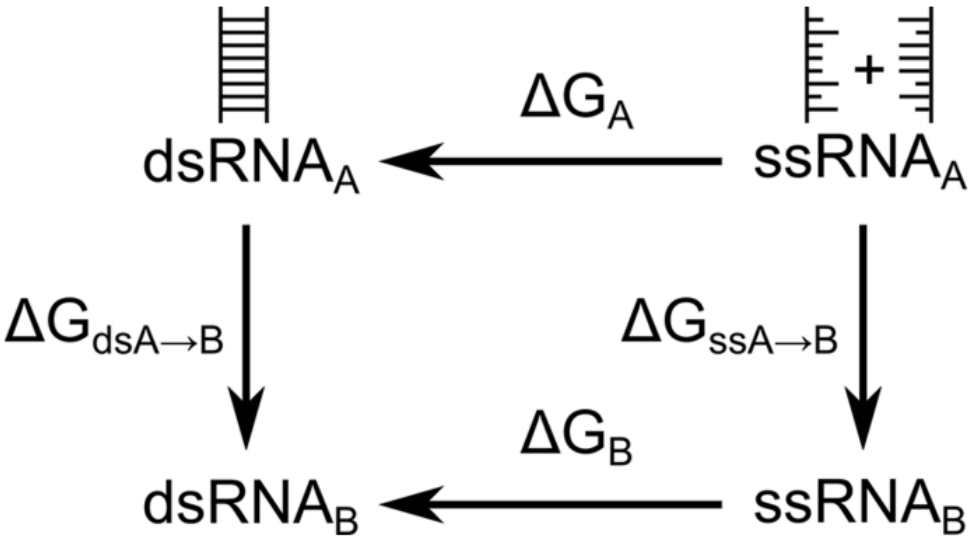


Figure 1-12: Thermodynamic cycle for the calculation of the binding free energies of ss-RNA into ds-RNA. By calculating $\Delta G_{ssA\rightarrow B}$ and $\Delta G_{dsA\rightarrow B}$, the difference in binding free energies can be obtained as $\Delta G_A - \Delta G_B = \Delta G_{dsA\rightarrow B} - \Delta G_{ssA\rightarrow B}$. Reproduced with permission from Sakuraba *et al.*[92] Copyright 2015 American Chemical Society.

Using the same method for a large range of sequences in 2019, Nishida, Sakuraba, Asai and Hamada determined the incremental energies for RNA secondary structure prediction[94] similar to the work of Tinoco *et al*. in 1973.[86] The maximum absolute difference between their computationally and experimentally derived incremental free energies was 0.25 kcal/mol. In 2020, Sakuraba and Asai *et al*. also determined incremental energies for non-standard inosine-cytosine base pairs using the same approach.[95]

In 2021, Golyshev, Pyshnyi and Lomzov[96] used a computationally inexpensive but approximate method,[97] called the molecular mechanics Poisson-Boltzmann [Generalised Born] surface area approach (MM-PB[GB]SA) (Chapter 2.5), to derive incremental energies for ds-RNA and RNA:DNA binding. The authors demonstrated that ds-RNA experimental binding enthalpies, but not binding free energies and entropies, were well approximated for ds-RNA, though all three were poorly approximated for DNA:RNA heteroduplexes. Since their computational method differed more substantially from the experimental data than was observed for Nishida *et al*. with an alchemical method, the authors applied a correction. Much of the variance between the computational and experimental binding energies could be accounted for using a linear correction of the form $\Delta\chi_E = a\Delta\chi_C + b$ to the computational data. After calculating $\Delta\chi_E$ for all sequences according to this equation they then calculated incremental energies using multiple regression according to Equation 1-11. Their reported incremental enthalpies had standard errors that often overlapped with incremental enthalpies for ds-RNA obtained via experimental methods.

The linear correction of Golyshev *et al*. has a non-trivial effect on the incremental binding energies, however. According to Equation 1-11, a binding energy is given by a linear sum of its occurrence-weighted incremental energies, plus the helix initiation energies and symmetry terms. Ignoring symmetry and grouping the initiation terms for brevity, the correction $\Delta\chi_E = a\Delta\chi_C + b$ is then equivalent to (Equation 1-12).

[Eq. 1–12] $$\left(\sum j_i \Delta x_i + \Delta x_{init}\right)_E = a\left(\sum j_i \Delta x_i + \Delta x_{init}\right)_C + b$$

If a linear correction is used then the i^{th} incremental energy $(\Delta x_i)_E$ is equal to $a(\Delta x_i)_C$. The consequence of this is that the theoretical equality $\Delta G_E = \Delta H_E - T\Delta S_E$ for the i^{th} ds-stack only holds when $a_G\Delta G_C = a_H\Delta H_C - a_S T\Delta S_C$ where a_G, a_H and a_S are the slope coefficients of the linear corrections for free energy,

enthalpy and entropy respectively. This is not the case for their study and so the application of the linear correction by Golyshev *et al*. lead to the inequality $\Delta G_E \neq \Delta H_E - T\Delta S_E$ which is not compatible with the broadly accepted theory outlined in Chapter 1.2.1.

In addition to determining incremental energies for developing a nearest-neighbour model, molecular dynamics has been used to reconstruct free energy surfaces associated with stacking and pairing coordinates. In 1995 Norberg and Nilsson[98] reconstructed one-dimensional free energy surfaces for the unstacking of different ss-stacks (Figure 1-13), providing more detailed evidence of the higher stacking energies of purine bases. In general, ss-stacks involving only pyrimidines were characterised by a broad and shallow global minimum in the free energy along this coordinate whereas ss-stacks involving purines, especially adenines, were deeper and narrower. Since the free energy surface has a statistical meaning, this implies that pyrimidine-containing ss-stacks are more likely to exist in an open, or frayed, conformation than ss-stacks containing only purines.[16]

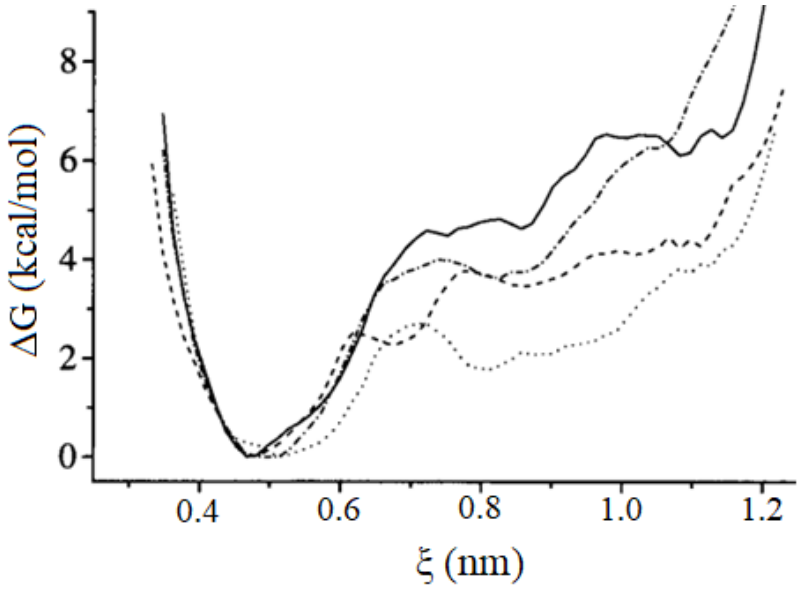


Figure 1-13: Free energy landscapes for four ss-stacks GA, GC, GG and GT as continuous, dashed, dashed-and-dotted and dotted lines respectively. Reproduced from Norberg and Nilsson, Copyright (1995), with permission from Elsevier

(License No.: 5559291250394).[98] ξ is the reaction coordinate distance and ΔG the free energy change.

Similar studies have been demonstrated for the base pairing coordinate. In 1999, Stofer, Chipot and Lavery restrained isolated base pairs in the Watson-Crick pairing conformation.[99] The reaction coordinate was defined as the distance between the N1 atom of the purine and the N3 atom of the pyrimidine and the free energy landscape along this one-dimensional pairing coordinate (Figure 1-14) for both A:T and G:C pairs was reconstructed. Both landscapes were characterised by deep and narrow wells at approximately 0.3nm separated from a local minimum at approximately 0.5nm by a free energy barrier. This local minimum arises due to the insertion of a water molecule between the unbinding pair, bridging the gap between the two with additional water molecules. From the plateau in the free energy landscape at longer distances, Watson-Crick pairing free energies of -4.3 and -5.8 kcal/mol were determined for A:T and G:C pairs respectively, corroborating the G:C pair's higher contribution to the stability of the double-helix.

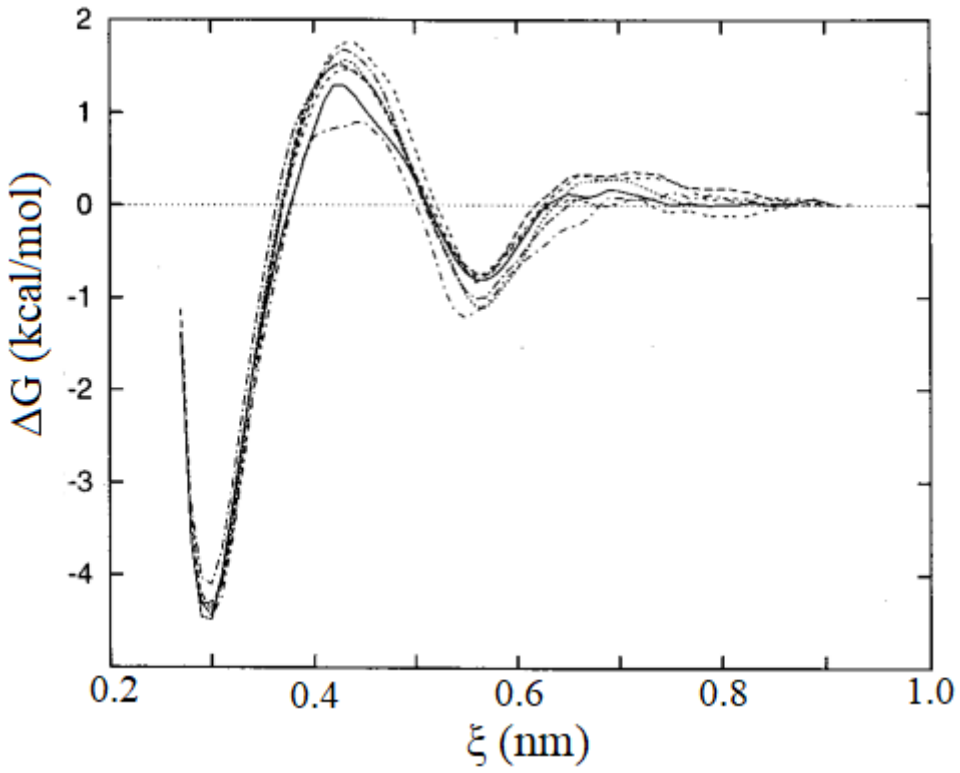


Figure 1-14: Free energy landscapes for the separation of A:T along the Watson-Crick hydrogen bonding coordinate. Each line is a replicate. ξ is the reaction coordinate distance and ΔG the free energy change. Reprinted with permission from Stofer *et al*. Copyright (1999) American Chemical Society.[99]

Since both pairing and stacking affect the free energy landscapes of unbinding base pairs, one-dimensional free energy landscapes cannot provide a complete picture of the unbinding process. Not only that, but studies have reported that the free energy landscape along one coordinate can be affected by sampling states associated with other coordinates. For example, in studies of the Watson-Crick pairing coordinate, a stacked state where two bases in a pair slip over one another has been demonstrated to affect the overall binding free energy of the pair, and so these states usually have to be manually excluded with additional restraints.[100] A more appropriate alternative, then, would be to reconstruct a two-dimensional free energy surface containing both the stacking and pairing coordinates, though this requires additional sampling and is therefore more resource intensive.

In 2017, Lindahl, Villa and Hess demonstrated the reconstruction of such a two-dimensional landscapes.[100] Defining the stacking coordinate as being between the centres of geometry of the 6-membered rings of paired nucleobases within a double-helix, and the pairing coordinate as being between the N1 and N3 atoms of purines and pyrimidines respectively, they produced a landscape with three minima (Figure 1-15). In their landscape, a global Watson-Crick bound state minimum at a low N1:N3 distance and an intermediate 6-ring distance is separated from the metastable stacked state at an intermediate N1:N3 and low 6-ring distance by a saddle point in the free energy landscape. The relative stability of the two states can be assessed, with the free energy of the stacked state having an approximately 10 k_BT higher free energy than the Watson-Crick bound state. In addition, a shallow minimum at high N1:N3 and 6-ring distances corresponds to an open state where the bases have slipped outside of the helix. By producing a more complete picture of the free energy landscapes associated with base pairing and stacking, the authors were able to determine that, when the stacked conformation was excluded, the total opening free energy would increase by approximately 1 k_BT .

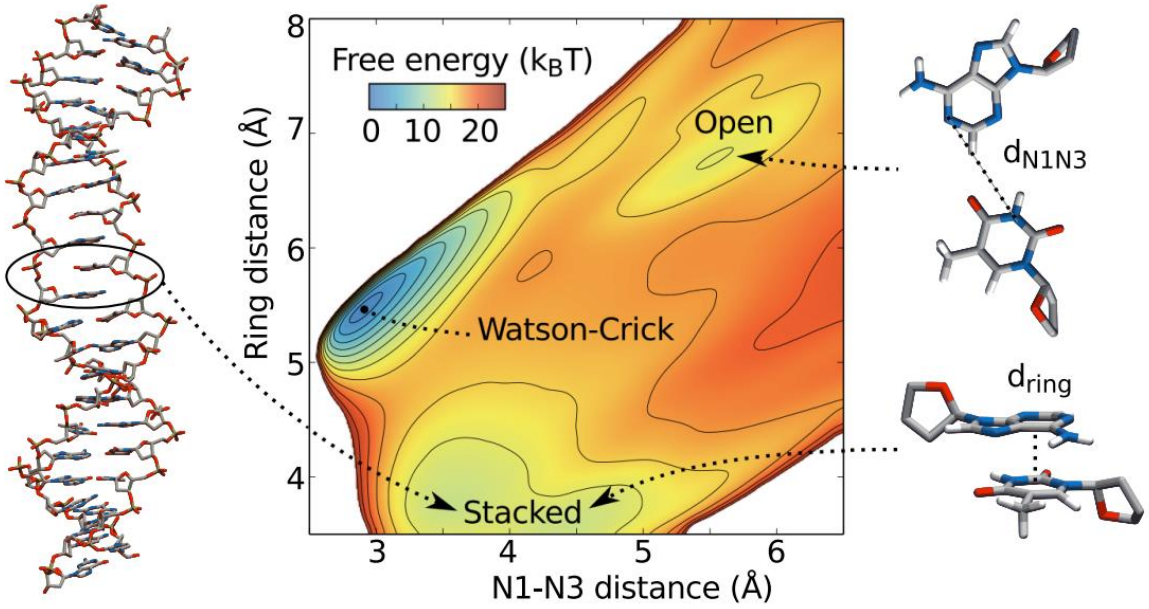


Figure 1-15: Two-dimensional free energy landscape for a ds-DNA base pair with a stacking coordinate on the *y* axis and a pairing coordinate on the *x* axis. Reproduced from Lindahl *et al* under a Creative Commons Attribution License.[100] Distance given in Ångstrom (Å) with 1Å = 0.1nm,

1.5.4 State of nearest-neighbour models for peptide nucleic acids

The nearest-neighbour model has been shown to predict the binding energies and entropies of DNA and RNA with high accuracy. The conclusions that can be made from the nearest-neighbour model are arguably more general since they are applicable regardless of sequence. Previous binding energy studies of peptide nucleic acid complexes have drawn conclusions based on direct comparisons between individual sequences, for example by comparing the same sequence in ds-DNA and ds-PNA, but this doesn't necessarily allow comparisons to be made between these molecules in general. For example, in PNA:DNA heteroduplexes, sometimes enthalpy-entropy compensation is observed in that a higher binding enthalpy is mitigated by a higher binding entropy, but for other cases this is not observed. Developing a nearest-neighbour model for PNA-containing molecules would undoubtedly be a useful tool to predict and differentiate energies for the general case.

In 1998 Giesen *et al*. provided an empirical formula for calculating the melting temperatures of PNA:DNA duplexes using a linear model of their length, normalised pyrimidine content between zero and unity, and the corresponding ds-DNA melting temperature as evaluated by the ds-DNA nearest-neighbour model.[101] Fitting the model to 316 melting temperatures revealed an R^2 of 0.87 with 90% of the evaluated melting temperatures being within 5K of their experimentally-determined values. Similar results made by dividing binding parameters by sequence length were determined by Ratilainen *et al.*[102] who determined average standard binding free energies, enthalpies and entropies of -1.56 kcal/mol, -7.17 kcal/mol and -18.84 cal/K mol per base pair respectively. Though neither of these are nearest-neighbour models, they do enable useful general comparisons with ds-DNA. For example, for ds-DNA, Privalov *et al*. determined -1.22 and -9.65 kcal/mol standard per base pair binding free energies and enthalpies respectively.[47] Since the average per base pair enthalpy for PNA:DNA was greater,

but the average per base pair free energy was lower, than their equivalents for ds-DNA, this supports the hypothesis of enthalpy-entropy compensation in PNA:DNA.

There have additionally been direct attempts to develop a nearest-neighbour model for PNA:DNA hybridisation. Griffin and Smith authored one such model in 1998, determining incremental enthalpies between -5.8 and -11.9 kcal/mol and incremental free energies between -0.9 and -3.6 kcal/mol. These values cover a broader range than their ds-DNA equivalents (Table 1-4) but were able to accurately fit a total of 13 DNA:PNA melting temperatures.[103] However, and as highlighted by Giesen *et al.*,[101] their model was parameterised using only 11 melting temperatures over a very narrow range of sequence lengths, raising the possibility that the model is overfitted to this small sample of sequences and that its incremental energies are not representative of a more diverse set of PNA:DNA binding energies. The validity of a nearest-neighbour model for PNA:DNA heteroduplexes was also tested by Sugimoto *et al.* in 2001, who studied the differences in binding energies of different sequences with otherwise identical lengths and occurrences of unique ds-stacks.[104]

Despite both Griffin and Smith and Sugimoto publishing articles relating to nearest-neighbour models for PNA:DNA hybrids, publications of exact incremental energies based on a broader range of PNA:DNA experimental energies are lacking in 2023. Though a 2019 study by Ghosh and Sugimoto *et al.*[105] states that nearest-neighbour models have been used to predict the stabilities of "duplexes formed by peptide nucleic acid (PNA) and DNA" they reference only the Sugimoto *et al.* 2001[104] study which presents no new incremental energies and makes no predictions based on such. In contrast to the initial steps in developing these models that have been made for heteroduplexes, there had been no attempts to determine a nearest-neighbour model for ds-PNA prior to the current body of work.

1.6 Nucleic acids under external force

1.6.1 Two distinct loading directions

In ds-DNA there are two primary modes of melting under external force: shear and unzipping.[106] In the shear mode the direction of pulling is parallel to the longitudinal axis of the double-helix whereas in the unzipping mode the direction of pulling is normal to this axis. Experimentally, for shearing, this is achieved by pulling from the 5' or 3' terminus of one strand and the same terminus of the other strand in antiparallel ds-DNA. The unzipping mode is studied by pulling from the 5' terminus of one strand and the 3' terminus of the other strand in antiparallel ds-DNA (Figure 1-16).

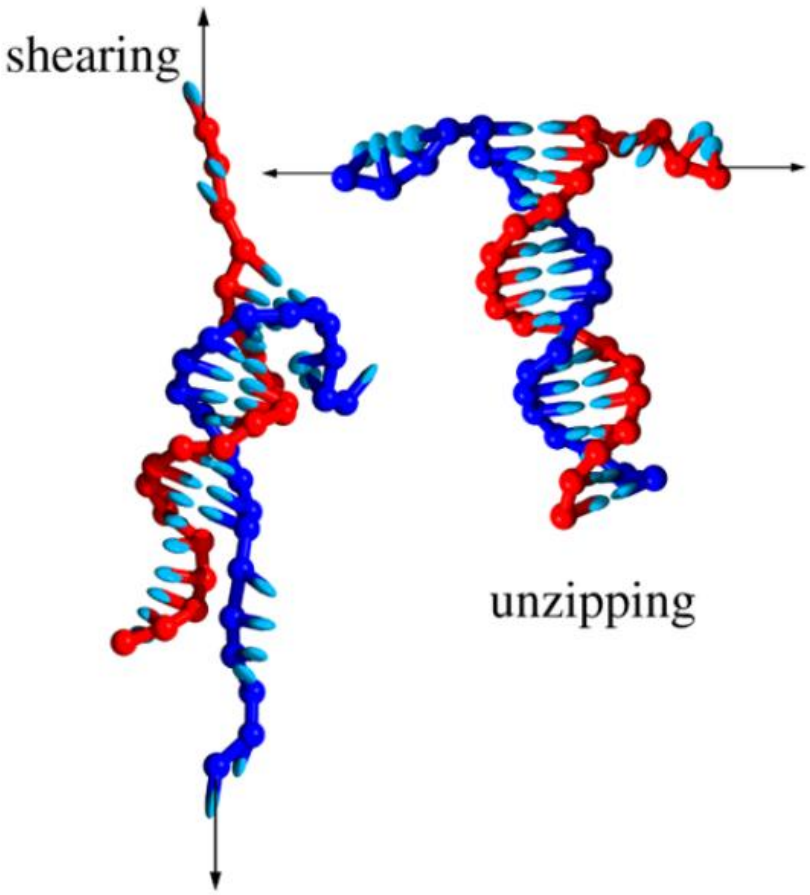


Figure 1-16: Cartoon of the two main modes of ds-DNA rupture reprinted with permission from Mosayebi *et al.* Copyright (2015) American Chemical Society.[106]

1.6.2 The de Gennes model for shearing short oligonucleotides

In 2001 Nobel laureate Pierre-Gilles de Gennes produced a foundational theory to understanding the effect of force on ds-DNA in the shearing mode by arguing that, as the length of a ds-DNA primary sequence approached infinity, the rupture force would approach an asymptotic value.[107] Two years earlier, in 1999, Strunz *et al*. had already demonstrated that the ds-DNA rupture force was a non-linear increasing function of sequence length[108] and the de Gennes model provided the basis for understanding this non-linearity. In 2008, Hatch *et al*. demonstrated results for oligonucleotides under 100 base pairs in length which were consistent with the de Gennes theory: rupture force increased linearly until approximately 20 base pairs but was already within 5% of the asymptotic rupture force by 32 base pairs.[109] In 2015, Mosayebi *et al*. reviewed this work and accounted for sequence composition by fitting the data of Hatch *et al*. to a model involving the average free energy per base pair, showing that the asymptotic rupture force increased with average base pair free energy.[106]

For short oligonucleotides, the de Gennes model therefore argues that the rupture force measured during shearing experiments of nucleic acids is an increasing function of length and average base pair free energy until a critical length at which the relationship flattens. This asymptote likely arises from the fact that, according to both Hatch *et al*. and Mosayebi *et al*., force is mostly distributed across approximately 10 base pairs near the pulling terminus during shearing.[106,109] Resultantly, according to the de Gennes model, an infinitely long ds-nucleic acid under shear force experiences loading on a finite number of terminal base pairs, hence leading to a length-independence in rupture force beyond a certain length.

1.6.3 Overstretched deoxyribonucleic acid

Several studies have demonstrated that, during the stretching of longer ds-DNA molecules, if the force on the molecule exceeds roughly 65 pN then the helix enters an ‘overstretched’ or S-form conformation with the distance per base pair being

approximately 1.7 times greater than in the B-form.[110,111,112,113] In 1994, for example, Lee *et al*. evidenced the S-form in 241 nm ds-DNA by observing that rupture forces were preceded by an elastic region in the force curve.[114] Danilowicz *et al*. argue that this is incompatible with the de Gennes theory[115] since the S-form has a higher length per base pair, resulting in a lower stacking energy and hence lower effective spring constant. For the S-form, then, force would theoretically be distributed across a greater number of base pairs than the terminal 10 pairs predicted by the de Gennes theory. For sequences longer than 500 base pairs, Danilowicz *et al*. demonstrated this incompatibility by observing that rupture force increased logarithmically with the loading rate but did not plateau, though the authors did note that they could not discount an asymptote for sequences longer than 10,000 base pairs.[115]

Though the overstretched form was observed in long chains this does not necessarily preclude overstretching in short chains should the force on such chains be able to exceed 65 pN. One of the limitations of both molecular dynamics and force spectroscopy experiments is that they each can only partly characterise the force-loading curve.[20,116] Notably, force spectroscopy is limited to lower loading rates,[117] whereas molecular dynamics is limited to higher rates. Since rupture force is a function of loading rate, this raises the possibility for overstretched ds-DNA conformations to be demonstrated in short sequences at high loading rates where the rupture force exceeds the 65 pN force of overstretching.

A 1999 molecular dynamics study by MacKerell Jr. and Lee[118] corroborated the 1994 experimental study by Lee *et al*.[114] that an elastic intramolecular region during shearing precedes a rapid intermolecular rupture event using a shorter nucleic acid. This indicated the existence of an overstretching conformation, and subsequent molecular dynamics studies demonstrated definitively its existence in oligonucleotides shorter than 30 base pairs at high loading rates (Figure 1-17).[117] This highlights an interesting use case for molecular dynamics in that, for some

molecules, it can explore conformations that are inaccessible to experimental setups.

Figure 1-17: Snapshots of a ds-DNA molecule during the process of shearing, with increasing force from left to right. Near the right end the extended S-form conformation is evident. Reproduced from Naserian-Nik *et al.*, Copyright (2013), with permission from RSC Publishing (License No.: 1360504-1).[117]

In addition to demonstrating the existence of the S-form in short oligonucleotides from structural analysis, MacKerell Jr. and Lee reconstructed the free energy landscape during shearing.[118] They showed that this elastic region is characterised by a near-constant free energy surface and that, at the maximum stretching distance, a steep and narrow barrier is encountered in the free energy surface which must be surmounted for rupture to occur. This elastic region represents the process of overstretching until a critical distance at which rupture occurs.

1.6.4 The length independence of unzipping

During shearing, the rupture force is dependent on length and loading rate. During unzipping, however, it has been argued that the rupture force is independent of both length and loading rate.[119] This is arguable since unzipping can be thought of as the

sequential breaking of single base pairs, and so any force accumulating on the molecule is constantly dissipated meaning the maximum force reached does not increase with length. Because base pairs break sequentially, it has also been demonstrated that these forces are directly sequence-dependent.[120,121]

According to the force spectroscopy models, however, the rupture force of each individual base pair in the sequence is theoretically still a function of the loading rate.[19] As it turned out, the process of unzipping was more nuanced than simply being entirely independent of length or loading rate, bringing it in line with the theoretical models. In 2001, Cocco, Monasson and Marko demonstrated that the rupture force was dependent on the sequence length and loading rate for rates below approximately 10^{-8} Ns^{-1}, but lost this dependence at intermediate loading rates of about 10^{-5} Ns^{-1}.[122] At even higher rates, the rupture force again became a function of loading rate only because the nucleic acid could not respond to the force before it became large. The authors reported a rate- and length-independent rupture force of approximately 12 pN, which is lower than the forces reported in shearing ds-DNA such as the 65 pN force needed to stretch B-form DNA into the S-form. This demonstrates that unzipping in ds-DNA requires less force to occur than shearing.

Force spectroscopy models have been fitted to unzipping studies of a DNA hairpin secondary structure in ss-DNA. During the process of unzipping, a nonlinear relationship between the unzipping time and force, expressed as a voltage, was observed, meaning the full curve could not be fitted by the Bell-Evans formula. Instead, a Dudko-Hummer-Szabo fit was conducted, revealing a transition barrier height of $\Delta G^{\ddagger}$ of 7.05 $\mathrm{kcal/mol}$.[123] Since, in the process of unzipping, base pairs rupture sequentially and the highest stability individual ds-stack has an incremental free energy of only 2.24 $\mathrm{kcal/mol}$ in the unified model (Table 1-4), then this result reveals the existence of a large transition barrier opposing base pair rupture that is much greater than the incremental free energies of the ds-stacks themselves.

1.6.5 State of peptide nucleic acid external force studies

Similarly to the state of PNA:DNA nearest-neighbour models, there are only limited studies on the rupture of PNA:DNA hybrids under shear load. For example, a study on the rupture of a PNA:DNA 6-mer revealed that even such short lengths formed stable complexes with detectable rupture forces. This contrasts with ds-DNA which cannot form stable strands under standard conditions for sequences shorter than 8 base pairs.[124] This higher stability during shearing was also noted by Dutta, Armitage and Lyubchenko in 2016 for a 10-mer sequence.[125]

Of potential interest to the overstretching of PNA:DNA hybrids is that rupture forces of approximately 65 pN for PNA:DNA were reported at low, experimentally-obtainable loading rates.[125] Assuming the existence of a PNA:DNA S-form, and assuming the transition to said form occurs at a similar stretching force of 65 pN,[110,111,112,113] this raises the possibility of studying overstretching in short PNA:DNA hybrids using loading rates available to atomic force microscopes.

Mirroring the absence of nearest-neighbour studies for ds-PNA, the literature search did not reveal any force spectroscopy studies for PNA homoduplexes.

1.7 Project aims

1.7.1 Gaps in the current literature

To establish a baseline for the single-molecular study of ds-PNA as a candidate bioadhesive using molecular dynamics, the extent of current understanding on its binding stabilities should be assessed. Theories to quantify the binding stability of biomolecular associations are well established from the equilibrium binding free energy[16] to rupture forces to non-equilibrium force spectroscopy models.[19,22,24] Computational methods to quantify these equilibrium and non-equilibrium quantities are well established. These computational tools, discussed at length in the

following chapter, form the foundation for the present work on the equilibrium and non-equilibrium binding stabilities of ds-PNA.

Regarding ds-DNA and ds-RNA, a clear, consistent research effort in deriving binding stability quantities is evident. A key development in equilibrium binding stability studies specific to nucleic acids is the decomposition of binding entropies, enthalpies and free energies into primary sequence parameters, as per the nearest-neighbour model.[79] Similar key developments in non-equilibrium studies for nucleic acids is the overstretching of ds-DNA for shear forces exceeding 65 pN[110,111,112,113] and the significantly lower peak forces during unzipping.[122] Extensions of the current literature on ds-DNA could involve simulations of overstretched ds-DNA in shorter oligonucleotides, due to the higher accessible loading rates, and derivation of incremental energies for synthetic nucleobases.

There have been attempts to replicate the equilibrium and non-equilibrium binding studies of DNA and RNA with PNA, for example in the derivation of incremental energies for PNA heteroduplexes[103] or in the decomposition of binding enthalpies by sequence length.[102] PNA heteroduplexes have been sheared, and peak forces near the overstretching regime of ds-DNA observed.[125] Notably absent from that study, however, is any discussion of overstretching in PNA heteroduplexes. Similarly absent from the literature are any studies of external forces, and any derivations of incremental energies, for ds-PNA.

In general, although there were a spur of studies recording the binding energies and entropies of PNA hetero- and homo-duplexes in the late 1990's and early 2000's,[75,76,101,102,103,104] interest has seemingly waned since. There are probably two reasons for this: firstly, interest in PNA's applications continues. For example, studies on duplex invasion wherein an ss-PNA binds to a ds-DNA duplex for a potential therapeutic purpose are reported, and so focus may have simply shifted.[49,50,70] Secondly the binding energies of PNA-containing duplexes were

largely published by the same personalities. For example, Peter E. Nielsen, who first developed the synthetic analogue, has appeared 15 times in the references at this point with only 2 sources dated after 2005. Nielsen's publications since then indicate a shift in focus towards applications, and where kinetic and thermodynamic data is available it is for individual sequences without mention of generalising data into a binding model.[126,127] Likewise, Sugimoto or Griffin and Smith, who studied the validity of a nearest-neighbour model[104] or developed one with a limited sample size,[103] have not pursued it in future research. Ghosh and Sugimoto's 2019 quote[105] that nearest-neighbour models have been used to "study the stability" of PNA-containing duplexes, without referencing any such model, may even indicate that they consider the matter closed.

The lower interest in ds-PNA may be because, whereas heteroduplexes have clear applications in antigene or antisense technologies,[68,70,71] the applications of ds-PNA are less clear. ds-PNA as a candidate bioadhesive as proposed in the present work is an example application with a literature search failing to reveal any prior suggestions of this.

1.7.2 Project aims

By addressing the literature, the research question, "Can molecular dynamics be used to assess and screen the single-molecular binding properties of a candidate bioadhesive?", can be answered. Firstly, the property of binding stability is assessed in terms of the binding energy, as was addressed in subchapter 1.2. Secondly, the equilibrium binding energies of ds-nucleic acids can be decomposed into primary sequence parameters according to the nearest-neighbour model, as was addressed in subchapter 1.5. Thirdly, the binding stability of nucleic acids under external force can be quantified by the force-loading curve or the underlying free energy landscape. Therefore, to answer the research question, the efficacy of molecular dynamics in deriving binding energies, force-loading curves and underlying free energy landscapes of the candidate bioadhesive ds-PNA must be assessed.

To quantify the binding stability of ds-PNA the present work therefore achieves the following goals:

(1) A complete nearest-neighbour model for predicting ds-PNA binding energies from primary sequence is developed. This is presented in Chapter 4 and addresses technical research question (1).
(2) The free energy landscapes associated with base pairing and stacking during shearing are derived. The contributions of different energy terms, such as torsional angle energies, during the processes of unzipping and shearing are presented. The force-loading relationship during shearing and unzipping at high loading rates is expressed. This is presented in Chapter 5 and addresses technical research question (2).
(3) A new computational model for ds-PNA allowing a broader force-loading curve to be derived is created. This is presented in Chapter 6 and addresses technical research question (3).

1.7.3 How the literature informed methods in the present work

To develop the nearest-neighbour model the method outlined by Golyshev, Pyshnyi and Lomzov in 2021[96] to use MM-PB[GB]SA to measure binding energies and entropies, is used given its accuracy and inexpensiveness. This was since, as discussed in criticisms of Griffin and Smith's work[101], a large number of sequences of different lengths would need to be tested, and alternative methods like alchemical free energy calculations, as used by Sakuraba, Asai and Kameda[92], were more resource intensive. Unlike Golyshev, Pyshnyi and Lomzov, however, a linear correction would not be used due to the $\Delta G \neq \Delta H - T\Delta S$ inequality it introduces for the incremental energies and entropies of binding.

Since the de Gennes' model of ds-DNA indicated that external force in shear loads mostly onto the terminal 10 base pairs,[107] this provided justification for the sequence selection used in this work. Namely, due to the available computational

power, a single ds-PNA sequence was selected for all studies of ds-PNA under external force. Since GTAGATCACT is the most well-characterised ds-PNA sequence in the literature[58,62,65,75,78,80] in terms of its binding properties it would be the most natural pilot sequence for all force studies, and by also being 10 base pairs long this selection is further justified by the suggestion from the de Gennes model that a sequence of this length would be capable of capturing the impact of loading on structure.

To extend the possible simulation timescales, a new 'coarse-grained' model of ds-PNA is developed in this work. This coarse-grained model is based on the Martini framework (Chapter 2.1) for other nucleic acids[128,129] and allows longer simulation timescales to be achieved by clustering atoms into beads to reduce the degrees of freedom of the simulation.

Finally, the results of molecular dynamics simulations are validated against binding energies and entropies obtained from the literature.

1.7.4 Outline of future chapters

Chapter 2 introduces the theory and methods of the computational approaches used in the current work. During molecular dynamics it is important for systems to be properly equilibrated and for computed structures to represent experimental data, so Chapter 3 presents equilibration and structural data including the helical parameters for ds-PNAs used in the present work. Chapter 4 presents the derivation of the nearest-neighbour model for ds-PNA using MM-GBSA. Chapter 5 presents external force studies on all-atom ds-PNA alongside base pairing and stacking energy landscapes during shearing. Chapter 6 presents the derivation of the Martini model for ds-PNA and a partial force-loading curve from the shearing of Martini ds-PNA including its overlap with all-atom simulations data. Chapter 7 discusses the significance of these results and their place within, and contributions to, the literature on equilibrium and external force studies of ds-PNA. In addition, Chapter

7 reflects on the extensibility of the present results as a useful metric on the stabilities of bioadhesives of different chemical makeup and on the macroscale.

1.8 References

[1] Anthony von Fraunhofer, J. Adhesion and Cohesion. *Int. J. Dent.* **2012**, *2012*, 951324
[2] Jin, Fan-Long.; Li, X.; Park, Soo-Jin. Synthesis and application of epoxy resins: A review. *J. Ind. Eng. Chem.* **2015**, *29*, 1-11
[3] Khanlari, S.; Dubé, M. Bioadhesives: A Review. *Macromol. React. Eng.* **2013,** *7*, 573-587
[4] Li, X.; Li, S.; Huang, X.; Chen, Y.; Cheng, J.; Zhan, Z. Protein-mediated bioadhesion in marine organisms: A review. *Marine Environmental Research* **2021**, *170*, 105409
[5] Lee, H.; Scherer, N.; Messersmith, P. Single-molecule mechanics of mussel adhesion. *Proc. Natl. Acad. Sci.* **2006**, *103*, 12999-13003
[6] Yan, S.; Verestek, W.; Zeizinger, H.; Schmauder, S. Characterization of Cure Behavior in Epoxy Using Molecular Dynamics Simulation Compared with Dielectric Analysis and DSC. *Polymers* **2021**, *13*, 3085
[7] Ninan, L.; Monahan, H.; Stroshine, R.; Wilker, J.; Shi, R. Adhesive strength of marine mussel extracts on porcine skin. *Biomaterials* **2003**, *24*, 4091-4099
[8] Jelić, A.; Sekulić, M.; Travica, M.; Gržetić, J.; Ugrinović, V.; Marinković, A.; Božić, A.; Stamenović, M.; Putić, S. Determination of Mechanical Properties of Epoxy Composite Materials Reinforced with Silicate Nanofillers Using Digital Image Correlation (DIC). *Polymers (Basel)* **2022**, *14*, 1255
[9] Bucior, I.; Scheuring, S.; Engel, A.; Burger, M. Carbohydrate-carbohydrate interaction provides adhesion force and specificity for cellular recognition. *J. Cell. Biol.* **2004**, *165*, 529-537
[10] Sundaram, M.; Kaliannagounder, V.; Selvaprithiviraj, V.; Suresh, M.; Biswas, R.; Vasudevan, A.; Varma, P.; Jayakumar, P. Bioadhesive, Hemostatic and Antibacterial *in Situ* Chitin-Fibrin Nanocomposite Gel for Controlling Bleeding and Preventing Infections at Mediastinum. *ACS Sustainable Chem. Eng.* **2018**, *6*, 7826-7840
[11] Allen, P.; Khaing, Z.; Schmidt, C.; Ellington, A. 3D Printing with Nucleic Acid Adhesives. *ACS Biomater. Sci. Eng.* **2015**, *1*, 19-26
[12] Duan, W.; Bian, X.; Bu, Y. Applications of Bioadhesives: A Mini Review. *Front. Bioeng. Biotechnol.* **2021**, *9*
[13] Tzagiollari, A.; McCarthy, H.; Levingstone, T.; Dunne, N. Biodegradable and Biocompatible Adhesives for the Effective Stabilisation, Repair and Regeneration of Bone. *Bioengineering (Basel)*. **2022**, *9*
[14] McDevitt, J.; Grigsby, W. Life Cycle Assessment of Bio- and Petro- Chemical Adhesives Used in Fiberboard Production. *J. Polym. Environ.* **2014**, *22*
[15] Van der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A.; Berendsen, H. GROMACS: Fast, Flexible and Free. *J. Comput. Chem.* **2005**, *26*, 1701-1718
[16] Stowe, K. An Introduction to Thermodynamics and Statistical Mechanics (2nd ed.). *Cambridge University Press* **2007**, ISBN 0-521-86577-3
[17] Laidler, K. A glossary of terms used in chemical kinetics, including reaction dynamics (IUPAC Recommendations 1996). *Pure & Appl. Chem.* **1996**, *1*, 149-192
[18] Bell, G. Models for the Specific Adhesion of Cells to Cells. *Science* **1978**, *200*, 618-627
[19] Evans, E.; Ritchie, K. Dynamic Strength of Molecular Adhesion Bonds. *Biophysical Journal* **1997**, *72*, 1541-1555
[20] Rico, F.; Russek, A.; González, L.; Grubmüller, H.; Scheuring, S. Heterogeneous and rate-dependent streptavidin-biotin unbinding revealed by high-speed force spectroscopy and atomistic simulations. *Proc. Natl. Acad. Sci.* **2019**, *116*, 6594-6601
[21] Adhikari, S.; Beach, K. Reliable extraction of energy landscape properties from critical force distributions. *Phy. Rev. Res.* **2020**, *2*, 023276
[22] Dudko, O.; Hummer, G.; Szabo, A. Intrinsic Rates and Activation Free Energies from Single-Molecule Pulling Experiments. *Phys. Rev. Lett.* **2006**, *96*, 108101
[23] Hane, F.; Attwood, S.; Neonenko, Z. Comparison of three competing dynamic force spectroscopy models to study binding forces of amyloid-β (1-42). *Soft Matter* **2014**, *10*, 1924-1930
[24] Friddle, R.; Noy, A.; De Yoreo, J. Interpreting the widespread nonlinear force spectra of intermolecular bonds. *Proc. Natl. Acad. Sci.* **2013**, *110*, 13573-13578
[25] Van der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A.; Berendsen, H. GROMACS: Fast, Flexible and Free. *J. Comput. Chem.* **2005**, *26*, 1701-1718

[26] Berman, H.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.; Weissig, H.; Shindyalov, I.; Bourne, P. The Protein Data Bank. *Nucleic Acids Research*, **2000**, *28*, 235-242
[27] Strout, D.; Scuseria, G. A quantitative study of the scaling properties of the Hartree-Fock method. *J. Chem. Phys.* **1995**, *102*, 8448
[28] Parr, R.; Craig, D.; Ross, I. Molecular Orbital Calculations of the Lower Excited Electronic Levels of Benzene, Configuration Interaction Included. *J. Chem. Phys.* **1950**, *18*, 1561-1563
[29] Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; de Groot, B.; Grubmüller, H.; MacKerell Jr, A. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nature Methods*, **2016**, *14*, 71-73
[30] MacKerell Jr., A. Empirical Force Fields for Biological Macromolecules: Overview and Issues. *J. Comput. Chem.* **2004**, *25*, 1584-1604
[31] Marky, L.; Breslaur, K. Calculating Thermodynamic Data for Transitions of any Molecularity from Equilibrium Melting Curves. *Biopolymers* **1987**, *26*, 1601-1620
[32] Schroeder, S.; Turner, D. Optical Melting Measurements of Nucleic Acid Thermodynamics. *Methods Enzymol.* **2009**, *468*, 371-387
[33] Goodman, J.; Attwood, D.; Kiely, J.; Coladas Mato, P.; Luxton, R. Modeling Peptide Nucleic Acid Binding Enthalpies Using MM-GBSA. *J. Phys. Chem. B.* **2022**, *126*, 9528-9538
[34] Janshoff, A.; Neitzert, M.; Oberdörfer, Y.; Fuchs, H. Force Spectroscopy of Molecular Systems – Single Molecule Spectroscopy of Polymers and Biomolecules. *Angew. Chem. Int. Ed.* **2000**, *39*, 3212-3237
[35] Wong, J.; Chilkoti, A.; Moy, V. Direct force measurements of the streptavidin-biotin interaction. *Biomol. Eng.* **1999**, *16*, 45-55
[36] Hinterdorfer, P.; Baumgartner, W.; Gruber, H.; Schilcher, K.; Schindler, H. Detection and localization of individual antibody-antigen recognition events by atomic force microscopy. *Proc. Natl. Acad. Sci.* **1996**, *93*, 3477-3481
[37] Ritzefeld, M.; Walhorn, V.; Anselmetti, D.; Sewald, N. Analysis of DNA interactions using single-molecule force spectroscopy. *Amino Acids* **2013**, *44*, 1457-1475
[38] Janshoff, A.; Neitzert, M.; Oberdörfer, Y.; Fuchs, H. Force Spectroscopy of Molecular Systems – Single Molecule Spectroscopy of Polymers and Biomolecules. *Angew. Chem. Int. Ed.* **2000**, *39*, 3212-3237
[39] Lyubchenko, Y. Amyloid misfolding, aggregation, and the early onset of protein deposition diseases: insights from AFM experiments and computational analyses. *AIMS Molecular Science* **2015**, *2*, 190-210
[40] Watson, J.; Crick, F. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature* **1953**, *171*, 737-738
[41] Humphrey, W.; Dalke, A.; Schulten, K. VMD – Visual Molecular Dynamics. *J. Molec. Graphics* **1996**, *14*, 33-38
[42] Lan, W.; Hu, Z.; Shen, J.; Wang, C.; Jiang, F.; Liu, H.; Long, D.; Liu, M.; Cao, C.; Structural investigation into physiological DNA phosphorothioate modification. *Scientific Reports* **2016**, *6*, 25737
[43] Spencer, M. The Stereochemistry of Deoxyribonucleic Acid. II. Hydrogen-Bonded Pairs of Bases. *Acta Cryst.* **1959**, *12*, 66-71
[44] Sinden, R. DNA Structure and Function (1st ed.). *Academic Press* **1994**, ISBN 0-12-645750-6.
[45] Khandelwal, G.; Bhyravabhotla, J. A Phenomenological Model for Predicting Melting Temperatures of DNA Sequences. *PLoS One* **2010**, *5*, e12433
[46] Privalov, P.; Monaselidze, D.; Mrevlishvili, G.; Magaldadze, V. "Intramolecular" heat of fusion of macromolecules. *J. Exptl. Theoret. Phys.* **1965**, *47*, 2073-2079.
[47] Privalov, P.; Ptitsyn, O. Determination of Stability of the DNA Double Helix in an Aqueous Medium. *Biopolymers* **1969**, *8*, 559-571
[48] Weng, Y.; Huang, Q.; Li, C.; Yang, Y.; Wang, X.; Yu, J.; Huang, Y.; Liang, X. Improved Nucleic Acid Therapy with Advanced Nanoscale Biotechnology. *Mol. Ther. Nucleic Acids* **2020**, *19*, 581-601
[49] Taskova, M.; Mantsiou, A.; Astakhova, K. Synthetic Nucleic Acid Analogues in Gene Therapy: An Update for Peptide-Oligonucleotide Conjugates. *ChemBioChem.* **2017**, *18*, 1671-1682
[50] Veedu, R.; Wengel, J. Locked Nucleic Acids: Promising Nucleic Acid Analogs for Therapeutic Applications. *Chemistry & Biodiversity* **2010**, *7*, 536-542
[51] Moulton, J. Using Morpholinos to Control Gene Expression. *Curr. Protoc. Nucleic Acid Chem.* **2006**, *27*, 4.30.1-4.30.24
[52] Wittung, P.; Nielsen, P.; Buchardt, O.; Egholm, M.; Nordén, B. DNA-like double helix formed by peptide nucleic acid. *Nature* **1994**, *368*, 561−563
[53] Wittung, P.; Eriksson, W.; Lyng, R.; Nielsen, P.; Nordén, B. Induced Chirality in PNA-PNA Duplexes. *J. Am. Chem. Soc.* **1995**, *117*, 10168-10173
[54] Nielsen, P. Peptide Nucleic Acid. A Molecule with Two Identities. *Acc. Chem. Res.* **1999**, *32*, 624−630.
[55] Nielsen, P. Peptide nucleic acid (PNA). A DNA mimic with a pseudopeptide backbone. *Chem. Soc. Rev.* **1997**, *2*, 73-78
[56] Sen, S.; Nilsson, L. Molecular Dynamics of Duplex Systems Involving PNA: Structural and Dynamical Consequences of the Nucleic Acid Backbone. *J. Am. Chem. Soc.* **1998**, *120*, 619-631

[57] He, W.; Hatcher, E.; Balaeff, A.; Beratan, D.; Gil, R.; Madrid, M.; Achim, C. Solution Structure of a Peptide Nucleic Acid Duplex from NMR Data: Features and Limitations. *J. Am. Chem. Soc.* **2008**, *130*, 13264-13273

[58] Tomac, S.; Sarkar, M.; Ratilainen, T.; Wittung, P.; Nielsen, P.; Nordén, B.; Gräslund, A. Ionic Effects on the Stability and Conformation of Peptide Nucleic Acid Complexes. *J. Am. Chem. Soc.* **1996**, *118*, 5544−5552

[59] Eriksson, M.; Nielsen, P. E. PNA-nucleic acid complexes. Structure, stability and dynamics. *Q. Rev. Biophys.* **1996**, *29*, 369−394

[60] Nielsen, P. E.; Egholm, M.; Berg, R. H.; Buchardt, O. Sequence-Selective Recognition of DNA by Strand Displacement with a Thymine-Substituted Polyamide. *Science* **1991**, *254*, 1497−1500

[61] Demidov, V.; Potaman, V.; Frank-Kamenetskil, M.; Egholm, M.; Buchard, O.; Sönnichsen, S.; Nielsen, P. Stability of peptide nucleic acids in human serum and cellular extracts. *Biochemical Pharmacology* **1994**, *48*, 1310-1313

[62] Sen, A.; Nielsen, P. On the stability of peptide nucleic acid duplexes in the presence of organic solvents. *Nucleic Acids Res.* **2007**, *35*, 3367-3374

[63] Shakeel, S.; Karim, S.; Ali, A. Peptide nucleic acid (PNA) – a review. *J. Chem. Technol. Biotechnol.* **2006**, *81*, 892-899

[64] Jasiński, M.; Miszkiewicz, J.; Feig, M.; Trylska, J. Thermal Stability of Peptide Nucleic Acid Complexes. *J. Phys. Chem. B* **2019**, *123*, 8168−8177

[65] Sen, A.; Nielsen, P. Unique Properties of Purine/Pyrimidine Asymmetric PNA·DNA Duplexes: Differential Stabilization of PNA· DNA Duplexes by Purines in the PNA Strand. *Biophys. J.* **2006**, *90*, 1329−1337

[66] Shakeel, S.; Karim, S.; Ali, A. Peptide nucleic acid (PNA) – a review. *J. Chem. Technol. Biotechnol.* **2006**, *81*, 892-899

[67] Toneli, R.; Purgato, S.; Camerin, S.; Fronza, R.; Bologna, F.; Alboresi, S.; Franzoni, M.; Corradini, R.; Sforza, S.; Faccini, A.; Shohet, S.; Marchelli, R.; Pession, A. Anti-gene peptide nucleic acid specifically inhibits MYCN expression in human neuroblastoma cells leading to cell growth inhibition and apoptosis. *Mol. Cancer Ther.* **2005**, *4*, 779-785

[68] Gupta, A.; Mishra, A.; Puri, N. Peptide nucleic acids: Advanced tools for biomedical applications. *J. Biotech.* **2017**, *259*, 148-159

[69] Akisawa, T.; Ishizawa, Y.; Nagatsugi, F. Synthesis of peptide nucleic acids containing a crosslinking agent and evaluation of their reactivities. *Molecules* **2015**, *20*, 4708-4719

[70] Montazersaheb, S.; Hejazi, M.; Charoudeh, H. Potential of Peptide Nucleic Acids in Future Therapeutic Applications. *Adv. Pharm. Bull.* **2018**, *8*, 551-563

[71] Singh, K.; Sridevi, P.; Singh, R. Potential applications of peptide nucleic acid in biomedical domain. *Engineering Reports* **2020**, *2*, e12238

[72] Wu, J.; Meng, Q.; Ren, H.; Wang, H.; Wu, J.; Wang, Q. Recent advances in peptide nucleic acid for cancer bionanotechnology. *Acta Pharmacologica Sinica* **2017**, *38*, 798-805

[73] Wu, J.; Zou, Y.; Li, C.; Sicking, W.; Piantanida, I.; Yi, T.; Schmuck, C. A Molecular Peptide Beacon for the Ratiometric Sensing of Nucleic Acids. *J. Am. Chem. Soc.* **2012**, *134*, 1958-1961

[74] Jeong, S.; Kim, J.; Jeong, S.; Bae, I.; Song, W. Evaluation of peptide nucleic acid-mediated multiplex real-time PCR kits for rapid detection of carbapenemase genes in gram-negative clinical isolates. *J. Microbiol. Methods* **2015**, *113*, 4-9

[75] Ratilainen, R.; Holmén, A.; Tuite, E.; Haaima, G.; Christensen, L.; Nielsen, P.; Nordén, B. Hybridization of Peptide Nucleic Acid. *Biochemistry* **1998**, *37*, 12331-12342

[76] Schwarz, F.; Robinson, S.; Butler, J. Thermodynamic comparison of PNA/DNA and DNA/DNA hybridization reactions at ambient temperature. *Nucleic Acids Research* **1999**, *27*, 4792-4800

[77] Jensen, K.; Ørum, H; Nielsen, P.; Nordén, B. Kinetics for Hybridization of Peptide Nucleic Acids (PNA) with DNA and RNA Studied with the BIAcore technique. *Biochemistry* **1997**, *36*, 5072-5077

[78] Totsingan, F. Synthesis and Applications of PNA and Modified PNA in Nanobiotechnology. **2007**. Ph.D. Thesis, University of Parma, Italy.

[79] SantaLucia Jr., J. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl. Aca. Sci.* **1998**, *95*, 1460-1465

[80] Sforza, S.; Haaima, G.; Marchelli, R.; Nielsen, P. Chiral Peptide Nucleic Acids (PNAs): Helix Handedness and DNA Recognition. *Eur. J. Org. Chem.* **1999**, *1999*, 197-204

[81] Privalov, P.; Ptitsyn, O. Determination of Stability of the DNA Double Helix in an Aqueous Medium. *Biopolymers* **1969**, *8*, 559-571

[82] Marmur, J.; Doty, P. Heterogeneity in Deoxyribonucleic Acids 1. Dependence on Composition of the Configurational Stability of Deoxyribonucleic Acids. *Nature* **1959**, *183*, 1427-1429

[83] Marmur, J; Doty, P. Determination of the Base Composition of Deoxyribonucleic Acid from its Thermal Denaturation Temperature. *J. Mol. Biol.* **1962**, *5*, 109-118

[84] Bunville, L.; Geiduschek, E.; Rawitscher, M.; Sturtevant, J. Kinetics and Equilibria in the Acid Denaturation of Deoxyribonucleic Acids from Various Sources. *Biopolymers* **1965**, *3*, 213-240

[85] Wells, R.; Larson, J.; Grant, R. Physicochemical Studies on Polydeoxyribonucleotides Containing Defined Repeating Nucleotide Sequences. *J. Mol. Biol.* **1970**, *54*, 465-497
[86] Tinoco, I.; Borer, P.; Dengler, B.; Levine, M.; Uhlenbeck, O.; Crothers, D.; Gralla, J. Improved Estimation of Secondary Structure in Ribonucleic Acids. *Nature New Biology* **1973**, *246*, 40-41
[87] Filimonov, V.; Privalov, P. Thermodynamics of Base Interaction in $(A)_n$ and $(A{\cdot}U)_n$. *J. Mol. Biol.* **1978**, *122*, 465-470
[88] Gotoh, O. Tagashira, Y. Stabilities of nearest-neighbor doublets in double-helical DNA determined by fitting calculated melting profiles to observed profiles. *Biopolymers* **1981**, *20*, 1033-1042
[89] Marky, L.; Breslauer, K. Calorimetric Determination of Base-Stacking Enthalpies in Double-Helical DNA Molecules. *Biopolymers* **1982**, *21*, 2185-2194
[90] Breslauer, K.; Frank, R.; Blöcker, H.; Marky, L. Predicting DNA duplex stability from the base sequence. *Proc. Natl. Acad. Sci.* **1986**, *83*, 3746-3750
[91] SantaLucia Jr., J.; Allawi, H.; Seneviratne, P. Improved nearest-neighbor parameters for predicting DNA duplex stability. *Biochemistry* **1996**, *35*, 3555-62
[92] Sakuraba, S.; Asai, K.; Kameda, T. Predicting RNA Duplex Dimerization Free-Energy Changes upon Mutations Using Molecular Dynamics Simulations. *J. Phys. Chem. Lett.* **2015**, *6*, 4348-4351
[93] Beveridge, D.; DiCapua, F. Free Energy Via Molecular Simulation: Applications to Chemical and Biomolecular Systems. *Annu. Rev. Biophys. Biophys. Chem.* **1989**, *18*, 431-492
[94] Nishida, S.; Sakuraba, S.; Asai, K., Hamada, M. Estimating energy parameters for RNA secondary structure predictions using both experimental and computational data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **2019**, *16*, 1645-1655
[95] Sakuraba, S.; Iwakiri, J.; Hamada, M.; Kameda, T.; Tsuji, G.; Kimura, Y.; Abe, H.; Asai, H. Free-Energy Calculation of Ribonucleic Inosines and Its Application to Nearest-Neighbor Parameters. *J. Chem. Theory Comput.* **2020**, *16*, 5923-5935
[96] Golyshev, V.; Pyshnyi, D.; Lomzov, A. Calculation of Energy for RNA/RNA and DNA/RNA Duplex Formation by Molecular Dynamics Simulation. *Molecular Biology* **2021**, *55*, 927-940
[97] Lee, M.; Salsbury, F.; Brooks, C. Novel generalized Born methods. *J. Chem. Phys.* **2002**, *116*, 10606-10614
[98] Norberg, J.; Nilsson, L. Potential of Mean Force Calculations of the Stacking-Unstacking Process in Single-Stranded Deoxyribodinucleoside Monophosphates. *Biophysical Journal* **1995**, *69*, 2277-2285
[99] Stofer, E.; Chipot, C.; Lavery, R. Free Energy Calculations of Watson-Crick Base Pairing in Aqueous Solution. *J. Am. Chem. Soc.* **1999**, *121*, 9503-9508
[100] Lindahl, V.; Villa, A.; Hess, B. Sequence dependency of canonical base pair opening in the DNA double helix. *PLoS Comput. Biol.* **2017**, *13*, e1005463
[101] Giesen, U.; Kleider, W.; Berding, C.; Geiger, A.; Ørum, H.; Nielsen, P. A formula for thermal stability (T_m) prediction of PNA/DNA duplexes. *Nucleic Acids Research* **1998**, *26*, 5004-5006
[102] Ratilainen, T.; Holmén, A.; Tuite, E.; Nielsen, P.; Nordén, B. Thermodynamics of sequence-specific binding of PNA to DNA. *Biochemistry* **2000**, *39*, 7781-7791
[103] Griffin, T.; Smith, L. An Approach to Predicting the Stabilities of Peptide Nucleic Acid:DNA Duplexes. *Analytical Biochemistry* **1998**, *260*, 56-63
[104] Sugimoto, N.; Satoh, N.; Yasuda, K.; Nakano, S. Stabilization Factors Affecting Duplex Formation of Peptide Nucleic Acid with DNA. *Biochemistry* **2001**, *40*, 8444-8451
[105] Ghosh, S.; Takahashi, S.; Ohyama, T.; Endoh, T.; Tateishi-Karimata, H.; Sugimoto, N. Nearest-neighbor parameters for predicting DNA duplex stability in diverse molecular crowding environments. *Proc. Natl. Acad. Sci.* **2020**, *117*, 14194-14201
[106] Mosayebi, M.; Louis, A.; Doye, J.; Ouldrige, T. Force-Induced Rupture of a DNA Duplex: From Fundamentals to Force Sensors. *ACS Nano* **2015**, *9*, 11993-12003
[107] De Gennes, P. Maximum pull out force on DNA hybrids. *Comptes Rendus de l'Académie des Sciences – Series IV – Physics* **2001**, *2*, 1505-1508
[108] Strunz, T.; Oroszlan, K.; Schäfer, R.; Güntherodt, H. Dynamic force spectroscopy of single DNA molecules. *Proc. Natl. Acad. Sci.* **1999**, *96*, 11277-11282
[109] Hatch, K.; Danilowicz, C.; Coljee, V.; Prentiss, M. Demonstration that the shear force required to separate short double-stranded DNA does not increase significantly with sequence length for sequences longer than 25 base pairs. *Phys. Rev. E.* **2008**, *78*, 011920
[110] Rief, M.; Clausen-Schaumann, H.; Gaub, H. Sequence-dependent mechanics of single DNA molecules. *Nature Structural Biology* **1999**, *6*, 346-349
[111] Smith, S.; Cui, Y.; Bustamante, C. Overstretching B-DNA: The Elastic Response of Individual Double-Stranded and Single-Stranded DNA Molecules. *Science* **1996**, *271*, 795-799
[112] Morfil, J.; Kühner, F.; Blank, K.; Lugmaier, R.; Sedlmair, J.; Gaub, H. B-S Transition in Short Oligonucleotides. *Biophysical Journal* **2007**, *93*, 2400-2409
[113] Ho, D.; Zimmermann, J.; Dehmelt, F.; Steinabach, U.; Erdmann, M.; Severin, P.; Falter, K.; Gaub, H. Force-Driven Separation of Short Double-Stranded DNA. *Biophysical Journal* **2009**, *97*, 3158-3167

[114] Lee, G.; Chrisey, L.; Colton, R. Direct Measurement of the Forces Between Complementary Strands of DNA. *Science* **1994**, *266*, 771-773
[115] Danilowicz, C.; Hatch, K.; Conover, A.; Ducas, T.; Gunaratne, R.; Coljee, V.; Prentiss, M. Study of force induced melting of dsDNA as a function of length and conformation. *J. Phys.: Condens. Matter* **2010**, *22*, 414106
[116] Sheridan, S.; Gräter, F.; Daday, C. How Fast Is Too Fast in Force-Probe Molecular Dynamics Simulations? *J. Phys. Chem. B.* **2019**, *123*, 3658-3664
[117] Naserian-Nik, A.; Tahani, M.; Karttunen, M. Pulling of double-stranded DNA by atomic force microscopy: a simulation in atomistic details. *RSC Adv.* **2013**, *3*, 10516-10528
[118] MacKerell Jr. A.; Lee, G. Structure, force, and energy of a double-stranded DNA oligonucleotide under tensile loads. *Eur. Biophys. J.* **1999**, *28*, 415-426
[119] Mishra, R.; Mishra, G.; Li, M.; Kumar, S. Effect of shear force on the separation of double-stranded DNA. *Phys. Rev. E.* **2011**, *84*, 032903
[120] Essevaz-Roulet, B.; Bockelmann, U.; Heslot, F. Mechanical separation of the complementary strands of DNA. *Proc. Natl. Acad. Sci.* **1997**, *94*, 11935-11940
[121] Bockelmann, U.; Essevaz-Roulet, B.; Heslot, F. Molecular Stick-Slip Motion Revealed by Opening DNA with Piconewton Forces. *Phys. Rev. Lett.* **1997**, *79*, 4489-4492
[122] Cocco, S.; Monasson, R.; Marko, J. Force and kinetic barriers to unzipping of the DNA double helix. *Proc. Natl. Acad. Sci.* **2001**, *98*, 8608-8613
[123] Dudko, O.; Hummer, G.; Szabo, A. Theory, analysis, and interpretation of single-molecule force spectroscopy experiments. **2008**, *105*, 15755-15760
[124] Cao, M.; Deng, L.; Xu, H. Study of PNA-DNA hybridization by AFM-based single-molecule force spectroscopy. *Colloids and Surfaces A: Physicochem. Eng. Aspects* **2015**, *470*, 46-51
[125] Dutta, S.; Armitage, B.; Lyubchenko, Y. Probing of miniPEGγ-PNA-DNA Hybrid Duplex Stability with AFM Force Spectroscopy. *Biochemisty* **2016**, *55*, 1523-1528
[126] Sen, A.; Nielsen, P. Hydrogen bonding versus stacking stabilization by modified nucleobases incorporate in PNA·DNA duplexes. *Biophysical Chemistry* **2009**, *141*, 29-33
[127] Liu, J.; Tian, S.; Nielsen, P.; Knoll, W. *In situ* hybridization of PNA/DNA studied label-free by electrochemical impedance spectroscopy. *Chem. Commun.* **2005**, 2969-2971
[128] Uusitalo, J.; Ingólfsson, H.; Akhshi, P.; Tieleman, D.; Marrink, S. Martini Coarse-Grained Force Field: Extension to DNA. *J. Chem. Theory Comput.* **2015**, *11*, 3932-3945
[129] Uusitalo, J.; Ingólfsson, H.; Marrink, S.; Faustino, I. Martini Coarse-Grained Force Field: Extension to RNA. *Biophysical Journal* **2017**, *113*, 246-256

Chapter 2: Theory and methods

2.1 Molecular mechanics forcefields for nucleic acids

2.1.1 All-atom peptide nucleic acid forcefields

The results of molecular dynamics simulations are directly dependent on the forcefield parameters of the studied molecule. For PNA, a study directly evaluated the effect of changing various terms, such as torsional force constants or atomic charges, on results.[1] Minor changes of $\pm 10\%$ to atomic charges or torsional force constants could elicit changes in the free energies of binding of PNA dimers up to around 4 $\mathrm{kcal/mol}$, though in most cases changes were smaller (around 1 $\mathrm{kcal/mol}$). These results reflect the importance of accurate parameterisation of forcefields prior to interpretation.

All-atom molecular mechanics forcefields atomically render molecules and represent each atom as a point in space with a charge and mass subject to bonded and non-bonded potentials between itself and other atoms. For peptide nucleic acids the latest developments in all-atom forcefields were released in 2018 and provided up-to-date parameters for the CHARMM and Amber forcefields.[2] Both forcefields utilise the same general potential energy function (Equation 2-1).[3,4]

[Eq. 2–1]
$$V = \sum_{bonds} k(b - b_0)^2 + \sum_{angles} k(\theta - \theta_0)^2 + \sum_{torsions} k[1 + \cos(n\phi - \delta)] + \varepsilon_{nonbonded}$$

Where k is a force constant, θ is an angle, b is a bond length, ϕ is a torsional angle, n is the dihedral multiplicity, δ is the phase shift and $\varepsilon_{nonbonded}$ is a sum of the van

der Waals and Coulomb interactions between that atom and all other atoms. $\varepsilon_{nonbonded}$ is discussed in Chapter 2.2.3 and 2.2.4.

CHARMM and Amber differ in that the CHARMM forcefield splits the torsional angle term into 'proper dihedrals' of the type seen in Equation 2-1 and 'improper dihedrals' which are harmonic functions similar to the bond and angle potentials,[3] though these improper functions can equally be achieved by proper dihedrals with a multiplicity of one. Both forcefields are capable of modelling a variety of biomolecules.[5,6,7,8] The selection of one over the other for the present work is discussed in Chapter 3.

2.1.2 The Martini nucleic acid forcefield

Coarse-grained molecular mechanics forcefields, such as the Martini forcefield,[9] represent molecules not as systems of atoms but as systems of atom clusters called beads. Martini forcefields enable faster computation times for equivalent systems to be achieved by reducing the degrees of freedom and by enabling timesteps typically 10-fold higher (20 fs) than those regularly used in all-atom simulations (2 fs). Martini forcefields are parametrised against all-atom data by matching the distributions of, for example, bonds or angles, against the distributions of those same bonds and angles in all-atom simulations.[10]

To enable the study of larger nucleic acids over longer durations, models based on the Martini framework have been developed. In these models, purines are represented by four beads and pyrimidines by three, and the phosphoribose backbone is represented by two uncharged beads for (deoxy)ribose and a charged phosphate bead (Figure 2-1).[10,11] Free energy landscapes obtained along the base pairing and stacking coordinates demonstrate that these models overestimate the energy between stacked bases and underestimate the pairing free energy.[10] This is

possibly partly a consequence of the Martini framework since it does not model directional hydrogen bonding, as in base pairing, at the expense of faster computation rate.[10] In addition, double-helical DNA and RNA is poorly approximated by the CG model, forcing the use of a harmonic elastic network to maintain a conformation similar to the all-atom structure.[10,11] In using an elastic network, however, it is not possible to study the rupture forces of double-stranded DNA or RNA using the current Martini models since the elastic network introduces artificial energy barriers to dissociation.

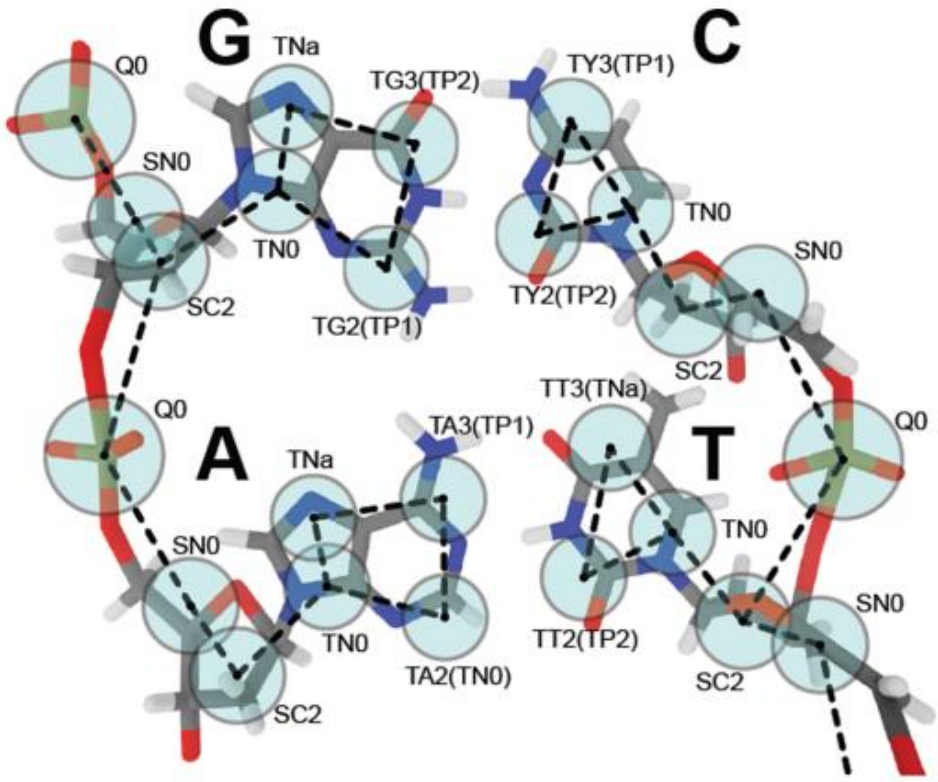


Figure 2-1: Beads, in translucent blue, of Martini DNA overlayed on top of all-atom structures. Bead types indicated by text. Dotted lines indicate bonds. Reproduced from Uusitalo *et al.*, Copyright (2015), with permission from the American Chemical Society under an ACS AuthorChoice usage agreement.[10]

2.2 Molecular mechanics parameter selection

2.2.1 Integrators

Integrators compute how positions and velocities evolve over discrete time. All molecular dynamics systems in the present work are conducted using Gromacs,[15] which implements a default 'leap-frog' integrator (Equation 2-2).

[Eq. 2–2a] $$\boldsymbol{v}\left(t+\frac{1}{2}\Delta t\right)=\boldsymbol{v}\left(t-\frac{1}{2}\Delta t\right)+\frac{\Delta t}{m}\boldsymbol{F}(t)$$

[Eq. 2–2b] $$\boldsymbol{r}(t+\Delta t)=\boldsymbol{r}(t)+\Delta t\boldsymbol{v}\left(t+\frac{1}{2}\Delta \mathrm{t}\right)$$

Where $\boldsymbol{v},\boldsymbol{r}$ and $\boldsymbol{F}$ are the velocity, position and force vectors. The velocity vector at time $t+\frac{1}{2}\Delta t$ is calculated by updating the velocity at time $t-\frac{1}{2}\Delta t$ using the force at time t. The position at time $t+\Delta t$ is calculated by updating the position at time t using the velocity at time $t+\frac{1}{2}\Delta t$. Since the velocity at time $t+\frac{1}{2}\Delta t$ is computed from the force at time t, which in molecular dynamics is solely due to the potential energy, and hence position at time $t,\boldsymbol{r}(t)$, then the positions update the forces which update the velocities which in turn update the positions, and the positions and velocities 'leap-frog' over one another.[12,13]

In the present work, this standard leap-frog integrator is used during equilibration, and for all Martini simulations, but for production all-atom simulations a stochastic integrator is used. A stochastic integrator is as accurate as the above integrator but includes a friction term that enables the efficient and theoretically rigorous implementation of temperature coupling,[14] as discussed in the succeeding subchapter.

2.2.2 Temperature and pressure coupling

Thermodynamics is a study of systems at equilibrium, and so for the accurate computations of energies like the free energy, molecular dynamics systems must be at equilibrium. For the calculation of the Gibbs free energy this involves the coupling of the system to an external bath at constant temperature and pressure. The temperature is calculated from the kinetic energy of the system (Equation 2-3).[12]

[Eq. 2–3] $$\frac{1}{2}nk_BT = E_{\text{kinetic}} = \frac{1}{2}\sum_i m_i v_i^2$$

Where n is the degrees of freedom of the system (for example, for a point in three-dimensional space there are 3 degrees of freedom from its translational x, y and z motion). m and v are the masses and velocities of the i^{th} atom and T is the temperature in Kelvin.

The coupling of this temperature to an external bath can be achieved via a Berendsen thermostat (Equation 2-4).[15] The Berendsen thermostat is suitable for equilibrating systems towards the target temperature because of its efficient implementation.

[Eq. 2–4] $$\frac{dT}{dt} = \frac{T_0 - T}{t}$$

For production runs, however, the Berendsen thermostat is unsuitable since it improperly suppresses fluctuations in the kinetic energy.[12] For accurate sampling of the system's kinetic energy, needed for theoretically rigorous production runs, more computationally expensive alternatives are implemented. In the present work, this is achieved using stochastic dynamics,[14] since the random motion of particles in the

system, which through Equation 2-3 can be used to express the temperature, results in friction. Though stochastic dynamics innately incorporates this, it can also be explicitly implemented in a temperature coupling algorithm should a non-stochastic integrator be used.[16,17] Alternatively, velocity can be rescaled at each step to account for this effect using a velocity rescaling thermostat.[18]

Similar to temperature coupling, pressure coupling is achieved by coupling the system to an external bath of constant pressure.[15] A Berendsen barostat (Equation 2-5) can be implemented though suffers from the same suppression of fluctuations as the Berendsen thermostat.[12]

[Eq. 2–5] $$\frac{dP}{dt} = \frac{P_0 - P}{t}$$

Parrinello-Rahman pressure coupling is used for production runs in the present work since fluctuations are accurately characterised.[19] Using a combination of a theoretically rigorous thermostat with the Parrinello-Rahman barostat allows quantities such as the free energy to be derived from production simulations since fluctuations in the system's energy, which contribute to the probability distribution of the system's energy from which the free energy is determined, are accounted for.

2.2.3 Van der Waals potentials

In both the CHARMM and Amber forcefields, there is a nonbonded component to the potential energy consisting of an uncharged van der Waals term and a charged Coulomb term. The van der Waals term is most often computed as a Lennard-Jones potential (Equation 2-6).[20]

[Eq. 2–6] $$V_{LJ} = 4\epsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right]$$

Where ϵ is the depth of an energy minimum, r is the distance and σ is the closest distance of approach. The Lennard-Jones potential consists of attractive $-(\sigma/r)^6$ and repulsive $(\sigma/r)^{12}$ contributions to the total potential energy, with σ being the distance closer than the equilibrium distance (at which the potential energy is minimised) wherein the attractive and repulsive terms exactly cancel out (Figure 2-2).[12]

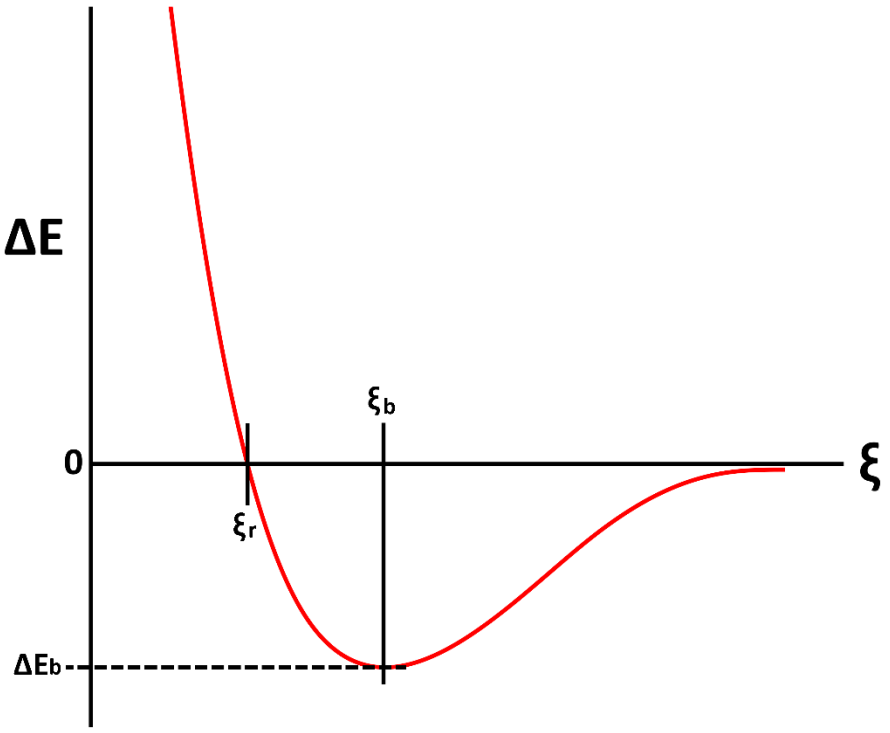


Figure 2-2: Schematic of a Lennard-Jones potential. ΔE is the energy change in arbitrary units and ξ is the reaction coordinate distance. ΔE approaches 0 at large ξ, representing the unbound state. ΔE_b is the energy change upon binding and is the minimum in the curve. ξ_b is the distance along ξ of the bound state and ξ_r is the distance along ξ at which the potential becomes repulsive (ΔE > 0).

To improve computational efficiency, molecular dynamics simulations often employ a cut-off which sets the energy to zero at a user-defined distance. To eliminate inconsistencies such as jumps in the potential energy to this zero point, potential switching functions are often employed that modify Equation 2-6 such that the potential energy smoothly transitions to zero by the cut-off point.[12,21]

2.2.4 Coulomb potentials

The second half of the nonbonded term is the Coulomb potential (Equation 2-7), which accounts for the energy between charged particles.[12]

[Eq. 2–7] $$V_{el} = \frac{q_i q_j}{4\pi\varepsilon_0 r}$$

Where ε_0 is the dielectric constant of a vacuum, known as the vacuum permittivity. q_i and q_j are the charges of the i^{th} and j^{th} interacting atoms. If an implicit solvent is used, meaning solvent atoms are not simulated and instead the solvent is treated as a homogeneous medium with a single dielectric constant, then ε_r is included in the denominator of Equation 2-7 to model the screening of the Coulomb potential by this medium. ε_r is also used in Martini simulations to model this screening and for these has an explicit value of 15.[22]

Like the van der Waals potential, the Coulomb potential can be smoothly tuned towards zero at a cut-off distance by modifying Equation 2-7 with a potential switch function. Instead of using a cut-off, the long-ranged Coulomb potential can also be modelled using a ‘particle mesh Ewald’ (PME)[23,24] which interpolates atomic charges over a grid. Long-ranged interactions are expensive to compute by direct summation, and the PME provides a comparatively inexpensive alternative since the

potential energy converges more rapidly.[12] The PME therefore prevents artefacts from cutting off the potential or from artificially smoothing towards zero at the cut-off point.

2.3 Free energy along physical coordinates

2.3.1 Steered molecular dynamics

Steered molecular dynamics (SMD) [25] is a method by which molecules can be forced along physical coordinates in molecular dynamics simulations. In an SMD simulation, a moving, typically harmonic, potential is applied to a bound complex such that, over time, it pulls one binding partner away from another along a specific reaction coordinate (Figure 2-3).[15,26,27] SMD traces demonstrate the existence of loading-rate dependent rupture forces consistent with force spectroscopy models.[28,29] In contrast to force spectroscopy, however, the loading rates used are typically orders of magnitude higher since lower loading rates require longer, resource intensive simulations.[30] This is problematic for the fitting of force spectroscopy models like the Bell-Evans model which fail in the limit of high loading rate.[31] In general, the fitting of any force spectroscopy model requires the characterisation of a ‘near-equilibrium’ regime at loading rates lower than those accessible using molecular dynamics, and if such a regime is not fitted the equilibrium free energy cannot be accurately derived.[32]

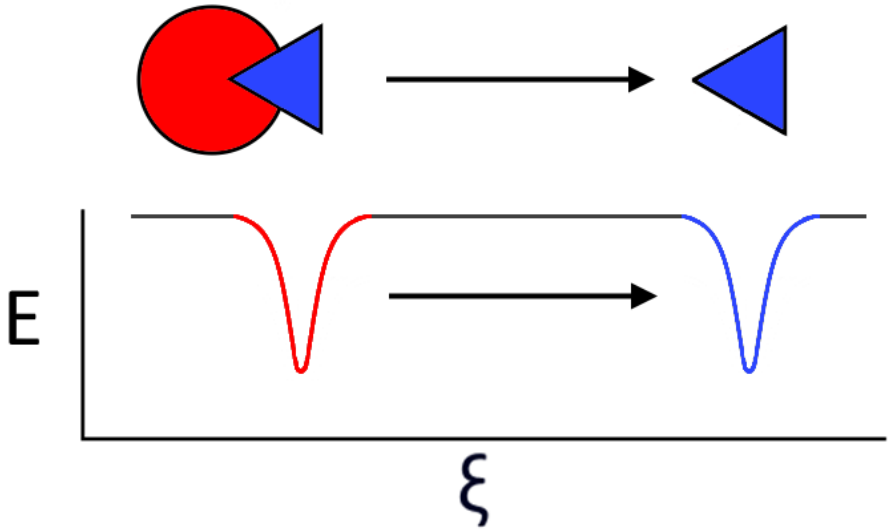


Figure 2-3: Schematic of steered molecular dynamics where a binding partner in blue is pulled from one in red using an external potential that moves from the position in red to the position in blue along the reaction coordinate ξ. The moving potential is represented by a minimum in the energy E.

2.3.2 Umbrella sampling

The free energy landscape underlying SMD simulations can be numerically solved using umbrella sampling (US).[33] This can overcome issues such as the inability to fit force spectroscopy models to high loading rate regimes since parameters such as barrier heights can be directly obtained from the free energy landscape.[34] During US, snapshots of the system during the unbinding process, for example at specific distances along the reaction coordinate, are obtained. These snapshots are used to sample the equilibrium free energy centred at different distances along the reaction coordinate, and the overlapping of these free energy estimates from these 'windows' allows reconstruction of the complete free energy landscape along the coordinate (Figure 2-4).

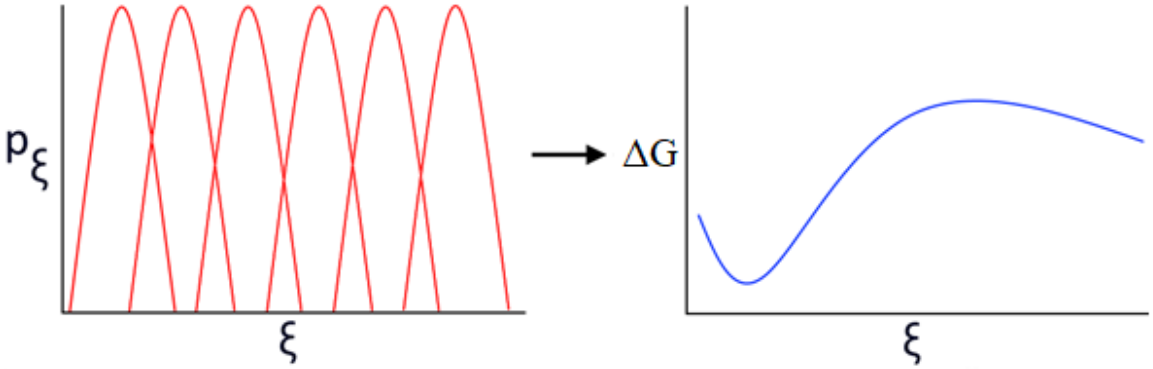


Figure 2-4: Umbrella sampling allows the free energy change ΔG along a reaction coordinate ξ to be reconstructed from the biased distributions in p_ξ. Each distribution is from a simulation centred at some distance in ξ and contains all distances sampled in that window. Since the free energy is related to these distributions as discussed in Chapter 1.2.3, it is possible to construct the free energy change along ξ, ΔG.

The free energy difference between two states, from Equation 1-8, is given by the ratio of partition functions, which are in turn derived from the probability distribution of energies at those states. In Figure 2-4, the probability distributions on the left are biased and need to be converted into an unbiased estimate for ΔG_ξ to be calculated, which can be done via the weighted histogram analysis method (WHAM).[35,36] The purpose of WHAM (Equation 2-8) is to convert biased distributions in p_ξ to unbiased distributions by determining the statistical uncertainty of unbiased distributions until the uncertainty converges on a minimum value. The unbiased distribution with the minimum uncertainty can then be used to calculate the free energy as described in Chapter 1.2.3.

[Eq. 2–8] $$p_{\xi\,\text{unbiased}} = \frac{\sum_i g_i^{-1} h_i}{\sum_j n_j g_j^{-1} \exp\left(\frac{1}{k_B T} w_j - f_j\right)}$$

Where g is the 'statistical inefficiency' of umbrella window i or j, h_i is the i^{th} histogram achieved by binning the continuous biased distributions in Figure 2-4 into discrete histograms, n_j is the number of samples of the j^{th} histogram, w_j is the biasing 'umbrella' potential used to maintain the system in the umbrella window and f_j is the unknown free energy of the j^{th} histogram. Sums are taken over all histograms.

Since both f_j and $p_{\xi_{\text{unbiased}}}$ are unknown these are iteratively solved until Equation 2-8 converges at values for them which minimise the statistical uncertainty.

2.3.3 Accelerated weight histograms

SMD simulations must assume that the chosen reaction coordinate is physically relevant for useful interpretations of its free energy landscape to be made. This method might be appropriate for unbinding systems with a physically relevant, narrow unbinding coordinate, but even in such cases SMD can force systems into improbable, physically irrelevant states.[37] An alternative method is the accelerated weight histogram (AWH) method.[38] In AWH, a target distribution centred at some distance along a coordinate ξ, or multiple coordinates, is estimated using an adaptive biasing potential g_ξ (Equation 2-9). The free energy estimate G_ξ with that target distribution p_ξ is, in turn, used to estimate the adaptive potential.

[Eq. 2–9] $$p_\xi = \frac{e^{-G_\xi + g_\xi}}{Z}$$

The free energy and the bias are both initially unknown and so must be iteratively solved. A probability histogram containing the biased probability distributions of the system's energy is maintained and this histogram is used to adaptively tune the

bias function.[37,38] A core advantage of AWH over US is that, by adaptively adjusting the bias function, improbable high-force states can be avoided, increasing simulation convergence. It is also practically easier to implement since a single simulation containing the entire reaction coordinate and history can be used, whereas US requires multiple sampling windows, and is therefore favoured over US in the present work.

2.4 Free energy along non-physical coordinates

2.4.1 Free energy perturbation

The binding free energy of two biomolecules doesn't have to be obtained by transformation along a physical coordinate. Computer simulations have the advantage in that atomic, non-bonded and bonded parameters can be tuned during simulation, effectively changing a molecule's identity or erasing it entirely. This kind of alchemical transformation can be used to: (i) determine the absolute binding free energy of two biomolecules by turning their interactions with one another off, or (ii) determine relative binding free energies, such as between two candidate drug molecules, by transforming one molecule into another.

According to the Zwanzig equation (Equation 2-10),[39] the free energy of state B can be obtained from a simulation taken at state A so long as the two states are similar enough that a simulation of one can sample the other.[40] When the two states are distinct, as might be the case for ds-DNA and ds-PNA, for example, a non-physical coordinate must be established which interpolates the bonded, non-bonded and atomic parameters of states A and B. During free energy perturbation (FEP) this non-physical coordinate connecting state A with state B is established using numerous intermediate states $0 < \lambda_N < 1$ with interpolated parameters where $\lambda_A = 0$ and $\lambda_B = 1$.[93] The intermediate states are not physically relevant but allow

the Zwanzig equation to be incrementally applied across them such that the difference in free energy between state A and B can be obtained.

[Eq. 2–10] $$\Delta G = -k_B T \ln \langle e^{-\frac{E_B - E_A}{k_B T}} \rangle_A$$

The angled brackets denote an average taken from a simulation at state A.

2.4.2 Thermodynamic integration

FEP allows the free energy difference between state A and B to be obtained using discrete intermediate states. An alternative method, thermodynamic integration (TI), allows the free energy difference to be computed by integration across a continuously changing λ (Equation 2-11).[41] Whilst FEP can be applied in equilibrium since each intermediate has a constant λ, TI technically cannot since λ is changing, meaning $d\lambda$ must be sufficiently small that a quasi-equilibrium is established.

[Eq. 2–11] $$\Delta G = \int_{\lambda=0}^{\lambda=1} \frac{\partial G}{\partial \lambda} d\lambda$$

2.4.3 Bennett acceptance ratio

A related method to FEP is the Bennett acceptance ratio (BAR).[42] Whereas FEP requires all high-probability states of state B to be sampled by state A (where state A and B can be intermediates), BAR samples over both state A and state B (Equation 2-22), requiring less strict overlap than is required for FEP.

$$\sum_{i=1}^{n_i} \frac{1}{1+e^{\left(\ln\left(\frac{n_i}{n_j}\right)+\frac{\Delta E_{ij}-\Delta G}{k_B T}\right)}} - \sum_{j=1}^{n_j} \frac{1}{1+e^{\left(\ln\left(\frac{n_j}{n_i}\right)+\frac{\Delta E_{ji}-\Delta G}{k_B T}\right)}} = 0 \quad \text{[Eq. 2–12]}$$

Where n_i and n_j are the number of samples i and j at state A and B respectively. Bennett arrived at this conclusion from Equation 1-8, which can equivalently be represented as Equation 2-13.[42]

$$\Delta G_{ij} = -k_B T \ln \frac{\langle M e^{\frac{-E_i}{k_B T}} \rangle_j}{\langle M e^{\frac{-E_j}{k_B T}} \rangle_i} \quad \text{[Eq. 2–13]}$$

Where angled brackets represent an average taken over all configurations. For example, the numerator in Equation 2-13 is an average of all $M \exp(-E_i/k_B T)$ taken at state j, meaning Equation 2-13 holds the same assumptions as the Zwanzig equation that state j can adequately sample the energies of state i and so must overlap. M is the metropolis function, $M = \min\{1, \exp(-x)\}$. Curly brackets denote a set, and min refers to the minimum value of this set which contains 1 and $\exp(-x)$. The purpose of M is to draw samples from the known Boltzmann distribution $\exp(-E_i/k_B T)$ at j. Bennett developed Equation 2-12, requiring less strict assumptions than the Zwanzig equation or Equation 2-13, by finding the value of M that minimised the variance in the free energy.[42]

2.5 Poisson-Boltzmann and Generalised Born approximations

Free energy differences obtained using US, AWH or FEP can be intensive since they require sampling of either the physical or non-physical coordinate connecting the two states. Molecular mechanics Poisson Boltzmann surface area (MM-PBSA) and molecular mechanics generalised Born surface area (MM-GBSA) allow binding

free energies to be approximated from end-points without simulating intermediate states, reducing resource consumption at the expense of accuracy. In both cases, the binding free energy is calculated in the same general way (Equation 2-14).[43,44,45]

[Eq. 2–14] $\Delta G = \Delta E_{MM} + \Delta G_{solv} - T\Delta S$

With ΔE_{MM} being calculated from the difference in bonded and non-bonded energies between the initial and final states (Equation 2-15) which could be, for example, ds-PNA versus two ss-PNAs.

[Eq. 2–15] $\Delta E_{MM} = \Delta E_{bonded} + \Delta E_{non-bonded}$

Both terms on the right-hand can be further subdivided and are ultimately all obtained from the molecular mechanics forcefield, hence the ‘MM’ in MM-PB[GB]SA. The PB and GB methods differ in the evaluation of ΔG_{solv}, which is divided into polar and non-polar terms (Equation 2-16).

[Eq. 2–16] $\Delta G_{solv} = \Delta G_{polar} + \Delta G_{non-polar}$

$\Delta G_{non-polar}$ is approximated from a linear function of the accessible surface areas of the initial and final states, hence the ‘SA’ in MM-PB[GB]SA (Equation 2-17).

[Eq. 2–17] $\Delta G_{non-polar} = mSA + b$

Where SA is the solvent-accessible surface area of the solute. In the unbinding of two nucleic acid strands, for example, it is likely to increase since the buried surface area between the two strands is exposed. The polar solvation energy ΔG_{polar} is calculated differently depending on whether the PB or GB method is used, with GB being itself an approximation of the PB equation.[43] The PB equation is computationally much more expensive to solve, and since prior research has shown comparative or even higher accuracy in resolving nucleic acid binding energies with the GB equation,[46] the GB approximation of ΔG_{polar} (Equation 2-17) is used in the present work.[47]

$$\Delta G_{polar} = \left(1 - \frac{1}{\epsilon}\right)\frac{1}{2}\sum_{ij}\frac{q_i q_j}{f_{GB}} \qquad \text{[Eq. 2–17]}$$

Where ϵ is the dielectric constant of the solvent, q is an atomic charge, and f_{GB} is a function of the Born radius, which is the spherically-averaged distance between an atom of the solute, such as a protein, and the solvent.[44] f_{GB} differs depending on the exact implementation of the GB equation used, with example implementations such as $igb = 5$[48] in the Gromacs molecular dynamics suite having well-documented applicability to proteins and nucleic acids. The purpose of f_{GB} is to reflect the screening of polar interactions between solute and solvent molecules by distance, charge, and other atoms in the solute.

2.6 References

[1] Bachmann, S.; Lin, Z.; Stafforst, T.; van Gunsteren, W.; Dolenc, J. On the Sensitivity of Peptide Nucleic Acid Duplex Formation and Crystal Dissolution to a Variation of Force-Field Parameters. *J. Chem. Theory Comput.* **2014**, *10*, 391-400

[2] Jasiński, M.; Feig, M.; Trylska, J. Improved Force Fields for Peptide Nucleic Acids with Optimized Backbone Torsion Parameters. *J. Chem. Theory Comput.* **2018**, *14*, 3603-3620

[3] MacKerell Jr., A.; Bashford, D.; Bellott, M.; Lunbrack, R.; Evanseck, J.; Field, M.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D.; Prodhom, B.; Reiher, W.; Roux, B.; Schlenkrich, M.; Smith, J.; Stote, R.; Straub, J.; Watanabe,

M.; Wiórkiewicz-Kuczera, J.; Yin, D.; Karplus, M. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B.* **1998**, *102*, 3586-3616
[4] Ponder, J.; Case, D. Force fields for protein simulations. *Adv. Prot. Chem.* **2003**, *66*, 27-85
[5] MacKerell Jr., A.; Banavali, N. All-atom empirical force field for nucleic acids: 1) parameter optimization based on small molecule and condensed phase macromolecular target data. *J. Comp. Chem.* **2000**, *21*, 86-104
[6] Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; de Groot, B.; Grubmüller, H.; MacKerell Jr, A. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nature Methods*, **2016**, *14*, 71-73
[7] Pérez, A.; Marchán, I.; Svozil, D.; Sponer, J.; Cheatham, T.; Laughton, C.; Orozco, M. Refinement of the AMBER Force Field for Nucleic Acids: Improving the Description of α/γ Confomers. *Biophys. J.* **2007**, *92*, 3817-3829
[8] Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. Comparison of multiple AMBER force fields and development of improved protein backbone parameters. *Proteins.* **2006**, *65*, 712-725
[9] Marrink, S.; Risselada, H.; Yefimov, S.; Tieleman, D.; de Vries, A. The MARTINI Force Field: Coarse Grained Model for Biomoleculer Simulations. *J. Phys. Chem. B.* **2007**, *111*, 7812-7824
[10] Uusitalo, J.; Ingólfsson, H.; Akhshi, P.; Tieleman, D.; Marrink, S. Martini Coarse-Grained Force Field: Extension to DNA. *J. Chem. Theory Comput.* **2015**, *11*, 3932-3945
[11] Uusitalo, J.; Ingólfsson, H.; Marrink, S.; Faustino, I. Martini Coarse-Grained Force Field: Extension to RNA. *Biophysical Journal* **2017**, *113*, 246-256
[12] Abraham, M.; Alekseenko, A.; Bergh, C.; Blau, C.; Briand, E.; Doijade, M.; Fleischmann, S.; Gapsys, V.; Garg, G.; Gorelov, S.; Gouaillardet, G.; Gray, A.; Irrgang, M.; Jalaypour, F.; Jordan, J.; Junghans, C.; Kanduri, P.; Keller, S.; Kutzner, C.; Lindahl, E. GROMACS 2023 Manual (Version 2023). *Zenodo*. **2023**. https://doi.org/10.5281/zenodo.7588711
[13] Berendsen, H.; van Gunsteren, W. Practical algorithms for dynamics simulations. *Molecular-Dynamics Simulation of Statistical-Mechanical Systems*. **1986**. 43-65.
[14] Goga, N.; Rzepiela, A.; de Vries, A.; Marrink, S.; Berendsen, H. Efficient Algorithms for Langevin and DPD Dynamics. *J. Chem. Theory Comput.* **2012**, *8*, 3637-3649
[15] Berendsen, H.; Postma, J.; van Gunsteren, A.; DiNola, J.; Haak, J. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **1984**, *81*, 3684-3690
[16] Hoover, W. Canonical dynamics: Equilibrium phase-space distributions. *Phys. Rev. A.* **1985**, *31*, 1695-1697
[17] Martyna, G.; Klein, M.; Tuckerman, M. Nosé-Hoover chains: The canonical ensemble via continuous dynamics. *J. Chem. Phys.* **1992**, *97*, 2635-2643
[18] Bussi, G.; Donadio, D.; Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **2007**, *126*, 014101
[19] Parrinello, M.; Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.* **1981**, *52*, 7182-7190
[20] Lennard-Jones, J. Cohesion. *Proceedings of the Physical Society*. **1931**, *43*, 461-482
[21] De Jong, D.; Baoukina, S.; Ingólfsson, H.; Marrink, S. Martini straight: Boosting performance using a shorter cutoff and GPUs. *Comp. Sci. Comm.* **2016**, *199*, 1-7
[22] Marrink, S.; Risselada, H.; Yefimov, S.; Tieleman, D.; de Vries, A. The MARTINI Force Field: Coarse Grained Model for Biomolecular Simulations. *J. Phys. Chem. B* **2007**, *111*, 7812-7824
[23] Hockney, R.; Goel, S.; Eastwood, J. Quiet High Resolution Computer Models of a Plasma. *J. Comp. Phys.* **1974**. *14.* 148-158.
[24] Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald. An Nlog(N) method for Ewald sums in large systems. *J. Chem. Phys.* **1993**. *98*. 10089-10092.
[25] Kamberaj, H. Molecular *Dynamics* Simulations in Statistical Physics: Theory and Applications. *Springer Nature* **2020**, ISBN 978-3-030-35701-6
[26] Park, S.; Schulten, K. Calculating potentials of mean force from steered molecular dynamics simulations. *J. Chem. Phys.* **2004**, *120*, 5946
[27] Isralewitz, B.; Gao, M.; Schulten, K. Steered molecular dynamics and mechanical functions of proteins. *Current Opinion in Structural Biology* **2001**, *11*, 224-230
[28] Evans, E.; Ritchie, K. Dynamic Strength of Molecular Adhesion Bonds. *Biophysical Journal* **1997**, *72*, 1541-1555
[29] Posch, S.; Aponte-Santamaría, C.; Schwarzl, R.; Karner, A.; Radtke, M.; Gräter, F.; Obser, T.; König, G.; Brehm, M.; Gruber, H.; Netz, R.; Baldauf, C.; Schneppenheim, R.; Tampé, R.; Hinterdorfer, P. Single molecule force spectroscopy data and BD- and MD simulations on the blood protein von Willebrand factor. *Data in Brief* **2016**, *8*, 1080-1087
[30] Sheridan, S.; Gräter, F.; Daday, C. How Fast Is Too Fast in Force-Probe Molecular Dynamics Simulations? *J. Phys. Chem. B.* **2019**, *123*, 3658-3664

[31] Dudko, O.; Hummer, G.; Szabo, A. Intrinsic Rates and Activation Free Energies from Single-Molecule Pulling Experiments. *Phys. Rev. Lett.* **2006**, *96*, 108101
[32] Noy, A.; Friddle, R. Practical single molecule force spectroscopy: How to determine fundamental thermodynamic parameters of intermolecular bonds with an atomic force microscope. *Methods* **2013**, *60*, 142-150
[33] Park, S.; Schulten, K. Calculating potentials of mean force from steered molecular dynamics simulations. *J. Chem. Phys.* **2004**, *120*, 5946
[34] You, W.; Tang, Z.; Chang, C. Potential Mean Force from Umbrella Sampling Simulations: What Can We Learn and What Is Missed? *J. Chem. Theory Comput.* **2019**, *15*, 2433-2443
[35] Hub, J.; de Groot, B.; van der Spoel, D. g_wham – A Free Weighted Histogram Analysis Implementation Including Robust Error and Autocorrelation Estimates. *J. Chem. Theory. Comput.* **2010**, *6*, 3713-3720
[36] Roux, B. The calculation of the potential of mean force using computer simulations. *Comp. Phys. Comm.* **1995**, *95*, 275-282
[37] Lindahl, V.; Lidmar, J.; Hess, B. Accelerated weight histogram method for exploring free energy landscapes. *J. Chem. Phys.* **2014**, *141*, 044110
[38] Lidmar, J. Improving the efficiency of extended ensemble simulations: The accelerated weight histogram method. *Phy. Rev. E.* **2012**, *85*, 056708
[39] Zwanzig, R. High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *J. Chem. Phys.* **1954**, *22*, 1420
[40] Beveridge, D.; DiCapua, F. Free Energy Via Molecular Simulation: Applications to Chemical and Biomolecular Systems. *Annu. Rev. Biophys. Biophys. Chem.* **1989**, *18*, 431-492
[41] Kästner, J.; Senn, H.; Thiel, S.; Otte, N.; Thiel, W. QM/MM Free-Energy Perturbation Compared to Thermodynamic Integration and Umbrella Sampling: Application to an Enzymatic Reaction. *J. Chem. Theory Comput.* **2006**, *2*, 452-461
[42] Bennett, C. Efficient Estimation of Free Energy differences from Monte Carlo Data. *J. Comput. Phys.* **1976**, *22*, 245-268
[43] Genheden, S.; Ryde, U. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opin. Drug Discov.* **2015**, *10*, 449-461
[44] Bashford, D.; Case, D. Generalized Born Models of Macromolecular Solvation Effects. *Annu. Rev. Phys. Chem.* **2000**, *51*, 129-152
[45] Onufriev, A.; Case, D. Generalized Born Implicit Solvent Models for Biomolecules. *Annu. Rev. Biophys.* **2019**, *48*, 275-296
[46] Golyshev, V.; Pyshnyi, D.; Lomzov, A. Calculation of Energy for RNA/RNA and DNA/RNA Duplex Formation by Molecular Dynamics Simulation. *Molecular Biology* **2021**, *55*, 927-940
[47] Goodman, J.; Attwood, D.; Kiely, J.; Coladas Mato, P.; Luxton, R. Modeling Peptide Nucleic Acid Binding Enthalpies Using MM-GBSA. *J. Phys. Chem. B.* **2022**, *126*, 9528-9538
[48] Onufriev, A.; Bashford, D.; Case, D. Exploring Protein Native States and Large-Scale Conformational Changes With a Modified Generalized Born Model. *Proteins* **2004**, *55*, 383-394

Chapter 3: Production and validation of CHARMM structures

3.1 Introduction

3.1.1 Forcefield selection

In the present work, numerous atomic structures of PNA were simulated. This requires the selection of an appropriate forcefield for these molecules, and a method by which the atomic structures of these molecules can be generated. This chapter thereby presents the forcefield selection using for all-atom simulations in the present study, the method of structure production, and finally routine validation procedures that are used throughout the rest of the study whenever molecular dynamics simulations are conducted.

Jasiński *et al*. presented the most recent forcefield parameters for peptide nucleic acids in 2018.[1] These parameters were developed for both the CHARMM and Amber all-atom molecular mechanics forcefields and, in both cases, had root mean square atomic deviations (RMSDs) for ds-PNA after 1 μs of approximately 0.12 nm. RMSD is a quantity reflecting the structural accuracy of a simulated biomolecule and measures the average distance of its atoms over time with their equivalent superimposed experimental crystal structure.[2] Typically, RMSDs of a couple Ångstrom (~ 0.2nm) are considered acceptable, though this depends on the use case and is somewhat arbitrary.[2] For Amber, the RMSD reported by Jasiński *et al*. was an order of magnitude better than for the previous Amber models, though for CHARMM the RMSD was reduced by only 0.01 nm for ds-PNA given the already low RMSDs of prior CHARMM models. Since these were the most recent available parameters and both had stable and low RMSD traces for ds-PNA, it was simply a matter of selecting which of the two, Amber or CHARMM, was most appropriate.

According to the authors, both Amber and CHARMM forcefields replicated the structure of PNA-containing helices well, though Amber reproduced the helical parameters of ds-PNA better than CHARMM. Contrastingly, CHARMM reproduced basic structural parameters like torsional angles more accurately than Amber as well as more accurately distinguishing helical parameters in PNA:DNA and PNA:RNA heteroduplexes. For the present study, CHARMM was selected as the molecular mechanics forcefield for all-atom studies for two reasons. Firstly, by more accurately representing heteroduplexes, the present work on ds-PNA could be extended to those systems. Secondly, since nucleic acids under external force are distorted from their equilibrium double-helices, accurately representing the P-form as in the Amber model would be less important than accurately reproducing basic structural parameters like torsional angles. This is since these basic structural parameters could influence the unbinding behaviour under external force.

3.1.2 Structure generation

Molecular dynamics simulations of ds-PNA require structure and topology files. The structure file and the forcefield are combined to produce a topology containing the force constants, equilibrium distances, and so forth for the potentials between atoms of the chosen structure. Structure files can be obtained from databases, such as the protein data bank (RCSB PDB).[3] An exact-match search for “peptide nucleic acid” in the RCSB PDB yielded 16 structures at the start of the current body of work, with a similar search for “PNA” yielding approximately 50. In the latter case, 39 of the submissions were either protein-nucleic acid complexes or false positives containing only proteins, and in both cases many of the PNA complexes were heteroduplexes or otherwise contained strand modifications such as lysine tags.

The proto-Nucleic Acid Builder (pNAB)[4] allows for the generation of ds-PNA structure files based on helical parameters inputted by the user, enabling a large number of structures with different sequences to be generated. This provided a more attractive alternative to the RCSB PDB since more total ds-PNA structures could be

acquired, additional modifications could be excluded, and sequences could be specified by the user. All simulated structures used in the current body of work, with the exception of one (PDB ID: 3MBS)[5] for validation purposes, were resultantly produced using initial helical parameters and a subsequent conformational search via the pNAB.

3.1.3 Equilibration runs

Before a production run, from which study outputs are collected, is conducted in molecular dynamics, equilibration runs are typically conducted. The purpose of equilibration is to ensure data is being collected under the desired conditions, for example constant temperatures and pressures representative of biological systems. In addition, thermodynamic quantities like the free energy are obtained from systems at equilibrium, when the free energy is related to the Boltzmann distribution.

All results from molecular dynamics simulations in the present work were obtained from equilibrated systems, with exact protocols being in the methods sections of the relevant chapters.

3.2 Methods

Molecular dynamics simulations were conducted on the ds-PNA 3MBS for the purpose of validating the simulation conditions used against previous data from the literature. This is because Jasiński *et al.*'s 2018 study[1] used the 3MBS[5] ds-PNA crystal structure to validate the structural accuracy of their forcefield parameters. A crystal structure of 3MBS was obtained from the RCSB PDB and imported into Gromacs using the CHARMM forcefield parameters developed by Jasiński *et al*. 3MBS was placed in a dodecahedral box and energy minimised using a steepest descent algorithm over 50,000 simulation steps before being solvated in spc216 water and minimised in the same manner as before. Solvated 3MBS was then

equilibrated at constant temperature using a velocity rescaling thermostat,[6] and then at constant pressure using a Berendsen barostat,[7] over 10 ns. 50 ns production runs were then obtained using a stochastic integrator[8] with pressure coupling handled by a Parrinello-Rahman barostat.[9]

During production runs the temperature and pressure were monitored. From production trajectories the RMSD between 3MBS and its crystal structure could also be obtained. The helical parameters of 3MBS were obtained and compared to the available literature.

3.3 Results

3.3.1 Temperature and pressure equilibration

Temperature (Figure 3-1) and pressure (Figure 3-2) traces of 50 ns 3MBS production runs demonstrate that temperature and pressure were well-coupled at 298 K and 1 bar respectively, indicating the system to be in temperature and pressure equilibrium. Consequently, the molecular dynamics parameters and the CHARMM forcefield discussed above can produce stable, well-equilibrated ds-PNA systems. Similar traces are obtained for all future molecular dynamics simulations for routine validation though are not shown.

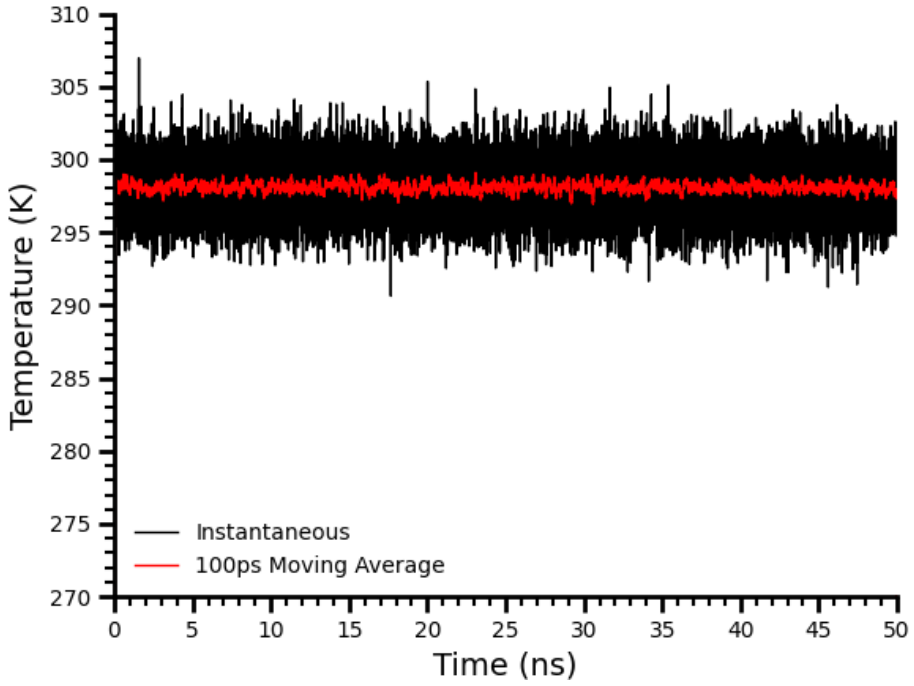


Figure 3-1: Temperature of a properly equilibrated simulation of the ds-PNA 3MBS taken over 50 ns.

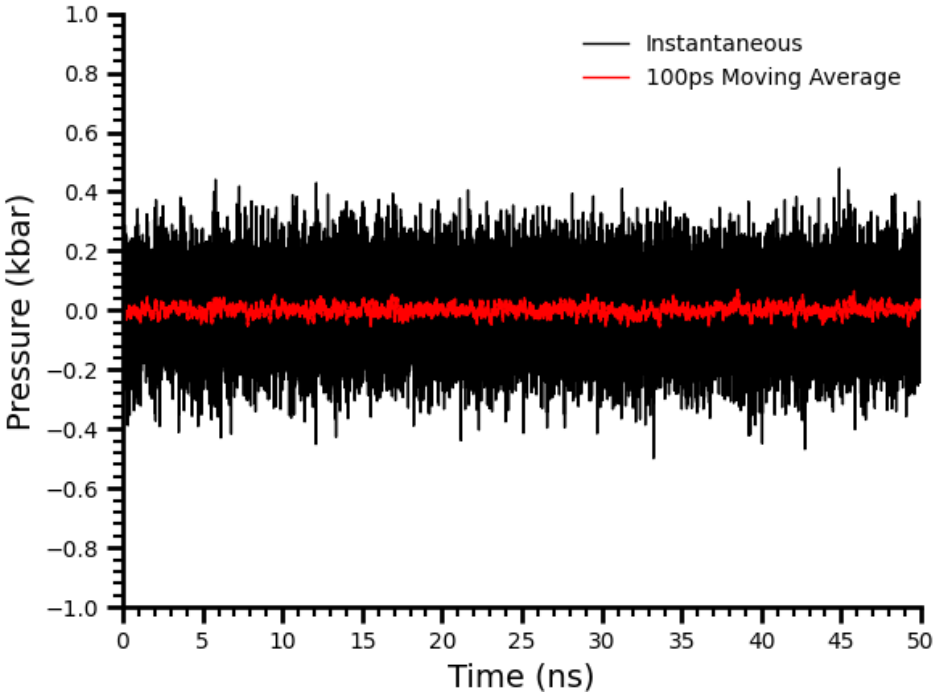


Figure 3-2: Pressure of a properly equilibrated simulation of the ds-PNA 3MBS taken over 50 ns.

3.3.2 RMSD and helical parameters

To confirm that simulation conditions used in the current work were representative of the literature, it was important that the structure of the ds-PNAs reproduced published structural parameters. Three 50 ns production replicates of the 3MBS right-handed helix produced an average RMSD of 0.124 nm (Figure 3-3). This is essentially identical to the value of 0.122 nm recorded by Jasiński *et al.*[1]

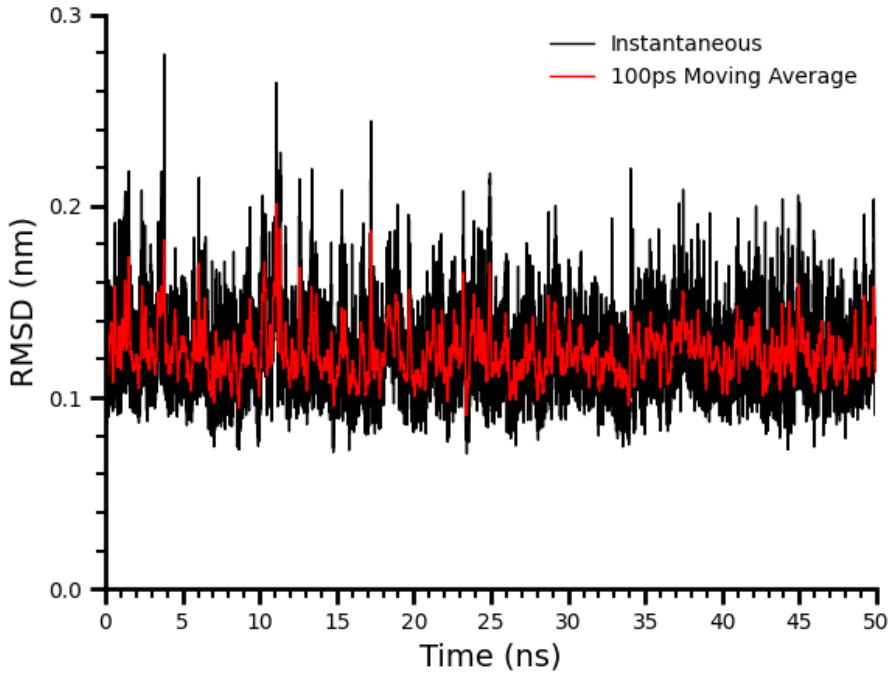


Figure 3-3: RMSD of a 3MBS replicate taken over one 50 ns simulation.

Similar agreement between the present study and the 2018 study by Jasiński *et al.* was seen in the helical parameters for right-handed 3MBS ds-PNA. The complete helical parameters (Figure 3-4) were measured over the three 50 ns production runs for all non-terminal base pairs, as described in the 2018 study, and averaged (Table 3-1).

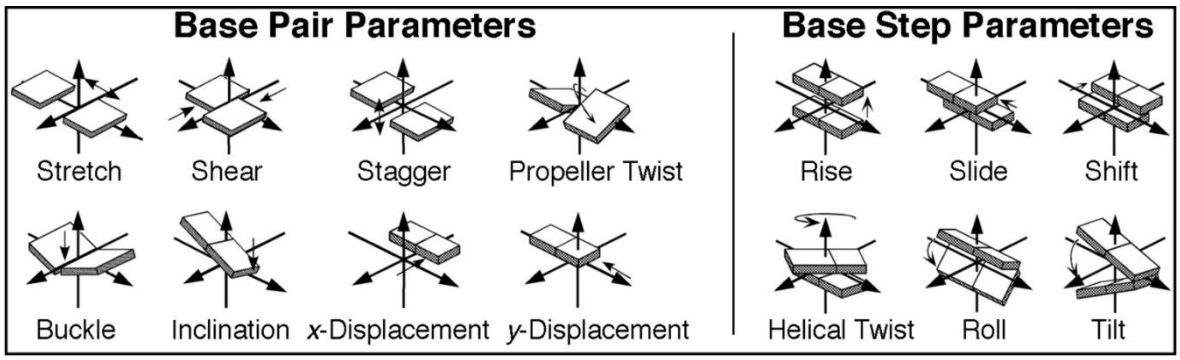


Figure 3-4: Schematic of different ds-nucleic acid parameters, reproduced from Ho and Carter.[10]

Table 3-1: Helical parameters for right-handed 3MBS from its published crystal structure[5] and using the Jasiński *et al.*[1] CHARMM 2018 forcefield in both their original publication and in the present work. ± indicate standard deviations.

Parameter	**Experimental**	**Jasiński *et al.***	**Present work**
Stretch (Å)	-0.10 ± 0.00	-0.11 ± 0.12	-0.06 ± 0.19
Shear (Å)	0.00 ± 0.10	0.00 ± 0.42	-0.03 ± 0.74
Stagger (Å)	0.10 ± 0.10	0.05 ± 0.34	0.08 ± 0.59
Propeller twist (◦)	-5.00 ± 3.60	-12.52 ± 6.48	-13.01 ± 7.22
Buckle (◦)	-0.20 ± 1.70	-0.01 ± 7.51	0.02 ± 8.49
Inclination (◦)	2.20 ± 11.30	15.59 ± 15.11	14.91 ± 13.08
x-displacement (Å)	-7.70 ± 0.80	-6.08 ± 1.92	-6.40 ± 2.13
y-displacement (Å)	-0.40 ± 1.40	0.00 ± 1.56	0.00 ± 1.89
Rise (Å)	3.20 ± 0.10	3.35 ± 0.27	3.34 ± 0.30
Slide (Å)	-2.50 ± 0.40	-1.77 ± 0.42	-1.81 ± 0.41
Shift (Å)	0.10 ± 0.30	0.00 ± 0.41	0.02 ± 0.37
Helical twist (◦)	19.50 ± 0.60	24.34 ± 3.38	24.58 ± 3.79
Roll (◦)	0.60 ± 3.70	6.29 ± 6.15	5.90 ± 7.51
Tilt (Å)	-0.40 ± 1.40	-0.00 ± 3.89	0.00 ± 3.44

Unsurprisingly, given that the same molecular mechanics forcefield was used, the helical parameters from Jasiński *et al.*'s 2018 publication were well-reproduced in

the present work. In general, a slightly larger standard deviation was measured over the base pairs in the three replicates though in all cases the standard deviations are overlapping. For both the present work and Jasiński *et al*.'s publication the inclination and roll angles are most different from the experimental work, being approximately an order of magnitude greater in both cases, though the low RMSDs with respect to the 3MBS crystal structure indicate that the difference in these angles is not crucial to replicating an accurate crystal structure. Deviation between the present work and the prior study is most likely since shorter 50 ns simulations were used in the present study due to the available computational resources, whereas 1 μs replicates were used in the 2018 study.

In general, these structural results indicate that the simulation protocols used in the present work accurately reproduce the available prior data and provide justification for the use of molecular dynamics simulations for the characterisation of ds-PNA complexes. Like temperature and pressure, RMSDs were regularly monitored throughout the present work and should be considered a routine part of each protocol.

3.4 References

[1] Jasiński, M.; Feig, M.; Trylska, J. Improved Force Fields for Peptide Nucleic Acids with Optimized Backbone Torsion Parameters. *J. Chem. Theory Comput.* **2018**, *14*, 3603-3620
[2] Maiorov, V.; Crippen, G. Size-independent comparison of protein three-dimensional structures. *Proteins* **1995**, *22*, 273-83
[3] Berman, H.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.; Weissig, H.; Shindyalov, I.; Bourne, P. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235-242
[4] Alenaizan, A.; Barnett, J.; Hud, N.; Sherrill, D.; Petrov, A. The proto-Nucleic Acid Builder: a software tool for constructing nucleic acid analogs. *Nucleic Acids Res.* **2020**, *49*, 79-89
[5] Yeh, J.; Pohl, E.; Truan, D.; He, W.; Sheldrick, G.; Du, S.; Ashim, C. The Crystal Structure of Non – Modified and Bipyridine – Modified PNA Duplexes. *Chem. – Eur. J.* **2010**, *16*, 11867-11875
[6] Bussi, G.; Donadio, D.; Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **2007**, *126*, 014101
[7] Berendsen, H.; Postma, J.; van Gunsteren, A.; DiNola, J.; Haak, J. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **1984**, *81*, 3684-3690
[8] Goga, N.; Rzepiela, A.; de Vries, A.; Marrink, S.; Berendsen, H. Efficient Algorithms for Langevin and DPD Dynamics. *J. Chem. Theory Comput.* **2012**, *8*, 3637-3649
[9] Parrinello, M.; Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.* **1981**, *52*, 7182-7190
[10] Shing, P.; Carter, M. DNA Structure: Alphabet Soup for the Cellular Soul. DNA Replication-Current Advances. *IntechOpen* **2011,** ISBN: 978-953-307-593-8

Chapter 4: Nearest-neighbour model for peptide nucleic acids

4.1 Introduction

Chapter 4 is in part a reproduction of the published article, “Modeling Peptide Nucleic Acid Binding Enthalpies Using MM-GBSA” produced as part of the present body of work.[1] Authorship rights for this article are jointly owned by the candidate J. Goodman, their supervisors D. Attwood, R. Luxton and J. Kiely, and P. Coladas Mato.

This chapter presents the development of the first nearest-neighbour model for ds-PNA. Using this model, the binding enthalpies of ds-PNAs can be predicted based on their primary sequences with a mean error of 8.74%. This model was developed by first defining a ‘benchmarking set’ of unique ds-PNA binding energies and entropies obtained from both the literature and from experimental data. The nearest-neighbour model was then developed in the following steps. Firstly, MM-GBSA was used to estimate the computational binding energies and entropies of the 10 sequences of the benchmarking set to determine whether MM-GBSA could accurately reproduce the experimental binding energies and entropies. Once it was determined to what extent MM-GBSA could reproduce these energies and entropies for the 10 sequences of the benchmarking set, an additional 39 sequences were simulated so that a total of 49 sequences were simulated. MM-GBSA was used to obtain binding energies and entropies for all 49 sequences, and these were then subject to a multiple linear regression model to decompose them, via the occurrences of each ds-stack, into incremental energies according to Equation 1-11. These incremental energies could then be reconstituted into binding energies of unknown sequences using only their primary sequence information.

By developing a nearest-neighbour model for ds-PNA, comparisons between the factors which contribute to the equilibrium stability, expressed by binding energy,

of ds-PNA and ds-DNA could be made. Quantification of these factors, expressed as the incremental enthalpies of ds-stacks, enables a more informed design of ds-PNAs with desired binding properties, and by comparison to ds-DNA an understanding of how the chemical structure of ds-PNA promotes higher stability in aqueous solutions can be developed.

4.2 Methods

4.2.1 Creating the benchmarking set

Seven ds-PNA sequences with binding energies and entropies from thermal melting experiments were compiled from the available literature.[2,3,4,5,6,7,8] Experimental conditions varied between the publications, namely regarding the concentration of sodium ions and the presence of a lysine tag. Hence, it was necessary either to determine whether this variance affected their published binding energies or entropies, or to exclude publications using conditions different to those used in the present work. Excluding, however, would have limited what was already a limited dataset of available ds-PNA binding energies and entropies. Since these would be necessary to benchmark the simulation data to determine the accuracy of the ds-PNA nearest-neighbour model, it was preferable to use all of the publications regardless of the experimental conditions if possible.

Since the concentration of sodium ions is consistently reported as having a negligible effect on binding energy in the literature, this factor was ignored.[2,9,10] To determine that lysine tagging did not affect binding energies, occurrences in the literature of the sequence GTAGATCACT, which appeared the most times and so could be used for comparing the publications, were categorised according to whether they used a lysine tag or not. The binding energies and entropies for unmodified sequences were all within one standard error of the mean binding energies and entropies for the tagged sequences (Table 4-1).

Table 4-1: Standard binding energies and entropies at 298 K for GTAGATCACT from the literature. $\pm$ indicates standard error.

Category	$-\Delta \boldsymbol{H}$ ($\mathbf{kcal\ mol^{-1}}$)	$-\Delta \boldsymbol{S}$ ($\mathbf{cal\ K^{-1}mol^{-1}}$)	$-\Delta \boldsymbol{G}$ ($\mathbf{kcal\ mol^{-1}}$)
All	86.92 ± 5.00	227.97 ± 14.48	18.99 ± 0.87
No lysine	89.36 ± 2.73	236.15 ± 8.28	18.85 ± 0.85
Lysine	85.45 ± 6.59	221.42 ± 18.83	19.70 ± 1.16

The seven unique ds-PNA binding energies and entropies were therefore grouped into a single data set regardless of the experimental conditions used, since lysine tagging had a negligible effect on these values.

To complete the benchmarking set, three additional ds-PNA binding energies and entropies were obtained through thermal melting experiments. These were purchased from Eurogentec at 95% purity (Cat. Number BA-PN010-005) as determined by high-performance reverse-phase and ion-exchange liquid chromatography. The three ds-PNA samples were diluted to a concentration of 1 μM in deionised water and were then melted via an electric heater in a water bath from which samples were taken at 2 K intervals between 293 and 343 K after 5-minute equilibrations at the target temperature. Using a Cary 60 UV-vis spectrophotometer, light with a wavelength of 260 nm was shone through a quartz cuvette containing the PNAs at the target temperature and the amount of absorbed light was measured. As temperature increased so did the absorbance, since the ss-PNAs produced during melting are hyperchromic (absorb more light) at 260 nm relative to ds-PNA.

For each of the three PNA sequences tested, three replicates were obtained. The absorbances for each PNA sequence were normalised between zero and one where $\alpha = 1$ corresponds to a sample containing only ds-PNA and $\alpha = 0$ corresponds to a sample containing only ss-PNA. The melting point was taken at $\alpha = 0.5$. The

equilibrium constant K was then derived according to the two-state assumption, meaning PNA was either ds-PNA or ss-PNA with no metastable states in-between. In general, this assumption allows the equilibrium constant to be derived according to Equation 1-9, though for self-complementary helices this reduces to Equation 4-1.[11]

[Eq. 4–1] $$K = \frac{\alpha}{2C(1-\alpha)^2}$$

Where C is the total strand concentration. The natural logarithm of the equilibrium constant was then plotted against the reciprocal temperature to derive standard enthalpies and entropies according to a linear form of the van't Hoff equation[12] as expressed in Equation 1-10. The binding free energy is then calculated according to the constant temperature and pressure equation $\Delta G = \Delta H - T\Delta S$.

The binding energies and entropies of the three ds-PNAs obtained from thermal melting experiments were combined with the seven unique ds-PNA binding energies and entropies from the literature to produce a benchmarking set consisting of ten unique ds-PNA binding energies and entropies.

4.2.2 Simulation parameters

Atomic ds-PNA structures were generated using the proto-Nucleic Acid Builder (pNAB).[13] A set of initial helical parameters[14,15,16] were obtained from published crystal or NMR structures. A weighted Monte Carlo conformational search was conducted to generate chemical structures based on these initial parameters and on thresholds such as bonding energies and interatomic distances set by the user. The structures generated by this were imported into the Gromacs program using the July 2021 edition of the CHARMM36m forcefield[17,18] and the N- and C- termini were

acetylated and amidated respectively. All molecular dynamics simulations in the present work were conducted using the open-source Gromacs software.

Simulated homoduplexes were energy minimised in vacuum over 50,000 simulation steps using a steepest-descent algorithm. Structures were then solvated in explicit waters in dodecahedral boxes using the Simple Point Charge (SPC) water model spc216.[19] The distance between any atom of the solute and the edge of the box was set to a minimum of 15 Å. Periodic boundary conditions[20] were enforced to imitate bulk solvent and solvated structures were energy minimised using steepest-descent minimisation over 50,000 simulation steps.

For each unique ds-PNA, three replicates were obtained from this point onwards. Each replicate was heated from 30 to 298K over 500 ps and Newton's equations of motion were integrated using a leap-frog integrator.[21,22] Temperature was weakly coupled to an external bath using a velocity rescaling[23] algorithm. Harmonic restraints of 24 kcal mol^{-1} $Å^{-1}$ were applied to non-hydrogen atoms whereas hydrogen atoms were constrained using a LINCS[24] algorithm. During this and for all future simulations the van der Waals interactions were handled using a switched cut-off scheme with switching to zero occurring from 10 to 12 Å. The electrostatic nonbonded interactions were treated using a particle mesh Ewald (PME)[25,26] algorithm with quartic interpolation and a grid spacing of 1 Å. The short-ranged component of the PME was computed to 10 Å and long-ranged components were handled using the fast Fourier transform library FFTW.[27]

After heating, position restraints were halved every 2 ns with the equations of motion and temperature coupling thereon handled by a stochastic integrator which incorporated friction and noise terms.[28] After 10 ns of simulation the non-hydrogen restraints were removed entirely and structures were equilibrated for 10 ns in the canonical ensemble. They were then equilibrated for a further 10 ns in the isobaric-isothermal ensemble with pressure coupling at 1 bar handled by a Berendsen

barostat.[29] 100 ns production runs were then carried out for each of the three replicas of each sequence using Parrinello-Rahman pressure coupling.[30] Simulation convergence was demonstrated by RMSD histograms and agreement between later MM-GBSA derived enthalpies, entropies and free energies as a function of simulation length (Appendix A.1).

Simulation conditions, such as the switching and cut-off distances, LINCS algorithm, and integrator, were chosen to closely reflect previous simulations of ds-PNA by Jasiński *et al.* whilst balancing the available computational resources.[3,31] The above molecular dynamics protocol can be visualised in the general process diagram (Figure 4-1). This general process is repeated in future chapters though exact parameters at each step may differ.

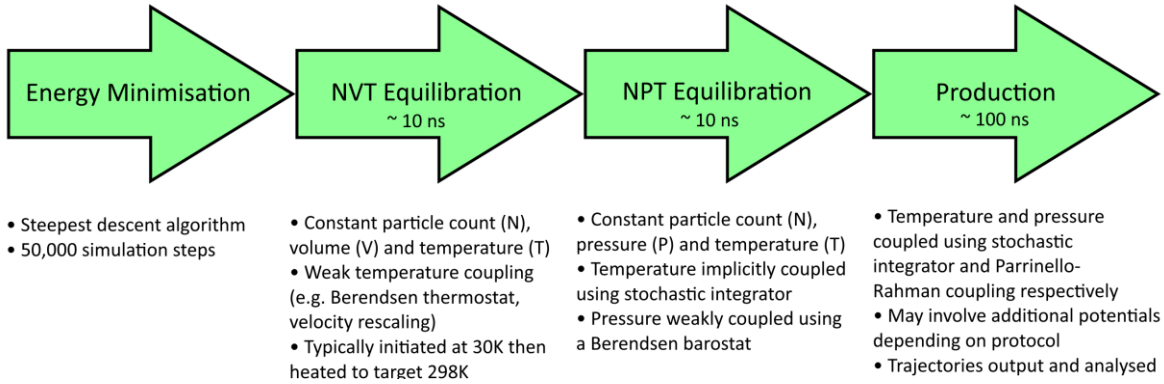


Figure 4-1: Process diagram for a general molecular dynamics simulation protocol as carried out in the present work.

4.2.3 Trajectory analysis using MM-GBSA

Trajectories were analysed by gmx_mmpbsa, which enables MM-GBSA estimations of enthalpy, entropy and free energy to be calculated.[32] The free energy of a molecule computed using MM-GBSA[33,34,35] is given by the average molecular mechanics potential energy E_{MM}, the solvation energy G_{solv} and the entropic term TS according to Equation 2-14. E_{MM}, which accounts for the bonded and non-

bonded energy terms according to Equation 2-15, was obtained for all PNAs from snapshots of the simulation trajectories in vacuum.

To evaluate ΔE_{MM} in the present work, a single-trajectory approach (STA) was used to save computation time. In this approach, only the ds-PNA is simulated, and the energies of the ss-PNAs are estimated based on this simulation. In the STA, the difference in bonded energy between ss-PNA and ds-PNA is zero. The STA therefore assumes that the conformation of a PNA strand in the duplex is identical to a free PNA strand in solution for the benefit of being computationally less expensive. To determine that this approach was valid, three multiple-trajectory approach (MTA) trials were conducted, meaning ss-PNAs were separately simulated to obtain their energies. For STA to be valid, the conformational enthalpy ΔH_{conf}, this being the difference between the enthalpy of ss-PNA determined from STA and MTA respectively, must be negligible. [36,37] For all three of the MTA trials, ΔH_{conf} was limited (~1% of the total energy) and hence considered negligible (Table 4-2).

Table 4-2: Enthalpies of ss-PNA as determined by STA and MTA. The conformational enthalpy ΔH_{conf} is the difference between the two.

Category	$-\boldsymbol{H_{ss-PNA,STA}}$ ($\mathbf{kcal\ mol^{-1}}$)	$-\boldsymbol{H_{ss-PNA,MTA}}$ ($\mathbf{cal\ K^{-1}mol^{-1}}$)	$-\boldsymbol{\Delta H_{conf}}$ ($\mathbf{kcal\ mol^{-1}}$)
CGATCG	657.43 ± 2.97	658.48 ± 1.62	1.05
GTAGATCACT	1027.47 ± 6.64	1038.52 ± 1.38	11.05
CATCTAGTGA	943.32 ± 6.86	951.88 ± 2.25	8.56

The solvation energy difference ΔG_{solv} from Equation 2-16 is the sum of non-polar and polar energies. In the present work, it was calculated using a continuum solvent model which approximates the explicit water solvent as a homogeneous medium with a dielectric constant of 78.5 at 298 K [38] The final contribution to the free

energy change according is then $T\Delta S$ which was evaluated in this study using a quasi-harmonic (QH) approximation[39], which is natively integrated into the functions of gmx_mmpbsa.[32] In the quasi-harmonic approximation, a covariance matrix $\boldsymbol{C}$ of atomic coordinates is obtained from the simulation trajectory. It is assumed that $\boldsymbol{C}$ is related to $\boldsymbol{H}$, this being a matrix of the second derivatives of the system's energy (Equation 4-2). $\boldsymbol{H}$ is then related to the system's energy surface through a Taylor series terminating at the second order, which is a harmonic approximation. In doing this, the quasi-harmonic approximation relates $\boldsymbol{C}$, which in turn relates to the fluctuations in atomic coordinates, with a harmonic approximation of the system's energy surface.

[Eq. 4–2] $$\boldsymbol{C} = \left(\frac{1}{k_B T}\boldsymbol{H}\right)^{-1}$$

A standard state correction was applied to the free energy differences to correct for the box volumes and thereby standardise the concentrations (Equation 4-3).

[Eq. 4–3] $$\Delta G^{\circ} = \Delta G - RT \ln\left(\frac{V_{\mathrm{com}}}{V^{\circ}}\right)$$

Where $V^{\circ} = 1661$ Å^3 and V_{com} is the volume of the simulation box.[40]

4.2.4 Multiple regression model

A nucleic acid duplex can be thought of as a series of overlapping ds-stacks as discussed in Chapter 1.4.2. This is known as the nearest-neighbour model.[41,42,43] The binding energy or entropy, generalised as χ, can be written in matrix form (Equation 4-4).

[Eq. 4–4]
$$\underbrace{\begin{pmatrix}\Delta\chi_1\\ \vdots\\ \Delta\chi_N\end{pmatrix}}_{\boldsymbol{\chi}} = \underbrace{\begin{pmatrix} j_1^1 & \cdots & j_v^1\\ \vdots & & \vdots\\ j_1^N & \cdots & j_v^N\end{pmatrix}}_{\boldsymbol{S}} \cdot \underbrace{\begin{pmatrix}\Delta\Delta\chi_1\\ \vdots\\ \Delta\Delta\chi_{v-2}\\ \Delta\chi_{init.GC}\\ \Delta\chi_{init.AT}\end{pmatrix}}_{\boldsymbol{\varepsilon}}$$

$\boldsymbol{\chi}$ is a $1 \times N$ column matrix containing either the mean enthalpies, entropies or free energies of hybridisation of PNA homoduplexes. $\boldsymbol{S}$ is an $N \times v$ matrix containing the occurrences j of all ds-stacks and the initiation terms. $\boldsymbol{\varepsilon}$ is a $1 \times v$ column matrix containing the unknown incremental energies or entropies. These unknowns are then evaluated by solving the linear least-squares problem which minimises the sum of the square of the residuals of $\boldsymbol{\chi}$ with respect to $\boldsymbol{S}$.

The multiple regression problem is generalised by weighting each observable $\Delta\chi$ by its reciprocal standard error (Equation 4-5).

[Eq. 4–5]
$$\boldsymbol{\sigma^{-1}} \cdot \boldsymbol{\chi} = \boldsymbol{\sigma^{-1}} \cdot \boldsymbol{S} \cdot \boldsymbol{\varepsilon}$$

Where $\boldsymbol{\sigma^{-1}}$ is an $N \times N$ diagonal matrix containing the reciprocal standard errors of each observable with non-diagonal elements of zero. The estimate $\boldsymbol{\varepsilon}$ obtained by solving the multiple linear regression problem thus contains the error-weighted least-squares estimates of the incremental energies or entropies.

Equation 4-5 was used in the present work to decompose the binding energies and entropies obtained by MM-GBSA into their incremental energies or entropies.

4.2.5 Hydrogen bonding analysis

The 100 ns production trajectories were analysed using gmx_hbond to compute the number of Watson-Crick hydrogen bonds between the base pairs over time. Hydrogen bonds were defined as bonds between the atom triplets involved in Watson-Crick bonding with distances between donor and acceptor of less than 0.325 nm and angles of hydrogen-donor-acceptor of less than 30 degrees. The purpose of hydrogen bonding analysis was to demonstrate that the incremental initiation enthalpies were corroborated by the structure and dynamics of ds-PNAs.

4.3 Results and Discussion

4.3.1 Benchmarking

The purpose of the benchmarking set (Table 4-3) was to determine how accurately MM-GBSA replicated binding energies and entropies obtained from thermal melting experiments. This benchmarking set consisted of 7 binding energies and entropies from the literature,[2-8] and 3 from optical melting experiments from the current work. The melting curves of the 3 ds-PNA sequences TAGCTA, AACGTT and CGATCG (Figure 4-2) obtained from optical melting experiments were determined to be 313.5 ± 0.4, 316 ± 0.4 and 319.7 ± 0.8K respectively. Full curves were processed according to Equation 4-1, and binding energies and entropies were obtained according to Equation 1-10. The standard errors of the binding enthalpies and entropies overlapped, whereas the binding free energy for CGATCG was more negative, suggesting it was stabilised relative to the other two as was corroborated by its higher melting temperature.

Table 4-3: Benchmarking set of experimental binding enthalpies, entropies and free energies (left to right). Standard errors obtained from the literature or averages of the literature. Standard errors from the present work are from triplicates.

ds-PNA Sequence	$-\Delta \boldsymbol{H}$ ($\mathbf{kcal\ mol^{-1}}$)	$-\Delta \boldsymbol{S}$ ($\mathbf{cal\ K^{-1}mol^{-1}}$)	$-\Delta \boldsymbol{G}$ ($\mathbf{kcal\ mol^{-1}}$)
	From the present work		
CGATCG	58.78 ± 4.09	151.42 ± 12.89	13.64 ± 0.25
AACGTT	52.05 ± 5.30	132.24 ± 16.45	12.64 ± 0.39
TAGCTA	55.08 ± 3.85	143.41 ± 11.95	12.34 ± 0.29
	From the literature		
GTAGATCACT[a,b,c,d,e,f]	86.92 ± 5.00	227.95 ± 14.48	18.99 ± 0.87
TGTTACGACT[g]	92.60 ± 5.70	240.00 ± 16.11	21.08 ± 1.00
AGGTAACCAG[c]	83.30 ± 2.80	216.60 ± 7.40	18.76 ± 0.60
AGTGAAGCAG[c]	82.15 ± 4.56	211.85 ± 12.86	19.02 ± 0.82
TGATCTAC[e]	60.90	224.70	13.10
GTAGATCACTGT[e]	97.10	253.80	21.40
GTAGATCACTGTCAC[e]	117.10	304.30	26.40

[a]Tomac, S.; Sarkar, M.; Ratilainen, T.; Wittung, P.; Nielsen, P.; Nordén, B.; Gräslund, A. Ionic Effects on the Stability and Conformation of Peptide Nucleic Acid Complexes. *J. Am. Chem. Soc.* **1996**, *118*, 5544-5552. [b]Sen, A.; Nielsen, P. On the stability of peptide nucleic acid duplexes in the presence of organic solvents. *Nuc. Acid. Res.* **2007**, *35*, 3367-3374. [c]Sen, A.; Nielsen, P. Unique Properties of Purine/Pyrimidine Asymmetric PNA-DNA Duplexes: Differential Stabilization of PNA-DNA Duplexes by Purines in the PNA Strand. *Biophys. J.* **2006**, *90*, 1329-1337. [d]Ratilainen, T.; Holmén, A.; Tuite, E.; Haaima, G.; Christensen, L.; Nielsen, P.; Nordén, B. Hybridization of Peptide Nucleic Acid. *Biochemistry*. **1998**, *37*, 12331-12342. [e]Totsingan, F. Synthesis and Applications of PNA and Modified PNA in Nanobiotechnology. Ph.D. Thesis, University of Parma, Parma, Italy, **2007**. [f]Sforza, S.; Haaima, G.; Marchelli, R.; Nielsen, P. Chiral Peptide Nucleic Acids (PNAs): Helix Handedness and DNA Recognition. *Eur. J. Org. Chem.* **1999**, 197-204. [f]Jasiński, M.; Miszkiewicz, J.; Feig, M. Trylska, J. Thermal Stability of Peptide Nucleic Acid Complexes. *J. Phys. Chem. B.* **2019**, *123*, 8168-8177.

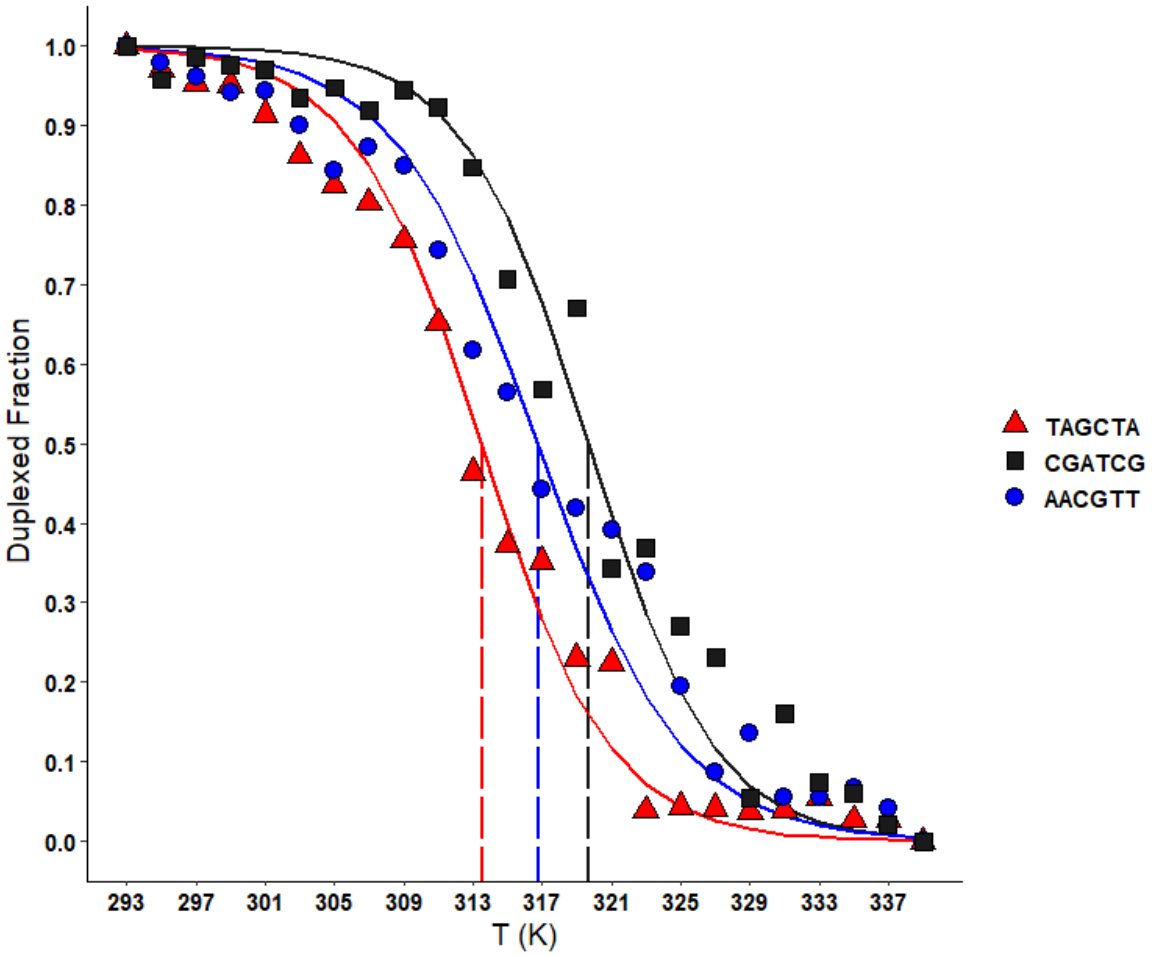


Figure 4-2: Normalised melting curves for the three ds-PNA hexamers TAGCTA, CGATCG and AACGTT as measured by optical melting. Melting points are shown by dashed vertical lines. Points represent the means of independent triplicates and were obtained at 2 K intervals.

Molecular dynamics simulations of all 10 sequences from the benchmarking set were conducted. Using MM-GBSA, their binding energies and entropies were determined by analysis of simulation trajectories. Means and standard errors were obtained from triplicates. Simple linear regression of the binding enthalpies of the 10 sequences obtained from MM-GBSA against those of the benchmarking set (i.e., from thermal melting experiments) produced an R^2 of 0.93 indicating that they were well-correlated (Figure 4-3). A high R^2 would be anticipated given that there were few sequences of the same length, meaning a positive correlation requires only that the simulation and reality both report, in general, a lower binding enthalpy for longer sequence lengths, which was what was observed. The mean absolute

difference, taken as the error of the prediction, between the binding enthalpies from MM-GBSA and from thermal melting experiments was 7.87 kcal/mol and the mean relative error was 10.01%.

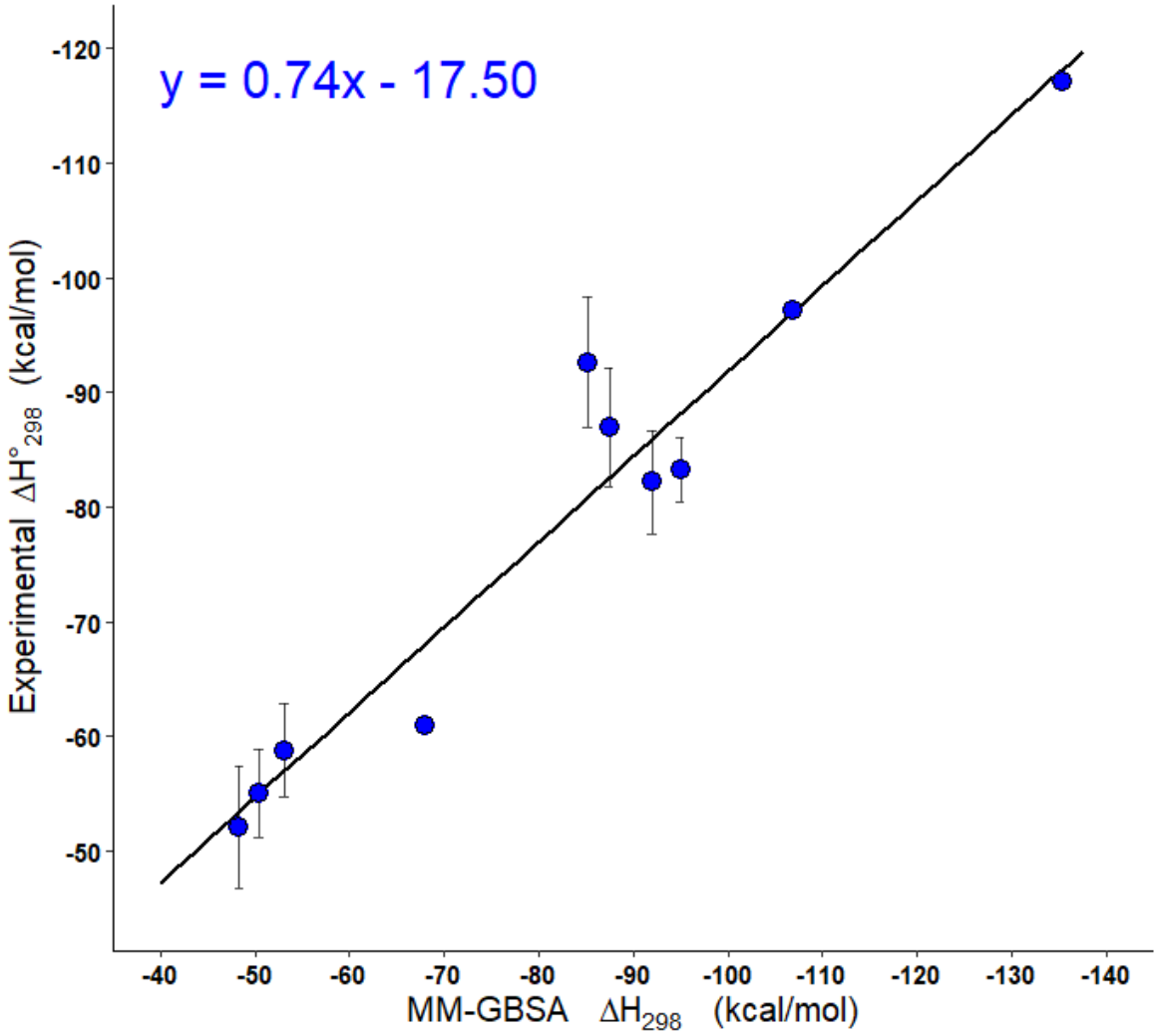


Figure 4-3: Correlation analysis for the binding enthalpy of ds-PNAs of the benchmarking set from thermal melting experiments and from MM-GBSA analysis.

The simple linear regression $y = ax + b$ between experimental and MM-GBSA binding enthalpies yielded a slope coefficient $a = 0.74$ and a shift coefficient $b = -17.50$. The large shift coefficient indicated a deviation between the thermal melting and MM-GBSA analysis. This was likely a combination both of assumptions made during simulation, such as in the MM-GBSA analysis'

approximation of solvation energy from a linear function of surface area, and of the small size of the benchmarking set since it was only 10 sequences long which could lead to the shift being determined by only a few points. Regardless, it could at least be said that, for the 10 sequences of the benchmarking set, MM-GBSA was a good predictor of their binding enthalpies due to its small relative error of 10% being in the range of experimental variance.

In contrast to the binding enthalpy, binding entropy and free energy were worse estimated though remained well correlated with R^2 of 0.84 and 0.86 respectively. The mean absolute and relative errors were $31.49\ \mathrm{kcal/mol}$ and 50.14% for $T\Delta S$ and $35.95\ \mathrm{kcal/mol}$ and 102.68% for ΔG (Figure 4-4). These errors are significantly higher and indicate that MM-GBSA could not accurately evaluate the binding entropy or free energy of ds-PNA with the simulation parameters and approach used in the present work. The obvious cause for this is the evaluation of the binding entropy since the binding enthalpy was well estimated and so the source of error for the binding free energy must be a result of the binding entropy term. Since this entropy term was determined by the QH approximation, this likely means that the assumption of harmonicity for the energy minimum of ds-PNA does not hold, and that atomic coordinates are better described by a complex energy surface at the energy minimum.

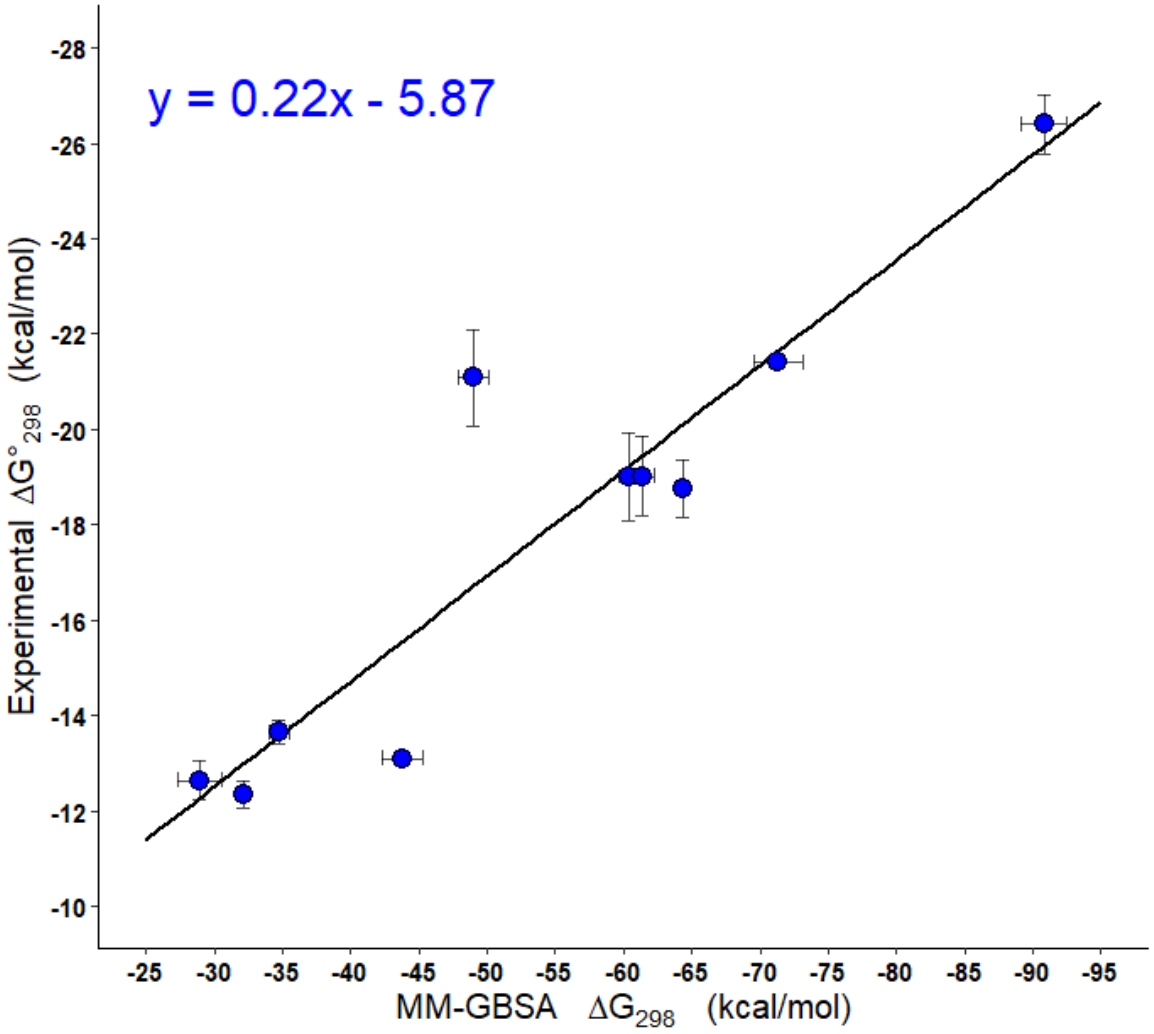


Figure 4-4: Correlation analysis for the binding free energy of ds-PNAs of the benchmarking set from thermal melting experiments and from MM-GBSA analysis.

According to the prior work of Golyshev *et al.*[44] a linear correction of the form $y = ax + b$ can reduce the absolute and relative errors between MM-GBSA and thermal melting estimates. Using the slope, $a = 0.74$, and shift, $b = -17.50$, coefficients for the binding enthalpy, reduced the relative and absolute error to 4.12 kcal/mol and 5.25% respectively. Similarly, using the slope $a = 0.22$ and shift $b = -5.87$ for the binding free energy reduced the absolute and relative errors to 1.16 kcal/mol and 6.53% respectively, representing an order of magnitude improvement. As discussed in Chapter 1.5, however, basing a nearest-neighbour model from linearly corrected MM-GBSA data has a non-trivial effect on the validity of the incremental binding energies and entropies by introducing the

inequality $\Delta G \neq \Delta H - T\Delta S$. Since the purpose of the present work is to quantify the factors that influence the binding stability of ds-PNAs, it was deemed improper to correct these factors using the slope and shift coefficients. This is since these coefficients represent computational error and are not physically relevant.

The correction of Golyshev *et al.*[44] is therefore not used in the present work and raw MM-GBSA estimates are used instead. Consequently, since the benchmarking set revealed raw binding free energy and entropy to be poorly approximated by the MM-GBSA method used, a nearest-neighbour model was developed only for the binding enthalpies of ds-PNA in the present work since the relative error of 10.01% was within the range of experimental variance.

4.3.2 The ds-PNA nearest-neighbour model

An additional 39 sequences were simulated using molecular dynamics and their binding energies and entropies estimated by MM-GBSA, producing a total of 49 simulations of unique ds-PNAs, each of which had three replicates. It is necessarily assumed that the errors of the raw MM-GBSA estimates for the benchmarking set are representative of the errors of the remaining 39 sequences. Trajectory analysis of all 147 simulations, containing 49 unique ds-PNA simulations, by MM-GBSA produced the binding enthalpies (Appendix A.2) which were used to solve the multiple linear regression problem of Equation 4-5. In solving this problem, incremental binding enthalpies for ds-PNAs were derived (Table 4-4).

Table 4-4: Incremental binding enthalpies for ds-PNAs as determined from MM-GBSA analysis. $\pm$ is the standard error of the mean.

Parameter	$-\Delta \boldsymbol{H}$ ($\mathbf{kcal\ mol^{-1}}$)	$-\langle \Delta \boldsymbol{H} \rangle$ ($\mathbf{kcal\ mol^{-1}}$)
AA:TT	8.65 ± 0.16	
AT:TA	8.28 ± 0.32	8.76 ± 0.31
TA:AT	9.34 ± 0.32	
CA:GT	10.31 ± 0.17	
GT:CA	8.90 ± 0.19	9.72 ± 0.46
CT:GA	10.64 ± 0.27	
GA:CT	8.91 ± 0.35	
CG:GC	11.71 ± 0.34	
GC:CG	8.69 ± 0.30	10.91 ± 1.12
GG:CC	12.32 ± 0.19	
Initiation (GC)	3.89 ± 0.35	
Initiation (AT)	2.27 ± 0.46	

All parameters from Table 4-4 were significant to $p < 10^{-5}$ in solving the multiple linear regression problem. Each ds-stack was categorised according to the number of Watson-Crick hydrogen bonds. These were six for GG, GC and CG; five for AG, GA, AC and CG; and four for AA, AT, and TA. On average, the incremental enthalpy became more negative as the number of hydrogen bonds increased, which was in line with expectations since these hydrogen bonds stabilise the duplex.

The nearest-neighbour parameters in Table 4-4 can be used to estimate the binding enthalpies of ds-PNA according to Equation 4-5, with the symmetry term excluded since it is zero for the binding enthalpy. These parameters were used to predict the binding enthalpies of the ds-PNAs of the benchmarking set (Table 4-5). These then had a mean relative error of 8.74%, which is lower than the 10.01% mean relative error from direct comparison between MM-GBSA estimates and thermal melting

data. Since the nearest-neighbour model was based on a total of 49 sequences, this indicated that the accuracy of the prediction was increased by accounting for unrelated sequences and that the model was therefore generalised without a loss of accuracy.

Table 4-5: Experimental binding enthalpies ΔH_{exp} for ds-PNA against the nearest-neighbour binding enthalpies ΔH_{NN} predicted from the incremental enthalpies estimated in the current work. Relative difference given as a percentage.

Category	$-\Delta \boldsymbol{H_{exp}}$ ($\mathbf{kcal\ mol^{-1}}$)	$-\Delta \boldsymbol{H_{NN}}$ ($\mathbf{cal\ K^{-1}mol^{-1}}$)	**% Difference**
CGATCG	58.78 ± 4.09	53.40	9.15
AACGTT	52.05 ± 5.30	49.07	5.73
TAGCTA	55.08 ± 3.85	50.91	7.57
GTAGATCACT	86.92 ± 5.00	87.09	0.20
TGTTACGACT	92.60 ± 5.70	85.24	7.95
AGGTAACCAG	83.30 ± 2.80	92.28	10.78
AGTGAAGCAG	82.15 ± 4.56	89.95	9.49
TGATCTAC	60.90	67.55	10.92
GTAGATCACTGT	97.10	106.30	9.47
GTAGATCACTGTCAC	117.10	136.03	16.17

4.3.3 Comparison with ds-DNA

The incremental enthalpy of a PNA ds-stack is typically lower than its counterpart for DNA (Table 4-6)[43] with the exception of the GC:CG stack. The lower incremental enthalpies are unsurprising, given that previous literature has demonstrated that ds-PNA's binding enthalpies are often lower than those of ds-DNA for the same sequence.[2,4,5]

Tabled 4-6: Incremental binding enthalpies for ds-PNAs and ds-DNAs, from the unified model.[43] $\pm$ is the standard error of the mean.

Parameter	$-\Delta H_{DNA}$ (**kcal mol^{-1}**)	$-\Delta H_{PNA}$ (**kcal mol^{-1}**)
AA:TT	7.9	8.65 ± 0.16
AT:TA	7.2	8.28 ± 0.32
TA:AT	7.2	9.34 ± 0.32
CA:GT	8.5	10.31 ± 0.17
GT:CA	8.4	8.90 ± 0.19
CT:GA	7.8	10.64 ± 0.27
GA:CT	8.2	8.91 ± 0.35
CG:GC	8.6	11.71 ± 0.34
GC:CG	9.8	8.69 ± 0.30
GG:CC	8.0	12.32 ± 0.19
Initiation (GC)	-0.1	3.89 ± 0.35
Initiation (AT)	-2.3	2.27 ± 0.46

For ds-stacks with the same number of Watson-Crick hydrogen bonds, the difference in their incremental enthalpies is due to stacking. Within each cluster of ds-stacks with the same number of hydrogen bonds, the order of stabilities differs between DNA and PNA. In addition, whereas DNA ds-stacks with the same number of hydrogen bonds had incremental enthalpies typically within 1 or 2 kcal/mol of one another, this variance was higher in PNA, being up to 4 kcal/mol. This suggests that stacking in PNA plays a larger role in the stabilisation of the double helix. This is supported by the average P-form helix having a smaller rise per base pair,[16,45,46,47] indicating increased stacking interaction strength. It is important to note, however, that the 42 non-literature sequences were manually selected, and bias may have been introduced if many ds-stacks were accounted for by a disproportionately small cluster of sequences.

Arguably the most distinct differences between the incremental enthalpies of ds-PNA and ds-DNA are the initiation enthalpies. For ds-DNA, regardless of the terminal base pair makeup, these initiation enthalpies destabilise the helix. In ds-PNA, however, these initiation enthalpies are negative and are stabilising in both cases. Since the helix initiation energy indicates a baseline energy of association between the two strands and is therefore related to the sequence independent interactions between the backbones,[43] this suggests that ds-PNA's neutral backbone results in a significantly lower enthalpy barrier for the association between two strands. This is well supported by the structure of ds-PNA since, compared to ds-DNA, there is no electrostatic barrier to association from net like-charge repulsion between the backbones.

4.3.4 Hydrogen bonds corroborate incremental enthalpies

Like with ds-DNA, helix initiation with only terminal G:C pairs is preferable to helix initiation with at least one terminal A:T pair present. This indicates that terminal A:T pairs are likely to be less stable than terminal G:C pairs, which could have structural implications.

To demonstrate the connection between the incremental enthalpies of Table 4-4 with structure, a hydrogen bonding analysis was carried out. The number of Watson-Crick hydrogen bonds between terminal and 3rd-position, classed as internal, base pairs were tracked. The number of hydrogen bonds was expressed as a percentage of the total number of possible bonds, this being three for G:C and two for A:T base pairs. These percentages were averaged over the simulation durations and for simulations involving sequences of the same length (Table 4-7).

Table 4-7: Mean normalised number of hydrogen bonds expressed as a percentage of terminal and internal A:T and G:C base pairs of varying sequence lengths. ± is the standard deviation.

Length (base pairs)	**% H-Bonded**	
	A:T	**G:C**
	Terminal	
6	70.4 ± 27.1	83.3 ± 4.3
8	74.4 ± 12.0	88.4 ± 2.9
10	84.7 ± 13.2	89.1 ± 2.1
12	79.3 ± 10.6	83.2 ± 7.1
14	80.3 ± 7.8	84.1 ± 6.5
16	68.4 ± 29.4	88.4 ± 1.7
18	67.6 ± 20.6	82.9 ± 3.9
	Internal	
6	90.2 ± 2.4	95.7 ± 0.3
8	90.9 ± 0.5	94.0 ± 1.5
10	91.7 ± 0.9	95.3 ± 0.5
12	89.9 ± 1.6	96.0 ± 0.9
14	91.6 ± 0.7	96.0 ± 0.7
16	91.1 ± 0.4	95.0 ± 0.9
18	89.6 ± 2.4	95.8 ± 0.4

The percentage of unbroken bonds, from Table 4-7, is independent of sequence length and always, on average, higher for internal than terminal base pairs. Terminal A:T hydrogen bonds were broken more often than terminal G:C hydrogen bonds, though both spent most of their time in the bound state. Additionally, the standard deviation in the percentage of unbroken bonds was greater, most often by around 5-fold but in one case over 15-fold, for terminal A:T than G:C base pairs. This indicated that the terminal A:T hydrogen bond behaviour was characterised by fluctuations between bound and melted states.

To determine the nature of these fluctuations, the duration of the melting events was plotted against the occurrence, per nanosecond, of a melting event of that duration being initiated (Figure 4-5). A melting event was defined as an uninterrupted period of time with at least one hydrogen bond broken.

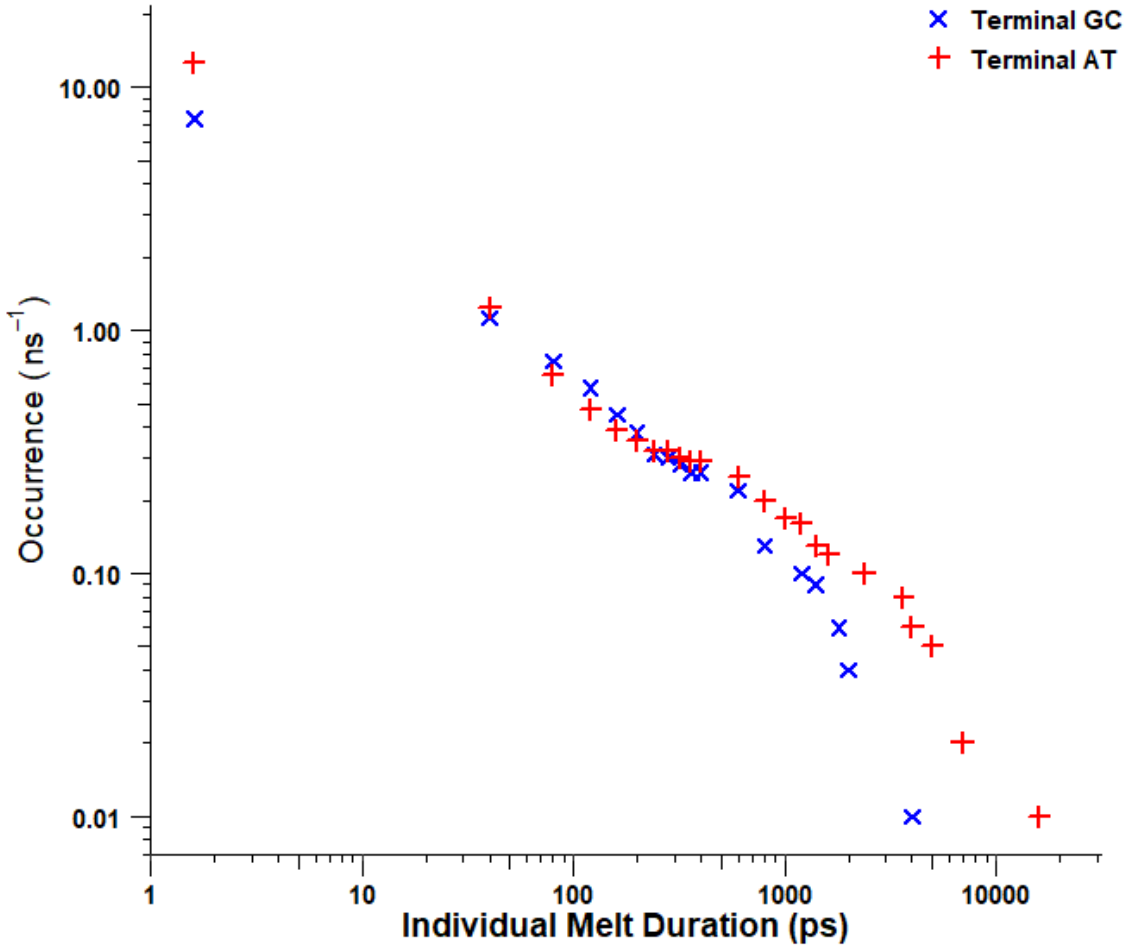


Figure 4-5: Occurrences per nanosecond of an individual melting event of at least a duration in picoseconds initiation for terminal G:C base pairs, in blue, and terminal A:T base pairs, in red.

From Figure 4-5 it was observed that, for both A:T and G:C terminal base pairs, melting events of a duration less than one nanosecond were similarly common. By contrast, melting events longer than this were more common for terminal A:T than G:C pairs. This indicated that the large deviations in the number of unbroken bonds

for the terminal A:T pairs were likely a consequence of long, persistent melting events as opposed to short-lived ones. Since long melting events of this manner may destabilise the duplex and increase its binding enthalpy, this may partly explain why the helix initiation enthalpy is greater if a terminal A:T pair is present.

4.4 Conclusions

These MM-GBSA studies of 49 ds-PNA sequences, which totalled to 147 100 ns production run simulations, effectively addressed the technical research question (1) "Can molecular dynamics be used to generate a model that allows binding properties to be predicted from structure alone, and what quantities contribute to the binding stability of the candidate bioadhesive?" with the following conclusions:

(1) The incremental binding enthalpies decrease with number of hydrogen bonds demonstrating that, as with other nucleic acids, G:C base pairs provide a greater contribution to helix stability.

(2) The increased variability of incremental binding enthalpy within a cluster of ds-stacks with the same number of pairs indicates that stacking interactions may play a more dominant role in base pair stability in ds-PNA than they do in ds-DNA.

(3) The negative helix initiation enthalpies show that, whereas association between strands must overcome an energetic barrier in ds-DNA due to net backbone repulsion, strand association does not present an enthalpic barrier for ds-PNA, which may drive association.

(4) The higher incremental enthalpy for helix initiation with A:T termini indicates that terminal A:T base pairs destabilise ds-PNA, which is explained via hydrogen bonding analysis as mostly arising from long, persistent melting events in excess of one nanosecond.

(5) The nearest-neighbour model developed for ds-PNA is capable of accurately predicting experimental binding enthalpies from primary sequence (structural) parameters alone with an error of 8.74%

4.5 References

[1] Goodman, J.; Attwood, D.; Kiely, J.; Coladas Mato, P.; Luxton, R. Modeling Peptide Nucleic Acid Binding Enthalpies Using MM-GBSA. *J. Phys. Chem. B.* **2022**, *126*, 9528-9538

[2] Tomac, S.; Sarkar, M.; Ratilainen, T.; Wittung, P.; Nielsen, P.; Nordén, B.; Gräslund, A. Ionic Effects on the Stability and Conformation of Peptide Nucleic Acid Complexes. J. Am. Chem. Soc. 1996, 118, 5544−5552

[3] Jasiński,́ M.; Miszkiewicz, J.; Feig, M.; Trylska, J. Thermal Stability of Peptide Nucleic Acid Complexes. J. Phys. Chem. B 2019, 123, 8168−8177

[4] Sforza, S.; Haaima, G.; Marchelli, R.; Nielsen, P. Chiral Peptide Nucleic Acids (PNAs): Helix Handedness and DNA Recognition. Eur. J. Org. Chem. 1999, 1999, 197−204

[5] Ratilainen, T.; Holmén, A.; Tuite, E.; Haaima, G.; Christensen, L.; Nielsen, P.; Nordén, B. Hybridization of Peptide Nucleic Acid. Biochemistry 1998, 37, 12331−12342

[6] Sen, A.; Nielsen, P. On the stability of peptide nucleic acid duplexes in the presence of organic solvents. Nucleic Acids Res. 2007, 35, 3367−3374

[7] Totsingan, F.Synthesis and Applications of PNA and Modified PNA in Nanobiotechnology. Ph.D. Thesis, University of Parma, Italy, 2007

[8] Sen, A.; Nielsen, P. Unique Properties of Purine/Pyrimidine Asymmetric PNA·DNA Duplexes: Differential Stabilization of PNA· DNA Duplexes by Purines in the PNA Strand. Biophys. J. 2006, 90, 1329−1337

[9] Eriksson, M.; Nielsen, P. E. PNA-nucleic acid complexes. Structure, stability and dynamics. Q. Rev. Biophys. 1996, 29, 369−394

[10] Nielsen, P. E.; Egholm, M.; Berg, R. H.; Buchardt, O. Sequence Selective Recognition of DNA by Strand Displacement with a Thymine-Substituted Polyamide. Science 1991, 254, 1497−1500

[11] Marky, L.; Breslaur, K. Calculating Thermodynamic Data for Transitions of any Molecularity from Equilibrium Melting Curves. *Biopolymers* **1987**, *26*, 1601-1620

[12] Schroeder, S.; Turner, D. Optical Melting Measurements of Nucleic Acid Thermodynamics. *Methods Enzymol.* **2009**, *468*, 371-387

[13] Alenaizan, A.; Barnett, J.; Hud, N.; Sherrill, D.; Petrov, A. The proto-Nucleic Acid Builder: a software tool for constructing nucleic acid analogs. *Nuc. Acid Res.* **2020**. *49*. 79-89.

[14] Yeh, J.; Pohl, E.; Truan, D.; He, W.; Sheldrick, G.; Du, S.; Achim, C. The Crystal Structure of Non-Modified and Bipyridine-Modified PNA Duplexes. *Chem. Eur. J.* **2010**. *16*. 11867-11875.

[15]Kiliszek, A.; Banaszak, K.; Zbigniew, D.; Wojciech, R. The first crystal structures of RNA-PNA duplexes and a PNA-PNA duplex containing mismatches - towards anti-sense therapy against TREDs. *Nuc. Acids Res.* **2015**. *44*. 1937-1943.

[16] He, W.; Hatcher E.; Balaeff, A.; Beratan, D.; Gil, R.; Madrid, M.; Achim, C. Solution Structure of a Peptide Nucleic Acid Duplex from NMR Data: Features and Limitations. *J. Am. Chem. Soc.* **2008**, *130*, 13264-13273.

[17] van der Spoel, D.; Lindahl, E.; Hess, B.; Groenhoe, G.; Mark, A.; Berendsen, H. GROMACS: Fast, Flexible and Free. *J. Comput. Chem.* **2005**, *26*, 701-1718.

[18] Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; de Groot, B.; Grubmüller, H.; MacKerell Jr., A. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat. Methods*. **2017**. *14*. 71-73.

[19] Berendsen, H.; Postma, J.; Gunsteren, W.; Hermans, J. Interaction Models for Water in Relation to Protein Hydration. *Intermolecular Forces*. **1981**. *14*. ISBN: 978-90-481-8368-5.

[20] Bekker, H.; Dijkstra, E.; Renardus, M.; Berendsen, H. An Efficient, Box Shape Independent Non-Bonded Force and Virial Algorithm for Molecular Dynamics. **1995**. *14*. 137-151.

[21] Hocknehy, R.; Goel, S.; Eastwood, J. Quiet High Resolution Computer Models of a Plasma. *J. Comp. Phys.* **1974**. *14.* 148-158.

[22] Berendsen, H.; van Gunsteren, W. Practical algorithms for dynamics simulations. *Molecular-Dynamics Simulation of Statistical-Mechanical Systems*. **1986**. 43-65.

[23] Bussi, G.; Donadio, D.; Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **2006**. *126*. 01401.
[24] Hess, B.; Bekker, H.; Berendsen, H.; Fraaije, J. LINCS: A Linear Constraint Solver for Molecular Simulations. *J. Comput. Chem.* **1997**. *18*. 1463-1472.
[25] Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald. An Nlog(N) method for Ewald sums in large systems. *J. Chem. Phys.* **1993**. *98*. 10089-10092.
[26] Essmann, U.; Perera, L.; Berkowitz, M.; Darden, T.; Lee, H.; Pedersen L. A smooth particle mesh ewald potential. *J. Chem. Phys*. **1995**. *103*. 8577-8592.
[27] Frigo, M.; Johnson, S. The Design and Implementation of FFTW3. *Proc. IEEE.* **2005**. *93*. 216-231.
[28] Goga, N.; Rzepiela, A.; de Vries, A.; Marrink, S.; Berendsen, H. Efficient Algorithms for Langevin and DPD Dynamics. *J. Chem. Theory Comput.* **2012**. *8*. 3637-3649.
[29] Berendsen, H.; Postma, J.; DiNola, A.; Haak, J. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **1984**. *81*. 3684-3690.
[30] Parrinello, M.; Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.* **1981**. *52*. 7182-7190.
[31] Jasiński, M.; Feig, M.; Trylska, J. Improved Force Fields for Peptide Nucleic Acids with Optimized Backbone Torsion Parameters. *J. Chem. Theory Comput.* **2018**, *14*, 3603-3620
[32] Valdés-Tresanco M.S.; Valdés-Tresanco M.E.; Valiente P.; Moreno E. gmx_MMPBSA: A New Tool to Perform End-State Free Energy Calculations with GROMACS. *J. Chem. Theory Comput.* **2021**. *17*. 6281-6291.
[33] Still, C.; Tempczyk, A.; Hawley, R.; Hendrickson, T. Semianalytical Treatment of Solvation for Molecular Mechanics and Dynamics. *J. Am. Chem. Soc.* **1990**. *112*. 6127-6129.
[34] Onufriev, A.; Case, D. Generalized Born Implicit Solvent Models for Biomolecules. *Annu. Rev. Biophys.* **2019**. *48*. 275-296.
[35] Bashford, D.; Case, D. Generalized Born Models of Macromolecular Solvation Effects. *Annu. Rev. Phys. Chem.* **2000**. *51*. 129-152.
[36] Jawad, B.; Poudel, L.; Podgornik, R.; Ching, W.. Thermodynamic Dissection of the Intercalation Binding Process of Doxorubicin to dsDNA with Implications of Ionic and Solvent Effects. *J. Phys. Chem. B.* **2020**. *124*. 7803-7818.
[37] Jawad, B.; Poudel, L.; Podgornik, R.; Steinmetz, N.; Ching, W. Molecular mechanism and binding free energy of doxorubicin intercalation in DNA. *Phys. Chem. Chem. Phys.* **2019**. *21*. 3877-3893.
[38] Gonçalves, P.; Stassen, H. Calculation of the free energy of solvation from molecular dynamics simulations. *Pure Appl. Chem.* **2004**. *76*. 231-240.
[39] Janežič, D.; Venable, R.; Brooks, B. Harmonic analysis of large systems. III. Comparison with molecular dynamics. *J. Comp. Chem.* **1995**. *16*. 1554-1566.
[40] General, I. A Note on the Standard State's Binding Free Energy. *J. Chem. Theory Comput.* **2010**. *6*. 2520-2524.
[41] SantaLucia Jr., J.; Allawi, H.; Seneviratne, P. Improved Nearest-Neighbor Parameters for Predicting DNA Duplex Stability. *Biochemistry.* **1996**. *35*. 3555-3562.
[42] Xia, T.; SantaLucia Jr., J.; Burkard M.; Kierzek R.; Schroeder S.; Jiao X.; Cox C.; Turner D. Thermodynamic Parameters for an Expanded Nearest-Neighbor Model for Formation of RNA Duplexes with Watson-Crick Base Pairs. *Biochemistry*. **1998**. *37*. 14719-14735.
[43] SantaLucia Jr., J. A unified view of polymer, dumbbell and oligonucleotide DNA nearest-neighbor thermodynamics. *Biochemistry*. **1998**. *95*. 1460-1463.
[44] Golyshev, V.; Pyshnyi, D.; Lomzov, A. Calculation of Energy for RNA/RNA and DNA/RNA Duplex Formation by Molecular Dynamics Simulation. *Molecular Biology* **2021**, *55*, 927-940
[45] Nielsen, P. Peptide Nucleic Acid. A Molecule with Two Identities. *Acc. Chem. Res.* **1999**, *32*, 624−630.
[46] Nielsen, P. Peptide nucleic acid (PNA). A DNA mimic with a pseudopeptide backbone. *Chem. Soc. Rev.* **1997**, *2*, 73-78
[47] Sen, S.; Nilsson, L. Molecular Dynamics of Duplex Systems Involving PNA: Structural and Dynamical Consequences of the Nucleic Acid Backbone. *J. Am. Chem. Soc.* **1998**, *120*, 619-631

Chapter 5: Atomic simulations of ds-PNA under external force

5.1 Introduction

Technical research question (2) “Can molecular dynamics be used to assess the single-molecular properties of a candidate bioadhesive under external force, and what quantities contribute to this?” can be in part answered by deriving the relationship between energy and a reaction coordinate associated with an external force. Knowing how ds-PNA energies are affected by external forces could aid in the design of other nucleic acid analogues which optimise these energies along an unbinding coordinate. For example, if ds-PNA were exploited in a nucleic acid adhesive, as has been demonstrated for ds-DNA and as its intrinsic higher stability makes it suitable for,[1] then its response to an external force is critical to predicting and optimising its function.

These all-atom molecular dynamics simulations are partly validated since enthalpy was shown to be well-approximated by the molecular mechanics potential energy in Chapter 4. In molecular dynamics simulations, force is conservative, meaning it is only a function of the potential energy gradient and atomic positions, and therefore an accurate characterisation of potential energy implies an accurate characterisation of force. An additional validation step was taken, however, for the ds-PNA sequence studied by determining its binding free energy using AWH. AWH was additionally used to construct free energy landscapes for base pairing and stacking, as has been done for ds-DNA,[2] at different distances during shearing to evaluate the effect of external force on the pairing and stacking energies.

To study ds-PNA under external force the ds-PNA sequence GTAGATCACT, with its complementary strand implied, was selected to be the pilot sequence for all-atom SMD simulations using Jasiński *et al.*’s CHARMM parameters.[3] This was since it: (i) was the most well-studied ds-PNA in the literature, (ii) was a suitable length for

studying shearing according to the de Gennes model, and (iii) had the most accurately evaluated binding enthalpy from the MM-GBSA study in Chapter 4, indicating that the molecular mechanics forcefield, which largely determines this enthalpy, was well-parameterised. Using SMD, rupture forces at varying loading rates were obtained. Averages were then conducted over forces, displacements, van der Waals, electrostatic and torsional angle energies so that the behaviour during shearing and unzipping could be derived and generalised.

5.2 Methods

5.2.1 Defining the unbinding coordinates

An atomic antiparallel P-form structure of the ds-PNA GTAGATCACT was selected for the SMD simulations. At the C-terminus, the pulling atom is the terminal carbonyl carbon atom, and at the N-terminus it is the terminal nitrogen atom (Figure 5-1).

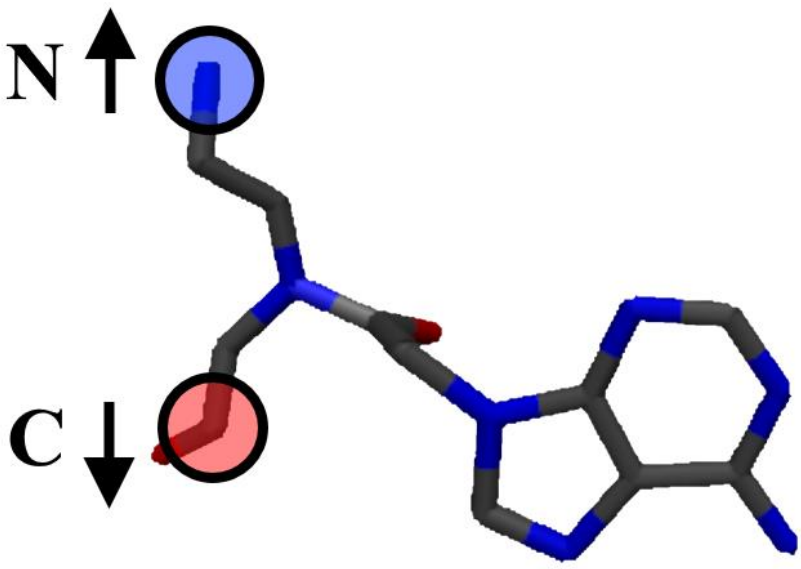


Figure 5-1: Atomic structure of a PNA monomer showing only heavy atoms.

Since the ds-PNA used is antiparallel, each double-stranded terminus is composed of an N- and a C- terminus from the two individual strands. For shearing, the two single strands are pulled away from one another along the helical axis defined by the distance between the N-terminal pulling atoms of each strand, whereas for unzipping the two single strands are pulled away from one another along a perpendicular axis defined by the distance between an N-terminal pulling atom of one strand and the C-terminal pulling atom of the other (Figure 5-2).

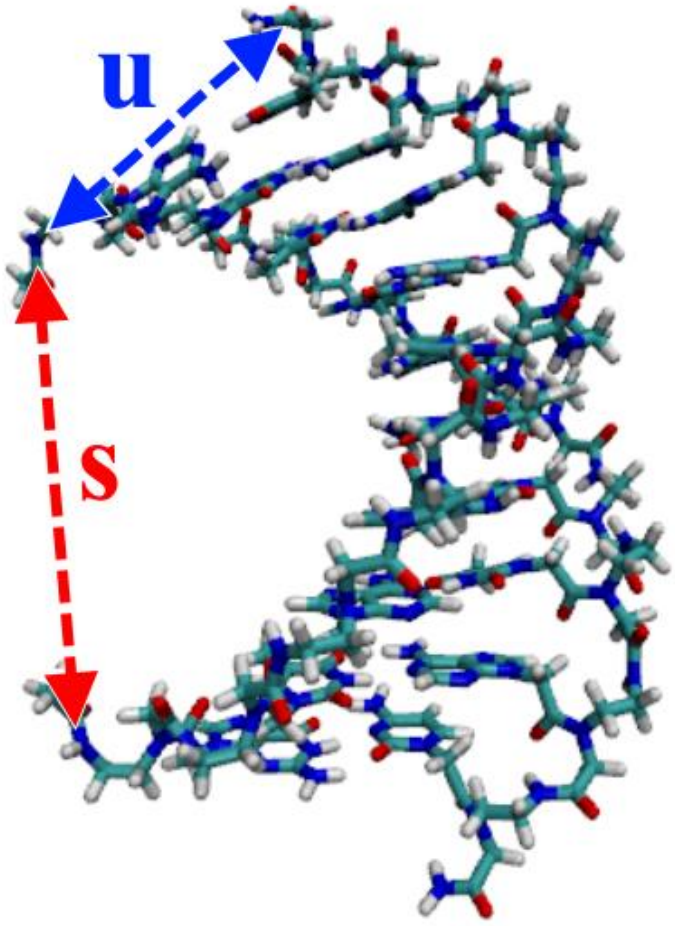


Figure 5-2: The unzipping (u) and shearing (s) reaction coordinates in blue and red respectively.

For both unzipping and shearing, the reaction coordinates are aligned with the longest axis of a cuboidal simulation box since, in Gromacs, if the pulling distance exceeds half the box length, then the simulation crashes.[4,5]

5.2.2 Equilibration with restraints

The same all-atom structure of GTAGATCACT ds-PNA generated previously via the pNAB was imported into Gromacs and placed in a cuboidal simulation box with short x and y axes and a long z axis with which the reaction coordinates as defined in Figure 5-2 were aligned. The N- and C- termini were acetylated and amidated respectively. The structure was energy minimised in vacuum over 50,000 steps using a steepest descent algorithm before being solvated in explicit spc216 water.[6] The system was then again energy minimised over 50,000 steps of the same steepest descent algorithm.

After minimisation, the structure was heated from 30 to 298 K over 500 ps with position restraints and Newton's equations of motion were integrated using a leap-frog integrator.[7,8] Temperature was weakly coupled to an external bath using a velocity rescaling[9] algorithm. Harmonic restraints of 24 kcal mol^{-1} $\mathrm{Å}^{-1}$ were applied to non-hydrogen atoms whereas hydrogen atoms were constrained using a LINCS[10] algorithm. The van der Waals interactions were handled using a switched cut-off scheme with switching to zero occurring from 10 to 12 Å. The electrostatic nonbonded interactions were treated using a particle mesh Ewald (PME)[11,12] algorithm with quartic interpolation and a grid spacing of 1 Å. The short-ranged component of the PME was computed to 10 Å and long-ranged components were handled using the fast Fourier transform library FFTW.[13]

After heating, position restraints were halved every 2 ns, being removed entirely at 10 ns, with the equations of motion and temperature coupling hereon handled by a stochastic integrator which incorporated friction and noise terms.[14] Unlike with the MM-GBSA study, position restraints on the pulling atoms were kept such that the reaction coordinate remained aligned with the longest box vector. The positional and velocity information at the end of this stage was used as the starting point to separately equilibrate 50 independent systems.

Each of the 50 systems was equilibrated in the same manner as with the MM-GBSA study. That is, they were equilibrated for 10 ns at constant temperature followed by 10 ns with pressure coupling at 1 bar handled by a Berendsen barostat.[15] These equilibrated systems were then used as the starting points for the production simulations, and 3 replicates would be obtained for each of the 50 separately equilibrated systems at each of the 4 loading rates (Figure 5-3).

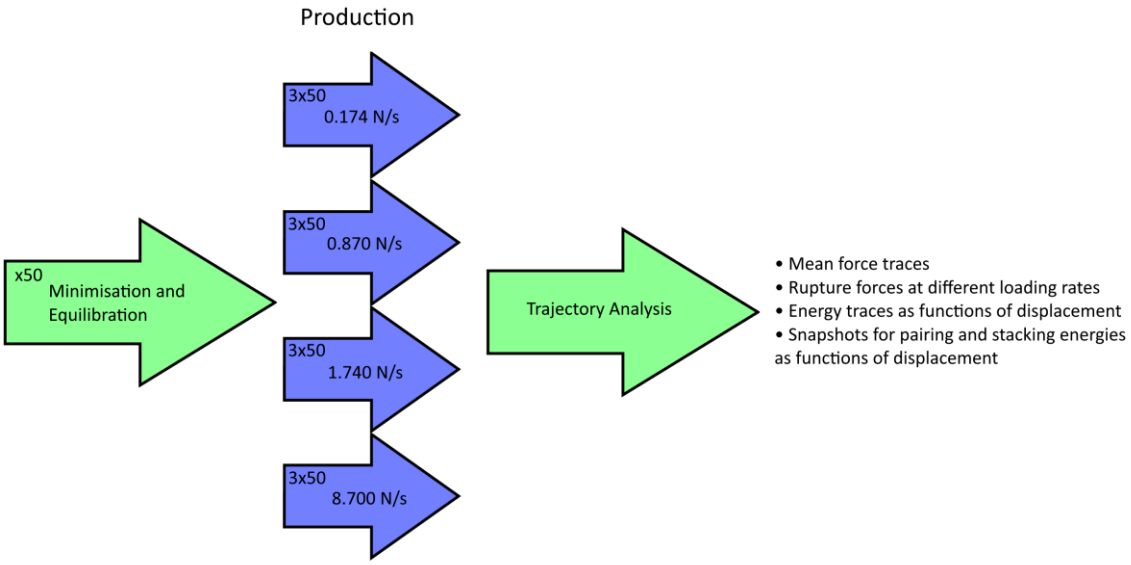


Figure 5-3: Process diagram for the general simulation protocol used in these SMD simulations. Numbers in the top left arrow indicate the number of replicates.

For the above process diagram, technical replicates are obtained at the production stage from 50 separately equilibrated systems. At the production stage, external force is provided using SMD. This process diagram is applied to both the unzipping and shearing coordinates.

5.2.3 Force constant selection

To ensure that the force constant did not introduce bias or instabilities into the simulations it was important to demonstrate that the reaction trajectory was consistent for different values of the constant. This was done by conducting shearing SMD simulations, described in Chapter 5.2.4, at 1 m/s pulling velocities with force constants ½, twice and four times the magnitude of the 0.174 N/m constant used. This 0.174 N/m constant was initially selected since it was the average force constant of the available in-house SNL-10 AFM cantilevers, which are designed for soft biological samples, and consequently opens the possibility for the present work to be consistently extended to real-world experiments. The force constant was determined to be appropriate since reaction trajectories demonstrated a well-defined unbinding event and trajectories collapsed to the same line, determined by the retraction velocity, regardless of the constant used (Figure 5-4).

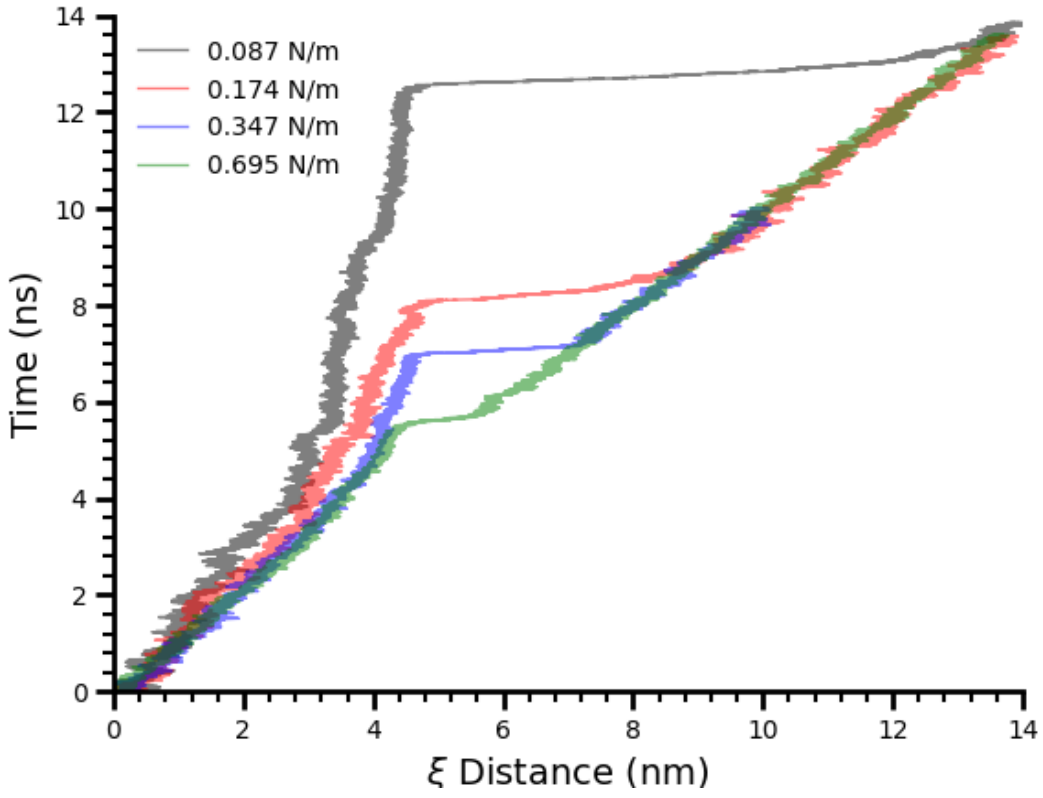


Figure 5-4: Time versus normalised reaction coordinate distance for the ds-PNA GTAGATCACT under shearing force

Lower spring constants take longer to rupture, but rupture occurs at the same displacement, indicating that the changing spring constant has no effect on the critical extension require for rupture to occur. In all cases, a well-defined unbinding step is evident, and unbound states collapse to the same line $y = x$ given by the retraction velocity of 1 m/s. Since it was the case that the force constant did not visually bias the trajectories, it was determined to be an appropriate force constant for the present work.

5.2.4 Steered molecular dynamics at variable loading rates

During the pulling simulations, the position restraint on the terminal atom of the complementary strand CATCTAGTGA was removed. A harmonic potential was added with an energy minimum at the initial starting distance between the two pulling atoms. The harmonic potential was then moved with a constant velocity along the vector connecting the two atoms such that the distance between them would be increased. Velocities of 1, 5, 10 and 50 m/s were selected, with velocities lower than these being inaccessible with the computational resources available since lower velocities resulted in lower loading rates, requiring more total simulation time. These velocities corresponded with loading rates of 0.174, 0.870, 1.740 and 8.700 N/s respectively.

For each loading rate, 50 independently equilibrated pulling simulations with 3 replicates each were conducted for both shearing and unzipping. Simulations were terminated when the distance between the two pulling atoms was equal to half the longest box length. During these production runs, pressure was coupled using Parrinello-Rahman pressure coupling.[16] Each simulation trajectory outputted a force-time and distance-time graph, and the van der Waals and electrostatic energies between the strands, and the torsional energies of their backbones, were similarly tracked.

5.2.5 Defining the most probable rupture force

The most probable rupture force is obtained from the peak in the distributions of rupture forces obtained from all 150 runs at each loading rate. For shearing, this rupture force corresponds with a singular rupture event when the two strands separate, as will be demonstrated in Chapter 5.3. For unzipping, however, rupture is not a singular event but a series of rupture events for the individual base pairs, as was discussed with ds-DNA and as will again be demonstrated in the present work for ds-PNA. Resultantly, it is hard to define a 'rupture force' for the unzipping of the entire ds-PNA strand since there are multiple rupture events. In acknowledgement of this, a maximum force is defined in the present work instead, which for shearing is equal to the rupture force but for unzipping does not necessarily represent any particular rupture event.

5.2.6 Shearing distances for pairing and stacking AWH

During shearing, ds-PNA extends towards a normalised distance, between the two pulling atoms, of approximately 4.0 nm before the rupture event, as seen in Figure 5-4. In Chapter 5.3 the nature of this extension will be further investigated. As will be later demonstrated, this 'rupture' distance of 4 nm, in addition to another distance of 2.5 nm, which will be found to be associated with a transition region in Chapter 5.3, may be structurally significant to the shearing process.

To investigate the effect of shearing on base pairing and stacking energies, equilibrium AWH simulations are conducted. Since the conformations at 2.5 nm and 4 nm may be structurally significant to the shearing process, snapshots of these structures were obtained from the pulling simulations. Position restraints of 24 kcal mol^{-1} Å^{-1} were enforced on the pulling atoms of these snapshots, which were then re-equilibrated as before, producing equilibrated ds-PNA structures with 0 nm, 2.5 nm and 4.0 nm normalised shearing distances. Three replicates were obtained at each distance, and these replicates were used as the starting points for AWH production runs.

5.2.7 AWH pairing and stacking production runs

Terminal and internal G:C base pairs were studied at the 0, 2.5 and 4 nm extensions using AWH. The terminal base pairs were taken at the N-terminus of GTAGATCACT. The internal base pairs were taken as close to the 3rd-position, from the N-terminus, as possible, which for a G:C base pair was the 4th position. This decision was made to keep the definition of 'internal base pair' as close to the definition used in the MM-GBSA study as possible, but since GTAGATCACT did not have any 3rd-position G:C base pairs from either terminus the 4th position was the closest. Terminal and internal A:T base pairs were not studied since pulling from the A:T terminus was not studied, as this would have doubled the computational resource requirements.

Using AWH, the free energy surface at N1:N3 and 6-ring distances (Figure 5-5) were sampled for distances between 0.25 and 0.65 and 0.35 and 0.80 nm respectively, as was demonstrated in the previous literature for ds-DNA.[17] Since the N1:N3 distance is representative of the base pairing coordinate, and the 6-ring distance representative of the stacking coordinate,[17] the two-dimensional free energy surface produced by AWH characterises contributions to stability from both pairing and stacking. A cut-off value for the free energy of 100 kj mol^{-1} was used to prevent the system being forced into non-physical high-force conformations.

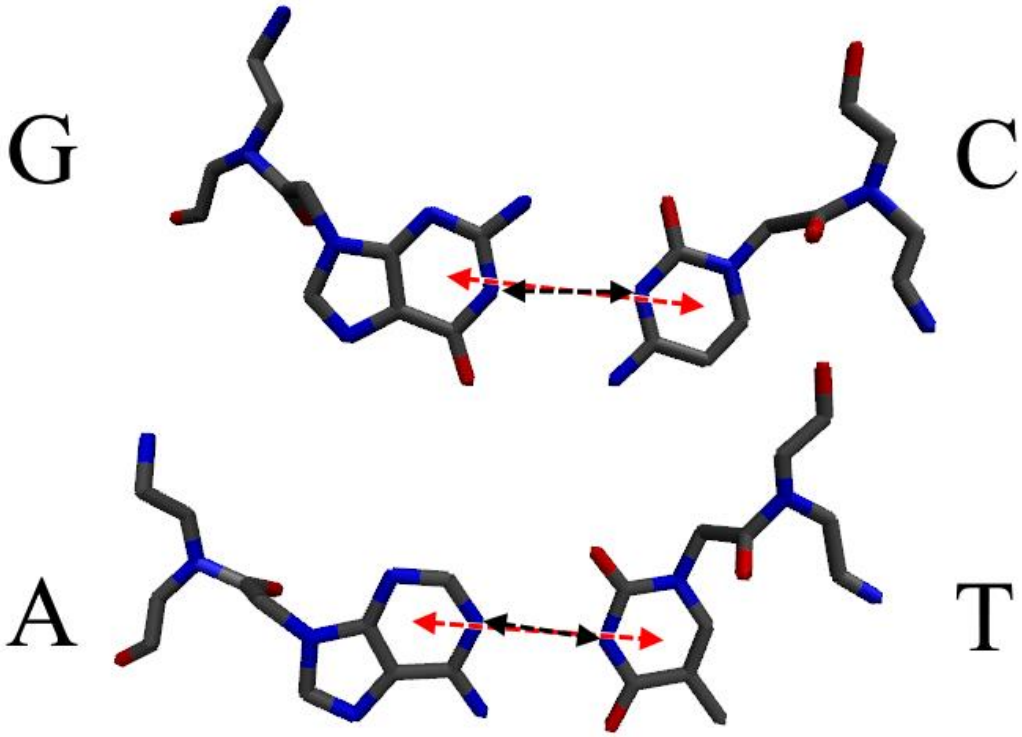


Figure 5-5: Reaction coordinate definitions for AWH simulations for G:C (top) and A:T (bottom) base pairs. 6-ring distance indicated by red arrows. N1:N3 distance indicated by black arrows.

5.2.8 AWH to determine binding free energy for validation

In addition to deriving free energy landscapes associated with pairing and stacking during shearing, AWH was used to determine the binding free energy of the ds-PNA GTAGATCACT. A reaction coordinate was defined as the distance between the two strands, which were pulled apart along a coordinate representative of neither shearing nor unzipping. This was done since, for both shearing and unzipping, the unbinding process results in an end-to-end alignment of the two strands and a very long unbinding coordinate between the two pulling termini, requiring very long box lengths. If the two strands are pulled apart in an effectively uniform manner along a coordinate perpendicular to their helical axis, with force applied to all atoms simultaneously, they do not align end-to-end and the box length in the pulling direction could be decreased approximately three-fold.

Deriving a free energy landscape along a very long reaction coordinate is computationally intensive, and by reducing the box length this limiting factor could be mitigated. This, however, means the free energy landscape is not representative of shearing or unzipping. However, since free energies are functions of state, the difference in the free energy (i.e., the binding free energy) between the bound and unbound states are theoretically equivalent regardless of the path taken. Since this reaction coordinate can derive the binding free energy as the difference in the free energy between bound and unbound states, even if intermediate states do not represent shearing or unzipping, it is a suitable and computationally achievable coordinate for further validation, alongside the results of Chapter 4, since the binding free energy of GTAGATCACT is well-reported in the literature.

To obtain the free energy along this coordinate, an external potential was applied to the centres of mass of all atoms of each strand, which were adaptively pulled apart along a reaction coordinate perpendicular to the longitudinal axis of the ds-PNA, for 100 ns. The free energy as a function of the normalised displacement between these centres of mass, obtained from 3 replicates, was plotted. The binding free energy was taken as the difference between the initial and final free energy at a normalised displacement of 4 nm.

5.3 Results and Discussion

5.3.1 Force-distance curves at variable loading rates

As hypothesised based on the prior literature for the shearing and unzipping of ds-DNA,[18,19,20] the shearing and unzipping coordinates for ds-PNA were similarly distinguished from one another. For shearing, a single rupture event occurs beyond a critical distance that rapidly dissipates the force, whereas for unzipping the force-distance traces demonstrate a continuous dissipation of force resulting in an approximately flat force-distance relationship after a sharp initial rise. The exception to this is that, at a loading rate of 8.700 N/s, the force-distance

relationship increases much more rapidly during unzipping, resulting in much higher maximum forces.

Averages over forces will be discussed in Chapter 5.3.3, whilst this subchapter highlights differences between individual force traces to demonstrate how these traces change appearance with increasing loading rate. For shearing (Figure 5-6), and as predicted by the force spectroscopy models, the peak force is an increasing function of loading rate but the extension at which the peak force occurs is approximately constant, being between 4 and 5 nm. After rupture, force returns to a baseline, though only for the lowest loading rate, 0.174 N/s, was this observed in the present work. This is since, when peak force was higher, force decreased over a longer distance after rupture and the maximum observable distance was 11 nm because of the box size.

In all shear force-distance curves a change from a low linearly increasing to a high linearly increasing gradient was observed at approximately 2.5 nm extensions. This was more prominent for lower loading rates, indicating a potential change in the dynamics of shearing beyond these extensions. The increased prominence of this observation at lower loading rates may be since, at higher loading rates, the rate at which force builds is too large for internal rearrangements in ds-PNA to occur. The issue of large loading rates on steered molecular dynamics simulations has been discussed in the literature. A similar observation, that under high retraction velocities the force does not propagate through the complex, and so intramolecular rearrangements cannot occur, was made.[21]

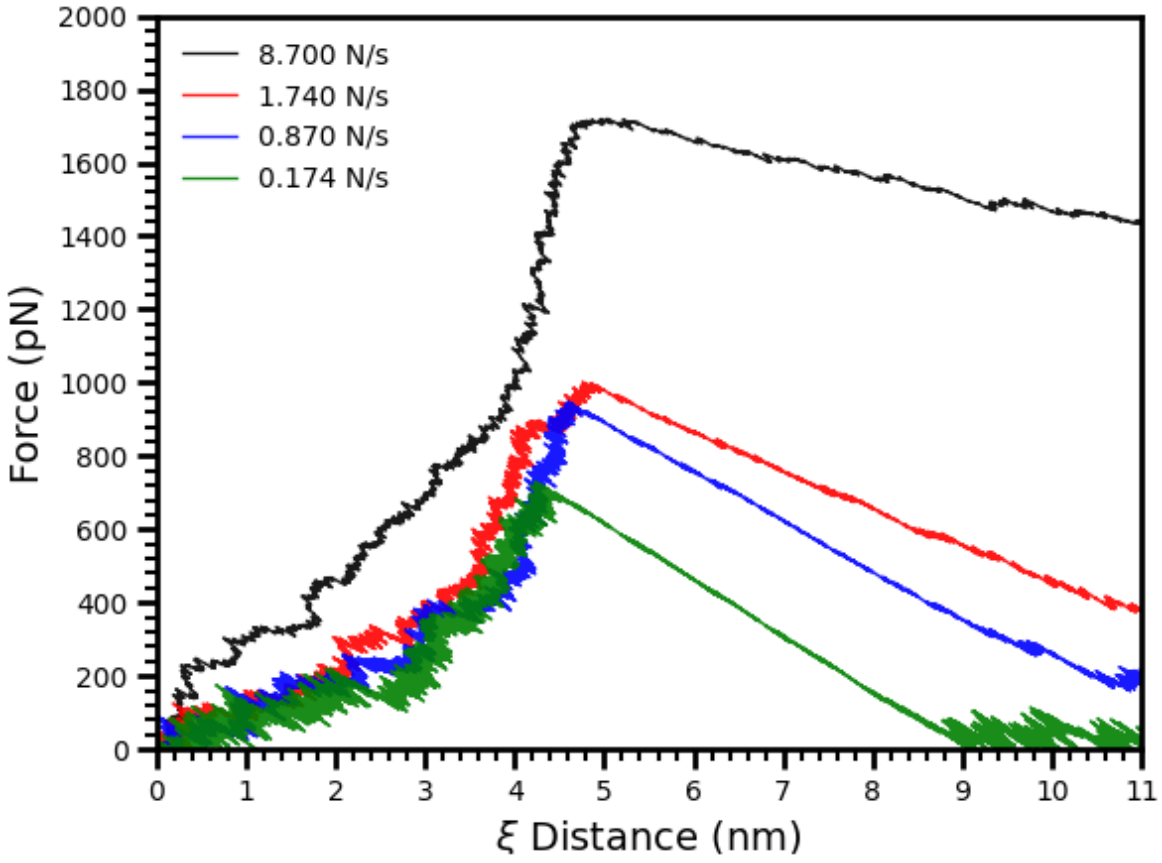


Figure 5-6: Example force-distance curves for the pulling of ds-PNA at 0.174, 0.870, 1.740 and 8.700 N/s in the shearing configuration in green, blue, red and black respectively.

In contrast to shearing, the maximum forces observed for unzipping are lower, peaking between 1100 and 1200 pN for the maximum pulling rate as opposed to between 1600 and 1800 pN for shearing. Force increases sharply on the strands until 0.6 nm displacements, after which it decreases for the lowest 2 loading rates before remaining approximately constant. This initial decrease occurs due to the first rupture event, corresponding to the rupture of the terminal base pairs, or a cluster of pairs, at the pulling terminus. For the 1.740 N/s loading rate, however, force decreases only slightly after the initial rupture before increasing again towards a higher, approximately flat region in the force trace. The 8.700 N/s trace is distinct from the other three in that force remains an increasing function of distance, excluding short flat regions. At this highest loading rate, visual inspection of the

trace would suggest that ds-PNA is unable to dissipate force through base pair rupture and therefore the trace does not stabilise but continuously increases.

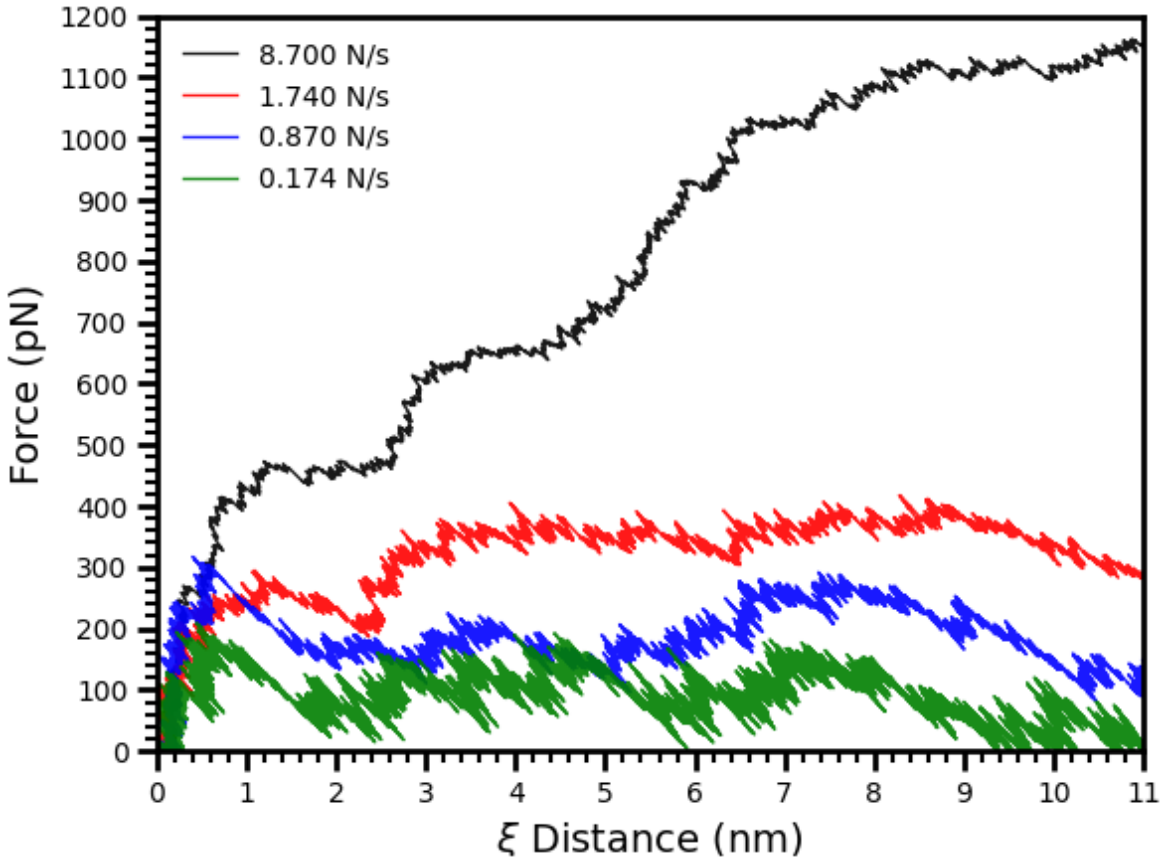


Figure 5-7: Example force-distance curves for the pulling of ds-PNA at 0.174, 0.870, 1.740 and 8.700 N/s in the unzipping configuration in green, blue, red and black respectively.

For both shearing and unzipping, the highest loading rate indicates a change in the dynamics of the system wherein the external potential dominates the internal attractive forces in ds-PNA. In shearing, this is evident in the occlusion of internal rearrangements (the gradient shift at 2.5 nm), and in unzipping this is evident in the inability of internal rearrangements i.e., base pair rupture, to mitigate the external forces.

5.3.2 Maximum force distributions

Example force-distance curves provide a visual demonstration of how loading rate affects the dynamics of ds-PNA under external force. For a more informed discussion, it is necessary to generalise these observations using averages or distributions. For shearing, the most probable rupture forces at 0.174, 0.870, 1.740 and 8.700 N/s loading rates are 1660.54, 1095.22, 971.77 and 747.84 pN respectively. Forces were distributed around these most probable values and histograms were approximately 500 pN wide at their base (Figure 5-8). According to the Bell-Evans force spectroscopy model, the most probable rupture force is a linear function of the natural logarithm of the loading rate,[22] though for the present work this is not the case. This reproduces the expected nonlinearity in force-loading curves at high loading rates and demonstrates the inapplicability of the Bell-Evans model to the force-loading curves obtainable for all-atom molecular dynamics simulations. This is because, in general, loading rates attainable using molecular dynamics are much higher than those used in the parameterisation and fitting of force spectroscopy models.[23]

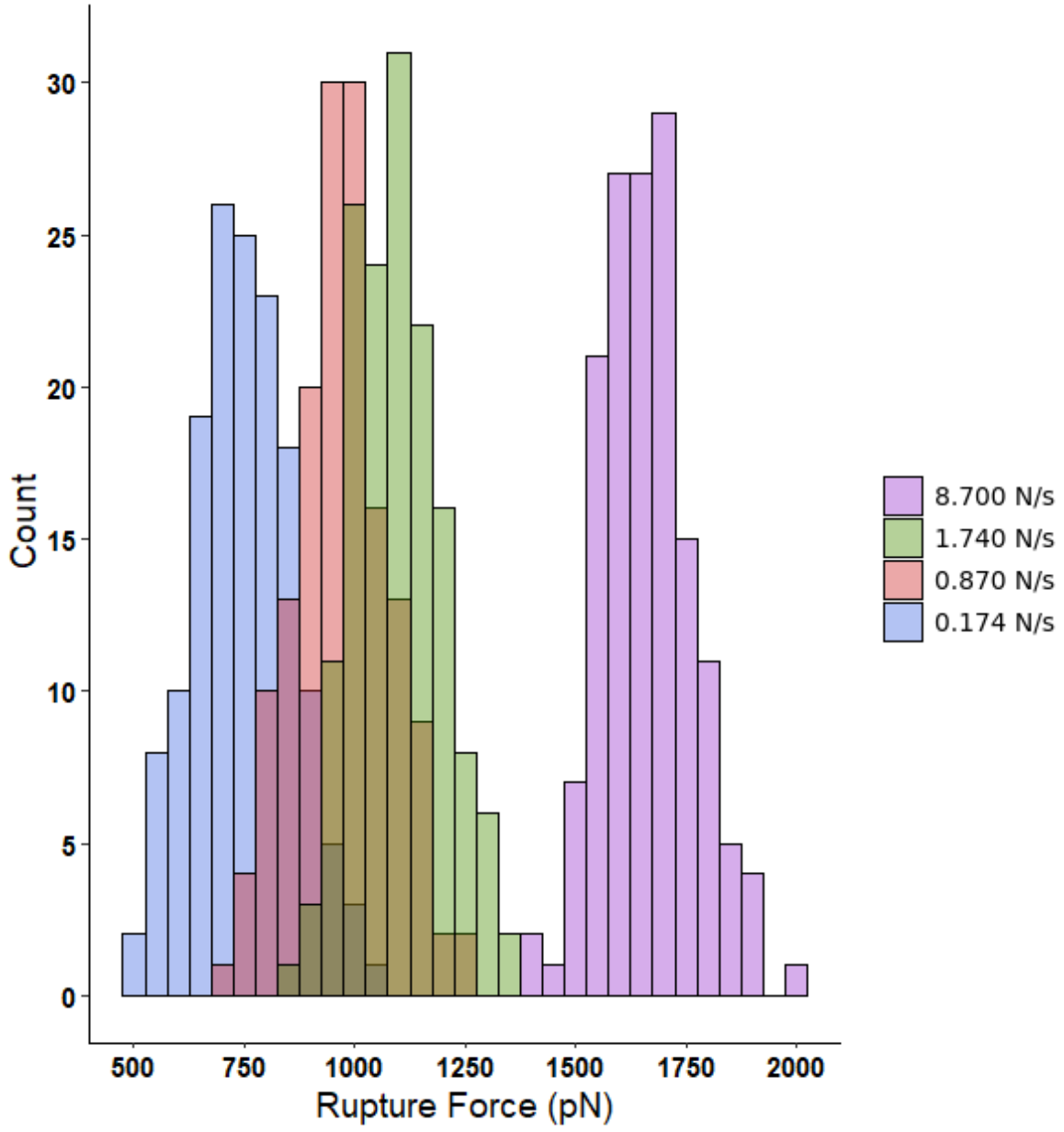


Figure 5-8: Histogram distributions of the shearing rupture force for the loading rates 0.174, 0.870, 1.740 and 8.700 N/s in blue, red, green and purple respectively. Overlapping histograms result in intermediate colouration.

Similarly to the shearing rupture force distributions, the maximum force distributions for unzipping (Figure 5-9) show a jump between the first and second highest loading rate, though the jump is visually more distinct in the unzipping case. The maximum force distributions are also narrower. This is likely since, whilst the rupture force of sheared ds-PNA corresponds to a singular rupture event, sampled once per trace, the maximum unzipping force is associated with a plateau in the

force trace and so the maximum force region is continuously sampled within the same trace, which could result in averaging and hence narrowing of the distributions.

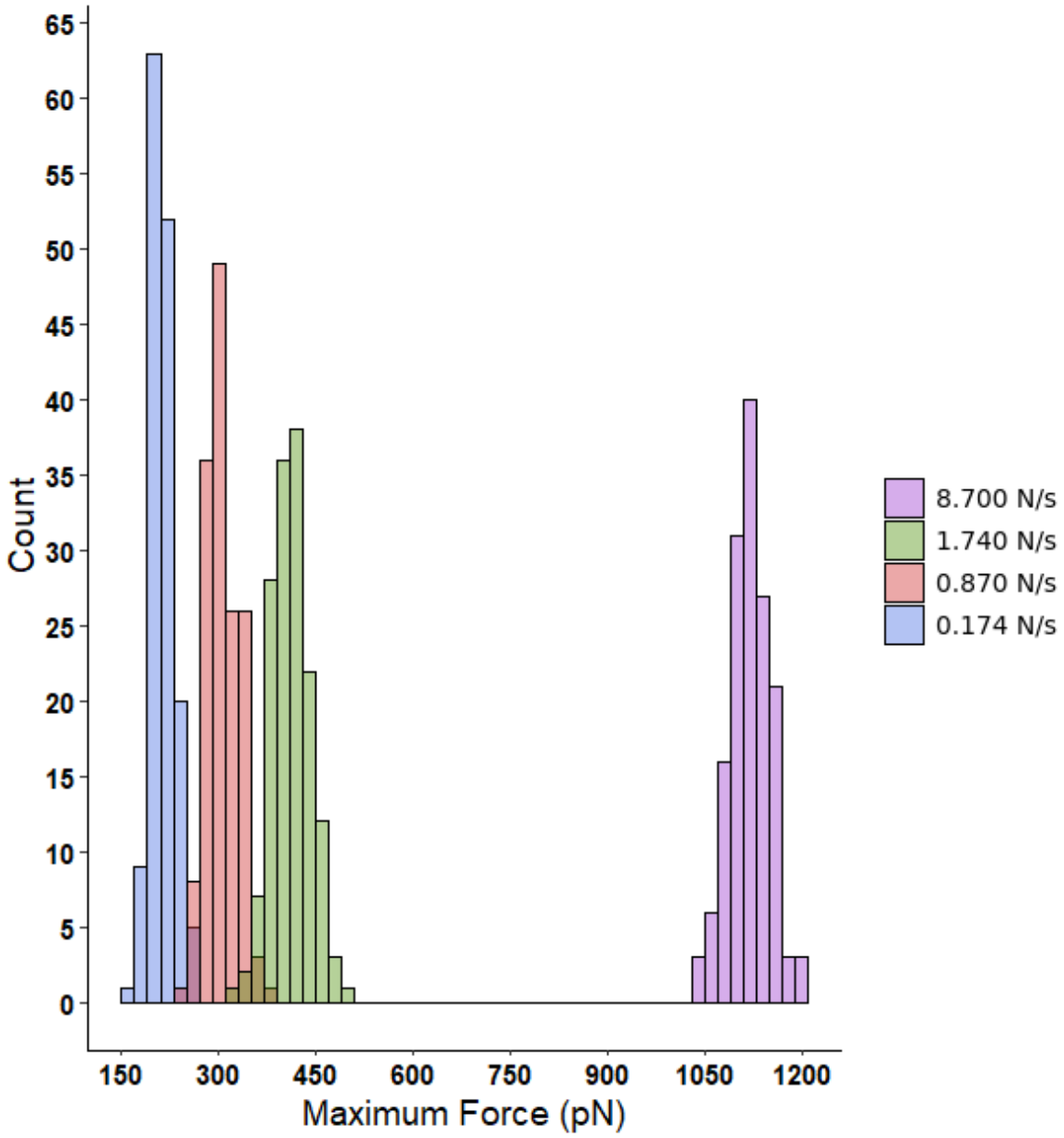


Figure 5-9: Histogram distributions of the unzipping maximum force for the loading rates 0.174, 0.870, 1.740 and 8.700 N/s in blue, red, green and purple respectively. Overlapping histograms result in intermediate colouration.

These maximum force distributions demonstrate that ds-PNA resists the application of external force more effectively in the shearing configuration, which has implications for loading configuration selection in technologies which maximise ds-PNA's resistance to bond rupture, for example as an adhesive.[1] In addition, these results agree with the prior observations that, for ds-DNA, the shearing maximum forces exceed the maximum forces recorded during unzipping.[18,24] It should be noted however that, since all-atom molecular dynamics simulations are limited to high loading rates, then these results reflect only the relationship between rupture force and loading rate in the high-force, highly non-equilibrium regime. At lower loading rates, nearer equilibrium conditions, it is plausible that rupture forces could converge.

5.3.3 Mean force and structure as functions of displacement

By averaging force-distance traces for a given loading rate it is possible to derive details in the unbinding process previously obscured due to fluctuations in the measured forces on a single trace or limited sampling. Since force only returns to the baseline for the example curves at the lowest loading rate, and since it has been suggested that velocities greater than 1 m/s may obscure internal rearrangements, then these mean force traces were produced only for the 0.174 N/s case (Figure 5-10).

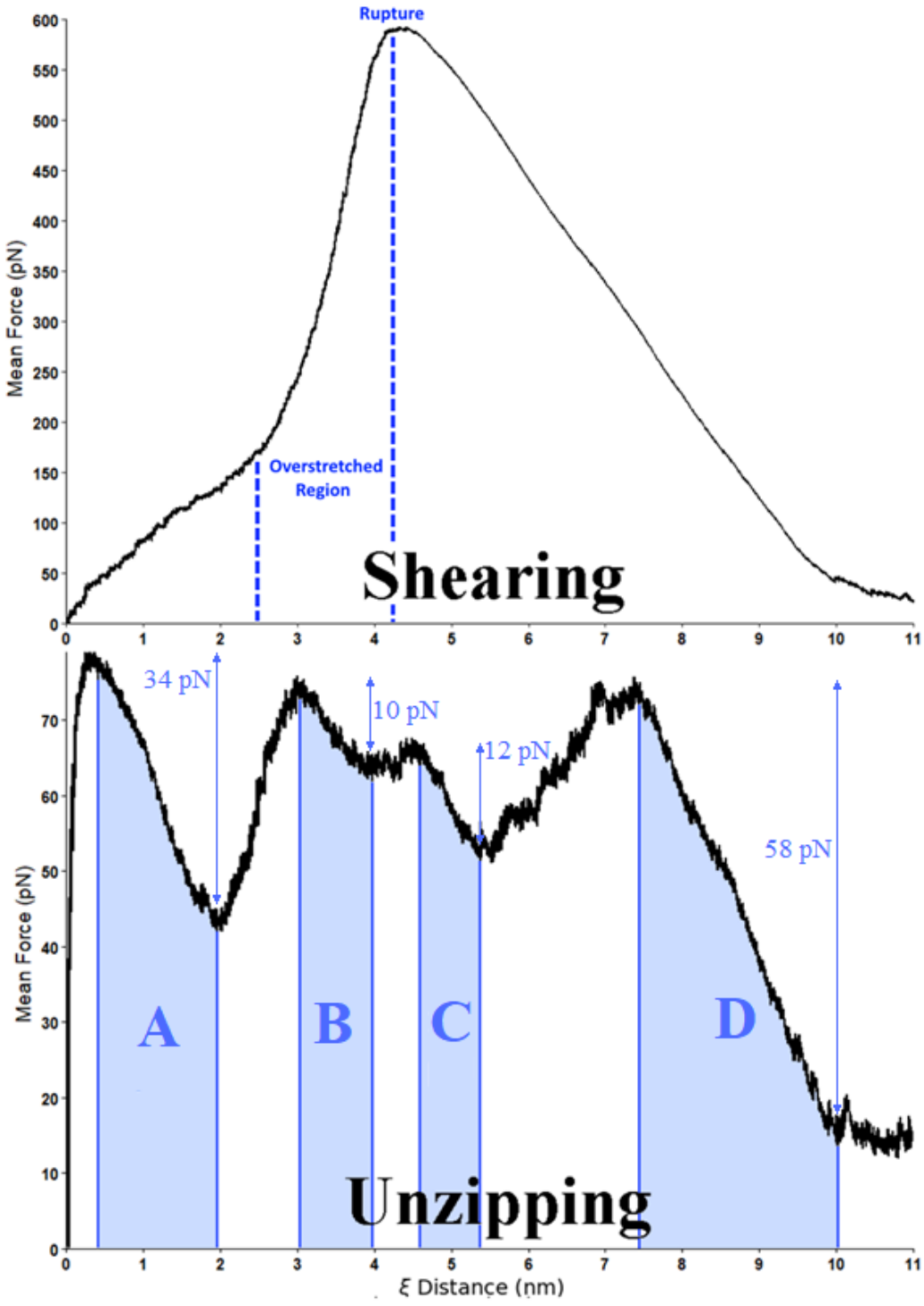


Figure 5-10: Mean force traces for shear (top) and unzipping (bottom) unbinding processes at 0.174 N/s. In the shear trace the overstretching region is indicated between the two blue dashed lines. The second dashed line is the rupture distance. In the overstretching trace the rupture events are indicated by the shaded regions A, B, C and D for the 1st, 2nd, 3rd and 4th rupture event respectively. The magnitude of each rupture event is indicated by the arrows in the direction of the mean force axis. These are 34, 10, 12 and 58pN for rupture events A, B, C and D respectively.

From Figure 5-10 it is evident that, for the shearing case, there is a rapid increase in gradient at approximately 2.5 nm extensions. The initial length of the energy-minimised ds-PNA structure along this reaction coordinate, representing the end-to-end distance between the two pulling atoms, is 2.7 nm. For B-form ds-DNA, an extension of 1.7 times the initial end-to-end distance represents stretching from the B-form to the S-form. [20,24] A similar factor, 1.9, connects the initial end-to-end distance in P-form ds-PNA in these pulling simulations with the distance at which the 2.5 nm gradient shift occurs, potentially indicating a similar transition from a double-helical P-form to an overstretched form of ds-PNA. By mapping the atomic structure during the shearing process (Figure 5-11) this overstretching can be demonstrated.

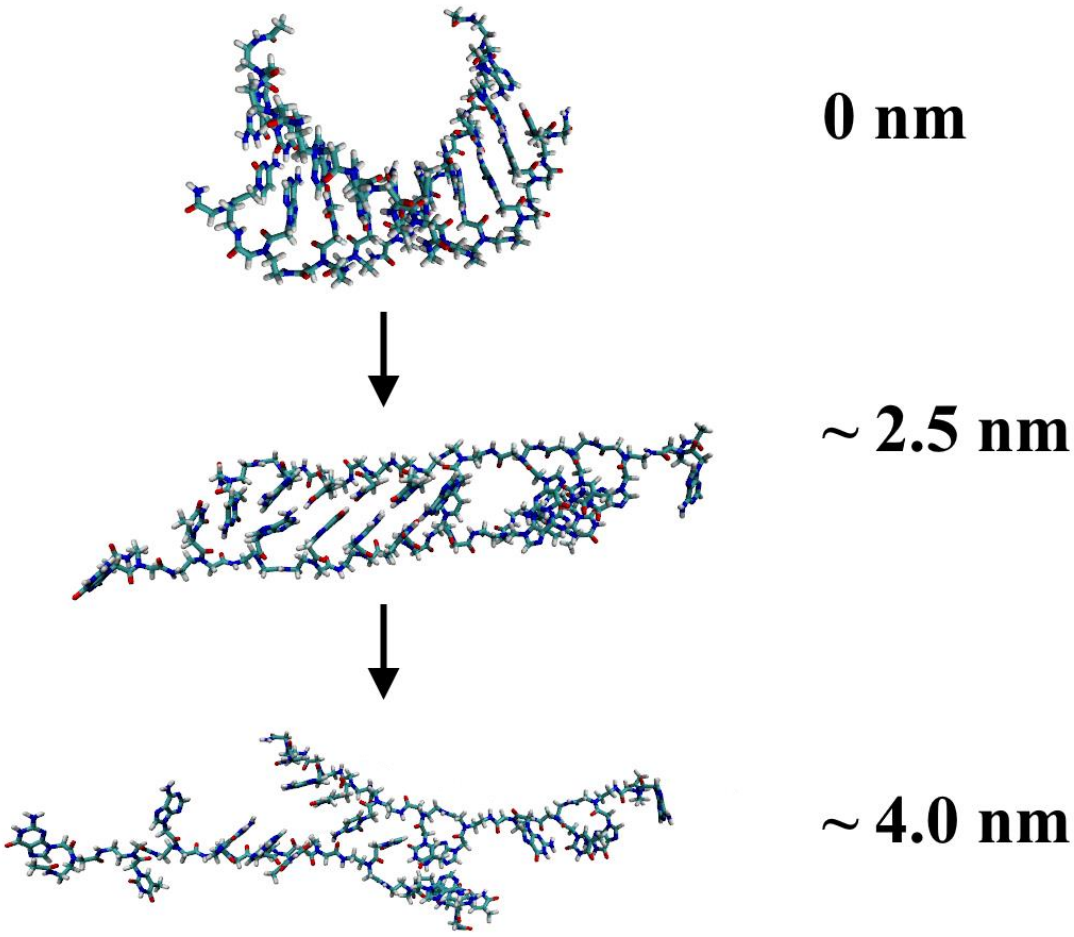


Figure 5-11: Atomic structure representations of ds-PNA under shearing force. Normalised 0 nm, 2.5 nm and 4.0 nm extensions are indicated.

Figure 5-11 shows that, at the 2.5 nm extension associated with the gradient shift, the ds-PNA adopts an overstretched ladder-like conformation similar to the S-form ds-DNA. In the overstretched form, the once-perpendicular base pairs become tilted in the pulling direction as seen in the middle pairs of the 2.5 nm image of Figure 5-11. This overstretched form demonstrates fraying, though remains stabilised by the tilted, complementary pairs in the middle of the ladder. At the 4 nm extension these complementary base pairs are now mostly gone, with the structure held together partly by non-specific associations between the bases immediately prior to the rupture event. These structural snapshots suggest that the sharp increase in the force gradient associated with overstretched ds-PNA is since extension can no longer be achieved by further unwinding of the double-helix. The transition to the overstretched form at this 2.5 nm extension is the justification for the selection of this extension for the later AWH study, as was discussed in Chapter 5.2.6.

For unzipping (Figure 5-12), averaging over the forces at 0.174 N/s has enabled individual rupture events to be resolved, which correspond to individual or sometimes doublets or triplets of base pairs. A sharp initial decrease in the force that bottoms out at 2 nm occurs when, usually, multiple base pairs rupture at once. This is rupture event A in Figure 5-10. Three other distinct peak-trough pairs are evident in the trace, with the largest difference between a peak and trough being the final rupture event, D. The heights of these events may be sequence dependent, though all individual base pair rupture forces are not resolved at the loading rate used since only 4 events are observed for a sequence with a length of 10 base pairs, demonstrating that, on average, between 2 and 3 base pairs rupture simultaneously. The largest rupture event, D, being the final indicates that the largest number of pairs are likely ruptured simultaneously at that point, suggesting that, below a minimum remaining length, the sequence is no longer able to resist the external potential through its cohesive forces and undergoes a rapid dissociation.

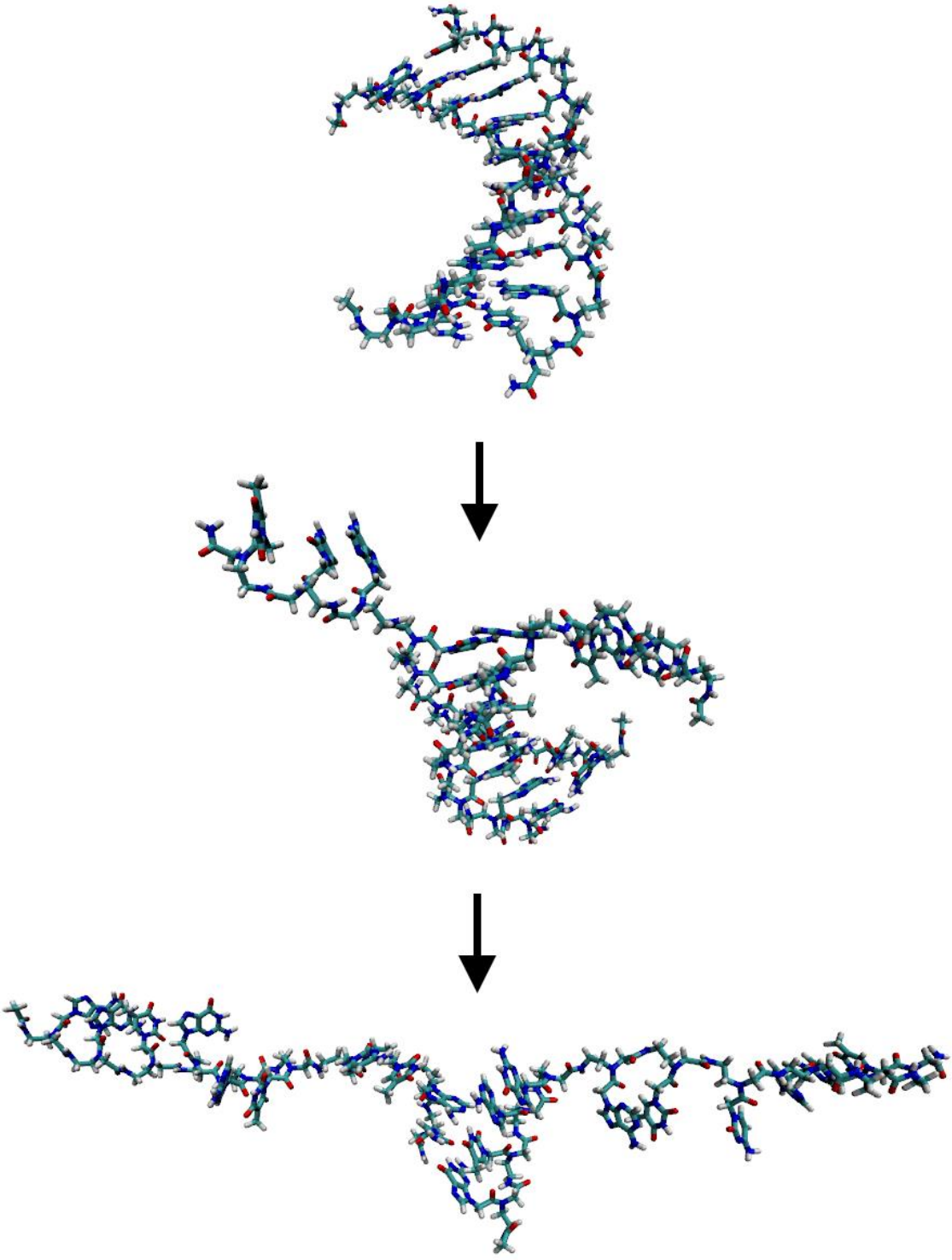


Figure 5-12: Atomic structure representations of ds-PNA under unzipping force. From top to bottom: equilibrated P-form structure, early unzipped structure after initial rupture event, late unzipped structure immediately prior to final rupture event.

The smallest two rupture events, the B and C, are near 10 and 12 pN respectively. The first event, A, at 34 pN, is approximately three-fold these values, whereas the final event, D, at 58 pN, is roughly five-fold. If the 2nd and 3rd rupture events are assumed to represent individual base pairs, then all ten base pairs can be accounted for. This is since the heights of the first and final rupture events are integer multiples (assuming slight variation depending on the position and composition of different pairs) that account for 3 and 5 bases respectively, totalling 10. This force trace is in slight disagreement with unzipping models which characterise unzipping as a base pair by base pair process,[25] since this is only likely the case for the 4th and 5th base pairs, given by the 2nd and 3rd rupture events, in the present work. Instead, these results indicate that, during unzipping, the P-form PNA resists the initial application of force until several pairs break at once. This is likely both a result of the higher inherent stability of ds-PNA compared to ds-DNA and the high loading rates testable in all-atom molecular dynamics simulations since the force increases to a large enough value to rupture several pairs very rapidly. At lower loading rates, unzipping may still be a base pair by base pair process.

5.3.4 Mean energies as functions of displacement

The structural changes, such as overstretching, discussed in the previous subchapter affect the unbinding process of ds-PNA. Similarly, inter-strand energies like van der Waals and Coulomb forces stabilise the double-helix and resist unbinding under the external potential. Also, intra-strand forces like the torsional forces of the backbone monomers may explain unbinding behaviour, particularly under shear force, given that torsional energies can be affected by unwinding of the double helix.

The inter-strand Coulomb energies over all 150 simulations at each loading rate were averaged for both shearing and unzipping (Figure 5-13). During shearing, the electrostatic forces of attraction between the two strands are unaffected for small extensions up to around 1.5 nm before increasing rapidly in a logistic-like manner that is steeper with increasing loading rate. The overstretched form of ds-PNA,

however, between 2.5 nm and the critical distance, occurring between 4 and 5 nm, experiences a rapidly increasing Coulomb energy. This suggests that whereas the initial unwinding of the helix does not decrease the electrostatic force of attraction, indicating that base pairing is still in a near-native, favourable conformation, the overstretching of the ds-PNA distorts these interactions. This is evident in the structural snapshots of ds-PNA, which demonstrate fraying in the overstretched form. At the critical distance at which rupture occurs, there are still significant Coulomb forces between the two strands, though since distance along the reaction coordinate increases rapidly at this point these forces likely arise from temporary interactions between the two strands as they slide past one another.

Contrastingly, during unzipping, the Coulomb energies increase in a mostly linear manner that is independent of the loading rate. Linearly increasing regions occur for displacements that correspond to rupture events in the unzipping mean force trace in Figure 5-10 and are indicated on Figure 5-13. These linearly increasing regions are separated by plateau regions, the least prominent of which is between the B and C rupture events. The Coulomb energy trace for the unzipping case thereby indicates that base pair rupture events result in a loading-rate independent increase in the Coulomb energy as complementary base pairs are pulled apart, with plateau regions indicating metastable distances prior to the rupture of the next pairs in line.

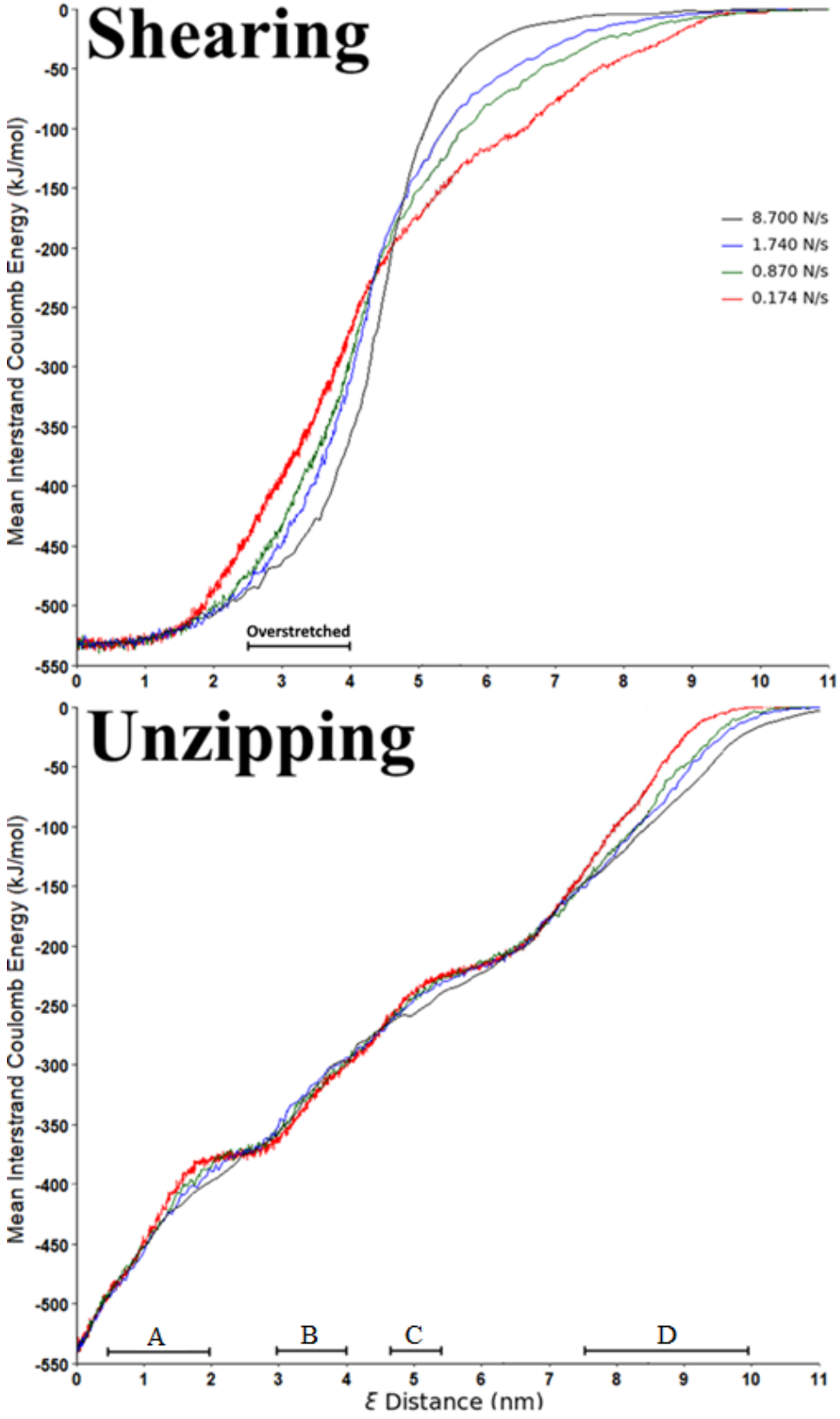


Figure 5-13: Mean inter-strand Coulomb energies for shearing (top) and unzipping (bottom) as a function of reaction coordinate distance. Regions of interest including the overstretching region in the shear trace and the four rupture regions in the unzipping trace are indicated.

A similar loading-rate dependence on energy traces during ds-PNA shearing is observed for the van der Waals energies between the strands (Figure 5-14). At low extensions, the van der Waals energy decreases towards a minimum that is deeper and shifted to the right for higher loading rates. At higher extensions, the van der Waals energy increases rapidly to the point of rupture and then levels off towards zero as strands slide past one another. The initial decrease in the van der Waals energy suggests that it initially acts to stabilise the double-helix during the shearing process, possibly due to the compression of the two strands such that non-specific forces of attraction are increased. This stabilisation is mitigated by an increase in the Coulomb energy. For example, for the lowest loading rate, the 2.5 nm overstretched form is associated with a $-37.7\ \mathrm{kJ/mol}$ stabilisation due to van der Waals energies, taken as the difference between the energy at 0 nm and 2.5 nm extensions, but a $+87.8\ \mathrm{kJ/mol}$ destabilisation due to Coulomb energies. The deepening of the well for higher loading rates may suggest that, under extreme shearing forces, the ds-PNA is compressed more extensively and these non-specific forces are further maximised, in part accounting for the higher shearing forces under higher loading rates.

For unzipping, a similar case to the Coulomb energy traces is observed, with the van der Waals energy being an approximately linearly increasing function of the reaction coordinate distance. Again, the van der Waals energy traces are not functions of the loading rate, and similarly to the Coulomb energy trace there are ‘plateaus’ in the curve that correspond to the displacements intermediate of two rupture events, though in the case of the van der Waals energy these are less prominent. Combined with the Coulomb energy trace, this trace suggests that unzipping is a process of approximately base-pair-by-base-pair rupture events wherein each base pair cluster, prior to its rupture, is associated with a plateau or near-plateau region in the Coulomb or van der Waals energy.

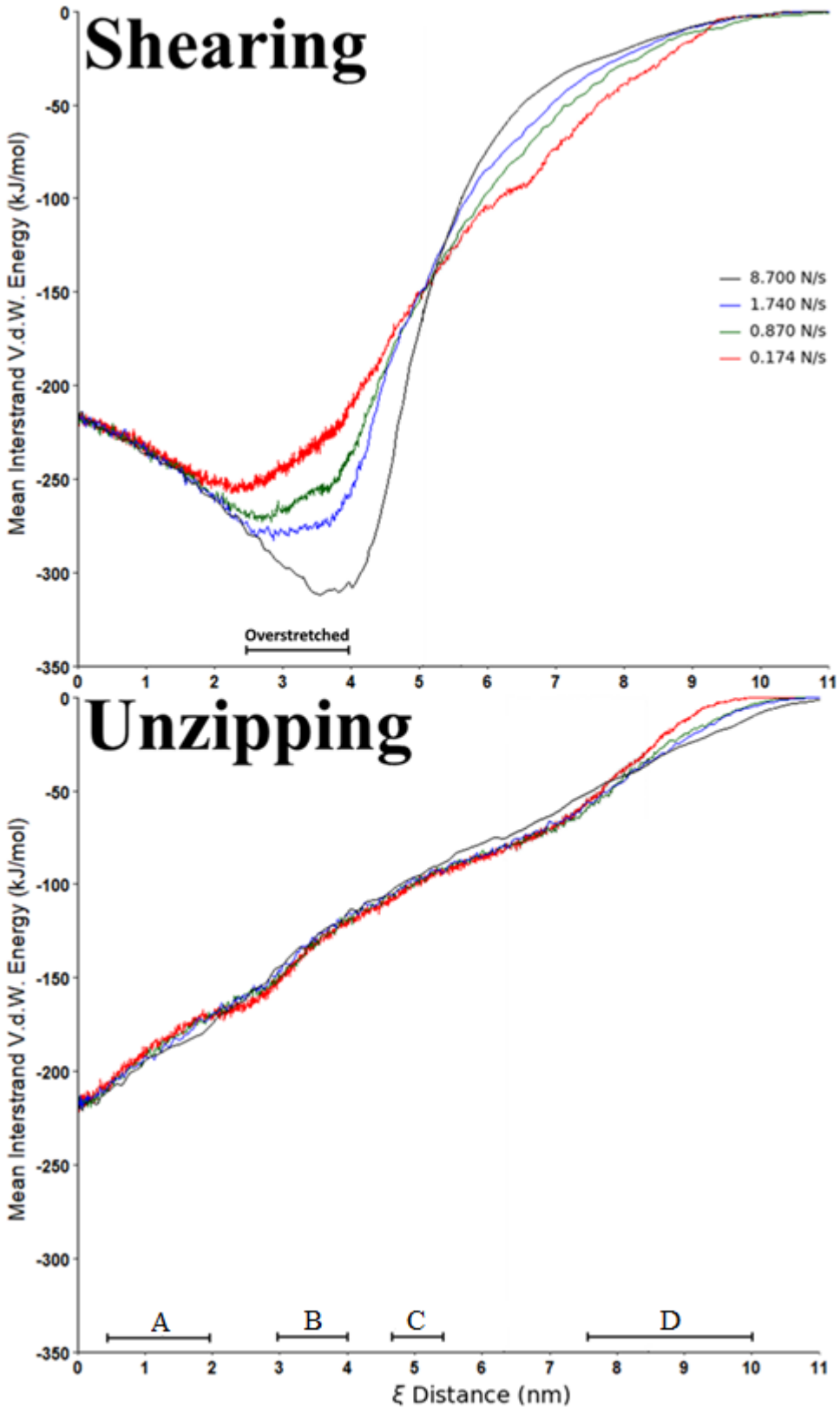


Figure 5-14: Mean inter-strand van der Waals energies for shearing (top) and unzipping (bottom) as a function of reaction coordinate distance. Regions of interest including the overstretching region in the shear trace and the four rupture regions in the unzipping trace are indicated.

Shearing and unzipping are similarly distinct with regards to their dihedral energy traces (Figure 5-15). These dihedral energies are associated with the torsional angles of the backbone,[4] and hence during unwinding, wherein the backbone is twisted, it is arguable that they would be expected to increase. During shearing, at low extensions, these dihedral energies are observed to increase slightly in a loading-rate independent manner. Since this region corresponds with the unwinding of ds-PNA towards an overstretched form, these higher energies reflect that change. Beyond the 2.5 nm overstretching distance, however, a loading rate dependence can be observed, and the dihedral energies peak rapidly towards a maximum between 4 and 5 nm at the point immediately prior to rupture. Of the three measured energy traces, the dihedral energies during shearing provide the most obvious visual indication of when rupture is likely to occur since the gradient peaks and then rapidly flattens before decreasing at the same pulling displacements at which the maximum force is reached.

For the three lower loading rates, the unzipping dihedral energy traces, as with the van der Waals and Coulomb energies, are mostly not functions of the loading rate. By contrast, for the highest loading rate, the maximum dihedral energy is distinctly higher than the maxima for the previous rates. This may partly explain the large jump in the most probable maximum force for the highest loading rate during unzipping, as seen in Figure 5-9. Since the dihedral energy reaches a significantly larger value when unzipping under the highest loading rate, this may indicate that greater twisting forces are being applied to the backbone of ds-PNA. This could mean that, at the highest loading rates, the process of unzipping is more complex than a base pair by base pair decomposition. Instead, the higher dihedral energies indicating more twisting of the helix could indicate a change in the dynamics of unzipping that may explain the jump in the rupture force at the highest rate.

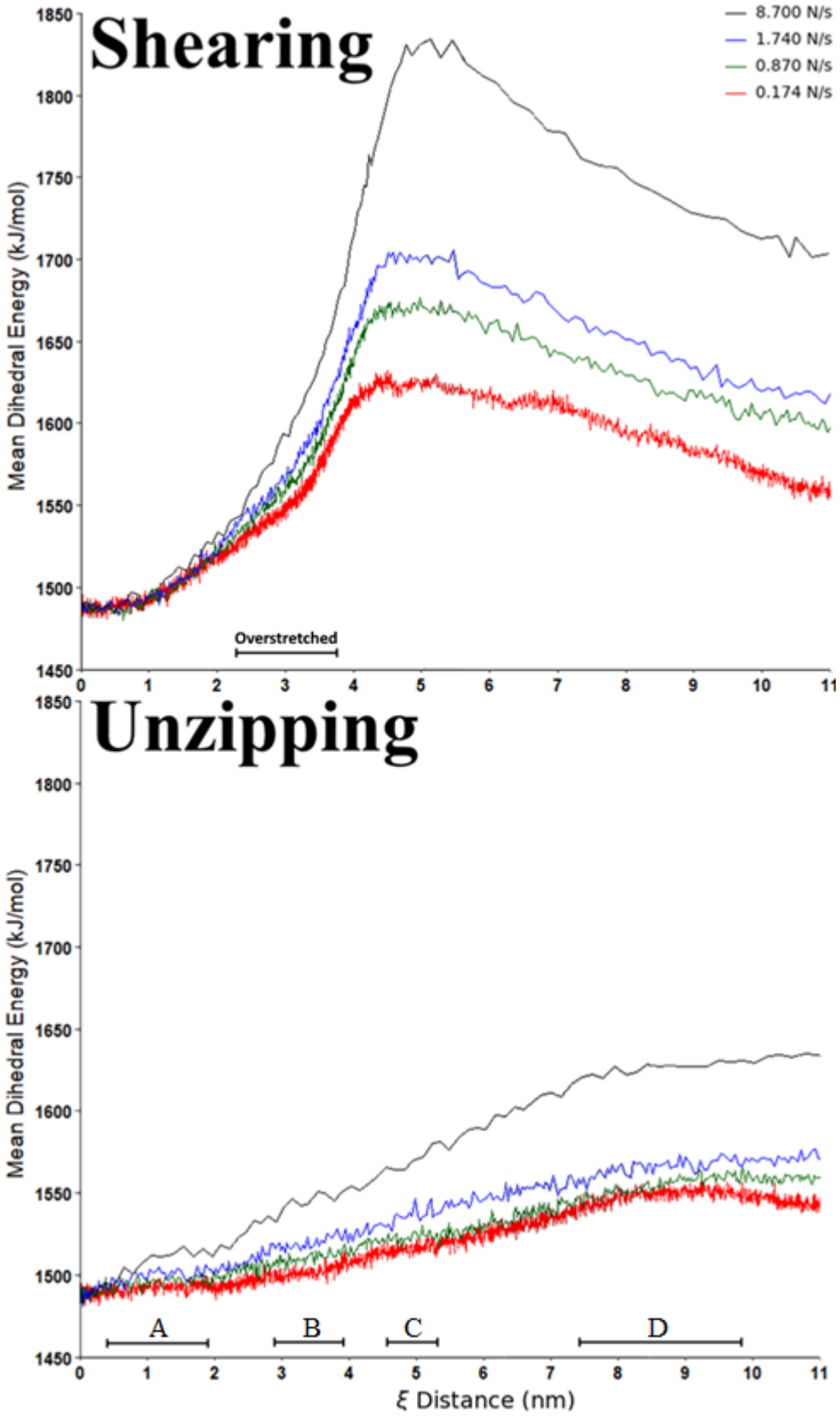


Figure 5-15: Mean ds-PNA dihedral energies for shearing (top) and unzipping (bottom) as a function of reaction coordinate distance. Regions of interest including the overstretching region in the shear trace and the four rupture regions in the unzipping trace are indicated.

In general, these energy traces suggest that not only are shearing and unzipping distinct with regard to their force traces, but that these force traces can be explained in terms of energies. For shearing, an initial van der Waals stabilisation, possibly as a result of strand compression, accommodates for rising forces during unwinding of the double-helix to the overstretched form. The overstretched form, relative to P-form PNA, is characterised by increased backbone torsional and Coulomb energies, but is partly stabilised by increased non-specific forces of attraction, possibly due to compression. Further stretching causes the attractive Coulomb and van der Waals forces to rapidly decrease as base pairs unbind and slide past one another, and dihedral energies rise sharply as tensile force increases in the direction of the now parallel backbones. The final rupture of the ds-PNA occurs immediately after a peak in the dihedral energies, making this peak a useful predictor of when rupture is likely to occur.

For unzipping, by contrast, the approximately linearly increasing energies in all three cases represent the fact that unbinding is a continuous process consisting of multiple rupture events, as opposed to a singular event as in shearing. The exception to this is that, under the highest loading rate, the dihedral energy trace suggests twisting of the backbones, which could explain the large jump in the rupture force distributions at the highest rates since the dynamics of dissociation are different. Additionally, whereas the van der Waals and Coulomb energies have distinct plateaus in their traces that correspond to base pair clusters prior to their rupture, these are not present in the dihedral trace.

5.3.5 Base pairing and stacking landscapes during shearing

For shearing, as opposed to unzipping, force ‘builds’ on ds-PNA before a rupture event occurs. Using AWH, it was possible to produce two-dimensional base pairing and stacking landscapes to demonstrate the effect of shear force on pairing and stacking energies (Figure 5-16). These free energy landscapes were obtained under

equilibrium conditions wherein structures part-way through shearing were re-equilibrated but with their pulling atoms restrained at the intended shearing distance.

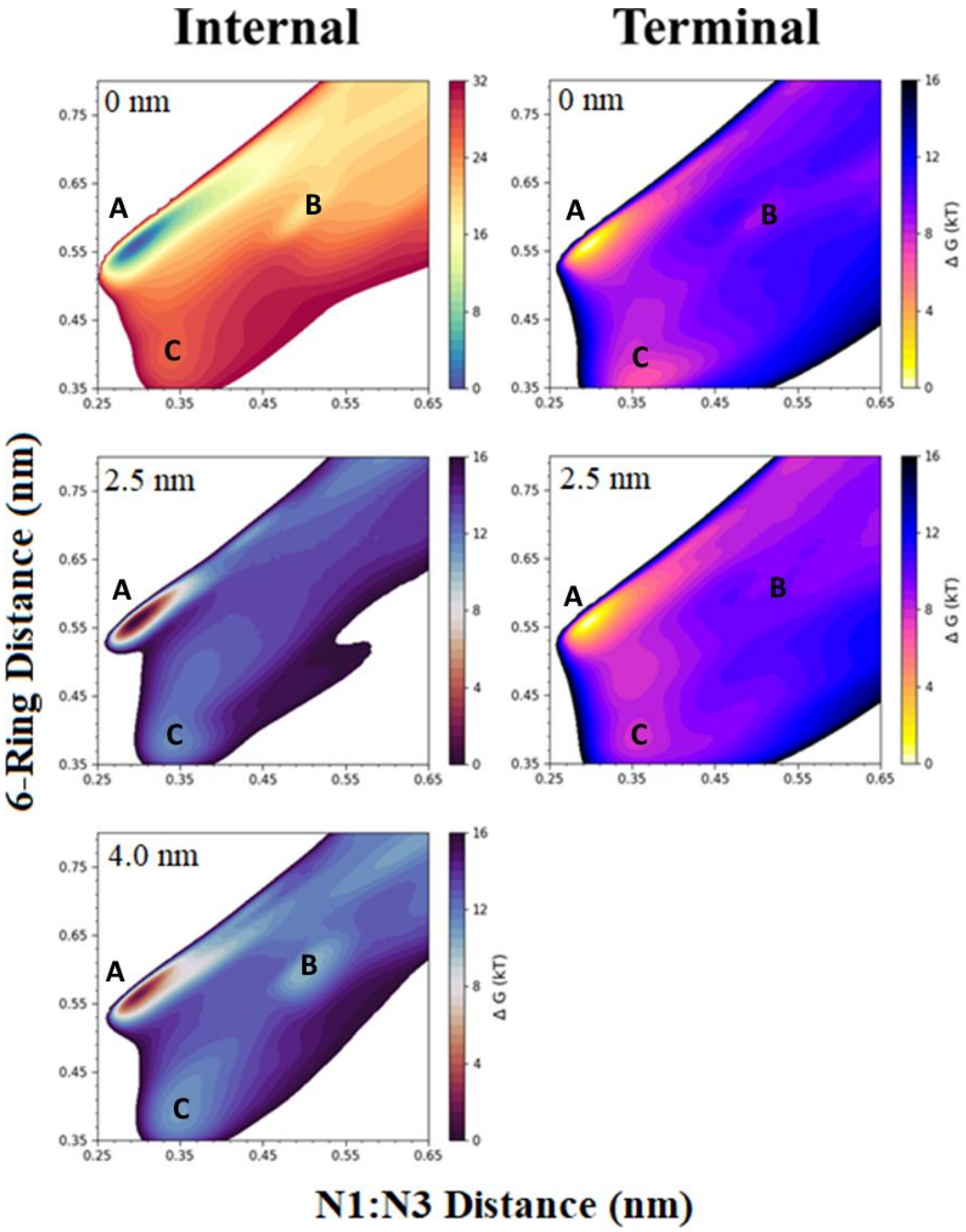


Figure 5-16: Two-dimensional free energy landscapes for internal and terminal base pairs as a function of shearing distances of 0, 2.5 and 4.0 nm from top to bottom. A, B and C refer to the Watson-Crick pairing, open and stacked states respectively. Colour bars indicate the free energy and the 6-ring distance and N1:N3 distance are on the y and x axes respectively.

In Figure 5-16, recolouring was used to indicate the change in scale for internal base pairs at 2.5 and 4.0 nm extensions. Terminal base pair landscapes were unobtainable at 4.0 nm extensions since terminal base pairs were fully ruptured at this point and both distances were outside the coordinate ranges in Figure 5-16.

In P-form ds-PNA, at 0 nm extensions, there are three main populated regions in the terminal base pair free energy landscape:

(A) A Watson-Crick pairing energy minimum at 0.30 nm N1:N3 distance and 0.55 nm 6-ring distance.

(B) An open region at high N1:N3 and 6-ring distances with shallow minima at, for example, a 0.50 nm and 0.60 nm N1:N3 and 6-ring distance respectively. In the open region, one base has slipped outside of the helical axis.

(C) A stacking region at 0.35 nm for both distances separated from the Watson-Crick state by a saddle point. In the stacking region, base pairs have slipped on top of one another within the helical axis.

For the terminal pair at 0 nm extensions, the energy cost of moving from A to C over the saddle point is approximately 8 k_BT. For the internal pair at 0 nm, the stacking state C has an energy cost more than double this and is not a local minimum. This suggests that base pair stacking is less favourable for the internal than terminal base pair. During stacking, two previously paired bases shift out of plane to sit on top of one another. At the terminus, this may disrupt nearby bases less, since there is no other pair on top of the stacked base pair, whereas for internal bases stacking increases the effective height of that base pair within the sequence which may disrupt adjacent pairs through steric effects.

At the 2.5 nm extensions there is a global flattening of the free energy landscape for both terminal and internal base pairs. The open region around B becomes more accessible in both cases but particularly for the terminal base pair. Since the 2.5 nm extension is representative of overstretched ds-PNA, this corroborates structural observations, since snapshots of overstretched ds-PNA demonstrate fraying of the terminal base pairs. For internal base pairs the energy landscape is arguably more similar to a terminal base pair free energy landscape than its own at 0 nm. In this case, the entire free energy landscape is flatter, with the energy cost of entering stacking or open conformations being approximately half relative to the 0 nm case for the internal base pair. This is particularly evident when the original scale is used (Figure 5-17).

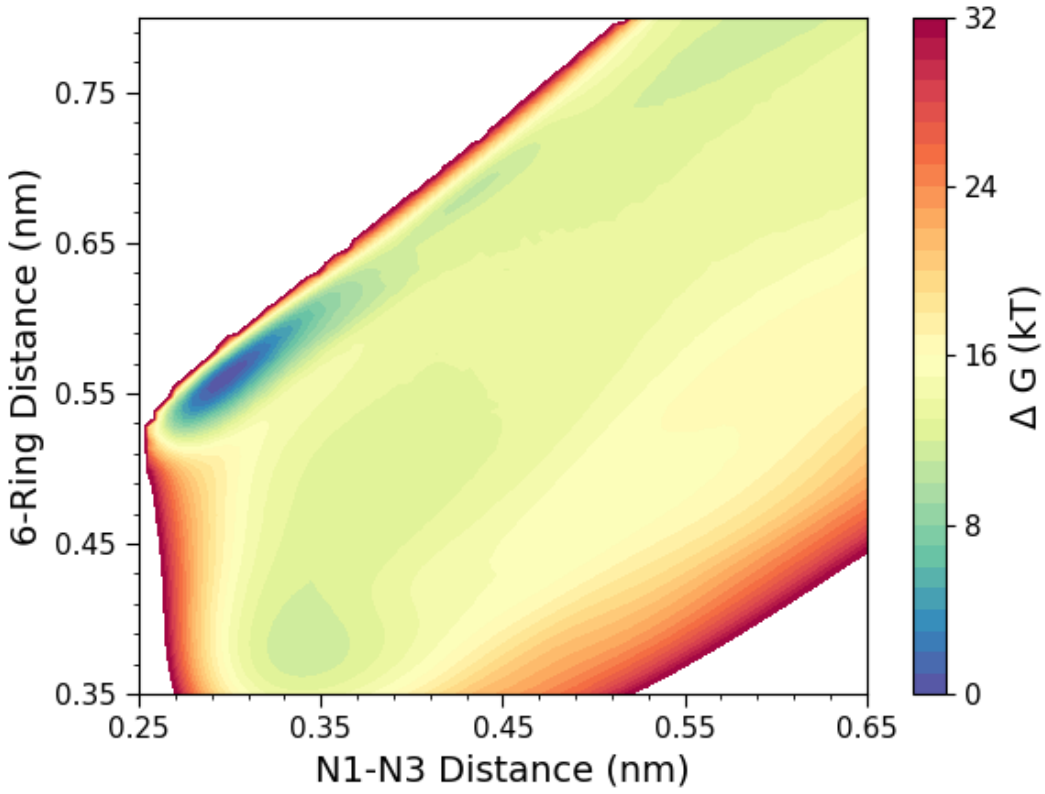


Figure 5-17: Two-dimensional free energy landscapes for an internal base pair in 2.5 nm overstretched ds-PNA.

At an extension of 4 nm, prior to the rupture event, the free energy landscape experiences further flattening. In particular, a metastable minimum within the open state B becomes more energetically favourable. This may be since that, immediately prior to the point of rupture, previously internal base pairs have become effectively terminal due to the relative displacement of the strands. In such a case, fraying may be observed, which could result in the stabilisation of the metastable open state B. In general, the effect of shearing on internal and terminal base pairs is to flatten free energy landscapes such that the open state B is more stable, though for the internal base pairs a more global flattening is observed that also decreases the energy cost of populating the stacking state C.

5.3.6 Binding free energy

The free energy landscape along a perpendicular reaction coordinate between the centres of mass of both strands was determined (Figure 5-18). The binding free energy was determined to be $\Delta G = -23.71$ kcal/mol, taken as the negative of the difference between the initial and final free energy from Figure 5-18. Applying the standard state correction to the binding free energy from Equation 5-3 yielded $\Delta G = -32.15$ kcal/mol. The literature average of the binding free energy of the ds-PNA GTAGATCACT presented in Chapter 4 was -18.99 kcal/mol, meaning the molecular dynamics estimate, using the CHARMM molecular mechanics forcefield, significantly stabilised the helix by a factor of 1.69 relative to experimental data. A similar stabilisation factor of 1.72 when using this forcefield was observed for the binding free energy for the ds-PNA TGTTACGACT in prior research.[26] These factors represent a reasonable quantitative agreement between experimental and computational data, though equally demonstrate a stabilisation that should not be discounted given that it may result in changes to the dynamics of dissociation discussed in this chapter relative to experimental data.

In general, however, since the binding free energy was a reasonable quantity well within the expected order of magnitude, and since these simulations demonstrated

structural transitions, such as overstretching, that are consistent with experimental results for other nucleic acids and which are visible in the force curves, it is arguable that the CHARMM molecular dynamics forcefield provides a reasonably accurate assessment of the shearing and unzipping of the ds-PNA GTAGATCACT.

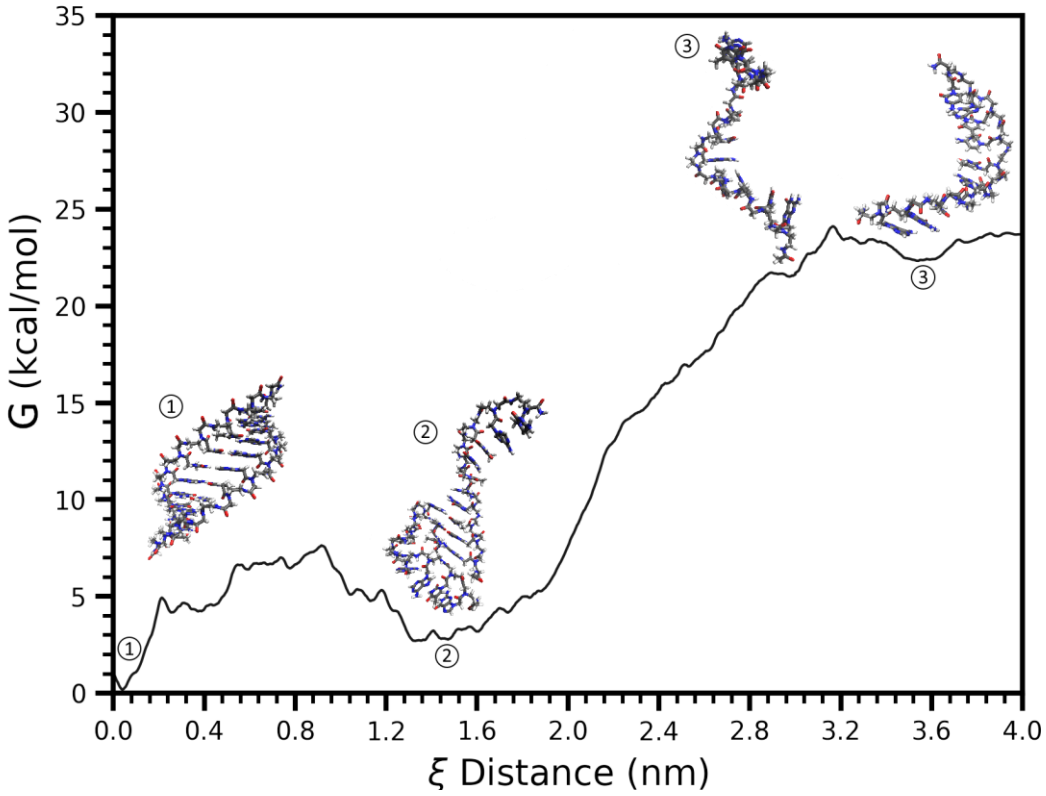


Figure 5-18: Free energy landscape for the unbinding of the ds-PNA GTAGATCACT along a perpendicular coordinate representative of neither shearing nor unzipping. Structures at points (1), (2) and (3) are overlayed on the curve and the regions from which these structures were obtained are indicated.

5.4 Conclusions

These all-atom molecular dynamics simulations provide a structural and energetic understanding of how the application of external force affects factors associated with the stability of ds-PNA. Technical research question (2), “Can molecular dynamics be used to assess the single-molecular properties of a candidate

bioadhesive under external force, and what quantities contribute to this?", is addressed through the following conclusions:

(1) For the unzipping coordinate, the primary contribution to the stability of ds-PNA under external force comes from individual or clusters of base pairs. These rupture events can be depicted in both the mean force and van der Waals or Coulomb energy traces.

(2) For the unzipping coordinate the van der Waals and Coulomb energies between strands increase near-linearly as a function of reaction coordinate distance and are near-independent of loading rate.

(3) For the shearing coordinate, unwinding and extension of the ds-PNA stabilises it under external force.

(4) For the shearing coordinate, a dip in the van der Waals energy at low displacements stabilises the double-helix and opposes the increasing torsional and Coulomb energies.

(5) For the shearing coordinate, most of the resistance to the external force is contributed to by the overstretched state, which was observed in ds-PNA for the first time, since the force-extension gradient is steepest for this region.

(6) Torsional energies play a large role in shearing, since shearing necessarily results in the unwinding and then stretching of the backbone torsional angles.

(7) For ds-PNA, larger maximum forces are observed during shearing than unzipping, indicating this to be the most resistive reaction coordinate.

Whereas under equilibrium conditions, the nearest-neighbour model of Chapter 4 suggests stability can largely be predicted from pairing and stacking energies alone,

stability under external force is more complex since the reaction coordinate affects which factors contribute to stability.

5.5 References

[1] Allen, P.; Khaing, Z.; Schmidt, C.; Ellington, A. 3D Printing with Nucleic Acid Adhesives. *ACS Biomater. Sci. Eng.* **2015**, *1*, 19-26

[2] Lindahl, V.; Villa, A.; Hess, B. Sequence dependency of canonical base pair opening in the DNA double helix. *PLoS Comput. Biol.* **2017**, *13*, e1005463

[3] Jasiński, M.; Feig, M.; Trylska, J. Improved Force Fields for Peptide Nucleic Acids with Optimized Backbone Torsion Parameters. *J. Chem. Theory Comput.* **2018**, *14*, 3603-3620

[4] Van der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A.; Berendsen, H. GROMACS: Fast, Flexible and Free. *J. Comput. Chem.* **2005**, *26*, 1701-1718

[5] Abraham, M.; Alekseenko, A.; Bergh, C.; Blau, C.; Briand, E.; Doijade, M.; Fleischmann, S.; Gapsys, V.; Garg, G.; Gorelov, S.; Gouaillardet, G.; Gray, A.; Irrgang, M.; Jalaypour, F.; Jordan, J.; Junghans, C.; Kanduri, P.; Keller, S.; Kutzner, C.; Lindahl, E. GROMACS 2023 Manual (Version 2023). *Zenodo*. **2023**. https://doi.org/10.5281/zenodo.7588711

[6] Berendsen, H.; Postma, J.; Gunsteren, W.; Hermans, J. Interaction Models for Water in Relation to Protein Hydration. *Intermolecular Forces*. **1981**. *14*. ISBN: 978-90-481-8368-5.

[7] Hocknehy, R.; Goel, S.; Eastwood, J. Quiet High Resolution Computer Models of a Plasma. *J. Comp. Phys.* **1974**. *14.* 148-158.

[8] Berendsen, H.; van Gunsteren, W. Practical algorithms for dynamics simulations. *Molecular-Dynamics Simulation of Statistical-Mechanical Systems*. **1986**. 43-65.

[9] Bussi, G.; Donadio, D.; Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **2006**. *126*. 01401.

[10] Hess, B.; Bekker, H.; Berendsen, H.; Fraaije, J. LINCS: A Linear Constraint Solver for Molecular Simulations. *J. Comput. Chem.* **1997**. *18*. 1463-1472.

[11] Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald. An Nlog(N) method for Ewald sums in large systems. *J. Chem. Phys.* **1993**. *98*. 10089-10092.

[12] Essmann, U.; Perera, L.; Berkowitz, M.; Darden, T.; Lee, H.; Pedersen L. A smooth particle mesh ewald potential. *J. Chem. Phys.* **1995**. *103*. 8577-8592.

[13] Frigo, M.; Johnson, S. The Design and Implementation of FFTW3. *Proc. IEEE.* **2005**. *93*. 216-231.

[14] Goga, N.; Rzepiela, A.; de Vries, A.; Marrink, S.; Berendsen, H. Efficient Algorithms for Langevin and DPD Dynamics. *J. Chem. Theory Comput.* **2012**. *8*. 3637-3649.

[15] Berendsen, H.; Postma, J.; DiNola, A.; Haak, J. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **1984**. *81*. 3684-3690.

[16] Parrinello, M.; Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.* **1981**. *52*. 7182-7190.

[17] Lindahl, V.; Villa, A.; Hess, B. Sequence dependency of canonical base pair opening in the DNA double helix. *PLoS Comput. Biol.* **2017**, *13*, e1005463

[18] Cocco, S.; Monasson, R.; Marko, J. Force and kinetic barriers to unzipping of the DNA double helix. *Proc. Natl. Acad. Sci.* **2001**, *98*, 8608-8613

[19] De Gennes, P. Maximum pull out force on DNA hybrids. *Comptes Rendus de l'Académie des Sciences – Series IV – Physics* **2001**, *2*, 1505-1508

[20] Danilowicz, C.; Hatch, K.; Conover, A.; Ducas, T.; Gunaratne, R.; Coljee, V.; Prentiss, M. Study of force induced melting of dsDNA as a function of length and conformation. *J. Phys.: Condens. Matter* **2010**, *22*, 414106

[21] Sheridan, S.; Gräter, F.; Daday, C. How Fast Is Too Fast in Force-Probe Molecular Dynamics Simulations? *J. Phys. Chem. B.* **2019**, *123*, 3658-3664

[22] Evans, E.; Ritchie, K. Dynamic Strength of Molecular Adhesion Bonds. *Biophysical Journal* **1997**, *72*, 1541-1555

[23] Hane, F.; Attwood, S.; Neonenko, Z. Comparison of three competing dynamic force spectroscopy models to study binding forces of amyloid-β (1-42). *Soft Matter* **2014**, *10*, 1924-1930

[24] Morfil, J.; Kühner, F.; Blank, K.; Lugmaier, R.; Sedlmair, J.; Gaub, H. B-S Transition in Short Oligonucleotides. *Biophysical Journal* **2007**, *93*, 2400-2409

[25] Mishra, R.; Mishra, G.; Li, M.; Kumar, S. Effect of shear force on the separation of double-stranded DNA. *Phys. Rev. E.* **2011**, *84*, 032903

[26] Jasiński, M.; Miszkiewicz, J.; Feig, M.; Trylska, J. Thermal Stability of Peptide Nucleic Acid Complexes. *J. Phys. Chem. B* **2019**, *123*, 8168−8177

Chapter 6: Coarse-grained model for ds-PNA

6.1 Introduction

The purpose of a Martini model is to enable higher simulation times to be conducted on the same molecular system for the same duration of real time. This is achieved firstly by clustering atoms to reduce the degrees of freedom of the system. Secondly, higher simulation timesteps can be used since Martini was parametrised for simulation timesteps between 20 and 40 fs.[1] These timesteps are partly accessible as a result of the larger masses of the beads, with recent literature suggesting that using the smallest type of Martini bead can limit timesteps to 25 fs.[2]

Martini beads fall into the tiny (T), small (S) or normal size categories. Martini beads also have 4 main interaction categories which are themselves split into subcategories. These main categories are charged (Q), polar (P), nonpolar (N) and apolar (C). A bead's 'type' is given by both its size and its category. The categories are used to form an interaction matrix (Table 6-1)[1] with levels of nonbonded interaction between all subcategories based on interactions between the atomically defined chemical groups the subcategories were developed from. These interactions are modelled using Lennard-Jones potentials. For the charged (Q) category there are additional Coulomb interactions.

Table 6-1: Martini interaction matrix. Roman numerals are the depths of the wells of the Lennard-Jones potentials for the different bead types. O = 5.6 kJ/mol, I = 5.0 kJ/mol, II = 4.5 kJ/mol, III = 4.0 kJ/mol, IV = 3.5 kJ/mol, V = 3.1 kJ/mol, VI = 2.7 kJ/mol, VII = 2.0 kJ/mol, IX = 2.0 kJ/mol. Reprinted with permission from Marrink *et al.*[1] Copyright (2007) American Chemical Society.

		Q				P					N				C				
	sub	da	d	a	0	5	4	3	2	1	da	d	a	0	5	4	3	2	1
Q	da	O	O	O	II	O	O	O	I	I	I	I	I	IV	V	VI	VII	IX	IX
	d	O	I	O	II	O	O	O	I	I	I	III	I	IV	V	VI	VII	IX	IX
	a	O	O	I	II	O	O	O	I	I	I	I	III	IV	V	VI	VII	IX	IX
	0	II	II	II	IV	I	O	I	II	III	III	III	III	IV	V	VI	VII	IX	IX
P	5	O	O	O	I	O	O	O	O	O	I	I	I	IV	V	VI	VI	VII	VIII
	4	O	O	O	O	O	I	I	II	II	III	III	III	IV	V	VI	VI	VII	VIII
	3	O	O	O	I	O	I	I	II	II	II	II	II	IV	IV	V	V	VI	VII
	2	I	I	I	II	O	II	II	II	II	II	II	II	III	IV	IV	V	VI	VII
	1	I	I	I	III	O	II	II	II	II	II	II	II	III	IV	IV	IV	V	VI
N	da	I	I	I	III	I	III	II	II	II	II	II	II	IV	IV	V	VI	VI	VI
	d	I	III	I	III	I	III	II	II	II	II	III	II	IV	IV	V	VI	VI	VI
	a	I	I	III	III	I	III	II	II	II	II	II	III	IV	IV	V	VI	VI	VI
	0	IV	IV	IV	IV	IV	IV	IV	III	III	IV	IV	IV	IV	IV	IV	IV	V	VI
C	5	V	V	V	V	V	V	IV	IV	IV	IV	IV	IV	IV	IV	IV	IV	V	V
	4	VI	VI	VI	VI	VI	VI	V	IV	IV	V	V	V	IV	IV	IV	IV	V	V
	3	VII	VII	VII	VII	VI	VI	V	V	IV	VI	VI	VI	IV	IV	IV	IV	IV	IV
	2	IX	IX	IX	IX	VII	VII	VI	VI	V	VI	VI	VI	V	V	V	IV	IV	IV
	1	IX	IX	IX	IX	VIII	VIII	VII	VII	VI	VI	VI	VI	VI	V	V	IV	IV	IV

All interaction levels except for IX have a closest distance of approach σ in the Lennard-Jones potential of 0.47 nm. IX has a σ of 0.62 nm.[1] This parameter effectively determines how closely beads will associate and so a higher distance means beads will repel one another from further away. The lowest interaction level IX is reserved for interactions between charged beads (i.e., ions) and the least polar solvents (e.g., long chain hydrocarbons). At the other end, level O is reserved for solid compounds, such as ionic lattices, or to represent hydration shells around charged beads. In peptide nucleic acids at neutral pH there are no explicit charges, and so Coulomb potentials are not present in Martini simulations of these molecules.

In addition to the nonbonded potentials, Martini models use bonded potentials. Bonds and angles between beads are represented by harmonic potentials and so covalent bond rupture is not modelled since an infinite force is encountered at an infinite displacement. Torsional angles are generally represented by trigonometric proper dihedral functions, with multiple energy minima determined by the multiplicity (for example, a multiplicity of 2 has 2 minima), or improper harmonic functions with a single energy minimum.[1] In addition to these general forms, Gromacs makes available other forms for the bonded interactions such as the restricted bending angle potential (Equation 6-1).[3,4] This potential is useful when the angle between three beads, determined from atomic distributions of those clusters, is near 180°,[3] since a singularity that causes simulation termination arises in the general form of the torsional potential when three beads become ‘collinear’. The restricted potentials prevent this collinearity from occurring and are used in the present work.

[Eq. 6–1] $$V_{RB} = \frac{1}{2} k \frac{(\cos\theta - \cos\theta_0)^2}{\sin^2\theta}$$

In Equation 6-1, a cosine-based angle potential (as opposed to harmonic) with an energy minimum at θ_0 and force constant k, is divided by $\sin^2\theta$. The denominator produces a very repulsive region near to 180°.

In this chapter, a Martini model for PNA is developed with the hypothesis that such a model would allow lower loading rates in pulling simulations to be reached since these lower loading rates require higher simulation times with existing hardware. This Martini model for PNA is in part based on the Martini model for DNA. There have been previous coarse-grained models for nucleic acids but the majority of them are unable to simulate interactions with other biomolecules and therefore lack versatility.[5] In addition, features such as sequence specificity, crucial for the function of nucleic acids as genetic carriers of information, can be lost.[6] The Martini forcefield is a particularly advantageous mode of developing a coarse-grained model since it uses a universal framework of beads. Each bead's nonbonded interactions with others are already determined from the interaction matrices or Coulomb potentials, and so newly parameterised molecules need only to have their bead types selected and bonded parameters determined. These new additions will then belong to a growing set of Martini models[7,8,9] with compatible topologies that can interact with one another. For PNA this is particularly important since PNA:DNA heteroduplexes are the major target of new PNA technologies.[10,11,12] Though there have been cases where new bead types have been developed for particular molecule classes,[13] the standard types are used in the present work since native compatibility with other nucleic acids was determined to be more important given the aforementioned significance of PNA:DNA heteroduplexes.

In the Martini models of DNA[5] and RNA,[14] the nucleobase beads are of the tiny (T) size, the effect of which is to reduce the σ to 0.32 nm in the Lennard-Jones potential between other tiny (T) beads. The same goes for small (S) beads, though to a lesser extent with $\sigma = 0.43$ nm. Tiny (T) beads interact with small (S) and regular beads using $\sigma = 0.43$ nm. Small (S) beads interact with regular beads using the standard $\sigma = 0.47$ nm.[1] The purpose of the tiny (T) beads for the nucleobases is to allow the

ring systems to pack into their flat, coplanar arrangement, since the rise of 0.332[15] nm in B-form DNA is smaller than the closest distance of approach in normal-sized beads. In addition to this, tiny (T) beads typically model only 2 atoms, as opposed to the usual 3.[5,14] Since these nucleobase bead arrangements are well described in DNA[5] and RNA,[14] and the chemical structure of the nucleobases in PNA is identical, their bead sizes and types are used for PNA as well.

For DNA, the authors described numerical instabilities arising for ds-nucleic acids, which were simulated with 10 fs timesteps, without the use of an elastic network.[5] Consequently, a similar elastic network is used in the present work. In addition, a Morse potential model of hydrogen bonding,[16] similar to the Lennard-Jones potential and natively integrated into Gromacs,[3] is used in the present work for the development of the Martini PNA model. This is since directional hydrogen bonding is lacking in current Martini nucleic acids.[5,14] The Martini PNA model is hence similar to the Martini DNA or RNA models in that it uses the same philosophy of tiny (T) bead types for the nucleobase rings but differs in its backbone parameterisation and the use of additional Morse potentials to capture hydrogen bonding. The development, validation, and application of this model for the modelling of the force-loading curve is discussed in this chapter.

6.2 Methods

6.2.1 Equilibration and recommended simulation parameters

Energy minimisations and equilibrations were conducted for all-atom reference simulations (CHARMM) in the same manner as described for previous all-atom simulations. That is, minimisations were conducted using steepest descent algorithms with 50,000 steps in both vacuum and in solvent, after which systems were gradually warmed to the target standard temperature of 298 K under position restraints. These restraints were gradually released over 10 ns and the systems were equilibrated under constant temperature, and then constant temperature and pressure, conditions for 50 ns each. Newton’s equations of motion were integrated

using a stochastic integrator[17] and van der Waals and Coulomb interactions were handled using a switched cut-off scheme from 10 to 12 Å and a PME[18,19] to 10 Å with long-ranged components handled by FFTW[20] respectively.

For the coarse-grained simulations, parameters were adjusted since specific values for parameters like the cut-off distances for van der Waals and Coulomb interactions are needed for the correct implementation of the Martini forcefield.[1] Minimisation was conducted using a steepest descent minimisation in vacuum and in solvent over 50,000 steps in the same manner as with the all-atom simulations. In the original Martini paper, different parameters to what are used in the present work are used since more recent research demonstrated improvements with a new set of parameters.[21] Namely, a standard leap-frog integrator, [22,23] as opposed to stochastic, is used and temperature and pressure are coupled using velocity rescaling and Parrinello-Rahman pressure coupling respectively. The main difference between the Martini molecular dynamics parameters and the all-atom parameters are in the calculation of van der Waals and Coulomb interactions. Van der Waals interactions were switched to zero at 11 Å using a potential modifier so that the potential smoothly approaches zero at this distance without jumping, as occurs without a potential shift at such a low cut-off distance. Coulomb interactions were treated using a reaction field which smoothly switches to zero at 11 Å, again removing artifacts that arise from jumping to zero at a short cut-off distance.[24] These parameters were carried forth into the production simulations.

During the parameterisation of Martini PNA, different simulations including isolated backbone monomers, isolated base pairs in solution (e.g., A:T), short ss-PNA sequences (e.g., TTTT), and finally whole sequences of GTAGATCACT: CATCTAGTGA, were conducted. Dodecahedral simulation boxes were used and minimum distances between solute atoms or beads and box edges of 1.5 nm were enforced. During SMD simulations, a cuboid box was used with an extended axis in the pulling direction.

6.2.2 Defining the beads

Since the nucleobase bead types were already determined from the DNA and RNA models (Table 6-2), only the bead types for the backbone needed to be determined. For the nucleobases, 2 or 3 atoms are clustered into individual tiny (T) beads, with 4 beads per purine and 3 beads per pyrimidine. For the PNA model, these beads are named, where the name is used for clarity and is separate from the type, from NU1 to NU4 for purines and from NU1 to NU3 for pyrimidines where the NU is shorthand for ‘nucleobase’. The name of the nucleobase is prefixed in lowercase before the bead name, for example ‘g’ for Guanine. Backbone beads were selected such that the nucleobase was connected to the backbone through a central bead, allowing the planarity of the ring systems to be enforced using torsional and angle potentials above and below the ring. Backbone beads are labelled BB1, BB2 and BB3 from the N- to the C- terminus, where BB is short for ‘backbone’ (Figure 6-1).

Table 6-2: Bead names and types for the nucleobases Guanine (g), Adenine (a), Cytosine (c) and Thymine (t). Atom numbers are the ring positions with side-chain atoms indicated by asterisks.

Name	**Type (Table 6-1)**	**Atoms (Fig. 1-8)**
gNU1	TN0	9, 4
gNU2	TP1	3, 2, N2*
gNU3	TP2	1, 6, O6*
gNU4	TNa	5, 7, 8
aNU1	TN0	9, 4
aNU2	TN0	3, 2
aNU3	TP1	1, 6, N6*
aNU4	TNa	5, 7, 8
cNU1	TN0	1, 6
cNU2	TP2	3, 2, O2*
cNU3	TP1	5, 4, N4*
tNU1	TN0	1, 6
tNU2	TP2	3, 2, O2*
tNU3	TNa	5, C5*, 4, O4*

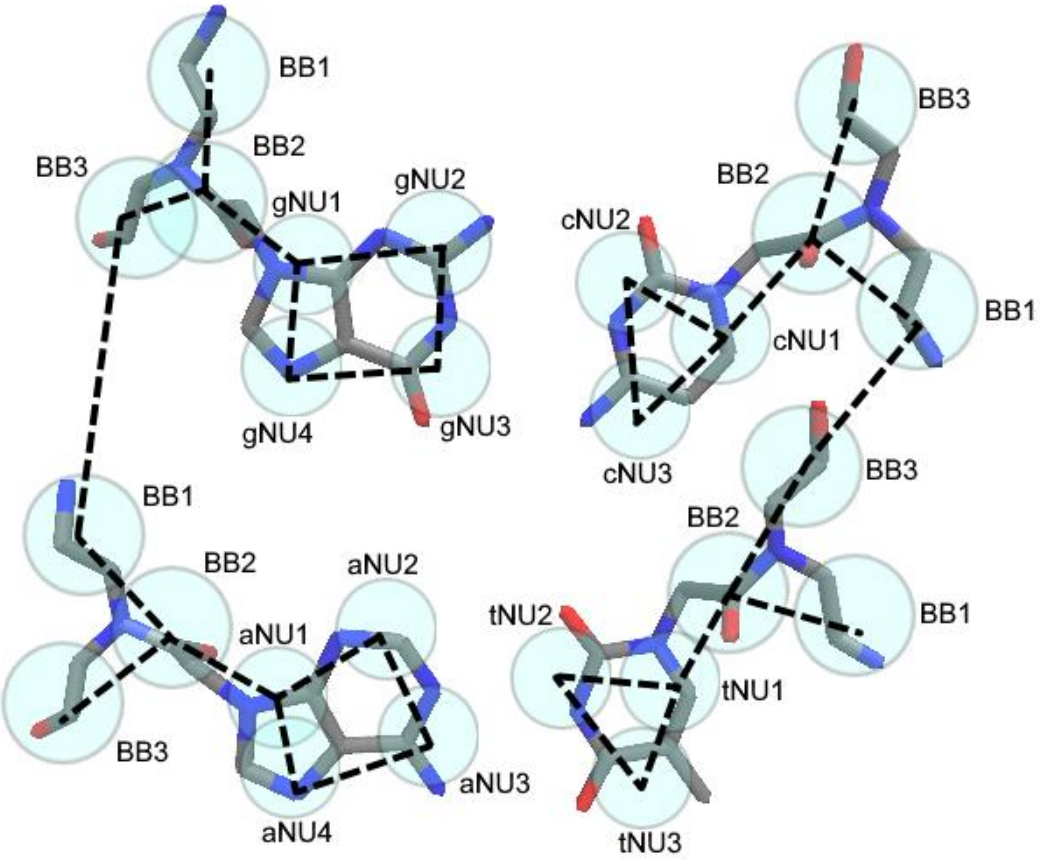


Figure 6-1: Bead assignments for the Martini model of PNA. Translucent blue circles represent beads connected by dotted lines.

BB2 is a cluster of four atoms including the tertiary, central nitrogen atom of the backbone and the complete carbonyl methylene linker. BB1 and BB3 by contrast form each ‘half’ of the amide bond. Since capping was used in the all-atom simulations, end-effects are limited, and the atom clusters that BB1 or BB3 represent can be considered the same regardless of position so long as it is acknowledged that the Martini PNA model is explicitly capped at the termini. Under these conditions, whilst BB2 is chemically similar to dimethyl ethanamide, BB1 and BB3 are always chemically similar to ethyl amine and acetaldehyde respectively.

To determine BB bead types, the partitioning free energies for backbone monomers were calculated between pure water and octanol, and pure water and

dichloromethane, for both the coarse-grained and all-atom cases. These partitioning free energies provide a metric of relative solubilities in solvents with different polarities, and matching the partitioning free energies of atomic simulations with Martini simulations is the primary method by which bead types are selected since these relative solubilities are in turn a metric of the solute's polarity.[5,14] Production simulations were conducted on backbone monomers that had been solvated with either water, dichloromethane or octanol. The all-atom solvent topologies were those provided by the CHARMM36m forcefield,[25] whereas solvent topologies from the Martini forcefield[21] were used in the coarse-grained simulations. The partitioning free energies are obtained for both, and bead types adjusted until the free energies obtained via Martini most closely match those obtained via CHARMM.

To obtain partitioning free energies, production simulations of backbone monomers in the different solvents were conducted. During these production simulations, the monomers were decoupled from the solvent in discrete steps, and the free energy of this de-solvation process obtained by the Bennet acceptance ratio (BAR)[26], which is implemented in Gromacs using gmx_bar.[3] That is, Equation 2-12 was applied over discrete alchemical steps $0 \leq \lambda \leq 1$. State A, at $\lambda = 0$ was defined such that all interactions between the backbone and solvent were 'switched on' and state B, at $\lambda = 1$ was defined such that all interactions with the solvent were switched off, which corresponds to the vacuum state. From this, it is possible to obtain the partition free energy between the two solvents through a thermodynamic cycle (Equation 6-2).[5]

[Eq. 6–2] $$\Delta G_{s_1 \to s_2} = \Delta G_{s_1 \to \emptyset} - \Delta G_{s_2 \to \emptyset}$$

Where $\Delta G_{s_1 \to s_2}$ is the partitioning free energy between solvent 1 s_1 and solvent 2 s_2 and $\emptyset$ is the vacuum state. The exact number of discrete steps in decoupling as used for prior studies on DNA were replicated in the present work.[5] For the Martini

backbones, since there are no charges, the equilibrated and solvated monomer's interactions with the solvent were decoupled over 11 evenly split simulations wherein the van der Waals interactions between the monomer and solvent were gradually switched off. Each simulation was 100 ns long and 20 fs timesteps were used. Gromacs was used to implement the soft-core parameters sc_alpha = 0.5 and sc_power = 1, as was done for DNA.[5] These soft-core potentials are necessary since, when $\lambda \approx 0$, corresponding to the near-vacuum state, atoms can approach very closely, resulting in large fluctuations in potential energy. Soft-core potentials solve this by preventing the atoms from approaching too closely, hence preventing singularities from superposition from occurring.[3, 27]

For the all-atom simulations, since there are charged interactions, Coulomb interactions must also be decoupled from the solvent, and this was achieved over 6 evenly spaced 20 ns simulations. Coulomb interactions are decoupled first since oppositely-charged atoms do not repel one another through the Coulomb potential, resulting in atom superposition without the van der Waals force. The van der Waals interactions were then decoupled over 11 evenly spaced 20 ns steps, with both stages using 2 fs timesteps and hence the total simulation times being accommodatingly lower. Soft-core potentials with the same parameters as with the Martini backbones were again used.

Partitioning free energies were obtained from three replicates of these decoupling processes in different solvents for both all-atom and Martini backbone monomers. Martini bead types were changed in an iterative process until the partitioning free energies most accurately reflected the all-atom partitioning free energies. Initial guesses on the backbone types were made based on their chemical similarity to the reference chemical structures against which the Martini standard bead types were defined.[1]

6.2.3 Fitting distributions

Whereas non-bonded interactions, given by the bead types, are determined by partitioning free energies, bonded interactions are determined by matching the distributions of bonds, angles and dihedrals from coarse-grained with all-atom simulations. 10 short ss-PNA sequences (Appendix A.3) were selected based on the protocol outlined by the prior literature,[5] and 50 ns production simulations were carried out on all-atom representations of these sequences using the CHARMM36m[25,28] forcefield. The molecular dynamics parameters, outlined in Chapter 6.2.1, namely stochastic integration and switched cut-off and PME schemes for van der Waals and Coulomb interactions respectively, were used for these all-atom simulations and three replicates were obtained for each ss-PNA. The centres of geometry of the clusters of atoms representing each bead were tracked over these runs and the distances, angles and dihedral angles between different clusters were exported as reference distributions.

These reference distributions were used to set initial distances, angles, force constants and multiplicities for Martini PNA. 50 ns production runs with these initial guesses were then conducted using the Martini molecular dynamics parameters outlined in Chapter 6.2.1, and the distributions were compared against the reference distributions. The bond, angle and dihedral parameters were iteratively refined until the all-atom distributions were well-reproduced by the coarse-grained simulations, which was determined by visual inspection as was done in the prior literature.[5] Since this process was done by manual refinement, it is not necessarily the case that the final distributions were the distributions which maximised the overlap between the all-atom and coarse-grained data, and user bias is present in the selection. However, the same general process of manual refinement was used for all distributions to ensure a cohesive methodology was undertaken. Parameters were ‘walked up’ towards increasing overlap between distributions until overshooting, after which the previous parameters for a specific bond, angle or dihedral were selected as the final values for bonded terms in Martini PNA.

6.2.4 Accelerated weight histograms for Morse potentials

The lack of directional hydrogen bonds in Martini has presented problems for the development of accurate nucleic acid secondary structures in the past. To solve this, Uusitalo *et al*. specifically tuned the nonbonded interactions between the beads involved in base pairing in Martini DNA from the values in Table 6-1, namely by increasing the interaction strength between complementary beads and decreasing the strength between non-complementary beads.[5] Free energy landscapes along base pairing and stacking coordinates indicated that these parameters enabled the stacking coordinate's shape to be well-approximated but its stability to be overestimated. Crucially, however, their model of Martini DNA underestimates the stability of base pairing, assumes a much broader minimum, and places this minimum at closer base-pairing distances. The authors argue that since Martini potentials are spherically symmetric, these broad minima are necessary to better approximate stacking, but given that the two-dimensional free energy landscapes for ds-PNA base pairs in Figure 5-14 demonstrate that the Watson-Crick bound state makes the largest contribution to base pair stability, it was decided that improving the accuracy of Watson-Crick pairing at the expense of stacking was more appropriate for ds-PNA.

In the present work, Morse potentials were introduced between the bases in the complementary base pairs of ds-PNA to address these issues. Unlike the harmonic potentials used to define most covalent bonds in molecular dynamics simulations, the Morse potential is anharmonic and, at infinite displacements, has zero force. In that manner, the Morse potential is similar to the Lennard-Jones potential and is typically applied to represent covalent bond stretching.[3] Compared to the Lennard-Jones potential, the Morse potential overestimates the attractive part of the potential curve and decays faster towards zero, with the two being identical at equilibrium.[29] Using Morse potentials for the hydrogen bonding interaction has been described in the literature in the past[30] though was found to be unnecessary in the all-atom case since hydrogen bonds were already well-approximated by the van der Waals and Coulomb potentials.[31] Since hydrogen bonds are not well approximated by these

potentials in Martini,[5] practically by the very definition of the forcefield in that it lacks hydrogen atoms,[1] this is not the case in the present work.

The Morse potential energy between two atoms i and j is determined by its well depth D, in $\mathrm{kJ/mol}$, equilibrium distance r_0 and steepness β, in $\mathrm{nm^{-1}}$ (Equation 6-3).

[Eq. 6–3] $$V_{ij} = D\left(1 - e^{-\beta(r-r_0)}\right)^2$$

Well depth, equilibrium distance and steepness parameters for the Morse potentials used in Martini PNA base pairing were optimised by matching the base pairing and stacking free energy landscapes of all-atom reference simulations of isolated A:T and G:C base pairs. A virtual site was introduced between the beads NU2 and NU3 to represent base pairing distance. A virtual site is a point in space inheriting its positional data from other atoms that does not interact with any other atoms unless explicit pair potentials are introduced.[3] This virtual site was positioned along the bond connecting NU2 and NU3 and was equidistant of them.

50 ns long all-atom reference simulations were conducted on pairs using these virtual sites. AWH with a $100\ \mathrm{kj\ mol^{-1}}$ cut-off was used to map the free energy as a function of the distance between the two virtual sites. If additional restraints are not used to force rings to remain coplanar, they are free to rotate in space and the reaction coordinate reflects a mostly stacking coordinate.[5,32] Consequently, to capture the pairing coordinate as well, the base pairs were aligned such that the pairing coordinate, defined as the line connecting the two virtual sites of coplanar rings, was aligned with the z axis. Movement along this z axis was unrestrained, but position restraints with force constants of $418.4\ \mathrm{kcal/mol}$ were applied to movement along the x and y axes. For both pairing (restrained) and stacking

(unrestrained) coordinates, three 50 ns replicates were used to obtain reference CHARMM free energy landscapes for both A:T and G:C base pairs.

These reference free energy landscapes were compared to those obtained via Martini. In the Martini simulations, the free energy landscapes of pairing and stacking coordinates were obtained also obtained using AWH, though using the Martini molecular dynamics parameters of Chapter 6.2.1. 8 different simulations, each consisting of 3 replicates, were conducted. Firstly, Martini A:T and G:C base pairing free energy landscapes were obtained using the restraints as described in the previous paragraph, and then Martini A:T and G:C base stacking free energy landscapes were obtained for the unrestrained case. The remaining 4 simulations then repeated these but with an additional Morse potential, to represent hydrogen bonding, introduced between the virtual sites (Figure 6-2). Initial guesses for the potential were based on the all-atom reference free energy landscapes and were then iteratively optimised. CHARMM reference curves had three major regions: the bound state, the transition region, and the unbound state. The Martini Morse potentials were optimised such that the energy of all three of these states was as accurately modelled as possible. If it was not possible to reproduce all three states accurately, an appropriate compromise such as an overestimation of the transition state and an underestimation of the unbound state, would be selected.

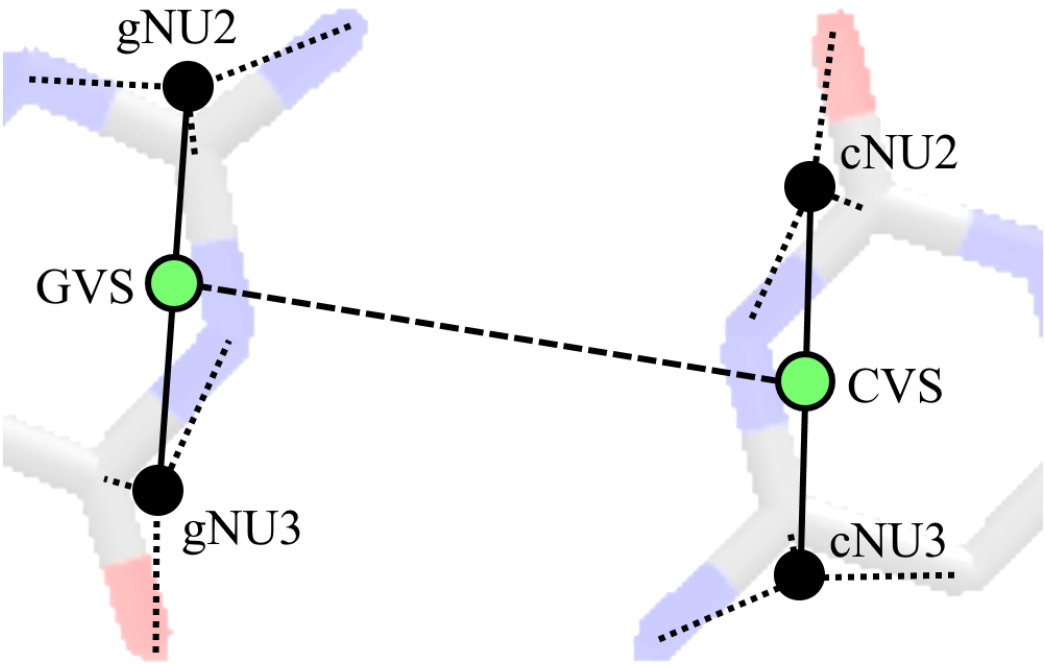


Figure 6-2: Virtual site (VS) in green introduced between guanine (G) and cytosine (C). Black dots are Martini beads and green dots are the virtual sites. The dashed line shows the Morse potential, representing hydrogen bonding, between the sites.

During the process of force-induced rupture, it is possible for strands to become displaced relative to one another, meaning complementary base pairs could arise between bases which are distant in the initial structure but which are brought into proximity by directional pulling (Figure 6-3). To account for this, during later external force studies of ds-PNA, Morse potentials between distant and complementary bases in the pulling direction are explicitly defined. That being said, the results of all-atom simulations on the shearing of ds-PNA indicated that it is not a process of gradual relative strand displacement but of stretching and then rapid displacement. In particular, Figure 5-16 shows that the Watson-Crick bound state for internal base pairs remains the most stable state at up to 4 nm displacements, demonstrating that, during the process of overstretching, the most highly populated conformations are those wherein internal base pairs remain in their initial complementary configurations. It is consequently the case that the majority of the dynamics of shearing may be modelled without explicitly defining hydrogen bonds between distant but complementary bases, with these hydrogen bonds likely mostly

occurring temporarily as strands slide past one another near the critical rupture distance. Regardless, these distant Morse potentials were still defined since the additional computational load is limited for the short sequences tested in the present work (for GTAGATCACT, only 8 additional Morse potentials need to be defined to account for possible hydrogen bonding from bases displaced in the pulling direction).

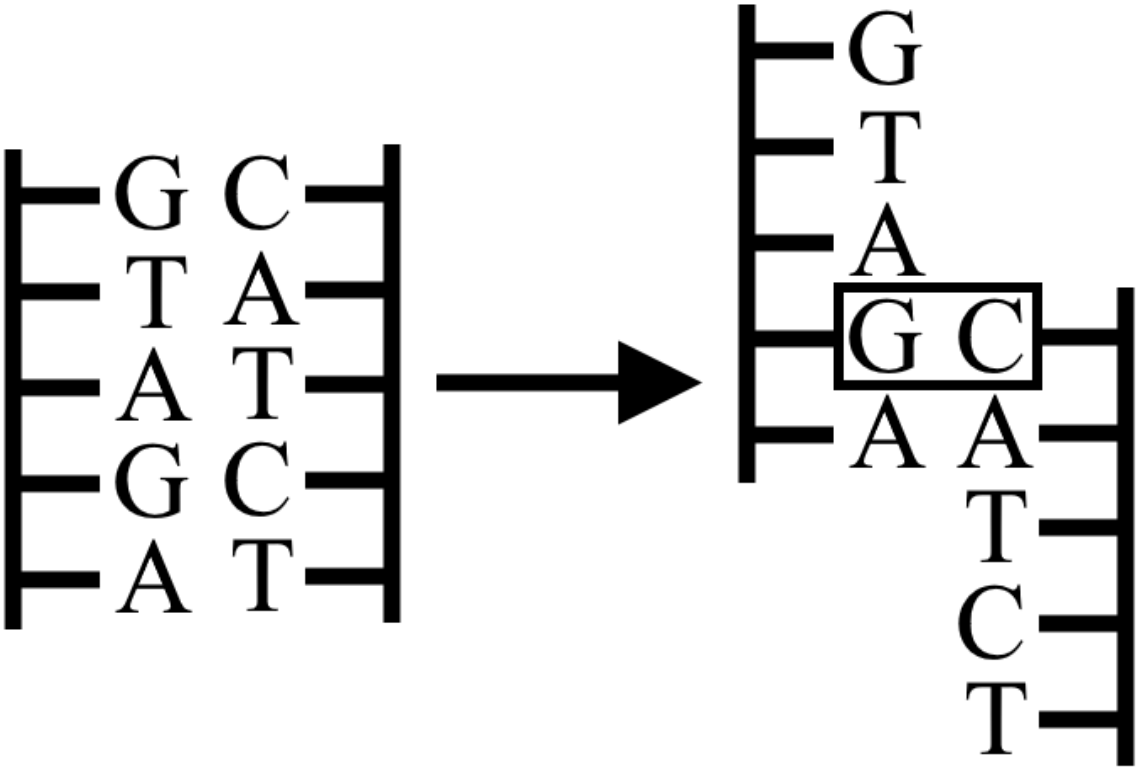


Figure 6-3: Schematic demonstrating the necessity of defining explicit Morse potentials between distant bases which are not in association in the bound state. Complementary base pair in displaced strand indicated by rectangle.

6.2.5 Defining the elastic network

Uusitalo *et al*. describe the necessity of an elastic network to replicate the structure of ds-DNA from the Martini model even when all-atom bond, angle and dihedral distributions are well-replicated by Martini distributions. In their elastic network, harmonic bonds are introduced between beads and other, distant beads in the structure. The equilibrium distance of these harmonic bonds was determined by their equilibrium distances from all-atom simulations.[5]

In the present work, a similar elastic network is used. The purpose of the Martini ds-PNA model developed in the present work is to study bond rupture events, and the study of these rupture events is not possible when an elastic network is used. So, the current work presents the Martini model of ds-PNA as a choice between accurate structure, in using an elastic network, and the possibility of external force studies, in not using it. It would be theoretically possible to use an anharmonic elastic network to enable both, though in the inclusion of an anharmonic network of additional potentials the magnitude of rupture forces would be vastly overestimated, as will be demonstrated by rupture forces obtained in the absence of such a network, and so the harmonic network was used due to its simpler implementation. Since the harmonic network is necessary to extend simulation timesteps due to numerical instabilities in its absence,[5] external force studies without it were limited to 10 fs timesteps, and simulations with it were numerically stable with 20 fs timesteps.

The elastic network was introduced in a similar manner as was initially proposed for proteins,[33] with each bead being connected via harmonic bonds to all other beads within a given distance, which for the present work was chosen to be 0.6 nm since this is approximately equal to 2 base pairs in either direction from a bead. This shorter distance was opted for given that ds-PNA is flexible, and so base pairs distant from one another should not be connected via rigid bonds so that their positions are not highly correlated. Force constants for the harmonic bonds of 1000, 200, 50, 10 and 0 were tested, and the average RMSD over 100 ns production

simulations for each force constant of a Martini GTAGATCACT ds-PNA was compared against that of an all-atom reference model.

6.2.6 Steered molecular dynamics of ds-PNA

A Martini representation of the ds-PNA GTAGATCACT was placed in a simulation box with a minimum distance of 1.5 nm between any atom of the solute and a box edge, and oriented such that the line connecting the N-terminal BB1 atoms of each strand was aligned exactly with the *z* axis of the simulation box, which was extended in that direction. Energy minimisation of vacuum and then solvated structures, using Martini water, was conducted over 50,000 simulation steps before the systems were gradually heated from 30 to 298 K under position restraints. These restraints were switched to zero over 10 ns, excluding the N-terminal BB1 atoms, before being equilibrated for 50 ns each under constant temperature and then pressure using velocity rescaling and Berendsen temperature and pressure couplers respectively. Production simulations were then obtained by removing the BB1 restraint from the N-terminus of the complementary strand AGTGATCTAC and applying a moving external harmonic potential in the direction of increasing reaction coordinate distance, defined as the distance between the N-terminal BB1 atoms. The loading rates (8.7, 1.7, 0.87 and 0.17 N/s) used in the CHARMM simulations of Chapter 5 were then applied over 50 replicates of 3 independently equilibrated Martini simulations. In addition to these, 50 replicates each of these 3 independently equilibrated systems were conducted at loading rates of 0.087, 0.017 and 0.0017 N/s. Only 1 simulation was conducted for each independently equilibrated system for a 0.00017 loading rate, since at this loading rate elapsed real times nearing a week were needed for each production run.

The force on ds-PNA was monitored over all simulations, allowing the force-loading relationship of Martini ds-PNA to be established for the tested loading rates.

6.3 Results and Discussion

6.3.1 Bead definitions for the PNA backbone

As discussed in 6.2.2, the backbone bead types need to be defined for Martini PNA such that the partitioning free energies between different solvents are accurately approximated. Bead types are assigned based on chemical similarity, and so since BB1, BB2 and BB3 are chemically similar to ethyl amine, dimethyl ethanamide, and acetaldehyde respectively, then initial guesses for the bead types were made to match these chemical groups using the bead type guidelines outlined in Marrink *et al.*'s original publication on the Martini force field.[1] For BB1, the closest bead type would likely be P1 (1-propanol) or Nda (1-butanol) to represent the decreased hydrophilicity of ethyl amine relative to P2 (ethanol). For BB2, the closest bead type would likely be of similar hydrophilicity, with an upper bound being P2 given that P3 is based on a related but more hydrophilic chemical structure, methyl formamide. For BB3, Na would be an appropriate initial guess since the closely related propionaldehyde is an example of an Na bead.

Initial guesses of P1, P1 and Na for BB1, BB2 and BB3 were made, and the partition free energies between water and dichloromethane and water and octanol for both CHARMM (all-atom) and Martini representations of the backbone were calculated using BAR. This initial guess underestimated both partition free energies and so given that BB2 could theoretically be represented by a P2 bead, the experiment was repeated with this assignment (Table 6-3).

Table 6-3: Partition free energies calculated according to Equation 5-5 for the transfer of individual backbone monomers) from pure water to dichloromethane (DCM) or pure water to octanol. $\pm$ is the standard error from three replicates.

Forcefield	$\Delta G_{water \rightarrow DCM}$ **(kcal/mol)**	$\Delta G_{water \rightarrow octanol}$ **(kcal/mol)**
CHARMM	3.06 ± 0.35	2.33 ± 0.23
Martini	3.92 ± 0.85	1.97 ± 0.91

Table 6-3 demonstrates that, using a bead assignment of P1, P2 and Na for BB1, BB2 and BB3, the partition free energies between water and dichloromethane and water and octanol agreed, within one standard error, of these same energies obtained from the all-atom simulations. In general, the backbone of Martini PNA slightly overestimates the solubility of the backbone in octanol and underestimates it in dichloromethane. The standard errors for the partition free energies are higher in the Martini case. This is possibly because, in Martini, the clustering of atoms into beads means there are fewer total interactions between solvent and solute meaning that the solvation energies are accounted for by fewer total molecules. Resultantly, fluctuations in the positions of individual molecules may have a larger effect on the measured partition free energy, resulting in the larger standard errors.

Given this agreement, bead assignments of P1, P2 and Na for BB1, BB2 and BB3 were selected as the optimal assignments for Martini PNA backbones. The first and latter of these were later updated after distributions were fitted since the distance between BB1 and BB3 was, on average, lower than the 0.47 nm σ of the Lennard-Jones potential between regular-sized beads. This led to difficulties in matching the all-atom distribution of the BB1, BB2, BB3 angle, but after updating both BB1 and BB3 to the small (S) type beads SP1 and SNa, this difficulty was resolved.

6.3.2 Bond, angle and dihedral distributions

Once all bead types are assigned, it is necessary to model the bonded potentials between them. This involves explicitly defining bond, angle and dihedral force constants and equilibrium points such that the distributions of bonds, angles and dihedrals obtained using the Martini PNA accurately match the distributions obtained using CHARMM PNA. The CHARMM distributions of bonds, angles and dihedral angles between the centres of geometry of clusters of atoms representing Martini beads were obtained from all-atom simulations of the ss-PNA sequences in Appendix A.3. These distributions were then matched by conducting equivalent

simulations on Martini representations of these same sequences. These distributions were matched (Figure 6-4) to the all-atom distributions by iteratively adjusting force constants, equilibrium bond distances and angles, and multiplicities, until the distributions overlapped.

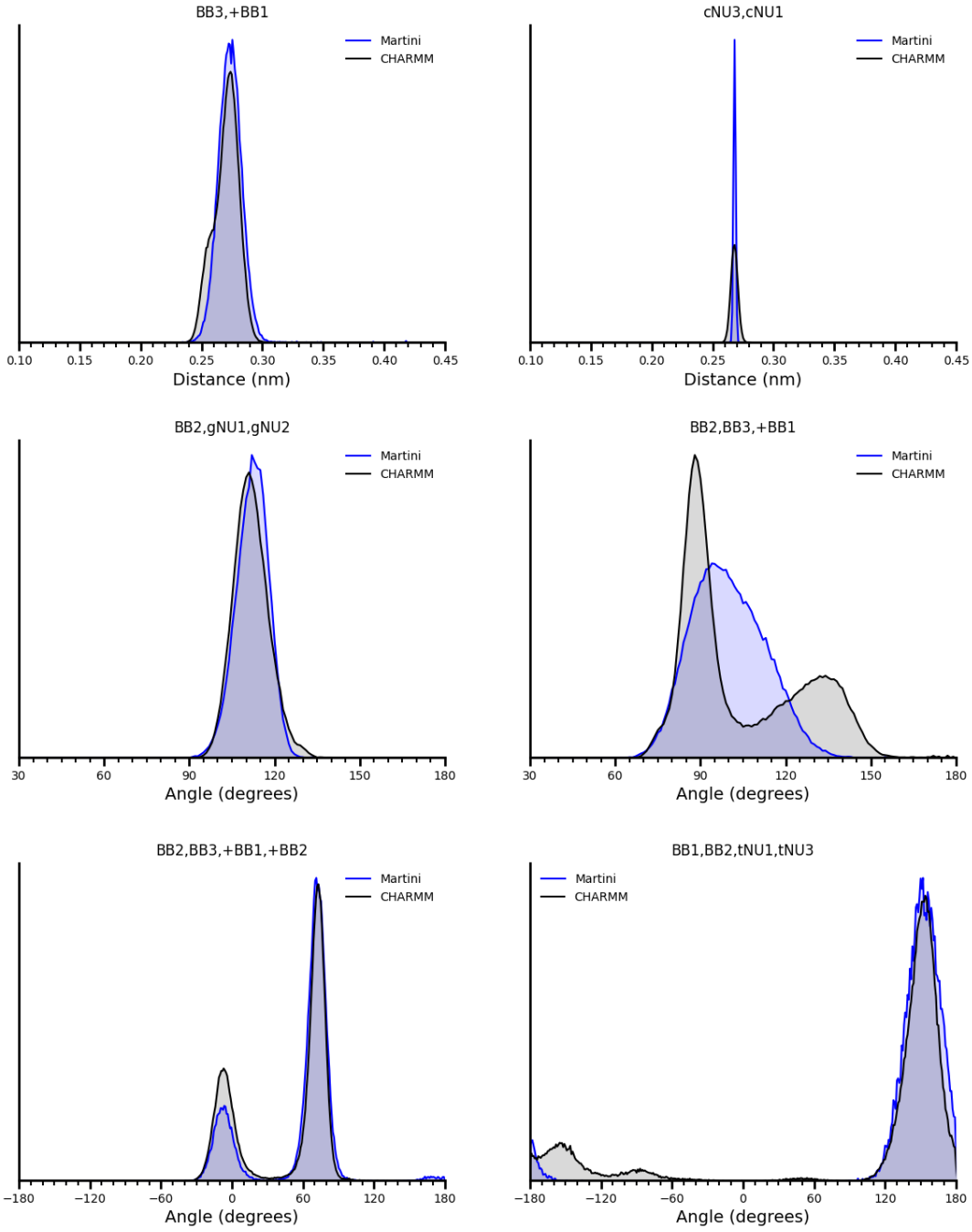


Figure 6-4: Example bond, angle and dihedral (top to bottom) distributions for CHARMM atom clusters and Martini beads. CHARMM distributions are in black and Martini are in blue.

A full set of all 68 distributions is included in Appendix A.4. Figure 6-4 demonstrates that, whereas in some cases distributions have well-defined unimodal peaks, and can be fitted with simple harmonics, as is the case for example for the BB2, gNU1, gNU2 distribution, other distributions are multimodal or otherwise more complex. For example, in the BB2, BB3, +BB1 distribution, where the + refers to the BB1 bead belonging to the next monomer in sequence, two peaks are present. These peaks were found to be independent of the position in sequence and, for any individual simulation, the angle between the centre of geometry of the clusters of atoms representing BB2, BB3, +BB1 was distributed between these two peaks, though the cause of this is not known. One possibility is that a better bead assignment, for example clustering atoms in another way, could reduce this multimodality, though given that this is the only angle distributed in this manner this suggests that, in all other cases, the bead assignment is appropriate. The approach to CHARMM distributions like the BB2, BB3, +BB1 angle which cannot be completely fitted with a simple harmonic angle was to select an equilibrium point at an angle between the two peaks determined by the ratio of their heights, thereby weighting the distribution in favour of the more populous peak. In addition, lower force constants were used in such cases so that the distribution was also smeared in the direction of the smaller peak. An issue with this approach, however, is that the relatively unpopulated intermediate angles or distances are overrepresented in the Martini model.

In addition, Figure 6-4 demonstrates the narrow Martini peaks for the bond distance cNU3, cNU1. This was the case for all bonds between nucleobase beads and is intentional. The narrow distribution arises from the fact that nucleobase bead bond distances are not fitted using a simple harmonic but instead a constraint algorithm which periodically resets the bond length to an equilibrium value. This same methodology was used in the Martini models for DNA and is the recommended approach to parameterising ring systems with Martini beads given that constraints better reflect the rigidity of the ring systems.[5]

For the dihedral distributions in the bottom row of Figure 6-4, a narrow and well-defined multimodal distribution can be seen in the left panel. In the right panel, a distribution mostly defined by a single, broader peak is present. To fit the distributions, torsional angle potentials with varying multiplicities can be overlayed on top of one another, with narrower peaks achieved using higher force constants. Alternatively, improper dihedral potentials, these being just simple harmonic potentials, can be used to fit unimodal distributions. In general, as few torsional angle potentials as possible were used in the parameterisation of the Martini model, and when distributions were mostly dominated by an individual peak corresponding to almost all of the area, then these were fitted with improper dihedrals as is the case in the bottom right panel of Figure 6-4.

By fitting distributions, it was possible to derive a full set of Martini PNA bonded parameters (Table 6-4) which, in combination with the nonbonded parameters determined by the bead types SP1, P2, SNa for the backbone beads BB1, BB2 and BB3, and for the nucleobases as presented in Table 6-2, allows the construction of Martini PNA topology files for conducting coarse-grained molecular dynamics simulations. Hydrogen bonding Morse potentials, as presented in Chapter 6.3.3, and elastic networks, in Chapter 6.3.4, can additionally be included for external force and equilibrium studies respectively.

Table 6-4: Parameters for all constraints, bonds, angles and dihedrals used in the final Martini PNA model. $\boldsymbol{f}$ is the function type, as implemented in the 2021 edition of Gromacs. $\boldsymbol{r_0}$ and $\boldsymbol{\theta}_0$ are the equilibrium distances and angles respectively. $\boldsymbol{k}$ is the force constant in $\mathrm{kJ/mol/nm^2}$, $\mathrm{kJ/mol/rad^2}$ or kJ/mol for the bonds, angles and dihedrals respectively. $\boldsymbol{m}$ is the multiplicity.

Constraint	$\boldsymbol{f}$	$\boldsymbol{r_0}$	
aNU1, aNU2	1	0.229	
aNU2, aNU3	1	0.266	
aNU3, aNU4	1	0.288	
aNU4, aNU1	1	0.162	
gNU1, gNU2	1	0.295	
gNU2, gNU3	1	0.295	
gNU3, gNU4	1	0.285	
gNU4, gNU1	1	0.161	
cNU1, cNU2	1	0.220	
cNU2, cNU3	1	0.285	
cNU3, cNU1	1	0.268	
tNU1, tNU2	1	0.217	
tNU2, tNU3	1	0.322	
tNU3, tNU1	1	0.265	
Bond	f	r_0	k
BB1, BB2	1	0.316	18000
BB2, BB3	1	0.341	5000
BB3, +BB1	1	0.275	30000
BB2, aNU1	1	0.288	25000
BB2, gNU1	1	0.290	25000
BB2, cNU1	1	0.286	25000
BB2, tNU1	1	0.290	25000
Angle	f	$\boldsymbol{\theta}_0$	k
BB1, BB2, BB3	1	77.0	40
BB2, BB3, +BB1	1	105.0	50

BB3, +BB1, +BB2	1	98.0	100	
BB1, BB2, aNU1	10	125.0	400	
BB3, BB2, aNU1	10	147.0	250	
BB1, BB2, gNU1	10	125.0	300	
BB3, BB2, gNU1	10	150.0	150	
BB1, BB2, cNU1	10	118.0	300	
BB3, BB2, cNU1	10	165.0	350	
BB1, BB2, tNU1	10	114.0	250	
BB3, BB2, tNU1	10	165.0	100	
BB2, aNU1, aNU2	10	110.0	250	
BB2, aNU1, aNU4	10	125.0	400	
BB2, gNU1, gNU2	10	110.0	200	
BB2, gNU1, gNU4	10	125.0	300	
BB2, cNU1, cNU2	10	92.0	240	
BB2, cNU1, cNU3	10	146.0	700	
BB2, tNU1, tNU2	10	97.0	150	
BB2, tNU1, tNU3	10	149.0	700	
aNU1, aNU2, aNU3	1	85.0	200	
aNU2, aNU3, aNU4	1	74.0	200	
aNU3, aNU4, aNU1	1	98.0	200	
aNU4, aNU1, aNU2	1	125.0	200	
gNU1, gNU2, gNU3	1	69.5	200	
gNU2, gNU3, gNU4	1	84.0	200	
gNU3, gNU4, gNU1	1	94.0	200	
gNU4, gNU1, gNU2	1	125.0	200	
cNU1, cNU2, cNU3	1	61.0	200	
cNU2, cNU3, cNU1	1	47.0	200	
cNU3, cNU1, cNU2	1	71.0	200	
tNU1, tNU2, tNU3	1	55.0	100	
tNU2, tNU3, tNU1	1	42.0	100	
tNU3, tNU1, tNU2	1	83.0	100	
Dihedral	f	$\boldsymbol{\theta}_0$	k	m

BB1, BB2, BB3, +BB1	9	-25.0	5	1
	9	-80.0	5	4
BB2, BB3, +BB1, +BB2	9	-140.0	11	1
	9	140.0	8	4
BB3, +BB1, +BB2, +BB3	9	-30.0	5	1
	9	90.0	2.5	6
BB1, BB2, aNU1, aNU2	2	-123.0	22	-
BB1, BB2, aNU1, aNU4	2	107.0	30	-
BB1, BB2, gNU1, gNU2	2	-115.0	20	-
BB1, BB2, gNU1, gNU4	2	105.0	30	-
BB1, BB2, cNU1, cNU2	2	-122.0	50	-
BB1, BB2, cNU1, cNU3	2	178.0	50	-
BB1, BB2, tNU1, tNU2	2	-123.0	22	-
BB1, BB2, tNU1, tNU3	2	152.0	25	-

In Table 6-4 restricted bending potentials, $f = 10$, were used for all angles connecting bases with the backbones. This was particularly important for pyrimidines, since high dihedral angles of 178.0 and 152.0 were present for BB1, BB2, cNU1, cNU3 and BB1, BB2, tNU1, tNU3 respectively. During parameterisation, these high torsional angles routinely led to numerical instability, which was largely mitigated by the application of restricted bending potentials.

6.3.3 Mimicking hydrogen bonding with a Morse potential

Both bonded and non-bonded terms had been parameterised according to all-atom data, but Martini PNA, as with other Martini models, lacked the directional hydrogen bonding crucial to the function of ds-nucleic acids. To address this, virtual sites as illustrated in Figure 6-3 were implemented for pre-equilibrated CHARMM structures of isolated A:T and G:C base pairs in water. Base pairing and stacking free energies between these virtual sites were then measured using AWH as a function of the distance between the sites, and these were used to produce reference free energy landscapes against which Martini simulations could be iteratively

refined. Martini simulations were conducted on coarse-grained representations, using the parameters presented in Table 6-4, of the same isolated A:T and G:C base pairs in water. The same free energy landscapes were obtained both with and without the introduction of a Morse potential between the virtual sites, and simulations were repeated with Morse potential depth, steepness and equilibrium distance modified (Table 6-5) until further changing these parameters only decreased agreement between the Martini and CHARMM free energy landscapes (Figure 6-5)

Table 6-5: Final equilibrium distance $\boldsymbol{r_0}$, well depth $\boldsymbol{D}$ and steepness $\boldsymbol{\beta}$ for Morse potentials between A:T and G:C virtual sites respectively.

Base pair	$\boldsymbol{r_0}$ $(\boldsymbol{nm})$	$\boldsymbol{D}$ $(\boldsymbol{kJ/mol})$	$\boldsymbol{\beta}$ $(\mathbf{nm^{-1}})$
A:T	0.440	6.50	29
G:C	0.405	7.00	25

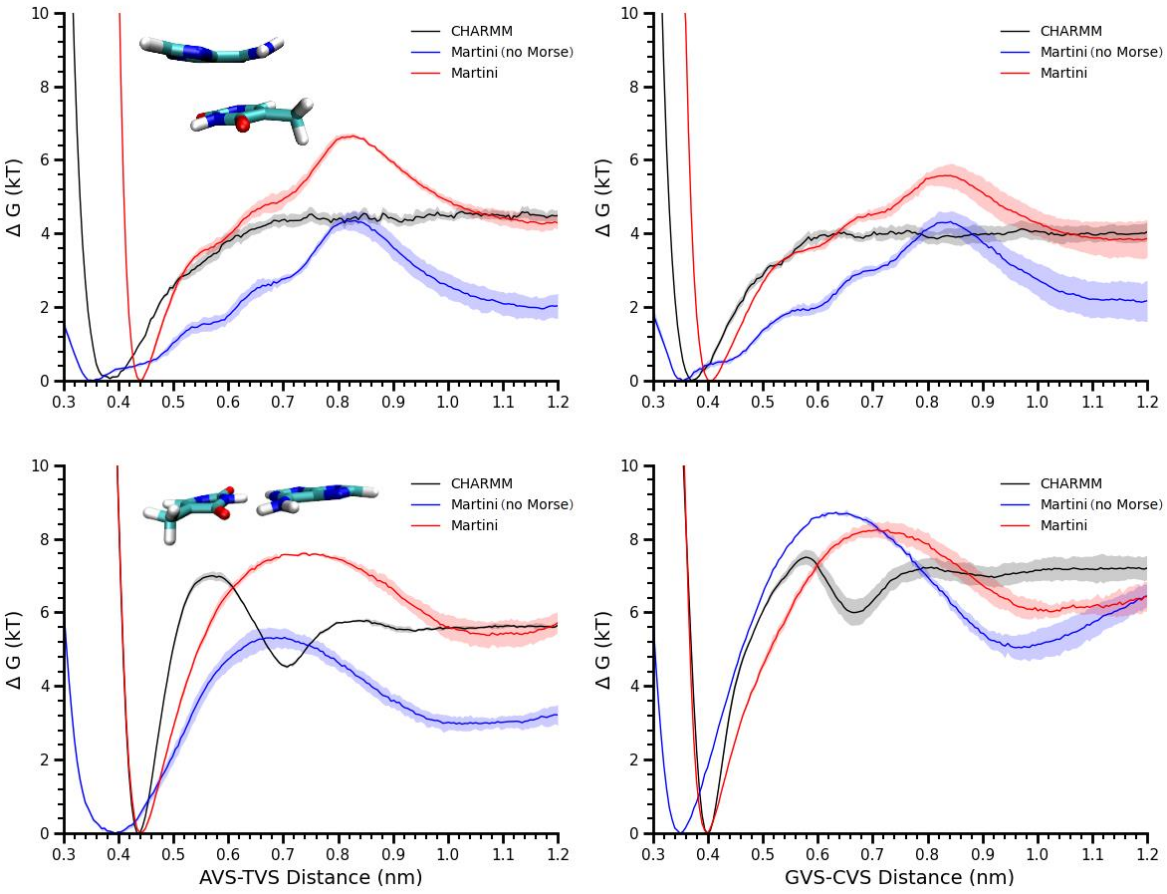


Figure 6-5: Free energy as a function of virtual site distance in ds-PNA. Cartoon insets represent the tested reaction coordinate which, for the top row, is the stacking coordinate, and for the bottom row is the pairing coordinate.

The CHARMM simulations in Figure 6-5 are in excellent agreement with the CHARMM simulations for free nucleobases determined by Uusitalo *et al.*[5] despite the reaction coordinate being slightly different (since a virtual site is used in the present work). In the present work, the CHARMM stacking coordinate has a broad minimum that gradually increases towards a plateau at 4 or 4.4 k_BT for A:T and G:C pairs respectively as the distance between the bases increases. The CHARMM pairing coordinate has a narrower, steeply increasing minimum at a greater distance, with a secondary minimum present at approximately 0.7 nm due to the insertion of a water molecule between the unbinding base pairs. In the pairing coordinate, the higher stability of the G:C base pair is evident in free energy differences between bound and unbound states of 7.2 k_BT as opposed to 5.6 k_BT for the A:T pair. For DNA, similar free energy differences between the bound and unbound states of 5.5

and 6.6 k_BT were measured by Uusitalo *et al*.[5] for CHARMM simulations along the base-pairing coordinate for A:T and G:C pairs respectively.

The inclusion of a Morse potential, as seen in Figure 6-5, was found to improve the agreement between Martini and CHARMM ds-PNA base pairing free energies but the results on the stacking free energies were mixed. The original Martini model in blue represents the breadth, shallowness and position of the stacking minima more accurately. The Martini model including the Morse potential, in red, estimates the free energy difference between initial and final states more accurately than the original model without it. Both Martini models predict a transition region along the reaction coordinate. However, for the Martini model with the Morse potential in red, this transition region is especially overestimated. This could result in the overestimation of the stability of the base stacking configuration.

Compared to the stacking coordinate, the Morse potential was found to result in significant improvements in the agreement between the CHARMM and Martini simulations along the base pairing coordinate. Both the position and steepness of the minima were more accurately estimated, as shown in red in Figure 6-5. At very low displacements the Morse potential exactly replicates the CHARMM model. This is expected since the Morse steepness parameter was manually adjusted until its steepness reflected the steepness of the repulsive region. At intermediate displacements, the gradient in the free energy is shallower for the Morse potential than for CHARMM. In addition, the Morse potentials accurately estimate the height of a transition region in the pairing coordinate but overestimate the breadth and position of this transition region. Similarly to the accurate estimation of transition barrier height, the Morse potentials also estimate similar free energies for the unbound states, though more accurately for the A:T than G:C base pair.

The agreement between the CHARMM and Martini base pairing free energies is significantly improved when a Morse potential is implemented relative to prior

models, which modified the bead types involved in base pairing instead. The prior results of Uusitalo *et al*. demonstrate an underestimation of the base pairing free energy minima for both A:T and G:C pairs compared to the results presented in Figure 6-5.[5] For the base stacking free energy coordinates, however, the Morse potentials in Figure 6-5 are less accurate. That said, they do produce a similar curve to that observed by Uusitalo *et al*. for DNA using the modified bead types method[5] since both predict a large transition barrier not present in the CHARMM simulations. Using a Morse potential thereby increases the similarity of the base pairing Martini coordinate with the CHARMM reference model without significantly affecting the stacking coordinate. Hence, the use of a Morse potential, as opposed to using modified bead types, to represent Watson-Crick hydrogen bonding between complementary base pairs is preferred in the present work.

6.3.4 RMSD of ds-PNA with different elastic networks

The Martini ds-PNA model has been parameterised to account for nonbonded, bonded and now explicit hydrogen bonding interactions. The next step in parameterisation involved analysing how well the Martini ds-PNA replicated the CHARMM reference structures, as determined via RMSD. The RMSD of Martini beads were compared against the centres of geometry of the atom clusters representing those beads, in CHARMM, over 100 ns production simulations. To determine the contributions of elastic networks to the accurate replication of structure, RMSDs were obtained using these networks also (Figure 6-6). In the elastic networks, each bead is bonded to all beads within 0.6 nm, excluding beads to which it is already bonded, with force constants of 1000, 200, 50, 10 and 0 $kJ/mol/nm^2$. 10 fs timesteps were used for the Martini simulations, and 2 fs timesteps for CHARMM simulations.

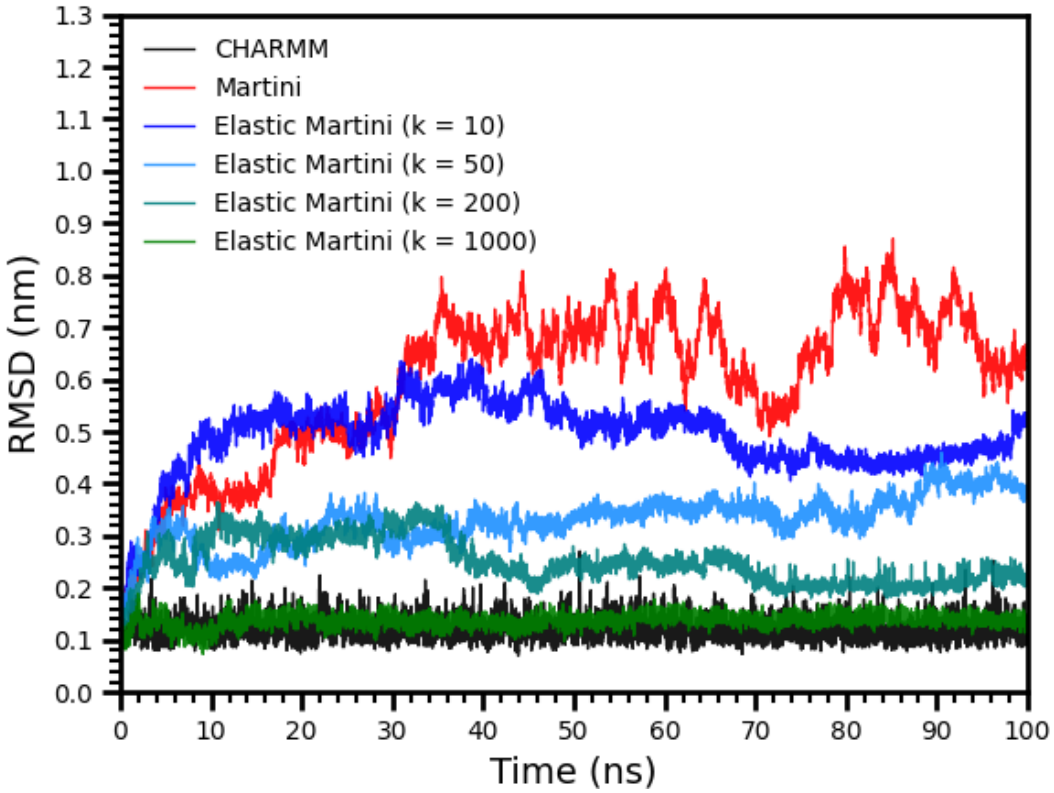


Figure 6-6: RMSD, in nm, of CHARMM and Martini simulations of the ds-PNA GTAGATCACT. Force constant, k, in units of $\mathrm{kJ/mol/nm^2}$.

In the absence of an elastic network, the structure of Martini ds-PNA deviates quickly from the stable CHARMM RMSD trace of between 0.1 and 0.2 nm. For the stiffest elastic network, $k = 1000$, the CHARMM structure is completely retained by the Martini simulation, with a slightly softer network at $k = 200$ still providing a very good approximation. Snapshots of final ds-PNA structures, taken after 100 ns of simulation (Figure 6-7), demonstrate the effect of increasingly releasing the elastic bond network.

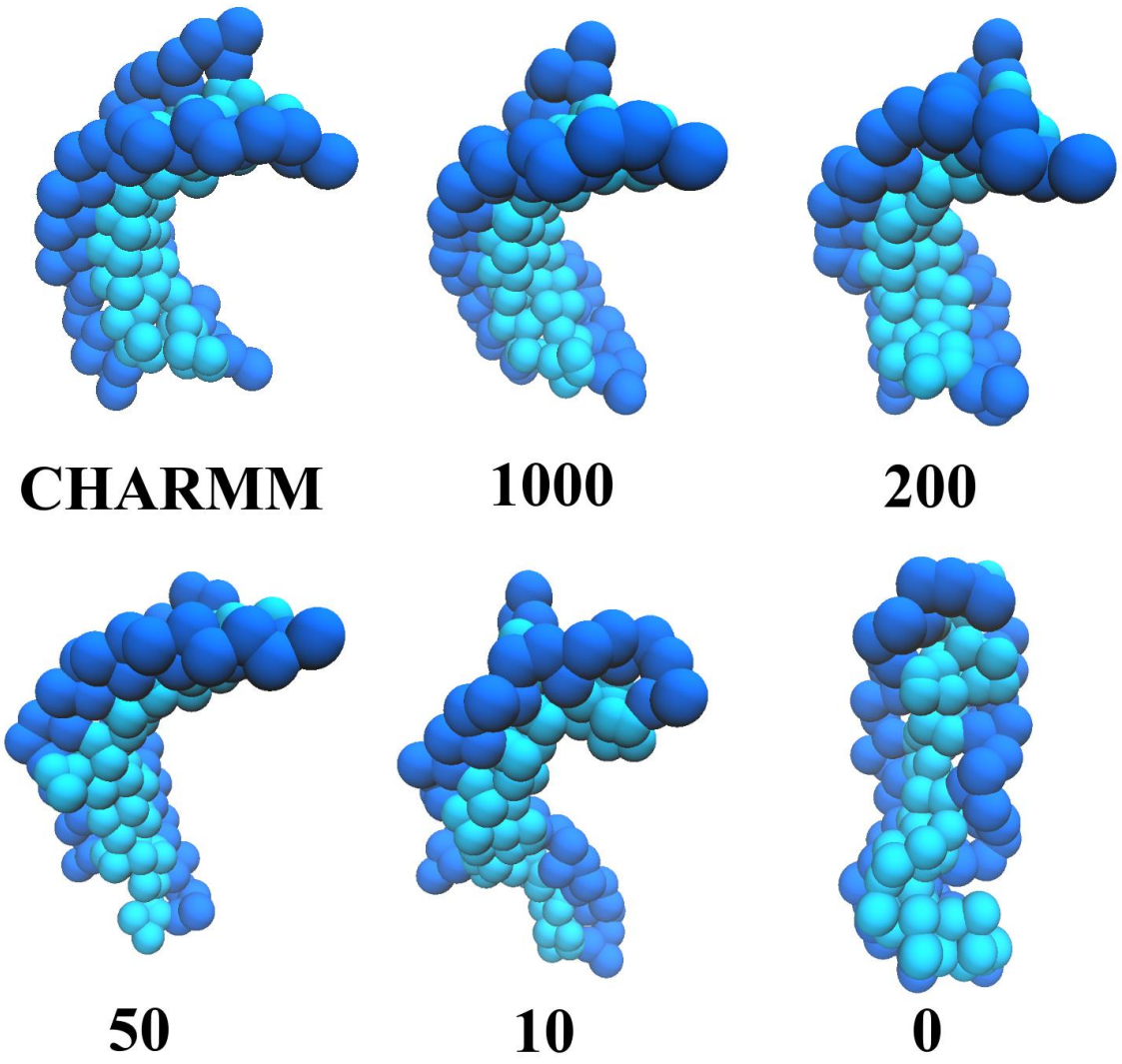


Figure 6-7: Increasingly softening the elastic network from force constants of 1000 to 0 $kJ/mol/nm^2$ increasingly affects the secondary structure of ds-PNA after 100 ns of simulation.

For the highest force constant, the structure at 100 ns is almost indistinguishable. A similarly good representation of the P-form helix is present when the force constant is reduced to 200 $kJ/mol/nm^2$. However, increasing structural artefacts are demonstrated with further softening of the elastic network, with the structure at 50 $kJ/mol/nm^2$ showing terminal base fraying in addition to bases slipping outside of the helical axis, and the structure at 10 $kJ/mol/nm^2$ having an overall structure very different from the P-form helix. Without the elastic network at all, however, no helix is present, which is expected since the bonds, angle and dihedral distributions

are parameterised from ss-PNA and these distributions can be replicated without necessarily reproducing the P-form helix, as demonstrated by comparison of Figure 6-7 with Figure 6-4. In the absence of the network, then, the ds-PNA minimises its potential energy by forming close associations with one another that maximise its forces of attraction. This results in a linear structure wherein the nucleobase ring systems are pressed ‘flat’ to one another to maximise stacking interactions, which is possible to do in Martini without negatively affecting pairing interactions given that the interactions of the Martini beads are spherically symmetric.[1]

6.3.5 Martini ds-PNA force spectra

The Martini ds-PNA model’s ability to extend the force-loading curve to simulation timescales inaccessible to the CHARMM model was tested using SMD along the shearing coordinate in the absence of the elastic network. Force spectra were obtainable at loading rates from 8.7 to 0.00017 N/s (Figure 6-8) and demonstrate the absence of a defined rupture event for loading rates beyond 0.17 N/s, likely since the high forces experienced prevent the relaxation of the system towards a baseline force within the box lengths used. At the lowest loading rates used, a clear rupture event like that observed for the CHARMM simulations is evident, though force traces indicate that rupture proceeds in two stages, particularly at the lowest rate.

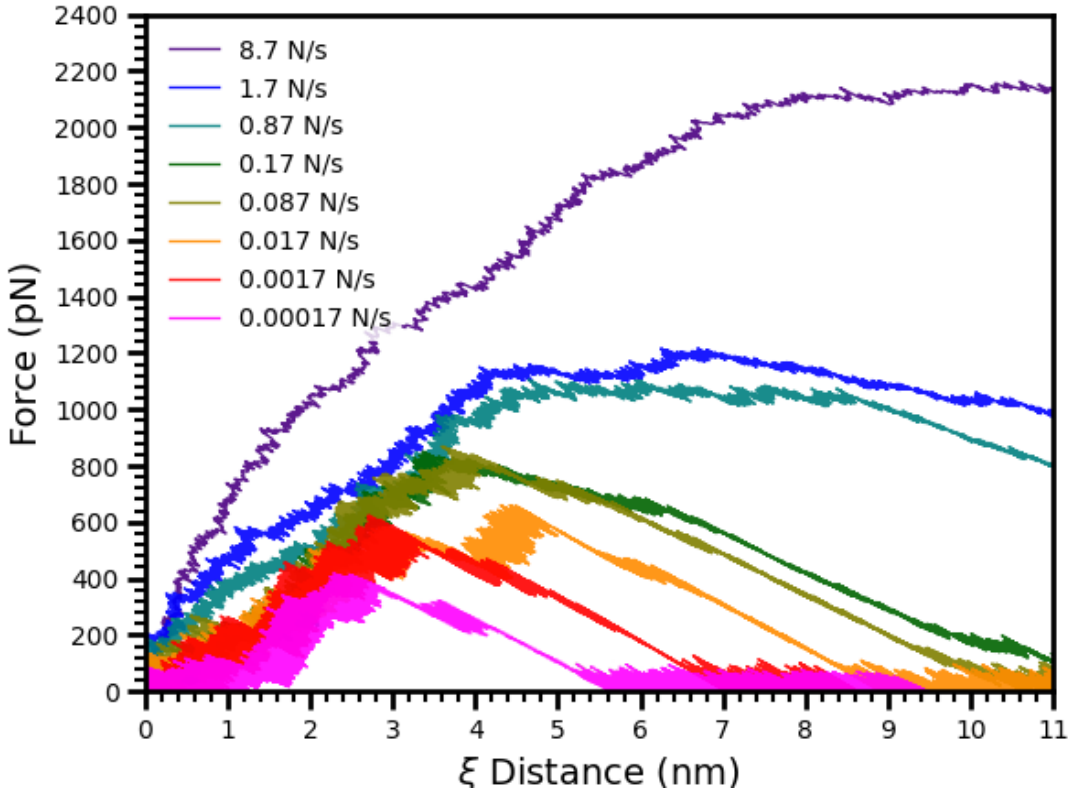


Figure 6-8: Example force spectra for the rupture of the ds-PNA GTAGATCACT modelled using Martini, with loading rates decreasing from dark purple to pink.

The secondary rupture event in the Martini model indicates the existence of a metastable state along the shearing coordinate despite the relative displacement of the two strands, which could suggest that the Martini model discriminates less with regard to non-complementary base pairing than the CHARMM model. For the all-atom simulations, well-defined rupture events are demonstrated only for loading rates below 1.7 N/s, which is an order of magnitude higher than with the Martini model. This suggests that the Martini model is more sensitive to high loading rates causing distinct rupture events to be occluded in the traces than with the CHARMM model. In addition, the two differ in that, when their mean force is plotted as a function of shearing distance (Figure 6-9), the Martini model lacks the distinct overstretching transition demonstrated in the CHARMM model.

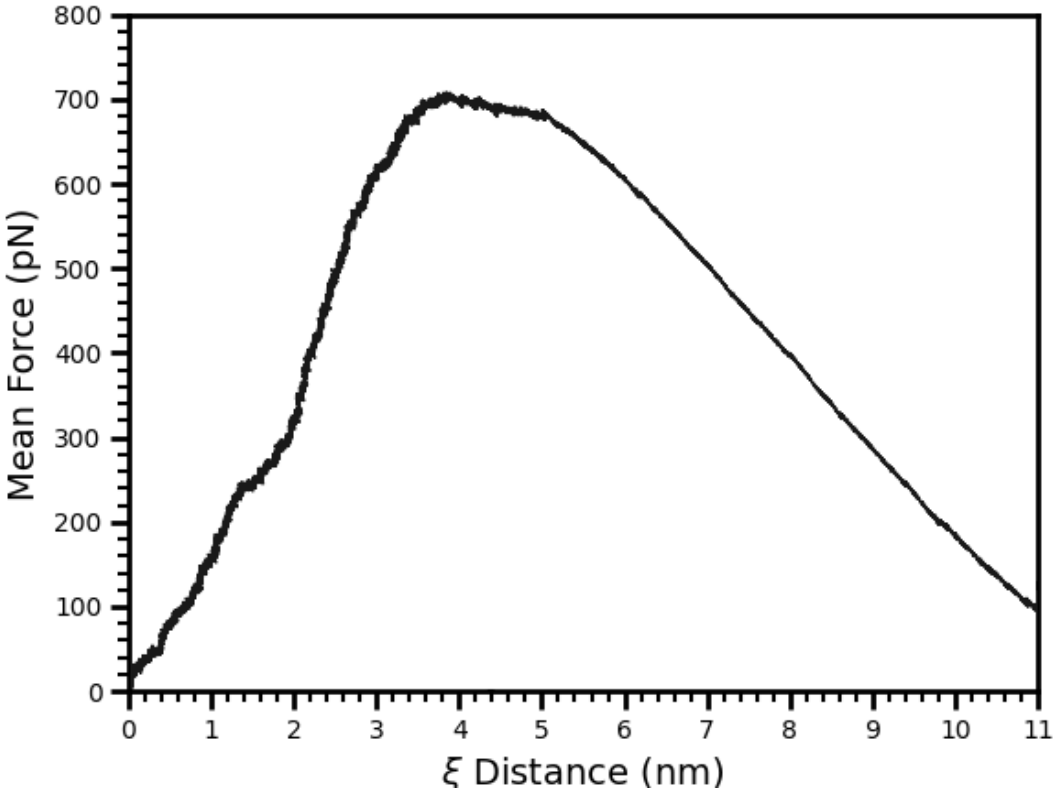


Figure 6-9: Mean force during shearing at 0.17 N/s, chosen for comparison with Figure 5-10, obtained from 150 total replicates of the Martini ds-PNA GTAGATCACT.

The absence of the overstretching transition in Martini ds-PNA is unsurprising, since without an elastic network, the equilibrium conformation of Martini ds-PNA is not double-helical but instead linear, meaning that when external force is applied longitudinal to the 'helical' axis there is no initial unwinding stage, as with CHARMM ds-PNA. In addition, the peak force in Martini ds-PNA exists over a 1.5 nm plateau region in the mean force curve, whereas in the mean force curve for CHARMM ds-PNA this distance is < 0.5 nm. Snapshots of Martini ds-PNA at 4 nm displacements (Figure 6-10) suggests this additional stabilisation at high force could be due to the formation of an ordered, non-complementary stacking structure that is not present in CHARMM ds-PNA, as seen in Figure 5-10. This ordered stacking structure could stabilise the strands as they slide past one another, in effect delaying

the final rupture event by the additional ~ 1 nm seen in comparison of Figures 5-9 and 6-9.

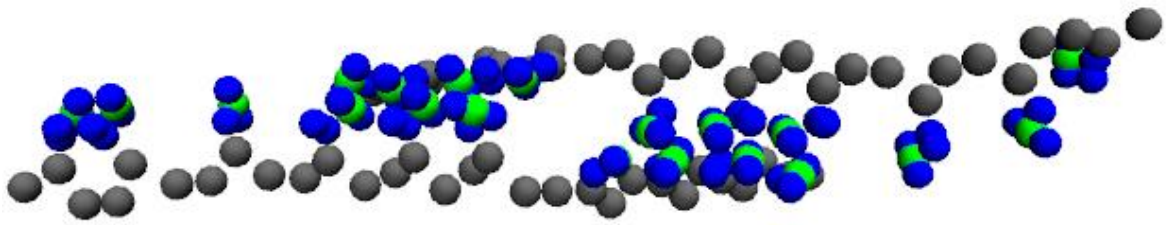


Figure 6-10: Martini ds-PNA bead representation demonstrating the existence of two ‘ordered stacking’ regions separated by a gap. Backbone, nucleobase, and virtual sites are grey, blue, and green respectively.

This ordered stacking structure at high displacements along the shearing coordinate corresponds to the metastable state in the force curves observed in Figure 6-8. It reinforces the conclusions that it exists as a result of the Martini ds-PNA not discriminating between non-complementary nucleobase associations since these stacked states are oriented in a non-pairing, and thereby non-specific, manner. This state is likely possible for the same reason that base pairing is poorly integrated into Martini; spherically symmetric potentials make coplanar stacking in the direction of the helix favourable since it can be achieved without significantly destabilising base pairing given that the pairing potential is spherically symmetric.[5]

6.3.6 The force-loading curve

The structure of Martini ds-PNA during loading in shear does not model the structure of CHARMM ds-PNA accurately, demonstrating the existence of an ordered stacking state at high force and a linearly parallel ‘helix’ at the initial displacement. However, it may still be possible for the force-loading curve to be approximated using this Martini model. This since, according to the nearest-neighbour model, nucleic acid stabilities can largely be accounted for by pairing and

stacking energies, which as per Figure 6-5 are well-replicated by Martini ds-PNA. Though other elements such as torsional potentials also influence the stability along the reaction coordinate, these are accounted for by matching the distributions of bonded potentials as described in Chapter 6.2.3. Under the assumption that pairing, stacking and bonded distributions in equilibrium can account for much of the behaviour of ds-PNA under external force, it is not unreasonable to suggest that the force-loading curve can be approximated by Martini (Figure 6-11). Contrastingly, however, if it is assumed that global structure contributes significantly towards the force-loading behaviour, then Martini ds-PNA cannot accurately model force-loading behaviour, or at the very least models it only as a result of error compensation, since global RMSDs in the absence of an elastic network are poorly replicated as in Figure 6-6.

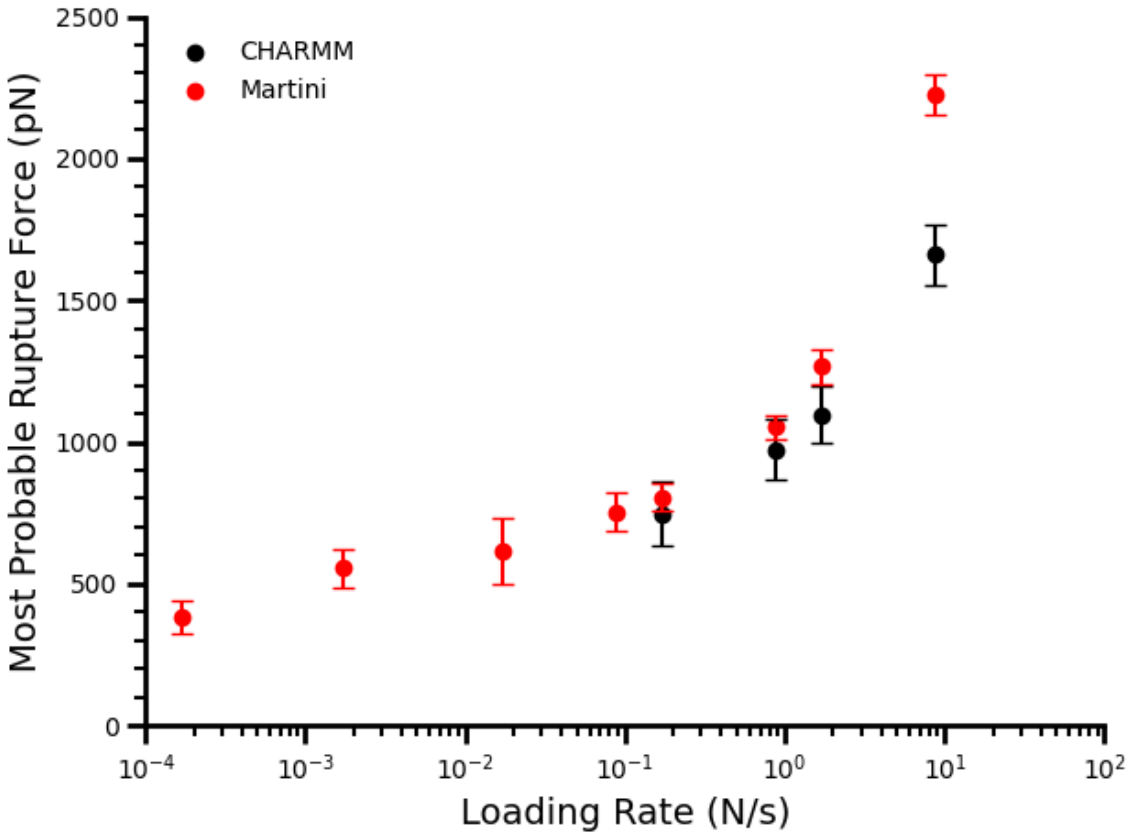


Figure 6-11: Martini ds-PNA force-loading curve in red, and the CHARMM ds-PNA force-loading curve in black.

At the highest loading rates, the most probable rupture force for Martini ds-PNA is equivalent to the maximum force, since no distinct rupture events were observed, as seen in Figure 6-8. In general, the Martini force-loading curve slightly overestimates the rupture force relative to the CHARMM curve, with this deviation increasing with loading rate. At the lowest loading rates accessible with the CHARMM model, however, the Martini and CHARMM models agree considerably with one another. The standard errors for the Martini model are lower than their equivalent CHARMM errors, demonstrating that the Martini model is more consistent in its evaluation of the most probable rupture force. This convergence indicates the Martini model replicates more similar shearing dynamics in repeat simulations than the CHARMM model arguably due to its decreased complexity limiting the model's ability to capture nuances in the unbinding pathway. Alternatively, since the maximum force is sampled over a greater distance, as demonstrated in the mean force trace of Figure 6-9, these lower errors could occur since the peak force region is more extensively sampled in each simulation than with CHARMM ds-PNA. Since the Martini model is an approximation of the all-atom model, which is itself an approximation of a real system, the lower errors are not a result of more accurate modelling of the real system and thereby do not reflect better convergence on some true value.

The loading rates used were orders of magnitude higher[34] than those typically needed to characterise the near-equilibrium regime of the force-loading curve, preventing the fitting of Figure 6-11 to force spectroscopy models since fits at these high loading rates would produce inaccurate assessments of thermodynamic parameters such as the binding free energy.[35] Despite these results not being fitted to force spectroscopy models, the force-loading curve in Figure 6-11 demonstrates that the Martini model replicates the expected linear-to-nonlinear transition in the force-loading behaviour at high loading rates. The Martini model succeeds in that it can characterise a broader force-loading curve than is attainable with the CHARMM model for the ds-PNA GTAGATCACT. In addition, it agrees considerably with the CHARMM model at the lowest rates attainable with CHARMM. However, it is stressed that this force-loading curve has been attained

for only a single sequence, and therefore it is not necessarily the case that all CHARMM ds-PNA force-loading curves would be well-replicated by Martini.

6.4 Conclusions

The Martini PNA model enables longer simulation timescales to be used and thus a broader force-loading curve can be produced. In theory, this force-loading curve could be combined with experimental data to produce a complete curve, enabling force spectroscopy models such as the Bell-Evans or Dudko-Hummer-Szabo model to be fitted and thereby to allow energy landscape properties to be modelled. The technical research question (3), “Can coarse graining be used to replicate properties observed in atomic molecular dynamics simulations?”, is addressed through the following conclusions:

(1) For the pilot GTAGATCACT sequence, the force-loading curve using Martini ds-PNA overlapped with the CHARMM simulations at the lowest loading rates achievable with CHARMM, though overestimated rupture forces at the highest loading rates. Consequently, Martini ds-PNA may be an appropriate simplistic model for rupture force prediction at these loading rates.

(2) Since the CHARMM model cannot be extended to lower rates, it is unknown whether the CHARMM and Martini force curves diverge at lower loading rates decrease.

(3) When an elastic network is used its RMSD agrees well with the RMSD from CHARMM simulations, as is the case for other Martini nucleic acid models.

(4) The model accurately evaluates base pairing free energies using a novel Morse potential representation of the hydrogen bond.

(5) Error compensation may result in the observed results since, in the absence of the elastic network, the secondary structure of ds-PNA is very poorly approximated despite the base pairing coordinate being well replicated, and hence the latter may be compensating for the former.

(6) Fitting of multimodal angles with simple harmonics, as seen in Figure 6-4, overestimates the occurrence of angles intermediate of the peaks.

The Martini ds-PNA model herein is the first coarse grained model of ds-PNA known to the candidate. Though it succeeds in accurately characterising base pair free energies, it is limited in its inaccurate evaluation of secondary structure in the absence of an elastic network, as with all Martini nucleic acid models. Since ds-PNA stabilities can be well-described by the nearest-neighbour model (Chapter 4) of pairing and stacking energies, and the pairing and stacking energies of Martini PNA bases are well-estimated by a Morse potential, this does not necessarily exclude Martini PNA from external force studies.

6.5 References

[1] Marrink, S.; Risselada, H.; Yefimov, S.; Tieleman, D.; de Vries, A. The MARTINI Force Field: Coarse Grained Model for Biomolecular Simulations. *J. Phys. Chem. B* **2007**, *111*, 7812-7824
[2] Fábián, B.; Thallmair, S; Hummer, G. Small ionic radii limit time step in Martini 3 molecular dynamics simulations. *J. Chem. Phys.* **2022**, *157*, 034101
[3] Van der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A.; Berendsen, H. GROMACS: Fast, Flexible and Free. *J. Comput. Chem.* **2005**, *26*, 1701-1718
[4] Bulacu, M.; van der Giessen, E. Effect of bending and torsion rigidity on self-diffusion in polymer melts: A molecular-dynamics study. *J. Chem. Theory Comput.* **2005**, *123*, 114901
[5] Uusitalo, J.; Ingólfsson, H.; Akhshi, P.; Tieleman, D.; Marrink, S. Martini Coarse-Grained Force Field: Extension to DNA. *J. Chem. Theory Comput.* **2015**, *11*, 3932-3945
[6] Maffeo, C.; Ngo, T; Ha, T.; Aksimentiev, A. A Coarse-Grained Model of Unstretched Single-Stranded DNA Derived from Atomistic Simulation and Single-Molecule Experiment. *J. Chem. Theory Comput.* **2014**, *10*, 2891-2896
[7] Alessandri, R.; Barnoud, J.; Gertsen, A.; Patmanidis, I.; de Vries, A.; Souza, P.; Marrink, S. Martini 3 Coarse-Grained Force Field: Small Molecules. *Adv. Theory Simul.* **2022**, *5*, 2100391
[8] Grünewald, G.; Punt, M.; Jefferys. E.; Vainikka, P.; König, M.; Virtanen, V.; Meyer, T.; Pezeshkian, W.; Gormley, A.; Karonen, M.; Sansom, M.; Souza, P.; Marrink, S. Martini 3 Coarse-Grained Force Field for Carbohydrates. *J. Chem. Theory Comput.* **2022**, *18*, 7555-7569
[9] Monticelli, L.; Kandasamy, S.; Periole, X.; Larson, R.; Tieleman, D.; Marrink, S. The MARTINI Coarse-Grained Force Field: Extension to Proteins. *J. Chem. Theory Comput.* **2008**, *4*, 819-834
[10] Montazersaheb, S.; Hejazi, M.; Charoudeh, H. Potential of Peptide Nucleic Acids in Future Therapeutic Applications. *Adv. Pharm. Bull.* **2018**, *8*, 551-563

[11] Singh, K.; Sridevi, P.; Singh, R. Potential applications of peptide nucleic acid in biomedical domain. *Engineering Reports* **2020**, *2*, e12238
[12] Gupta, A.; Mishra, A.; Puri, N. Peptide nucleic acids: Advanced tools for biomedical applications. *J. Biotech.* **2017**, *259*, 148-159
[13] Zhao, M.; Sampath, J.; Alamdari, S.; Shen, H.; Chen, C.; Mundy, C.; Pfaendtner, J.; Ferguson, A. MARTINI-Compatible Coarse-Grained Model for the Mesoscale Simulation of Peptoids. *J. Phys. Chem. B* **2020**, *124*, 7745-7764
[14] Uusitalo, J.; Ingólfsson, H.; Marrink, S.; Faustino, I. Martini Coarse-Grained Force Field: Extension to RNA. *Biophysical Journal* **2017**, *113*, 246-256
[15] Sinden, R. DNA Structure and Function (1st ed.). *Academic Press* **1994**, ISBN 0-12-645750-6.
[16] Costa Filho, R.; Alencar, G.; Skagerstam, B.; Andrade Jr., J. Morse potential derived from first principles. *Europhysics Letters* **2013**, *101*, 10009
[17] Goga, N.; Rzepiela, A.; de Vries, A.; Marrink, S.; Berendsen, H. Efficient Algorithms for Langevin and DPD Dynamics. *J. Chem. Theory Comput.* **2012**. *8*. 3637-3649.
[18] Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald. An Nlog(N) method for Ewald sums in large systems. *J. Chem. Phys.* **1993**. *98*. 10089-10092.
[19] Essmann, U.; Perera, L.; Berkowitz, M.; Darden, T.; Lee, H.; Pedersen L. A smooth particle mesh ewald potential. *J. Chem. Phys*. **1995**. *103*. 8577-8592.
[20] Frigo, M.; Johnson, S. The Design and Implementation of FFTW3. *Proc. IEEE.* **2005**. *93*. 216-231.
[21] De Jong, D.; Baoukina, S.; Ingólfsson, H.; Marrink, S. Martini straight: Boosting performance using a shorter cutoff and GPUs. *Comp. Sci. Comm.* **2016**, *199*, 1-7
[22] Hocknehy, R.; Goel, S.; Eastwood, J. Quiet High Resolution Computer Models of a Plasma. *J. Comp. Phys.* **1974**. *14.* 148-158.
[23] Berendsen, H.; van Gunsteren, W. Practical algorithms for dynamics simulations. *Molecular-Dynamics Simulation of Statistical-Mechanical Systems*. **1986**. 43-65.
[24] Tironi, I.; Sperb, R.; Smith, P.; van Gunsteren, W. A generalized reaction field method for molecular dynamics simulations. *J. Chem. Phys.* **1995**, *102*, 5451-5459
[25] Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; de Groot, B.; Grubmüller, H.; MacKerell Jr, A. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nature Methods*, **2016**, *14*, 71-73
[26] Bennett, C. Efficient Estimation of Free Energy differences from Monte Carlo Data. *J. Comput. Phys.* **1976**, *22*, 245-268
[27] Gapsys, V.; Seelinger, D.; de Groot, B. New Soft-Core Potential Function for Molecular Dynamics Based Alchemical Free Energy Calculations. *J. Chem. Theor. Comput.* **2012**, *8*, 2373-2382
[28] Jasiński, M.; Miszkiewicz, J.; Feig, M.; Trylska, J. Thermal Stability of Peptide Nucleic Acid Complexes. J. Phys. Chem. B 2019, 123, 8168−8177
[29] Lim T. The Relationship between Lennard-Jones (12-6) and Morse Potential Functions. *Zeitschrift für Naturforschung A.* **2003**, *58*, 615-617
[30] Fabiola, F.; Bertram, R.; Korostelev, A.; Chapman, M. An improved hydrogen bond potential: Impact on medium resolution protein structures. *Protein Sci.* **2002**, *11*, 1415-1423
[31] Hagler, A.; Huler, E.; Lifson, S. Energy Functions for Peptides and Proteins, I. Derivation of a Consistent Force Field Including the Hydrogen Bond from Amide Crystals. *J. Am. Chem. Soc.* **1974**, *96*, 5319-5327
[32] Norberg, J.; Nilsson, L. Potential of Mean Force Calculations of the Stacking-Unstacking Process in Single-Stranded Deoxyribodinucleoside Monophosphates. *Biophysical Journal* **1995**, *69*, 2277-2285
[33] Periole, X.; Cavalli, M.; Marrink, S.; Ceruso, M. Combining an Elastic Network With a Coarse-Grained Molecular Force Field: Structure, Dynamics, and Intermolecular Recognition. *J. Chem. Theory Comput.* **2009**, *5*, 2531-2543
[34] Friddle, R.; Podsiadlo, P.; Artyukhin, A.; Noy, A. Near-Equilibrium Chemical Force Microscopy. *J. Phys. Chem. C.* **2008**, *112*, 4986-4990
[35] Noy, A.; Friddle, R. Practical single molecule force spectroscopy: How to determine fundamental thermodynamic parameters of intermolecular bonds with an atomic force microscope. *Methods* **2013**, *60*, 142-150

Chapter 7: Discussion and future work

7.1 Discussion

7.1.1 To what extent has the research question been addressed?

The purpose of this thesis is to answer the central research question:

"Can molecular dynamics be used to assess and screen the single-molecular binding properties of a candidate bioadhesive?"

This was addressed by answering the three technical research questions using ds-PNA as the candidate bioadhesive:

(1) "Can molecular dynamics be used to generate a model that allows binding properties to be predicted from structure alone, and what quantities contribute to the binding stability of the candidate bioadhesive?"

(2) "Can molecular dynamics be used to assess the single-molecular properties of a candidate bioadhesive under external force, and what quantities contribute to this?"

(3) Can coarse graining be used to replicate properties observed in atomic molecular dynamics simulations?"

Technical research question (1) has two parts: (a) was a predictive model generated? and (b) what quantities contribute to stability as per the model? These were answered by the application of a nearest-neighbour model to molecular dynamics simulations of ds-PNA for the first time. The nearest-neighbour model produced in the present work was able to characterise the contributions to the binding stability of ds-PNA in equilibrium in terms of incremental stacking

enthalpies. Using this scheme, it was possible to calculate the binding enthalpies of ds-PNA sequences with an average error of 8.7% when compared against the enthalpies obtained from real-world experiments in the literature.

According to the model, the incremental binding enthalpies of ds-PNA were always negative, including the initiation terms, regardless of terminal identity. This contrasts with biologically occurring nucleic acids wherein helix initiation is enthalpically unfavourable.[1,2] This demonstrated that the association between two complementary ss-PNAs may be inherently enthalpically favourable, since helix initiation enthalpies were negative. This is likely since there is no net like-charge repulsion between the backbones.[3,4]

In addition to the helix initiation enthalpy, incremental enthalpies for ds-PNA were always more negative than their equivalent for ds-DNA except for the GC:CG ds-stack. This corroborates conclusions from the prior literature that the binding enthalpies of ds-PNAs are more negative than equivalent sequences of ds-DNAs.[4,5,6] It is plausible that selection bias could influence the magnitudes of these incremental enthalpies, which may explain the GC:CG discontinuity. However, it is also possible that the stacking enthalpy between base pairs plays a more significant role in ds-PNA since stacks with the same pairs, but different positions, are split between a higher and lower binding enthalpy member. This can be seen in Table 4-6. In such groupings, since the pair composition is the same, the difference in incremental enthalpies is due to stacking, and since in ds-DNA the incremental enthalpies of such groupings are closer together, this could suggest that stacking plays a larger role in differentiating ds-PNA stabilities than it does in ds-DNA.

The final quantity that affects the binding stability of the ds-PNA candidate bioadhesive, according to the model, is the terminal A:T pair. Whereas terminal G:C and A:T pairs both alternated between bonded and frayed states, for terminal A:T pairs longer, more persistent melting events were demonstrated. This suggested

that the terminal instability seen in the higher initiation enthalpy with terminal A:T pairs results from persistent melted states as opposed to short-lived ones.

The present work has thereby addressed technical research question (1) by demonstrating that the stability of ds-PNAs in equilibrium is a function of its sequence, expressed in terms of ds-stacks, and that stacks with more Watson-Crick hydrogen bonds, in general, contribute most to equilibrium stability. Additional information, such as the possibility of base stacking playing a larger role in ds-PNA stability, may also be interpreted. A key difference between ds-PNA and other nucleic acid binding enthalpies is evident in the higher stability floor due to the absence of like-charge repulsion. This is reflected in the negative helix initiation enthalpies. The nearest-neighbour model, which interprets the binding enthalpy of ds-PNA as per these assumptions, has an average error (8.7%) below typical experimental error.[7] Consequently, as per technical research question (1), a model has been produced that predicts a binding property of a candidate bioadhesive (ds-PNA) and the quantities associated with this stability have been expressed. However, technical research question (1) could have been more fully met had it been possible to resolve binding free energies with the method used, though this was not possible due to an inaccurate approximation of the binding entropy.

Technical research question (2) is answered if single-molecular properties of ds-PNA under external force can be resolved. By monitoring structure, forces, and energies during the application of external force via SMD along two primary ds-PNA unbinding coordinates, it was possible to resolve some of these single-molecular properties. The choice of reaction coordinate significantly affected the behaviour of ds-PNA under external force, with the shearing coordinate tolerating higher maximum forces prior to a rupture event.

The higher stability of ds-PNA along the shearing coordinate is possibly due to 3 factors: (i) an initial unwinding step wherein the P-form helix deforms into a linear

overstretched form, (ii) a compression of the strands indicated by a decrease in the van der Waals energy at intermediate displacements, and (iii) ‘storage’ of applied force in the ds-PNA backbone during stretching. The third point is suggested from a peaking torsional energy immediately prior to the point of rupture followed by a decline as ss-PNA strands rewind, indicating that potential energy is stored in the unwinding of the backbone. Since external force accumulates through compression and unwinding until a single, dramatic rupture event, the behaviour of ds-PNA under external shearing forces is unlike at equilibrium since it cannot simply be characterised based on individual base pair clusters as per the nearest-neighbour model. This single rupture event explains the larger maximum forces observed during shearing, since most of the external force is accumulated and then dissipated in a single event.

By contrast, during unzipping ds-PNA rupture involves several rupture events corresponding to single, or small clusters of, base pairs. This results in lower maximum forces and narrower distributions in the force histograms since the maximum force is repeatedly sampled in a single trace. In contrast to shearing, wherein the van der Waals, Coulomb and torsional energy traces were all functions of the loading rate, having deeper and broader wells, steeper exponential regions, and taller peak energies respectively, for unzipping these traces were mostly rate independent. The exception to this was the torsional energy. Particularly at the highest loading rate, torsional energy peaks at a higher value, which could partly explain a jump in the maximum force histograms at the highest rate. This jump could emerge if the dynamics of unzipping changed at the highest loading rate. Since the torsional energy trace is distinct at this rate, it is plausible that the dynamics of unzipping differ due to twisting forces, thereby explaining the jump in the rupture force.

Snapshots of ds-PNA during shearing were obtained at 0, 2.5 and 4.0 nm displacements and free energy surfaces reconstructed. These surfaces demonstrated the opening and flattening of the free energy landscapes for terminal and internal

base pairs respectively. At a 2.5 nm extension, for example, the free energy landscape for an internal G:C base pair flattens towards a landscape more like that of a terminal base pair than an internal pair at equilibrium. This indicates that the overstretching of ds-PNA could result in an internal-to-terminal like transition in the pairing and stacking energies of the sequence. Since the overstretched form of ds-PNA, as with ds-DNA, is approximately 1.7 times longer[8,9] than the equilibrium form, this flattening of the free energy landscape potentially arises from the increased distance between coplanar Watson-Crick base pairs.

Technical research question (2) was therefore addressed by demonstrated that the properties contributing to the stability of a ds-PNA candidate bioadhesive under external force are reaction coordinate and rate dependent. It was found that molecular dynamics could characterise and quantify these binding properties.

Technical research question (3) was answered by developing a novel coarse-grained model of ds-PNA based on the Martini framework. This question is successfully addressed if this novel model can replicate observables in the all-atom simulations. This model was produced by matching bond, angle, and dihedral distributions against all-atom data, and by using an elastic network in addition to this the all-atom structure is accurately reproduced under equilibrium conditions. However, for the purpose of external force studies, it is not possible to simultaneously use a harmonic elastic network and dissociate the molecules under external force given that the elastic network provides an unphysical contribution to the stability of the complex. Despite that, by accurately characterising base pairing free energy traces using a morse potential, the Martini ds-PNA force-loading curve overlaps with the CHARMM ds-PNA force-loading curve. This indicates that the Martini model may be acting as a simplistic, higher throughput method for modelling the single-molecular rupture of ds-PNA than the CHARMM model.

The **central research question** is then answered by stating that, since all technical research questions could be addressed, then it is the case that the single-molecular properties of a candidate bioadhesive, ds-PNA, were resolvable using molecular dynamics simulations. The choice to approach the problem of screening bioadhesives from the single-molecular perspective was based on the necessity that the approach be generalisable to other systems. This necessitated that all properties of the model, except for the specific coarse-grained model developed here for ds-PNA, arose from atomic information such that any new candidate bioadhesive can be tested in the same manner simply by changing the input atomic structure. It is acknowledged, however, that this relies on the assumption that single-molecular information is useful for the development of bioadhesives. An argument in favour of this usefulness is that, by characterising the single-molecular properties of a bioadhesive, observations at the macroscale can be interpreted in terms of this molecular understanding. In addition, by characterising the single-molecular behaviour this may ease future research since the non-specific contributions to stability that are a consequence of macromolecular properties such as composition or impurities may be more readily separated from the specific contributions made at the single-molecular level.

The present work has demonstrated that molecular dynamics can screen and assess a candidate bioadhesive based on single-molecular properties and so the **central research question** has been answered. In addition, it is believed that answering this central research question is of commercial or industrial significance to the development of bioadhesives as per the argument in the preceding paragraph.

7.1.2 Successes of the present work

The present work succeeds in advancing understanding of ds-PNA binding stabilities, as discussed in the previous subchapter. Significantly, the present work is the first time that the overstretched form of ds-PNA has been evidenced, and the overstretching transition occurs at a remarkably similar relative length to that

observed with ds-DNA,[8,9] suggesting uniformity in the overstretching transitions of nucleic acids with varying backbone character. Similar agreement between ds-DNA and ds-PNA unbinding dynamics are evidenced in the unzipping coordinate also, since unbinding in the unzipping coordinate is an approximate 'per base pair' process.[10,11] This observation suggests that mean force traces during ds-PNA unzipping could be used to predict sequence identities since integer multiples of single base pair rupture forces can be demonstrated. The 'reading' of nucleic acid primary sequences from unzipping mean force-distance traces has been demonstrated previously in ds-DNA.[10]

In the process of completing this thesis, two new tools were developed for the modelling of ds-PNAs. The first of these tools is the nearest-neighbour model for predicting ds-PNA binding enthalpies, which in theory enables a user to predict the binding enthalpy of a ds-PNA from its primary sequence alone. This tool can be used for sequence selection prior to time or financial investments by eliminating or selecting sequences from a study based on stability criteria. As an example, a study correlating the melting temperatures with binding enthalpies of ds-PNAs would benefit from using the nearest-neighbour model developed herein by allowing users to select ds-PNAs with well-distributed predicted binding enthalpies. In addition to the development of this nearest-neighbour model, the present work also highlighted the limitations of using a linear regression to correct MM-GBSA[7] data to experimental data by demonstrating that doing this results in the inequality $\Delta G \neq \Delta H - T\Delta S$ in the incremental enthalpies. This is a necessary consideration in the experimental design of future MM-GBSA studies on nucleic acids.

The second tool produced for the present work is the Martini model of ds-PNA, which to the awareness of the candidate is the first coarse-grained model of PNA ever developed. The model's bond, angle and dihedral distributions are modelled against those of the CHARMM model for ds-PNA of Jasiński *et al.*[12] Differently to prior Martini representations of nucleic acids, which use non-standard bead types to model complementary base pairing,[13,14] a Morse potential[15] between two virtual

sites is used. Relative to the non-standard bead type method, the pairing free energy is better approximated by the Morse potential. As with all Martini nucleic acid models, the absence of an elastic network significantly compromises the structure of the double-helix,[13,14] and so the Martini model for ds-PNA provides users with a tool for either external force or structural studies, but not both simultaneously.

The Martini ds-PNA model could be used, for example, to study long sequences of ds-PNA over timescales inaccessible to all-atom simulations, and since it is parameterised using standard Martini beads[16] it is innately compatible with coarse-grained models of proteins[17] and nucleic acids.[13,14] It can thereby be used to study PNA:DNA, PNA:RNA or PNA:protein complexes.

ds-PNA was used as an example candidate bioadhesive for the present work. This was so that the single-molecular methods of obtaining sequence-independent binding properties and other force-loading properties could be presented with an example. It is still worth discussing, however, how ds-PNA itself actually performed as a candidate bioadhesive. This thesis has detailed rupture forces for ds-PNA in shear between approximately 0.4 and 2.2 nN at loading rates from approximately 10^{-4} and 10 N/s. At the highest loading rate, the 2.2 nN rupture force is comparable to that of a single covalent bond,[18] though the nonbonded interactions between ds-PNA strands are advantaged over the covalent bond system in that they are reversible and sequence-specific. That being said, a rupture force for an entire ds-PNA complex comparable to a single covalent bond, and only at the highest loading rate, indicates that a ds-PNA bioadhesive would have a significantly lower bond density and thereby likely be outperformed, in terms of its strength, by an adhesive composed of covalently-bonded units such as an epoxy. However, since fibrin bioadhesives with single-molecular rupture forces in the sub-nanoNewton range are used in commercial applications, then it is likely that the rupture forces described in this thesis for ds-PNA may be sufficient for similar applications.[19]

7.1.3 Limitations of the chosen methodology

External force studies were conducted using the pilot sequence GTAGATCACT, and so it is assumed that ds-PNAs with different primary sequences behave similarly under external force. This limitation was imposed due to the availability of computational resources, with molecular dynamics simulations conducted in parallel using a 16 core AMD Ryzen 9 5950X processor with acceleration provided by a GTX 3090Ti GPU. Using this setup, simulation speeds of hundreds and thousands of nanoseconds per day were achievable for all-atom and coarse-grained simulations respectively. Distributed computing alternatives such as Microsoft Azure or distributed computing at the University of the West of England were pursued though were ruled out due to high costs and inferior computing power given that access to the in-house machine was unshared and unrestricted.

Since force-loading curves are constructed from histograms to determine the most probable rupture force, and typically hundreds of samples are used per histograms, this resulted in the single pilot sequence being selected.[20,21,22] However, it is likely that the same behaviours would be demonstrated with different primary sequences under external force and so the use of a pilot sequence may not be as limiting as initially suspected. The justification for this is since:

(1) For other nucleic acids, force-loading behaviours are conserved regardless of sequence identity, with overstretching below a critical force of 65 pN being consistently identified,[23,24,25] and for unzipping the existence of single base pair rupture events in the force traces of different sequences are demonstrated.[10,11]

(2) The backbone monomers of different sequences have identical chemistry regardless of nucleobase identity, making it unlikely for a changing primary sequence to affect the response of the backbone torsional angles to external force.

(3) The de Gennes model[26,27] suggests that a 10 base-pair sequence is long enough to capture the principal effects of shearing force on ds-PNA.

(4) The mean force trace during unzipping demonstrates no other rupture events than the integer multiple rupture events that correspond to the length of the pilot sequence. It may then be reasonable to suggest that the effect of primary sequence would simply be to change the height and position of these events, with this assumption additionally being supported by the sequential per base pair model of unzipping described for ds-DNA.

(5) Two-dimensional free energy landscapes for internal and terminal base pairs in the pilot ds-PNA sequence demonstrate the existence of all three populated regions present in an unrelated ds-DNA sequence.[28] This indicates the conservation of these properties not only for different sequence lengths, but between nucleic acids of different backbone identity. This suggests that individual base pair free energy landscapes are likely to express similar populated regions for different possible test sequence.

A limitation of using only computational data, apart from the thermal melting experiments conducted in Chapter 4, is that ‘a model is just a model’. With any computational study, experimental validation is important since the model can deviate from reality and error compensation may play a role in agreement even in the absence of deviation. Consequently, the present work is validated against experimental data structurally, enthalpically, on the basis of free energy and in its replication of force-loading behaviour consistent with theory:

(1) Structurally, the present work reproduces the base pairing and stepping parameters from previous literature,[1,12,29] as seen in Table 3-1. In addition, a low RMSD between CHARMM models of the 3MBS[29] ds-PNA structure and its experimentally obtained crystal structure is observed. Likewise, the Martini model of the ds-PNA GTAGATCACT has a similarly low RMSD

when compared against its equivalent CHARMM model, as discussed in Chapter 6.

(2) Enthalpically, the MM-GBSA studies used in the present work can predict experimental data with an error of 8.7% as discussed in Chapter 4.

(3) On the basis of free energy, the binding free energy of GTAGATCACT determined using AWH (-32.15 kcal/mol) is in reasonable agreement with the experimental data from Table 1-3 or Table 4-3 (-18.99 kcal/mol).

(4) Finally, force spectroscopy models predict a transition to a nonlinear force-loading curve at high loading rates, and this nonlinearity is demonstrated in the present work for both CHARMM and Martini models.

Beyond validation and the use of a single pilot sequence, the present work is limited in that lower loading rates were inaccessible with both the CHARMM and Martini models, preventing the fitting of force spectroscopy models. Fitting at these lower loading rates is useful since, at high loading rates, the force-loading curve covers only a far-from-equilibrium regime whereas the near-equilibrium regime must be characterised for accurate determination of parameters like the equilibrium free energy.[30] Initially, the present work intended to incorporate atomic force microscopy as a primary tool to characterising these lower loading rate regimes, but access to this equipment was lost within the first several months of the present work due to equipment failure. This equipment failure is what motivated the shift to a computational project since molecular dynamics enabled the replication of force spectroscopy experiments through simulation.

7.1.4 Position within the established literature

Previous literature on ds-PNA binding properties have, to the awareness of the candidate, exclusively focused on the derivation of the equilibrium binding properties of individual strands. This has included the acquisition of binding

entropies, enthalpies and free energies from thermal melting experiments,[4,5,6] the determination of the effect of mismatching on these stabilities,[31] and structural studies such as the induction of backbone chirality.[32]

The present work expands on this previous literature on ds-PNA binding properties in that it:

(1) Provides a model for the prediction of binding stability, in this case the binding enthalpy, as a function of ds-PNA primary sequence alone. This work is therefore the first time that conclusions about the binding stabilities of ds-PNA have been made from quantities conserved between unrelated ds-PNA sequences. In this case, these quantities are the incremental binding enthalpies.

(2) Demonstrates the application of this nearest-neighbour model to ds-PNA oligonucleotides for the first time.

(3) Details the first ever force-loading curves for ds-PNA. This includes the first time that shearing and unzipping have been differentiated for ds-PNA and the first time that individual base pair rupture events have been demonstrated in the force traces during unzipping of a PNA-based oligonucleotide.

(4) Demonstrates evidence of overstretching in ds-PNA for the first time.

(5) Describes how the effect of external force on ds-PNA can be understood in terms of van der Waals, Coulomb and torsional energy contributions.

In consideration of the above points, the present work is the first comprehensive study of the dynamics of ds-PNA molecules under external forces at varying loading rates and on different axes.

The present work belongs to a number of studies on the derivation of nucleic acid nearest-neighbour models using computational methods, such as the study by Nishida *et al*. which used BAR to predict RNA secondary structures.[33] In particular, the present work exists in contrast to the study of Golyshev *et al*. by arguing that their linear correction is an inappropriate tool to derive incremental energies from whole-sequence energies derived using MM-PB[GB]SA.[7] In addition, the present work used AWH in a manner inspired by Lindahl *et al.*[28] who determined two-dimensional pairing and stacking free energy landscapes for internal base pairs in ds-DNA. By extending the method of Lindahl *et al*. to free energy landscapes at 2.5 nm and 4 nm extensions along a shearing reaction coordinate, the present work demonstrates the applicability of AWH to the study of conformations driven far from the native state. Like Lindahl *et al*., the present work demonstrated the same three populated regions in these free energy landscapes.

The largest contribution to any individual computational method made by the present work is arguably its provision of an entirely new molecule type to the existing Martini forcefields. The present work details bead assignments and types for the construction of a coarse-grained model of ds-PNA for Martini and, in Table 6-4, provides all bond, angle and dihedral equilibrium positions, force constants, multiplicities and types for the implementation of a Martini ds-PNA model into Gromacs. The standard bead types presented in the original Martini paper[16] are utilised for direct compatibility with the other Martini models.[13,14,17] The present work's Martini model for ds-PNA utilises the same elastic network method described in the prior literature,[13,14] though differs from this existing body of work by implementing a Morse potential[15] for the more accurate derivation of base pairing free energy landscapes.

Finally, the present work validates the all-atom CHARMM model of ds-PNA developed by Jasiński *et al*.[12] first by demonstrating the reproducibility of their helical parameters and then by demonstrating that their model can capably reproduce binding enthalpies (Chapter 4) and free energies (Figure 5-18) from

experimental data. The present work then extends their CHARMM model by demonstrating its application to SMD simulations for the first time.

7.2 Future work

There are two main avenues by which the present work can be extended. Firstly, it can be extended by further characterising single-molecular properties. Secondly, it can be extended by scaling these observations into the macroscale and carrying out real-world experiments on a ds-PNA bioadhesive.

Regarding further work on the single-molecular properties:

(1) The partial force-loading curve in Figure 6-11 could be completed using atomic force microscopy. This would enable the near-equilibrium regime necessary for the fitting of force-spectroscopy models to be fitted,[30] and would therefore allow the apparent transition barrier heights and widths along the shearing coordinate to be studied. It would also provide an additional means of validation by demonstrating the degree of overlap between the experimental and computational partial force-loading curves. This was the initial intention of the present work, though could not be completed given the unavailability of the in-house atomic force microscope.

(2) SMD simulations with different ds-PNA sequences than GTAGATCACT could be conducted. This is since the present work relies on the assumption that the properties of ds-GTAGATCACT are transferable to other sequences. The validity of this assumption was argued for, however, in subchapter 7.1.3. Testing other pilot sequences could enable the generalisation of rupture forces as a function of primary sequence in the spirit of the nearest-neighbour model. This would involve the simulated pulling of a number of different primary sequences, the acquisition of their rupture forces and loading rates, and then solving a multiple linear regression model of the form in Equation 4-4. To the awareness of the candidate, the nearest-neighbour

model has never been applied in this manner and using it in such a way could be extremely beneficial in the design of nucleic acid bioadhesives[34] by allowing the prediction of force-loading curves as a function of primary sequence without experimentation.

Regarding further work on scaling beyond the single-molecular properties, the scalability of single-molecular interactions has been studied in the past.[35] An approximately linear scaling of the specific single-molecular interactions towards the mesoscale was identified, though there were also non-specific effects. The advantage of having already conducted the single-molecular experiments is that such an experiment on ds-PNA could be more readily achieved. This could involve measuring the rupture force between surfaces of known area functionalised with complementary ds-PNA at a known density and then deriving a relationship between ds-PNA density, single-molecular rupture force and multi-molecular rupture pressure.

In addition to adhering two flat surfaces with ds-PNA, a similar implementation to the DNA adhesive gels could be explored. In the DNA adhesives, beads were functionalised with complementary binding partners to form a colloid.[34] This enabled complex formation to be studied using optical measuring equipment, for example by fluorescently labelling the beads and then visualising them using fluorescence microscopy.

Currently, the present work on ds-PNA is far removed from the intended application as a candidate bioadhesive. However, since the single-molecular properties have been resolved, experiments on the macroscale can now be conducted on ds-PNAs and interpretations can be made with respect to their single-molecular behaviour. This protocol of characterising single-molecular properties using molecular dynamics is affordable and available given the contemporary broad accessibility of computational power and does not rely on specialist equipment.

Consequently, greater understanding of bioadhesives can be achieved by adopting a similar molecular dynamics approach, enabling quantities that contribute to the binding stability of these bioadhesives under external force to be derived. These quantities, in turn, may be used to inform experimental design or to screen bioadhesives for particular properties.

7.3 Conclusion

This thesis has presented advancements in the understanding of the binding stabilities of ds-PNA under equilibrium and external force. These understandings are both generally relevant to the development of ds-PNA technologies and of direct interest to the characterisation of other candidate bioadhesives using broadly available technology. In addition to the presentation of new knowledge on the single-molecular properties of ds-PNA, two new tools have been presented. The first of these is the nearest-neighbour model for ds-PNA. The second is the coarse-grained Martini model for ds-PNA. The nearest-neighbour model expedites sequence selection by allowing ds-PNA binding enthalpies to be predicted from primary sequence alone. The Martini model allows longer simulation timescales and larger simulation volumes to be studied and can replicate the rupture force vs loading rate behaviour of an all-atom representation of Martini ds-PNA for the high loading-rate regime.

The central research question, “Can molecular dynamics be used to assess and screen the single-molecular binding properties of a candidate bioadhesive?", was answered. Molecular dynamics was able to assess a candidate bioadhesive, this being ds-PNA, and a model was developed from these simulations such that ds-PNAs can be screened on the basis of their binding enthalpy. The nuance to this research question is that it assumes that understanding these single-molecule binding properties is useful to developing a candidate bioadhesive. It is argued here that this is the case since these properties are necessarily resolved for

distinguishments between single-molecular specific and macromolecular non-specific effects to be made in any candidate bioadhesive.

This thesis has generated new insights regarding the stabilities of ds-PNA at equilibrium and under external force and new tools that enable the rapid collection and interpretation of these insights. This thesis is of general interest to the development of PNA technologies by characterising their single-molecular behaviour using structural properties that are conserved regardless of sequence identity and by demonstrating how external force affects their binding properties. In addition, this thesis challenges previous assumptions by refuting the applicability of a linear MM-GBSA correction[7] for the derivation of the nearest-neighbour model and by introducing an alternative Morse parameterisation in the Martini model[13,14] to characterise base pairing free energy curves more accurately.

Finally, the methods used in this thesis may be extended to other bioadhesives which are missing single-molecular characterisations though their exact implementation may differ. For example, a protein bioadhesive does not adhere in the same manner as a nucleic acid and decomposing its binding properties into monomeric (amino acid) contributions, as per the nearest-neighbour model, may not be useful. The results of this thesis can be used to inform the development of ds-PNA bioadhesives by providing a single-molecular framework through which macroscopic observables may be understood.

7.4 References

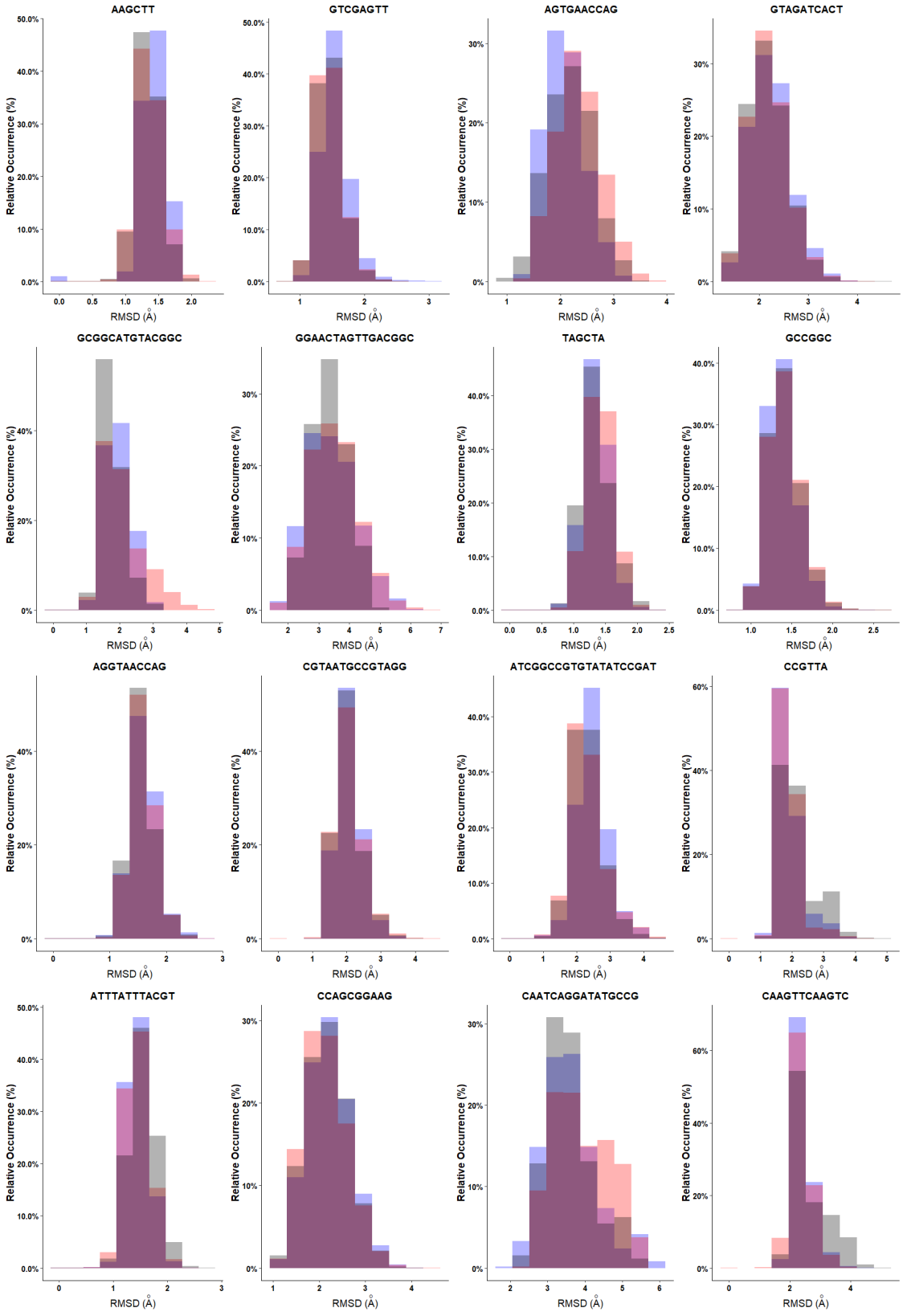
[1] SantaLucia Jr., J. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl. Aca. Sci.* **1998**, *95*, 1460-1465
[2] SantaLucia Jr., J.; Allawi, H.; Seneviratne, P. Improved nearest-neighbor parameters for predicting DNA duplex stability. *Biochemistry* **1996**, *35*, 3555-62
[3] Wittung, P.; Nielsen, P.; Buchardt, O.; Egholm, M.; Nordén, B. DNA-like double helix formed by peptide nucleic acid. *Nature* **1994**, *368*, 561−563
[4] Tomac, S.; Sarkar, M.; Ratilainen, T.; Wittung, P.; Nielsen, P.; Nordén, B.; Gräslund, A. Ionic Effects on the Stability and Conformation of Peptide Nucleic Acid Complexes. *J. Am. Chem. Soc.* **1996**, *118*, 5544−5552

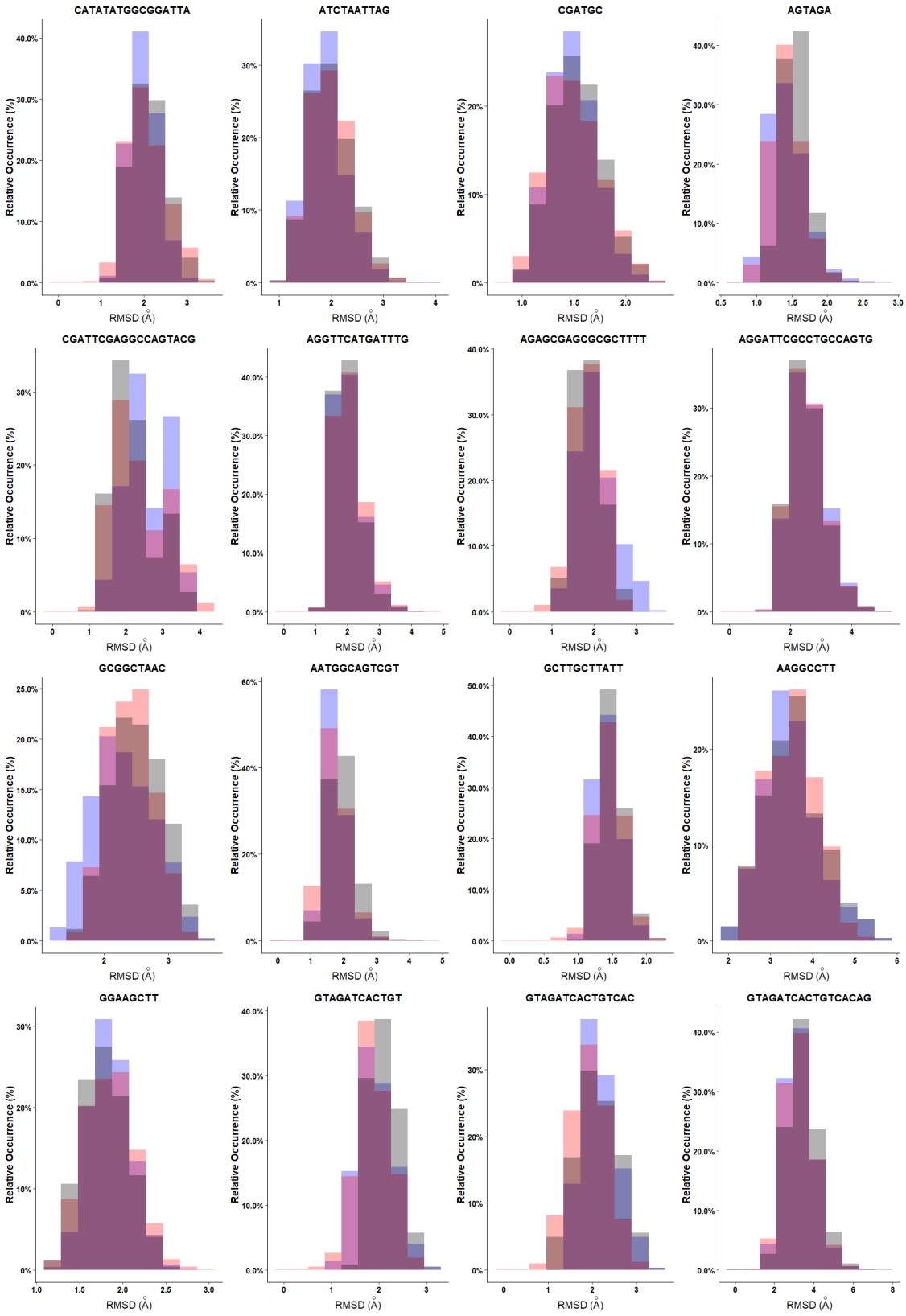
[5] Sen, A.; Nielsen, P. Unique Properties of Purine/Pyrimidine Asymmetric PNA·DNA Duplexes: Differential Stabilization of PNA· DNA Duplexes by Purines in the PNA Strand. *Biophys. J*. **2006**, *90*, 1329−1337
[6] Ratilainen, R.; Holmén, A.; Tuite, E.; Haaima, G.; Christensen, L.; Nielsen, P.; Nordén, B. Hybridization of Peptide Nucleic Acid. *Biochemistry* **1998**, *37*, 12331-12342
[7] Golyshev, V.; Pyshnyi, D.; Lomzov, A. Calculation of Energy for RNA/RNA and DNA/RNA Duplex Formation by Molecular Dynamics Simulation. *Molecular Biology* **2021**, *55*, 927-940
[8] Danilowicz, C.; Hatch, K.; Conover, A.; Ducas, T.; Gunaratne, R.; Coljee, V.; Prentiss, M. Study of force induced melting of dsDNA as a function of length and conformation. *J. Phys.: Condens. Matter* **2010**, *22*, 414106
[9] Morfil, J.; Kühner, F.; Blank, K.; Lugmaier, R.; Sedlmair, J.; Gaub, H. B-S Transition in Short Oligonucleotides. *Biophysical Journal* **2007**, *93*, 2400-2409
[10] Essevaz-Roulet, B.; Bockelmann, U.; Heslot, F. Mechanical separation of the complementary strands of DNA. *Proc. Natl. Acad. Sci.* **1997**, *94*, 11935-11940
[11] Bockelmann, U.; Essevaz-Roulet, B.; Heslot, F. Molecular Stick-Slip Motion Revealed by Opening DNA with Piconewton Forces. *Phys. Rev. Lett.* **1997**, *79*, 4489-4492
[12] Jasiński, M.; Feig, M.; Trylska, J. Improved Force Fields for Peptide Nucleic Acids with Optimized Backbone Torsion Parameters. *J. Chem. Theory Comput.* **2018**, *14*, 3603-3620
[13] Uusitalo, J.; Ingólfsson, H.; Akhshi, P.; Tieleman, D.; Marrink, S. Martini Coarse-Grained Force Field: Extension to DNA. *J. Chem. Theory Comput.* **2015**, *11*, 3932-3945
[14] Uusitalo, J.; Ingólfsson, H.; Marrink, S.; Faustino, I. Martini Coarse-Grained Force Field: Extension to RNA. *Biophysical Journal* **2017**, *113*, 246-256
[15] Costa Filho, R.; Alencar, G.; Skagerstam, B.; Andrade Jr., J. Morse potential derived from first principles. *Europhysics Letters* **2013**, *101*, 10009
[16] Marrink, S.; Risselada, H.; Yefimov, S.; Tieleman, D.; de Vries, A. The MARTINI Force Field: Coarse Grained Model for Biomolecular Simulations. *J. Phys. Chem. B* **2007**, *111*, 7812-7824
[17] Monticelli, L.; Kandasamy, S.; Periole, X.; Larson, R.; Tieleman, D.; Marrink, S. The MARTINI Coarse-Grained Force Field: Extension to Proteins. *J. Chem. Theory Comput.* **2008**, *4*, 819-834
[18] Lee, H.; Scherer, N.; Messersmith, P. Single-molecule mechanics of mussel adhesion. *Proc. Natl. Acad. Sci.* **2006**, *103*, 12999-13003
[19] Sundaram, M.; Kaliannagounder, V.; Selvaprithiviraj, V.; Suresh, M.; Biswas, R.; Vasudevan, A.; Varma, P.; Jayakumar, P. Bioadhesive, Hemostatic and Antibacterial *in Situ* Chitin-Fibrin Nanocomposite Gel for Controlling Bleeding and Preventing Infections at Mediastinum. *ACS Sustainable Chem. Eng.* **2018**, *6*, 7826-7840
[20] Florin, E.; Moy, V.; Gaub, H. Adhesion Forces Between Individual Ligand-Receptor Pairs. *Science* **1994**, *264*, 415-417
[21] Strunz, T.; Oroszlan, K.; Schäfer, R.; Güntherodt, H. Dynamic force spectroscopy of single DNA molecules. *Proc. Natl. Acad. Sci. USA* **1999**, *96*, 11277-11282
[22] Wong, J.; Chilkoti, A.; Moy, V. Direct force measurements of the streptavidin-biotin interaction. *Biomolecular Engineering* **1999**, *16*, 45-55
[23] Smith, S.; Cui, Y.; Bustamante, C. Overstretching B-DNA: The Elastic Response of Individual Double-Stranded and Single-Stranded DNA Molecules. *Science* **1996**, *271*, 795-799
[24] Morfil, J.; Kühner, F.; Blank, K.; Lugmaier, R.; Sedlmair, J.; Gaub, H. B-S Transition in Short Oligonucleotides. *Biophysical Journal* **2007**, *93*, 2400-2409
[25] Ho, D.; Zimmermann, J.; Dehmelt, F.; Steinabach, U.; Erdmann, M.; Severin, P.; Falter, K.; Gaub, H. Force-Driven Separation of Short Double-Stranded DNA. *Biophysical Journal* **2009**, *97*, 3158-3167
[26] De Gennes, P. Maximum pull out force on DNA hybrids. *Comptes Rendus de l'Académie des Sciences – Series IV – Physics* **2001**, *2*, 1505-1508
[27] Hatch, K.; Danilowicz, C.; Coljee, V.; Prentiss, M. Demonstration that the shear force required to separate short double-stranded DNA does not increase significantly with sequence length for sequences longer than 25 base pairs. *Phys. Rev. E.* **2008**, *78*, 011920
[28] Lindahl, V.; Villa, A.; Hess, B. Sequence dependency of canonical base pair opening in the DNA double helix. *PLoS Comput. Biol.* **2017**, *13*, e1005463
[29] Yeh, J.; Pohl, E.; Truan, D.; He, W.; Sheldrick, G.; Du, S.; Ashim, C. The Crystal Structure of Non – Modified and Bipyridine – Modified PNA Duplexes. *Chem. – Eur. J.* **2010**, *16*, 11867-11875
[30] Noy, A.; Friddle, R. Practical single molecule force spectroscopy: How to determine fundamental thermodynamic parameters of intermolecular bonds with an atomic force microscope. *Methods* **2013**, *60*, 142-150
[31] Jensen, K.; Ørum, H; Nielsen, P.; Nordén, B. Kinetics for Hybridization of Peptide Nucleic Acids (PNA) with DNA and RNA Studied with the BIAcore technique. *Biochemistry* **1997**, *36*, 5072-5077
[32] Wittung, P.; Eriksson, W.; Lyng, R.; Nielsen, P.; Nordén, B. Induced Chirality in PNA-PNA Duplexes. *J. Am. Chem. Soc.* **1995**, *117*, 10168-10173

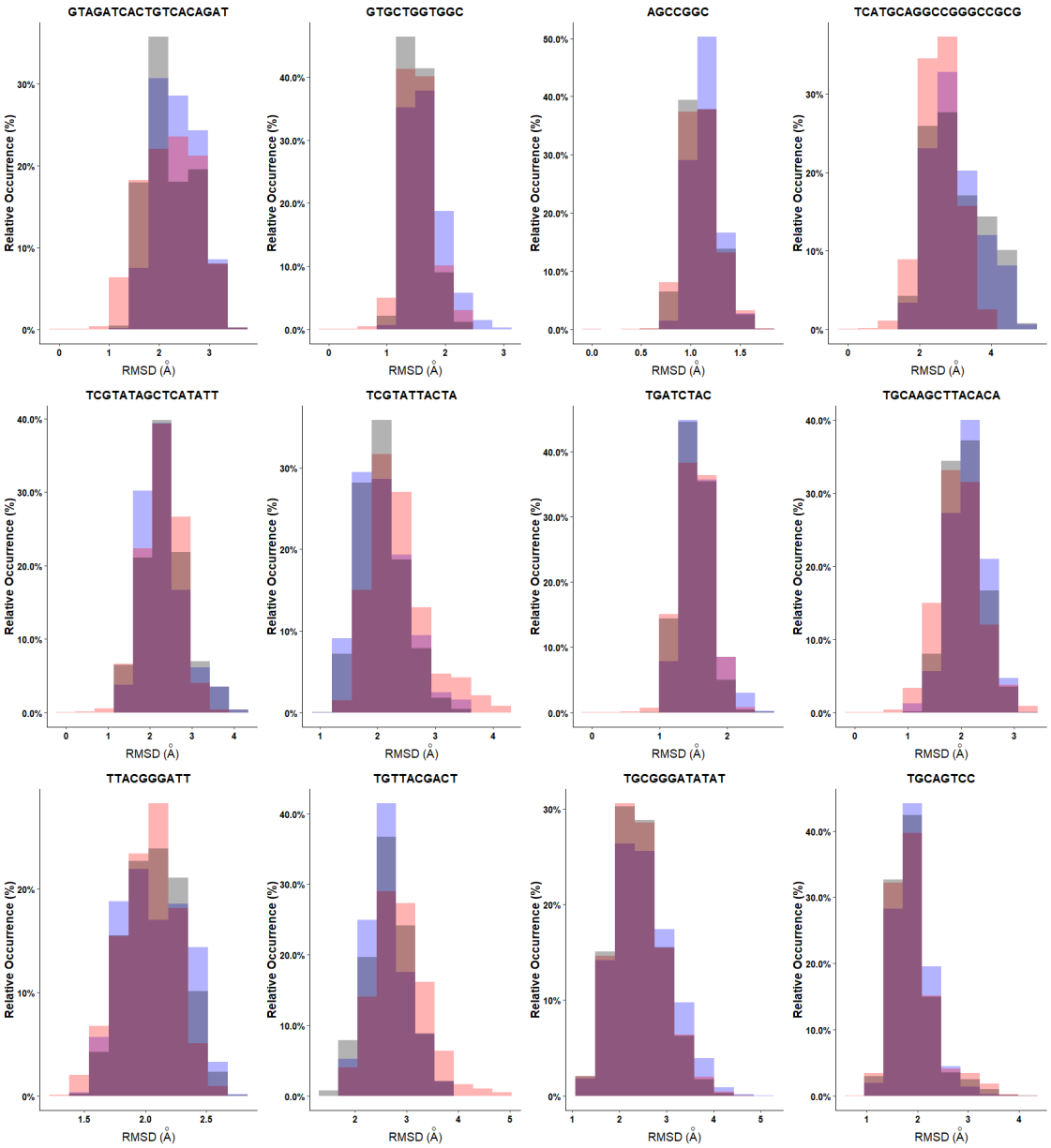
[33] Nishida, S.; Sakuraba, S.; Asai, K., Hamada, M. Estimating energy parameters for RNA secondary structure predictions using both experimental and computational data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **2019**, *16*, 1645-1655
[34] Allen, P.; Khaing, Z.; Schmidt, C.; Ellington, A. 3D Printing with Nucleic Acid Adhesives. *ACS Biomater. Sci. Eng.* **2015**, *1*, 19-26
[35] Raman, S.; Utzig, T.; Baimpos, T.; Shrestha, B.; Valtiner, M. Deciphering the scaling of single-molecule interactions using Jarzynski's equality. *Nature Communications* **2014**, *5*, 5539

Appendix

A.1 Chapter 4: Convergence metrics




CATATATGGCGGATTA
ATCTAATTAG
CGATGC
AGTAGA
CGATTCGAGGCCAGTACG
AGGTTCATGATTTG
AGAGCGAGCGCGCTTTT
AGGATTCGCCTGCCAGTG
GCGGCTAAC
AATGGCAGTCGT
GCTTGCTTATT
AAGGCCTT
GGAAGCTT
GTAGATCACTGT
GTAGATCACTGTCAC
GTAGATCACTGTCACAG
Relative Occurrence (%)
RMSD (Å)


GTAGATCACTGTCACAGAT
GTGCTGGTGGC
AGCCGGC
TCATGCAGGCCGGGCCGCG
TCGTATAGCTCATATT
TCGTATTACTA
TGATCTAC
TGCAAGCTTACACA
TTACGGGATT
TGTTACGACT
TGCGGGATATAT
TGCAGTCC
Relative Occurrence (%)
RMSD (Å)

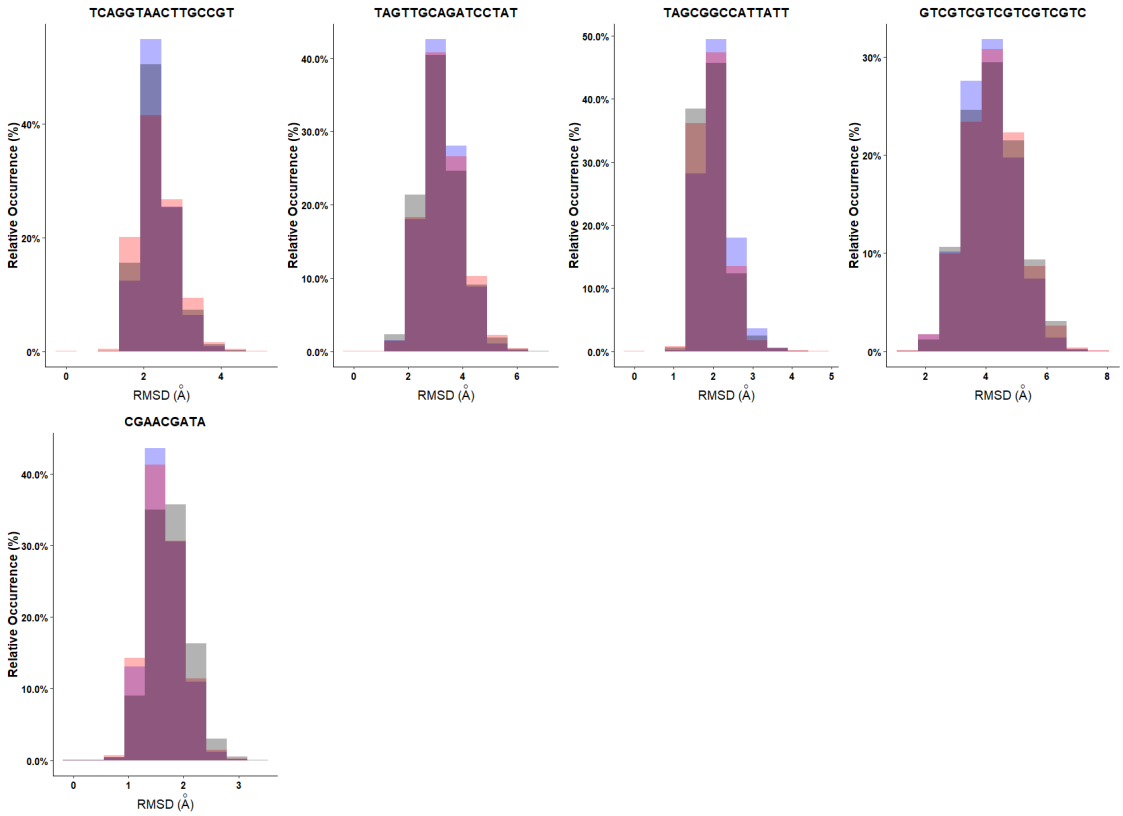


Figure A-1: RMSD distribution histograms for each sequence overlayed with its replicas. Independent runs are coloured differently. Sequence names are those of one strand and refer to a duplex whose opposing strand is its exact complement.

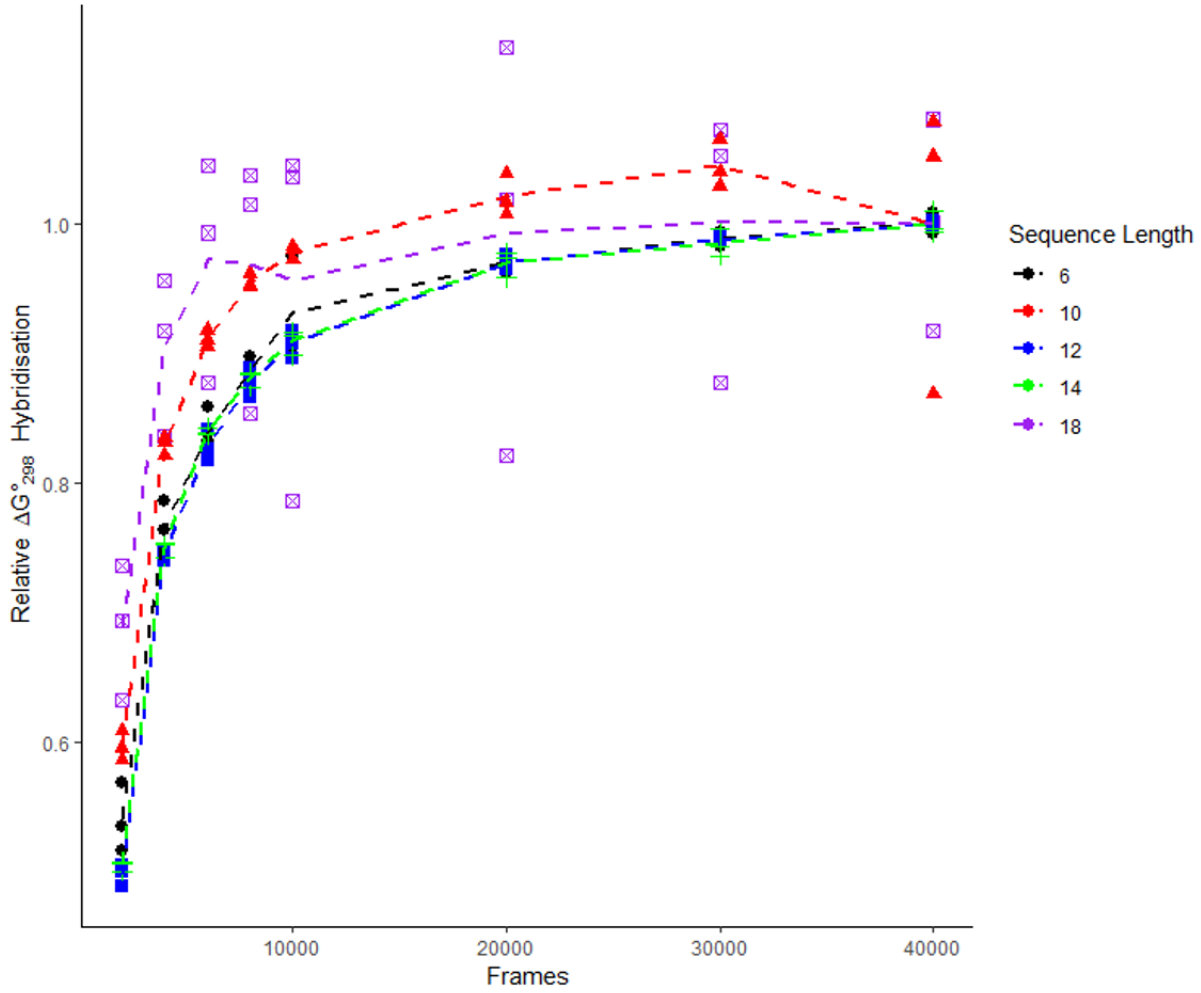


Figure A-2: Convergence of MM-GBSA analysis of binding free energy for sequences of different length against the number of frames used. 40,000 frames corresponded to 80 ns of simulation.

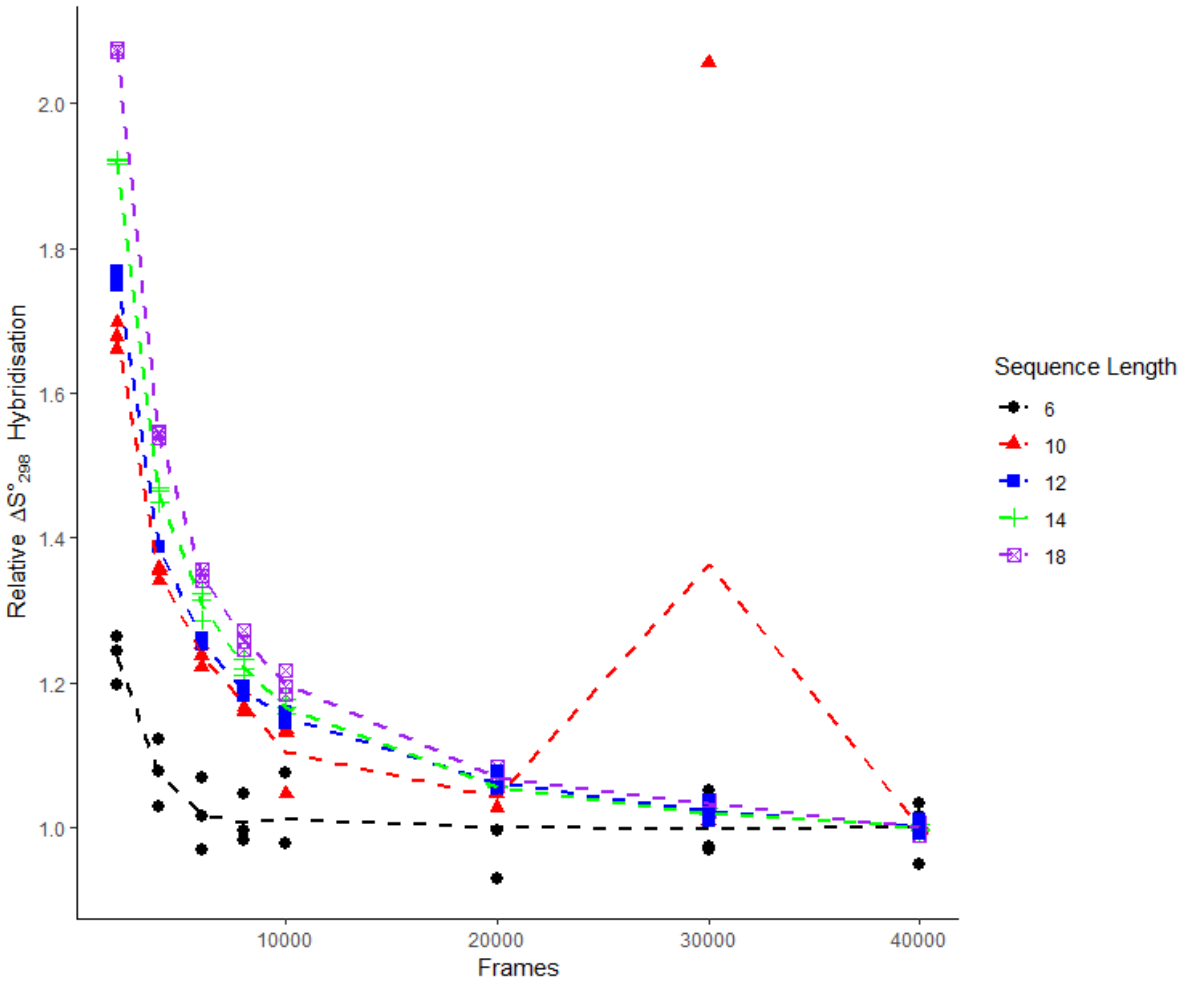


Figure A-3: Convergence of QH analysis of binding entropy for sequences of different length against the number of frames used. 40,000 frames corresponded to 80 ns of simulation.

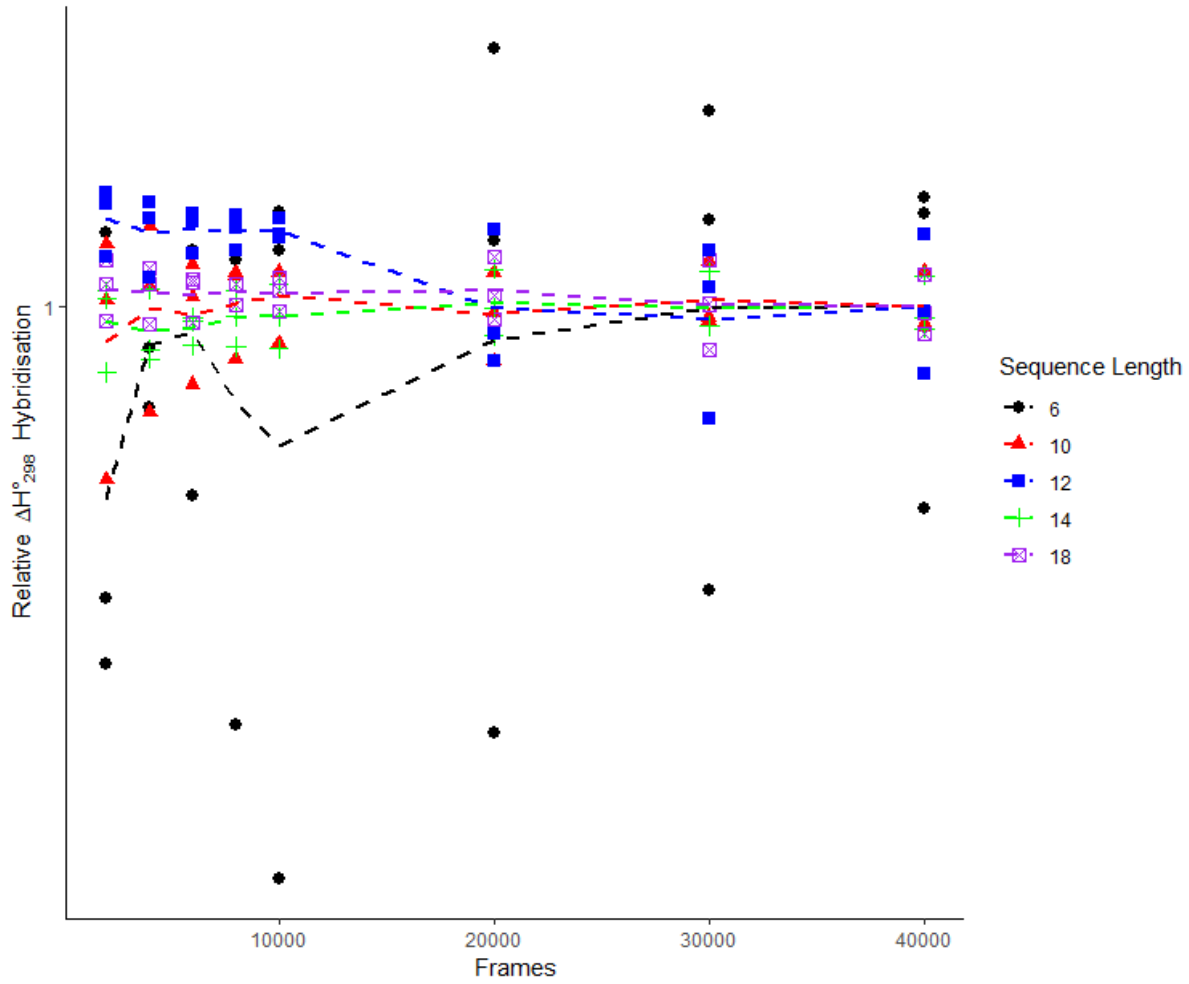


Figure A-4: Convergence of MM-GBSA analysis of binding enthalpy for sequences of different length against the number of frames used. 40,000 frames corresponded to 80 ns of simulation.

A.2 Chapter 4: Binding enthalpies of all 49 ds-PNAs

Table A-1: MM-GBSA binding enthalpies, entropies and free energies for all simulated homoduplexes. Standard errors obtained from triplicates.

ds-PNA Sequence	$-\Delta G^{o}_{298}$ (kcal mol^{-1})		$-\Delta H^{o}_{298}$ (kcal mol^{-1}))		$-\Delta S^{o}_{298}$ (cal K^{-1}mol^{-1})	
AACTAGATACAGTAATTA	96.87	± 1.45	161.00	± 2.63	227.03	± 0.81
AAGGCCTT	47.62	± 0.86	72.98	± 0.37	94.87	± 1.99
AATGGCAGTCGT	69.54	± 0.49	108.36	± 0.67	141.00	± 1.26
AGGATTCGCCTGCCAGTG	125.67	± 1.30	173.80	± 2.10	175.52	± 1.70
AGGTTCATGATTTG	88.73	± 0.22	125.46	± 0.41	134.44	± 0.69
AGTAGA	34.02	± 0.74	49.99	± 0.17	51.08	± 10.0
ATCTAATTAG	58.88	± 0.30	85.28	± 0.29	98.88	± 0.69
CAAGTTCAAGTC	74.06	± 1.16	106.08	± 0.97	118.31	± 2.05
CCAGCGGAAG	69.93	± 0.98	97.89	± 0.55	104.15	± 1.76
CGTAATGCCGTAGG	96.10	± 1.06	134.36	± 2.44	139.62	± 2.31
GCCGGC	42.46	± 0.27	57.80	± 0.58	58.96	± 12.9
GGAACTAGTTGACGGC	108.71	± 0.63	152.37	± 0.51	157.98	± 1.59
GTAGATCACT	60.34	± 0.62	87.45	± 0.66	101.19	± 0.97
GTCGAGTT	46.84	± 0.61	68.60	± 0.16	82.83	± 1.88
GTCGTCGTCGTCGTCGTC	118.60	± 0.65	168.97	± 0.56	180.82	± 1.25
TAGCGGCCATTATT	92.43	± 0.52	129.39	± 0.26	135.20	± 1.35
TAGTTGCAGATCCTAT	102.62	± 0.96	144.36	± 1.08	151.52	± 2.08
TCAGGTAACTTGCCGT	108.21	± 1.81	150.012	± 2.47	151.71	± 3.48
TGCAGTCC	50.05	± 1.28	72.56	± 1.03	85.30	± 3.49
TGCGGGATATAT	77.87	± 1.03	109.42	± 0.19	116.68	± 2.81
TGTTACGACT	49.00	± 1.17	85.23	± 0.84	131.83	± 2.92
CGAACGATA	50.94	± 0.75	78.99	± 0.05	103.27	± 2.50
ATTTATTTACGT	65.30	± 0.73	101.84	± 0.35	132.69	± 2.51
GCTTGCTTATT	63.95	± 0.73	95.03	± 0.13	114.10	± 3.69
TTACGGGATT	63.68	± 7.13	92.16	± 0.44	105.04	± 4.94
AGAGCGAGCGCGCTTTT	105.22	± 1.04	157.66	± 0.88	186.88	± 0.59
CAATCAGGATATGCCG	103.24	± 0.64	150.92	± 0.17	170.85	± 1.91
ATCGGCCGTGTATATCCGAT	129.83	± 0.51	189.45	± 0.40	211.47	± 1.17
TCGTATTACTA	64.01	± 1.27	95.44	± 0.76	115.27	± 3.10
CATATATGGCGGATTA	101.06	± 0.90	147.51	± 0.39	166.71	± 1.69
GTGCTGGTGGC	74.16	± 0.87	105.21	± 0.82	113.98	± 4.26
TGCAAGCTTACACA	85.38	± 1.83	126.30	± 0.09	147.82	± 7.02
CGATTCGAGGCCAGTACG	122.22	± 0.33	174.14	± 0.67	185.38	± 0.71
GCGGCATGTACGGC	92.61	± 0.57	134.69	± 0.30	151.78	± 1.92
TCATGCAGGCCGGCCGCG	127.66	± 0.80	181.31	± 1.93	191.14	± 2.87
TCGTATAGCTCATATT	95.53	± 1.59	141.82	± 0.34	166.18	± 3.58
GCGGCTAAC	56.61	± 0.79	83.37	± 0.35	98.97	± 1.62
CGATCG	34.72	± 0.75	53.08	± 0.21	69.94	± 2.04
AACGTT	28.94	± 1.57	48.34	± 0.58	73.48	± 5.34

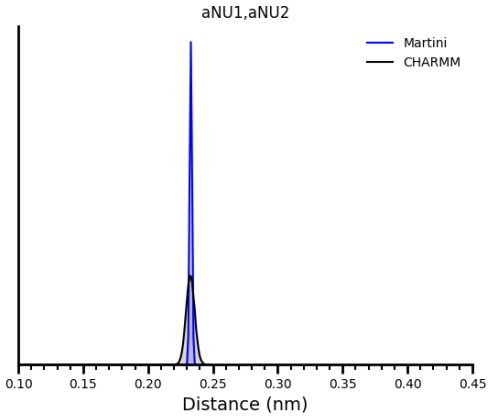
TAGCTA	32.13	± 0.49	50.40	± 0.97	69.62	± 2.14
AGGTAACCAG	64.33	± 0.03	94.98	± 0.53	112.32	± 1.94
AGTGAAGCAG	61.37	± 0.86	92.00	± 0.33	112.24	± 1.77
TGATCTAC	43.82	± 1.47	68.02	± 0.41	90.05	± 6.30
GTAGATCACTGT	71.33	± 1.78	106.82	± 0.69	129.15	± 1.20
GTAGATCACTGTCAC	90.83	± 1.67	135.41	± 0.47	160.28	± 1.17
GTAGATCACTGTCACAG	108.15	± 1.99	157.30	± 0.17	175.91	± 1.78
GTAGATCACTGTCACAGAT	117.26	± 0.46	173.53	± 0.65	200.04	± 5.23
GGAAGCTT	46.95	± 0.53	70.69	± 1.41	88.51	± 2.06
AGCCGGC	45.58	± 0.46	67.20	± 0.64	81.03	± 0.67

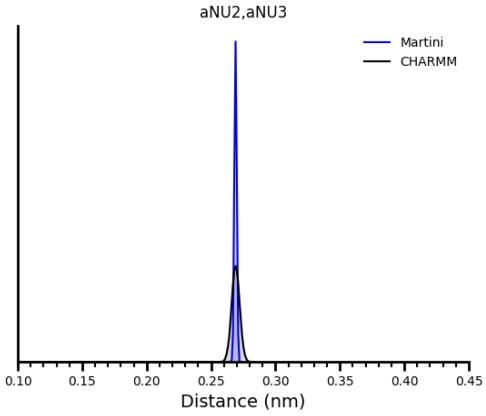
A.3 Chapter 6: ss-PNA sequences for bonded distributions

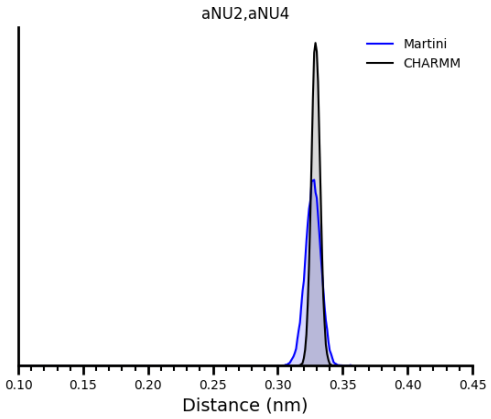
Table A-2: ss-PNA strands used to determine the bond, angle and dihedral distributions in all-atom and Martini PNA.

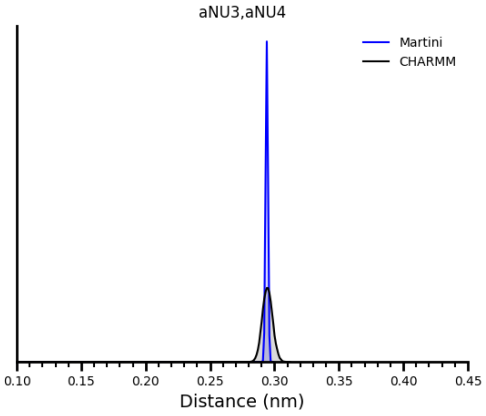
ss-PNA Sequence
AAAA
ACAC
AGAG
ATAT
CCCC
GCGC
GGGG
GTTG
TGGT
TTTT

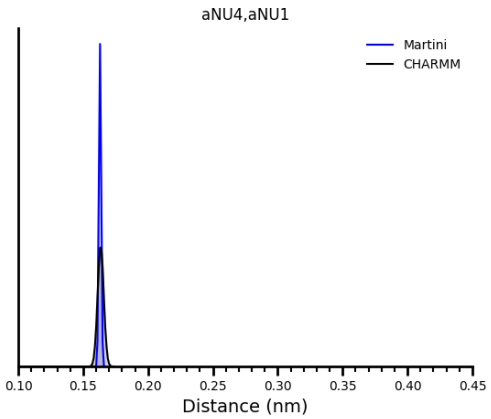
A.4 Chapter 6: Martini and CHARMM bonded distributions

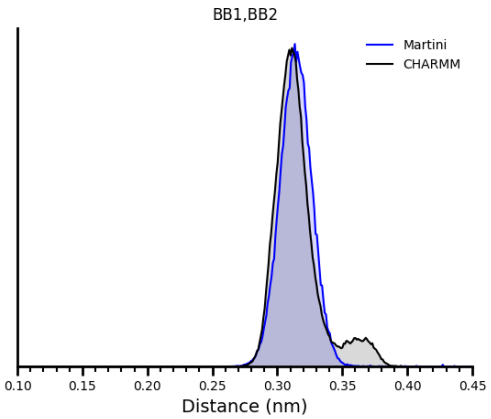


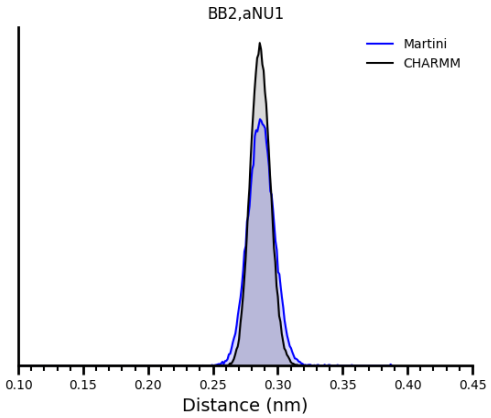


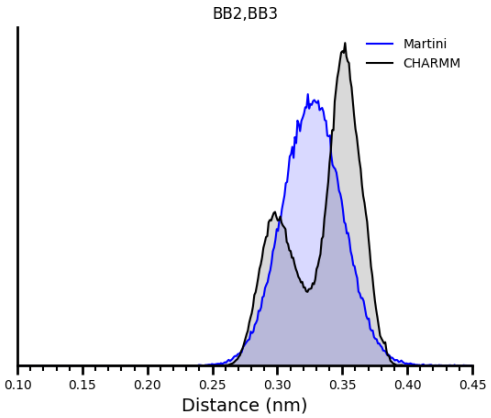


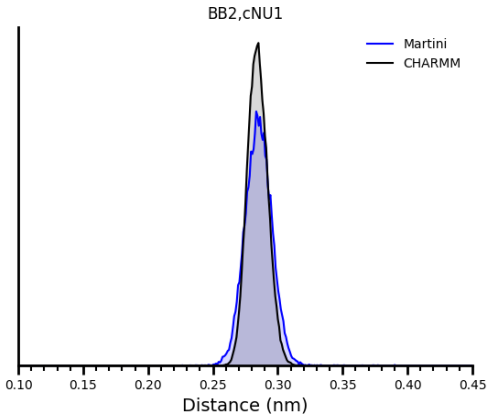


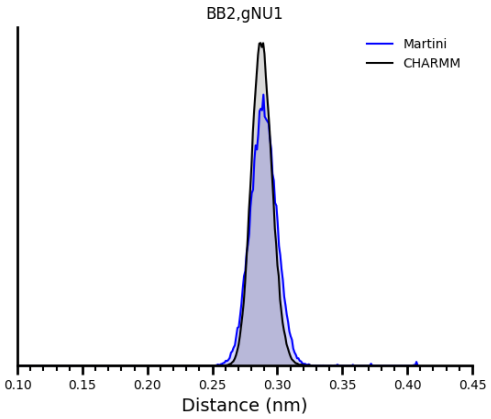

aNU4,aNU1
Martini
CHARMM
0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45
Distance (nm)

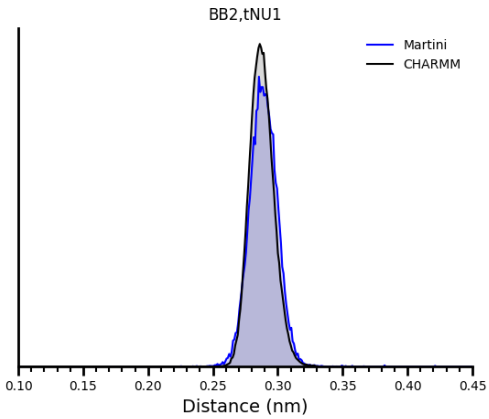

BB1,BB2
Martini
CHARMM
0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45
Distance (nm)

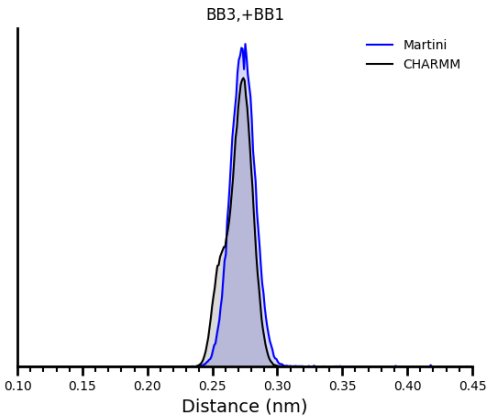

BB2,aNU1
Martini
CHARMM
0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45
Distance (nm)

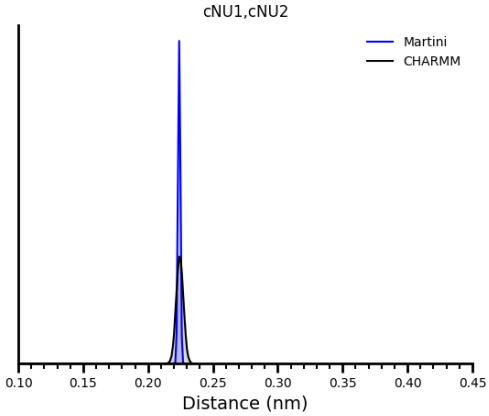

BB2,BB3
Martini
CHARMM
0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45
Distance (nm)

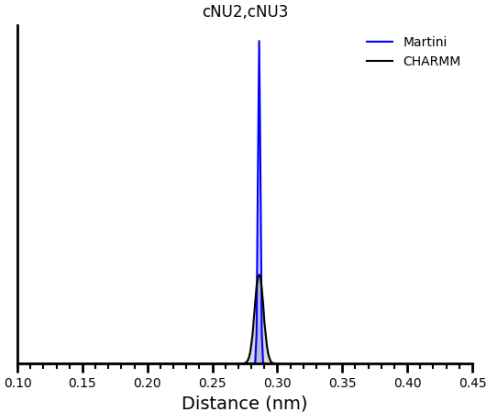

BB2,cNU1
Martini
CHARMM
0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45
Distance (nm)

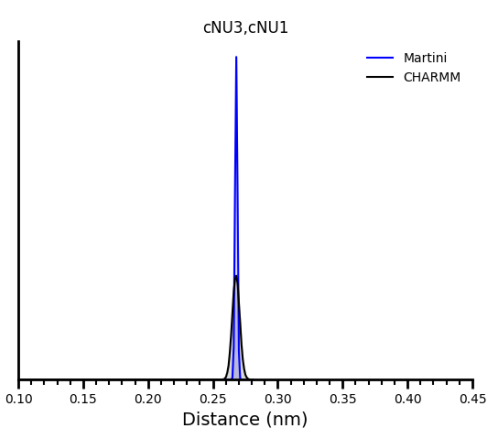

BB2,gNU1
Martini
CHARMM
0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45
Distance (nm)

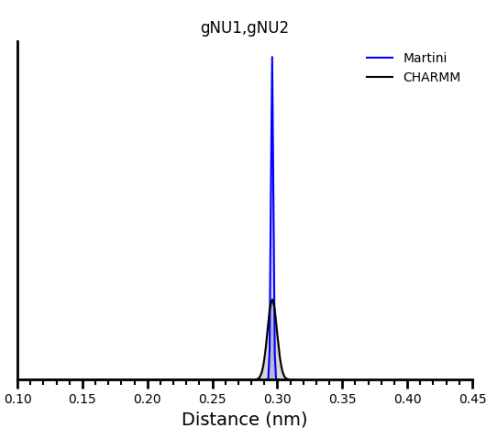

BB2,tNU1
Martini
CHARMM
0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45
Distance (nm)

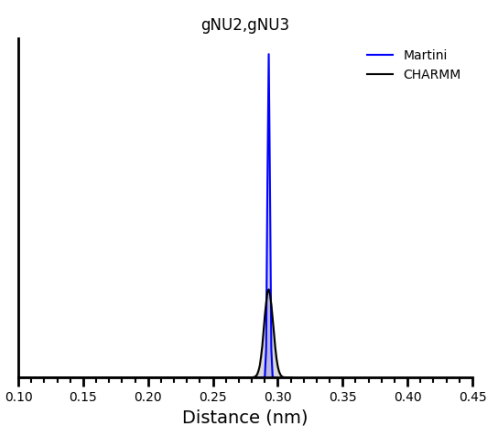

BB3,+BB1
Martini
CHARMM
0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45
Distance (nm)

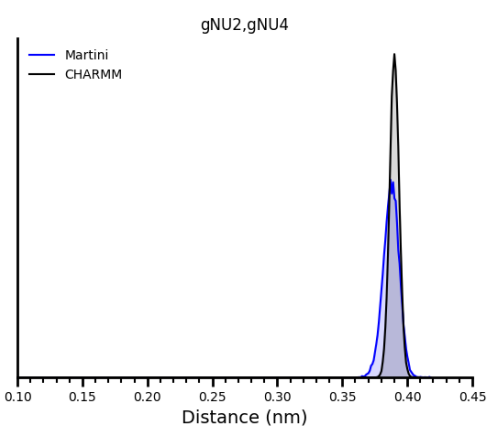

cNU1,cNU2
Martini
CHARMM
0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45
Distance (nm)


cNU2,cNU3
Martini
CHARMM
0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45
Distance (nm)


cNU3,cNU1
Martini
CHARMM
0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45
Distance (nm)


gNU1,gNU2
Martini
CHARMM
0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45
Distance (nm)


gNU2,gNU3
Martini
CHARMM
0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45
Distance (nm)


gNU2,gNU4
Martini
CHARMM
0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45
Distance (nm)

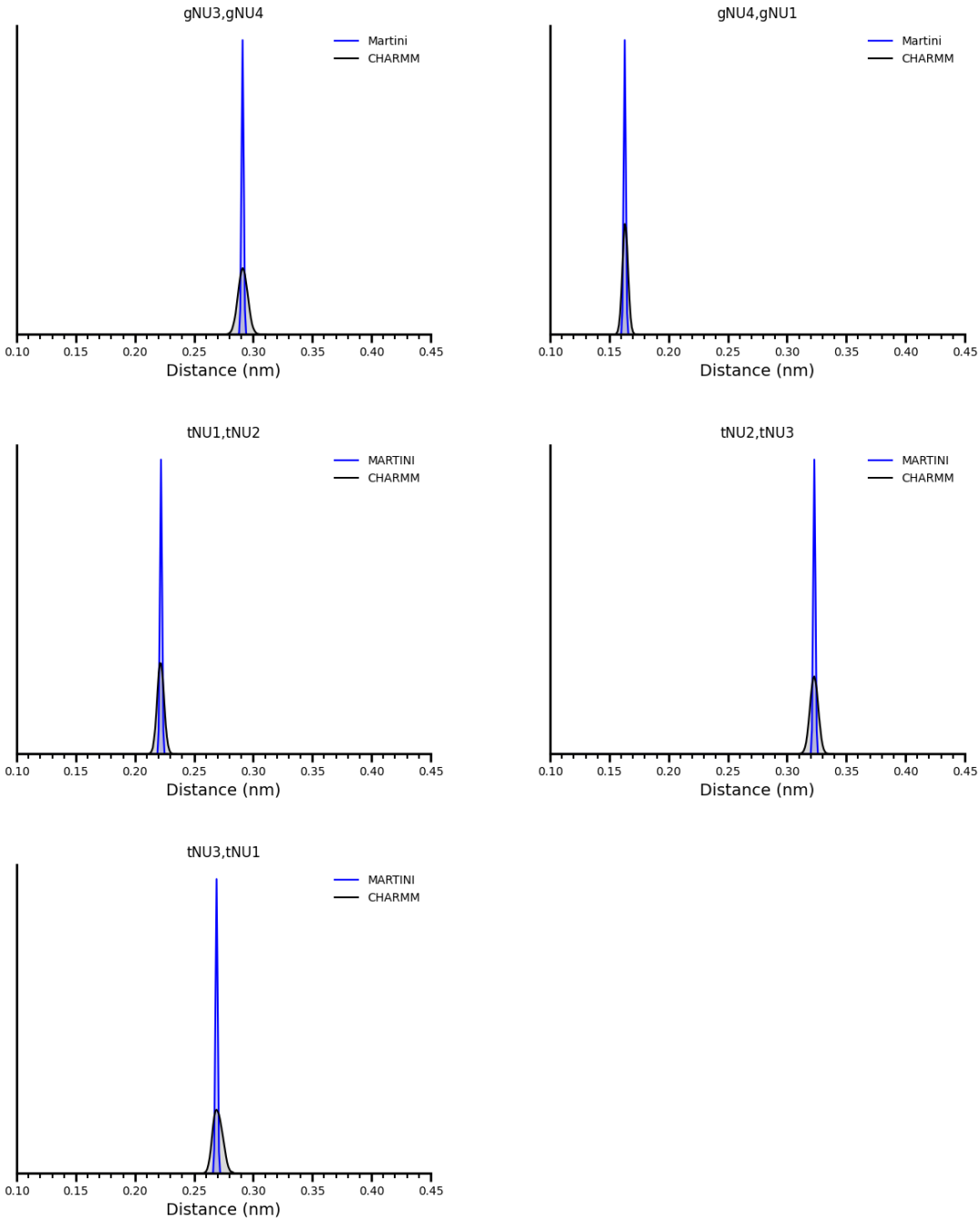
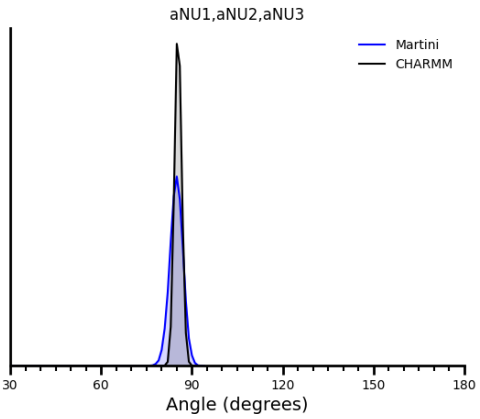
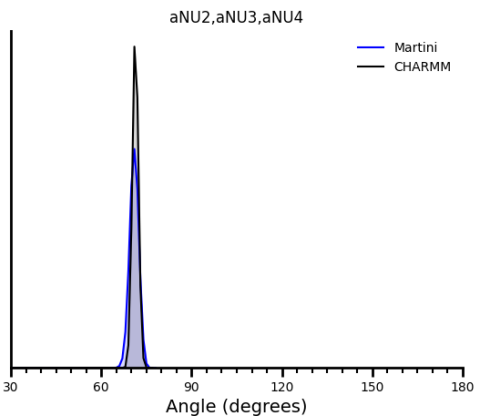
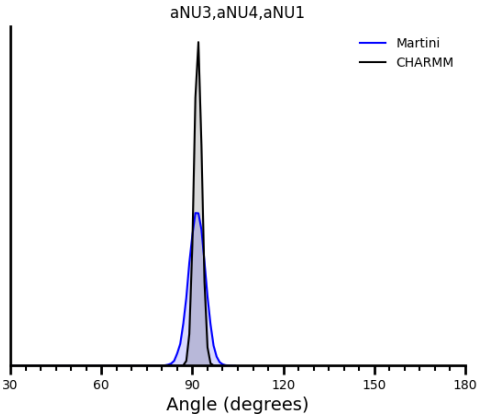
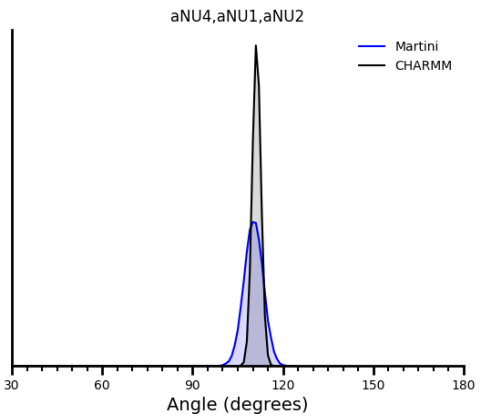


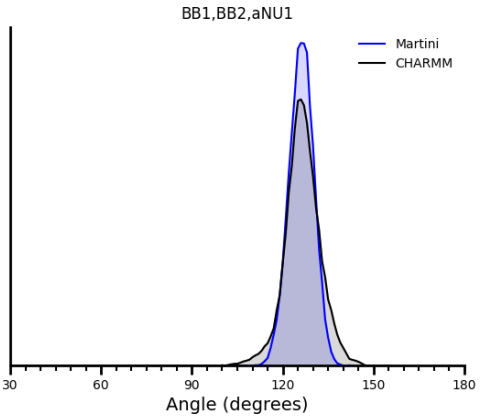
Figure A-5: Bond distributions of CHARMM and Martini simulations.

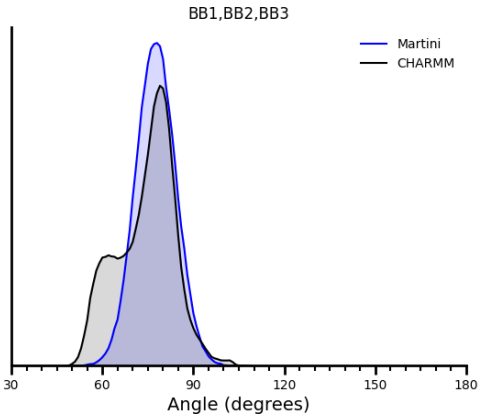

aNU1,aNU2,aNU3
Martini
CHARMM
30
60
90
120
150
180
Angle (degrees)

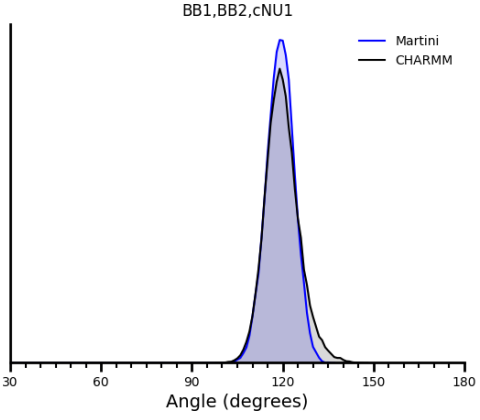

aNU2,aNU3,aNU4
Martini
CHARMM
30
60
90
120
150
180
Angle (degrees)

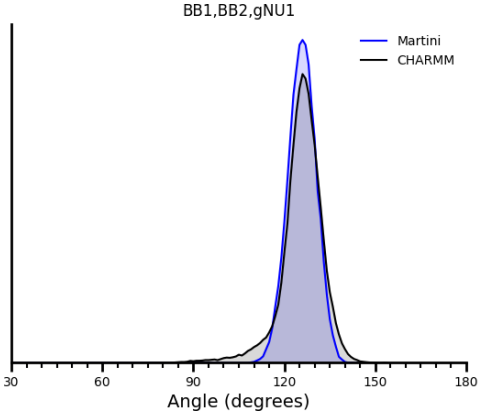

aNU3,aNU4,aNU1
Martini
CHARMM
30
60
90
120
150
180
Angle (degrees)

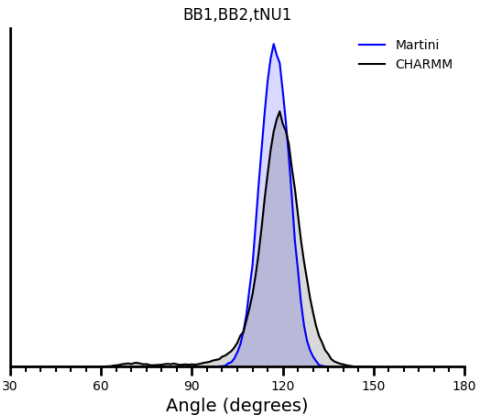

aNU4,aNU1,aNU2
Martini
CHARMM
30
60
90
120
150
180
Angle (degrees)

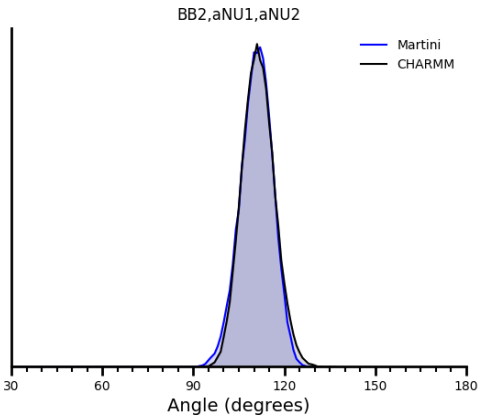

BB1,BB2,aNU1
Martini
CHARMM
30
60
90
120
150
180
Angle (degrees)

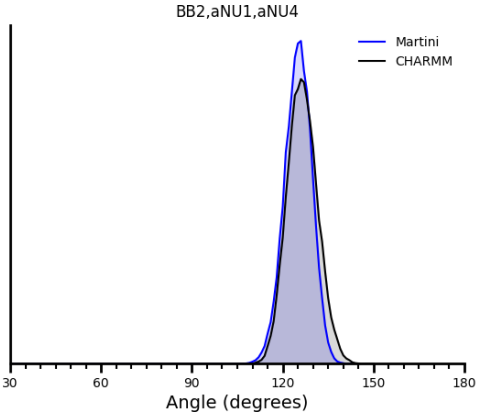

BB1,BB2,BB3
Martini
CHARMM
30
60
90
120
150
180
Angle (degrees)

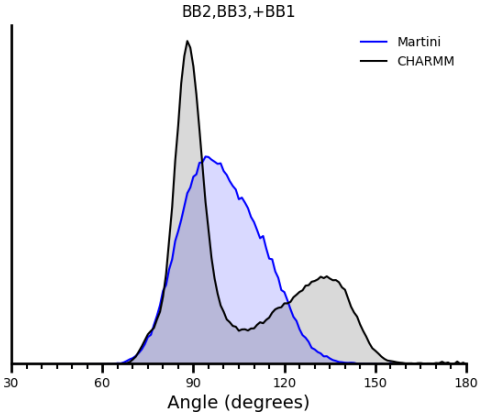

BB1,BB2,cNU1
Martini
CHARMM
30
60
90
120
150
180
Angle (degrees)

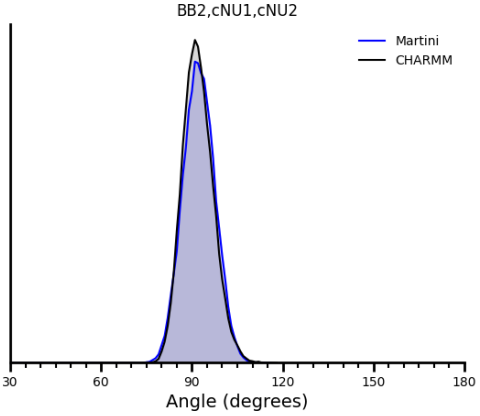

BB1,BB2,gNU1
Martini
CHARMM
30
60
90
120
150
180
Angle (degrees)

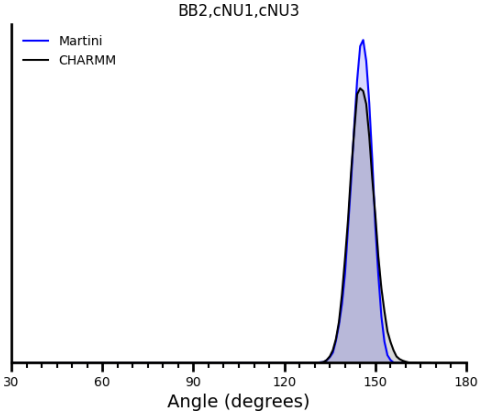

BB1,BB2,tNU1
Martini
CHARMM
30
60
90
120
150
180
Angle (degrees)

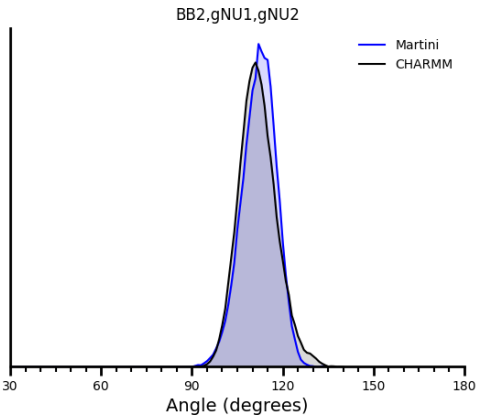

BB2,aNU1,aNU2
Martini
CHARMM
30
60
90
120
150
180
Angle (degrees)

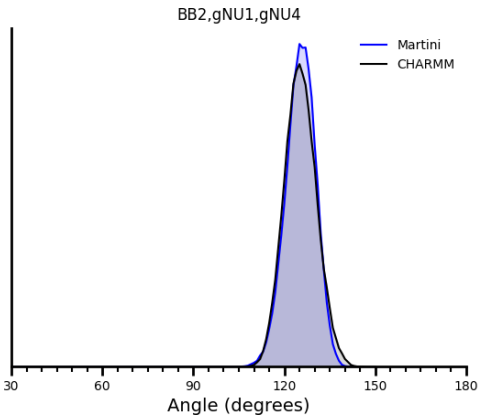

BB2,aNU1,aNU4
Martini
CHARMM
30
60
90
120
150
180
Angle (degrees)

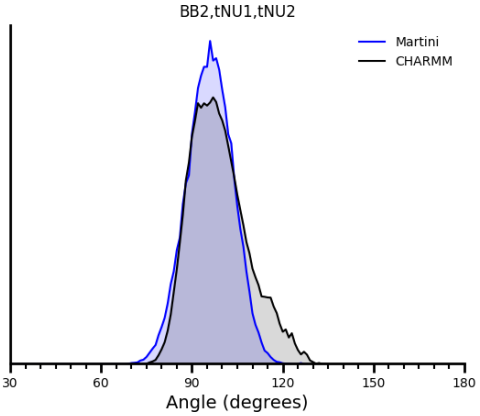

BB2,BB3,+BB1
Martini
CHARMM
30
60
90
120
150
180
Angle (degrees)

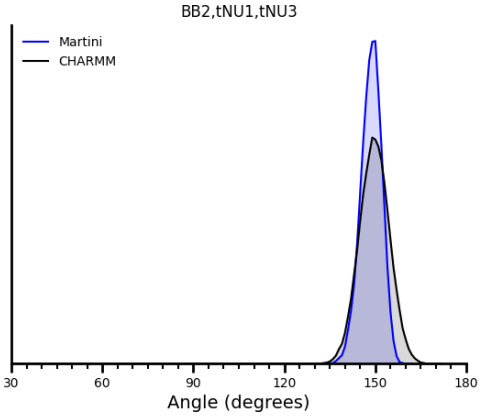

BB2,cNU1,cNU2
Martini
CHARMM
30
60
90
120
150
180
Angle (degrees)

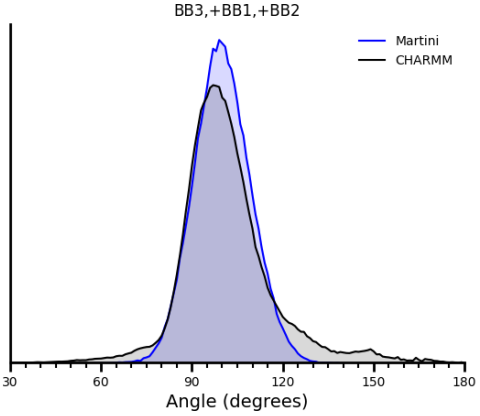

BB2,cNU1,cNU3
Martini
CHARMM
30
60
90
120
150
180
Angle (degrees)

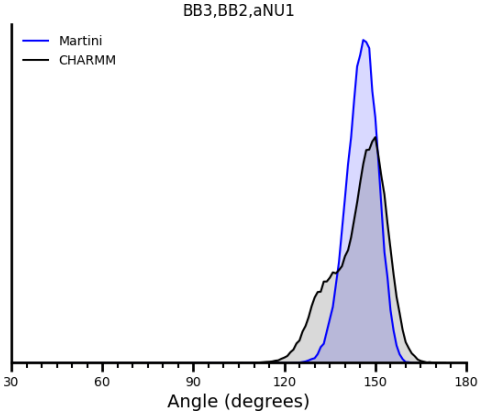

BB2,gNU1,gNU2
Martini
CHARMM
30
60
90
120
150
180
Angle (degrees)

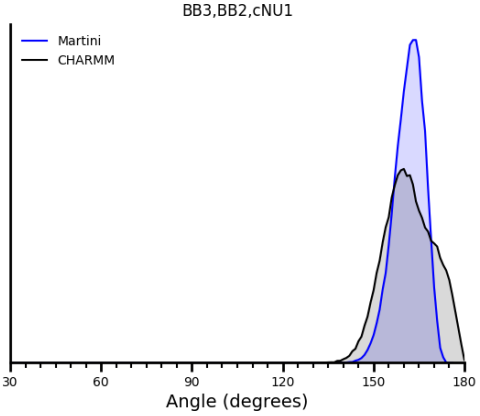

BB2,gNU1,gNU4
Martini
CHARMM
30
60
90
120
150
180
Angle (degrees)

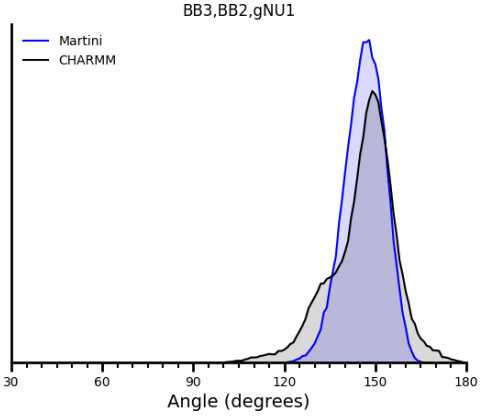

BB2,tNU1,tNU2
Martini
CHARMM
30
60
90
120
150
180
Angle (degrees)

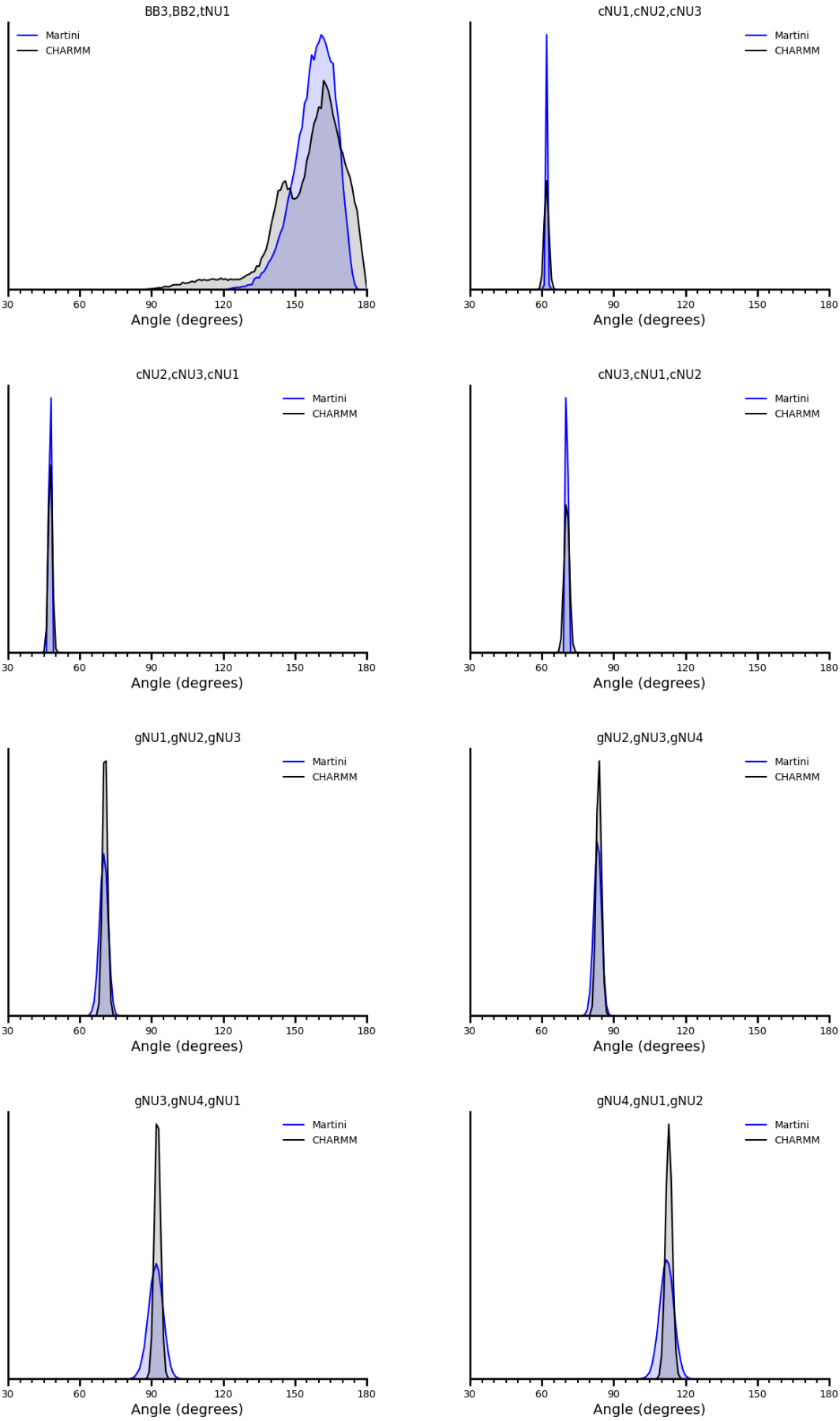

BB2,tNU1,tNU3
Martini
CHARMM
30
60
90
120
150
180
Angle (degrees)


BB3,+BB1,+BB2
Martini
CHARMM
30
60
90
120
150
180
Angle (degrees)


BB3,BB2,aNU1
Martini
CHARMM
30
60
90
120
150
180
Angle (degrees)


BB3,BB2,cNU1
Martini
CHARMM
30
60
90
120
150
180
Angle (degrees)


BB3,BB2,gNU1
Martini
CHARMM
30
60
90
120
150
180
Angle (degrees)


BB3,BB2,tNU1
cNU1,cNU2,cNU3
cNU2,cNU3,cNU1
cNU3,cNU1,cNU2
gNU1,gNU2,gNU3
gNU2,gNU3,gNU4
gNU3,gNU4,gNU1
gNU4,gNU1,gNU2
Martini
CHARMM
30
60
90
120
150
180
Angle (degrees)

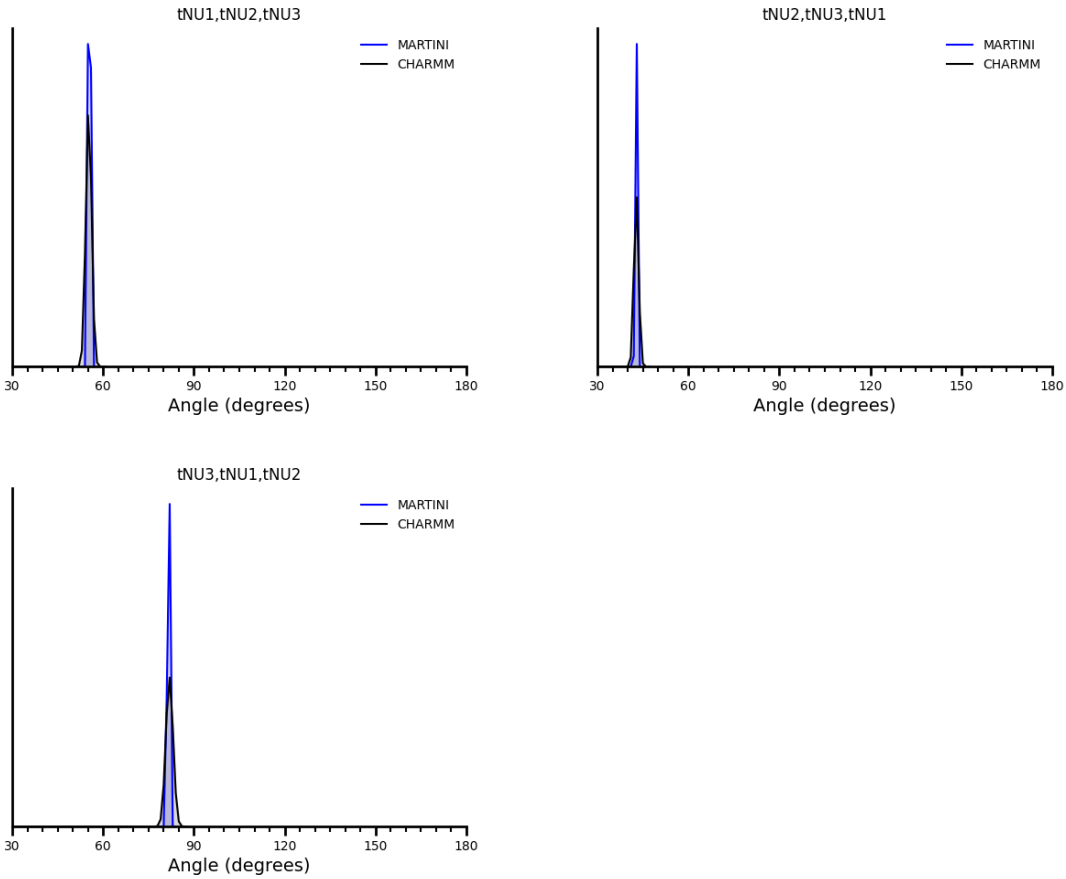
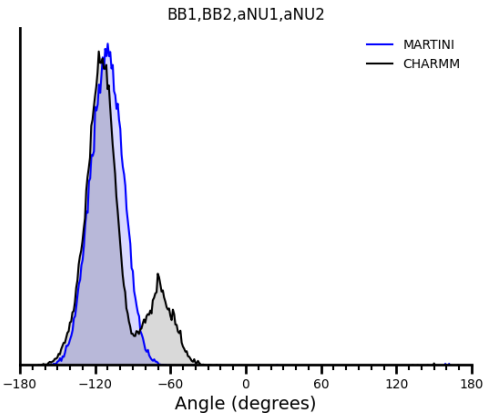
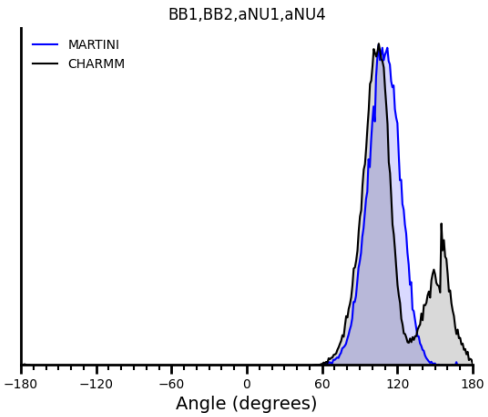
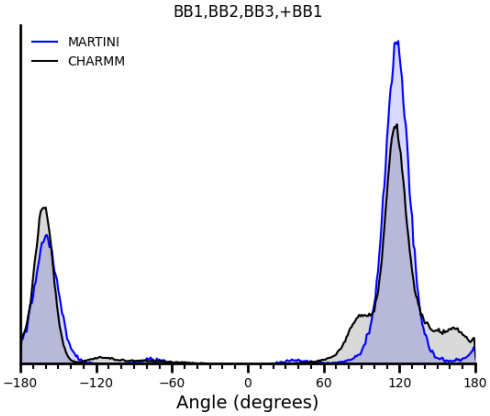
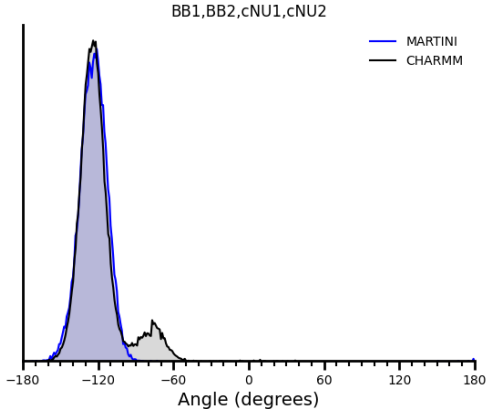


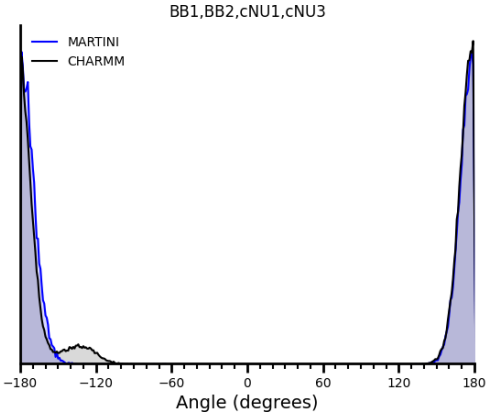
Figure A-6: Angle distributions of CHARMM and Martini simulations.

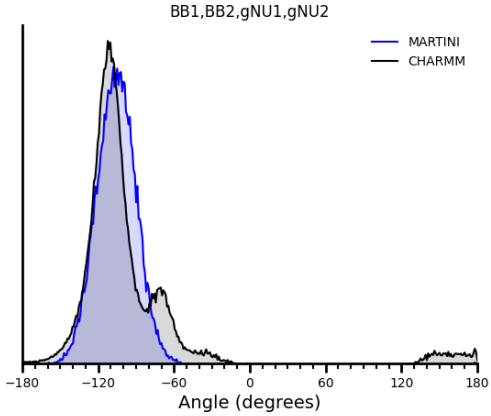

BB1,BB2,aNU1,aNU2
MARTINI
CHARMM
−180 −120 −60 0 60 120 180
Angle (degrees)

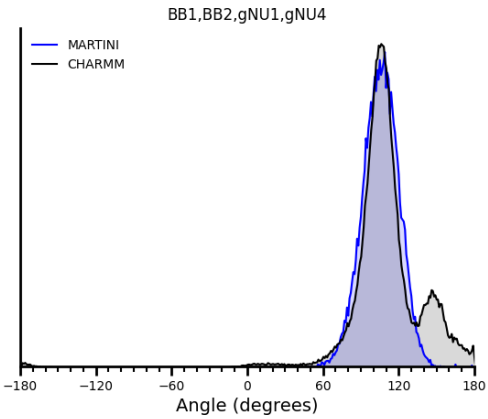

BB1,BB2,aNU1,aNU4
MARTINI
CHARMM
−180 −120 −60 0 60 120 180
Angle (degrees)

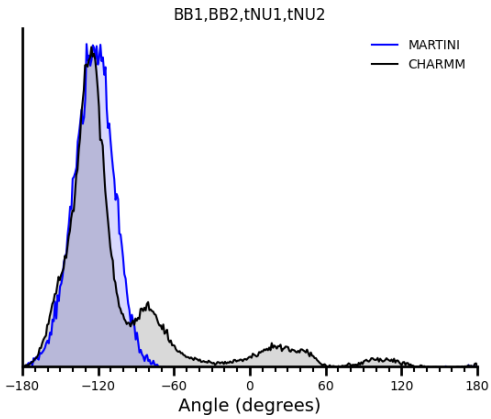

BB1,BB2,BB3,+BB1
MARTINI
CHARMM
−180 −120 −60 0 60 120 180
Angle (degrees)


BB1,BB2,cNU1,cNU2
MARTINI
CHARMM
−180 −120 −60 0 60 120 180
Angle (degrees)


BB1,BB2,cNU1,cNU3
MARTINI
CHARMM
−180 −120 −60 0 60 120 180
Angle (degrees)


BB1,BB2,gNU1,gNU2
MARTINI
CHARMM
−180 −120 −60 0 60 120 180
Angle (degrees)


BB1,BB2,gNU1,gNU4
MARTINI
CHARMM
−180 −120 −60 0 60 120 180
Angle (degrees)


BB1,BB2,tNU1,tNU2
MARTINI
CHARMM
−180 −120 −60 0 60 120 180
Angle (degrees)

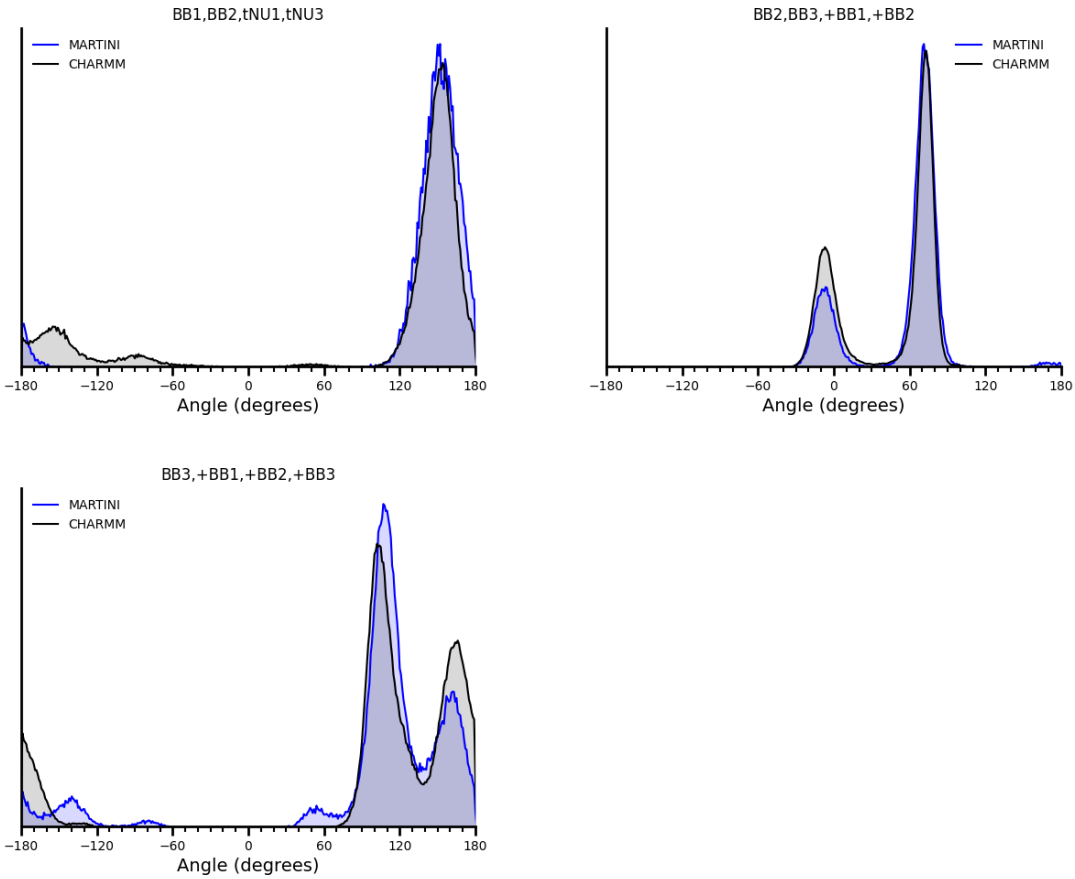


Figure A-7: Dihedral distributions of CHARMM and Martini simulations.

A.5 Published article: Modeling Peptide Nucleic Acid Binding Enthalpies Using MM-GBSA

THE JOURNAL OF PHYSICAL CHEMISTRY B
A JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

pubs.acs.org/JPCB

Article

Modeling Peptide Nucleic Acid Binding Enthalpies Using MM-GBSA

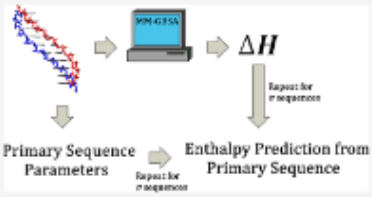
Jack Goodman,* David Attwood, Janice Kiely, Pablo Coladas Mato, and Richard Luxton

Cite This: J. Phys. Chem. B 2022, 126, 9528−9538

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The binding enthalpies of peptide nucleic acid (PNA) homoduplexes were predicted using a molecular mechanics generalized Born surface area approach. Using the nucleic acid nearest-neighbor model, these were decomposed into sequence parameters which could replicate the enthalpies from thermal melting experiments with a mean error of 8.7%. These results present the first systematic computational investigation into the relationship between sequence and binding energy for PNA homoduplexes and identified a stabilizing helix initiation enthalpy not observed for nucleic acids with phosphoribose backbones.



INTRODUCTION

The binding of nucleic acids is a sequence-dependent process facilitated primarily through the Watson−Crick pairing of complementary base pairs. This is not a property unique to naturally occurring nucleic acids such as DNA or RNA. It is therefore possible to produce nucleic acid analogue molecules which adhere in a sequence-specific manner but with modified properties. Peptide nucleic acid (PNA) is an analogue wherein the phosphoribose backbone of DNA or RNA has been entirely replaced. Instead, PNA has a backbone composed of repeating *N*-(2-aminoethyl)glycine monomers (Figure 1).[1]

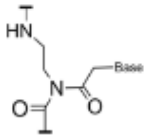


Figure 1. Schematic of an *N*-(2-aminoethyl)glycine monomer of a PNA backbone connected to a base through a carbonyl methylene linker.

The directionality of the backbone proceeds from an N-terminal amine to a C-terminal carboxyl, similar to a polypeptide, and PNA forms both stable homo- and heteroduplex structures in solution.[2,3]

PNA homoduplexes form P-form[1,3] double helices in solution. The P-form helix has a large 18 base-pair pitch and a wide 28 Å helical diameter (Figure 2). PNA/DNA heteroduplexes have a structure intermediate of the P- and B-forms, demonstrating that PNA is flexible enough to remain stable in its non-preferred helical conformation.[4]

Despite adopting non-preferred helical structures, PNA heteroduplexes are more thermally stable than DNA or RNA homoduplexes, though they are less stable than PNA homoduplexes.[3] Unlike DNA or RNA homoduplexes, PNA heteroduplex stability is largely independent of the concentration of salt in solution.[2,5] This is due to the PNA neutral backbone since the net repulsion between the negatively charged phosphoribose backbones in DNA or RNA is not present.[6] As a result, the PNA backbone does not need to associate with counterions to stabilize the double helix.[2]

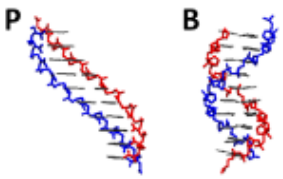


Figure 2. 10 base-pair structures of P-form PNA (left) and B-form DNA (right).

PNA heteroduplexes are also biostable since they are not substrates for nucleases and proteases.[5] The acidic conditions that would depurinate DNA or RNA homoduplexes also typically do not depurinate PNA.[7] Since PNA remains stable under a variety of conditions, it has attracted interest for applications ranging from antisense and antigene technologies to biosensing, genetic diagnostics, molecular biology, and medicinal chemistry. For example, PNA has been applied in vivo to silence gene transcription and translation.[8−10]

The PNA backbone can be modified to produce favorable properties in heteroduplexes. For example, the incorporation of lysine or arginine functional groups at the α carbon of the backbone increased the electrostatic forces of attraction between a PNA strand and a DNA complement.[11] Similar

Received: August 4, 2022
Revised: October 14, 2022
Published: November 14, 2022

https://doi.org/10.1021/acs.jpcb.2c05547
J. Phys. Chem. B 2022, 126, 9528−9538

modifications could be applied to PNA homoduplexes, with the advantage that identical chemistry can be used for both strands. Given the higher thermal and environmental stabilities of PNA relative to other nucleic acids with biotechnological applications, it is feasible that PNA technologies based on such modifications could have desirable properties.

PNA homoduplex technologies, however, must be preceded by a deeper understanding of the relation between the PNA structure and binding properties. Currently, no model exists that decomposes the binding energies of PNA duplexes into sequence parameters. All-atom molecular dynamics is an appropriate means to develop such a model since any number of PNA structures can be easily generated, and the expense of specialist equipment such as synthesizers can be avoided.

To this end, molecular mechanics generalized Born surface area (MM-GBSA) analysis of simulated PNA structures was conducted. Entropy was estimated using a quasi-harmonic (QH) approximation. The current force field parameters designed by Jasiński et al.[12] were implemented. The computational results were benchmarked against thermal melting data obtained both from this work and the available literature.

Using a previously developed method,[13] it was determined that nearest-neighbor decomposition of PNA binding enthalpies was possible. Entropy and free energy could also be obtained from experiments, though benchmarking the QH entropy revealed that it was a poorer estimate. Finally, it was found that sequence parameters for binding enthalpies could be understood in terms of the dynamics of PNA homoduplexes through hydrogen bonding analysis.

■ THEORY AND METHODS

Literature Analysis. Seven PNA homoduplexes with binding energies and entropies from thermal melting experiments were compiled from the literature.[2,3,14−18]

Experimental conditions used by the publications were inconsistent. Namely, the concentration of sodium ions and the use of a solubility-enhancing lysine tag were varied. Since prior research consistently agreed that PNA stability is negligibly affected[2,5,6] by a change in the concentration of sodium chloride, this was not considered critical.

To determine that lysine tagging did not affect the binding energies, the sequence GTAGATCACT, which appeared the most times in the published research, was categorized according to whether this tag was present or not. It was found that the binding energies and entropies for unmodified sequences were all within one standard error of the mean binding energies and entropies for sequences with the tag (Table S1).

Sequences were therefore grouped into a single data set regardless of these conditions.

Thermal Melting Experiments. Three PNA homoduplexes, CGATCG, AACGTT, and TAGCTA, were purchased from Eurogentec at 95% purity (Cat. Number BA-PN010-005) as determined by high-performance reverse-phase and ion-exchange liquid chromatography methods. Samples were diluted to a concentration of 1 μM in deionized water and were then melted via an electric heater in a water bath from which samples were taken at 2 K intervals after 5 min equilibrations at the target temperature. The absorbances of the PNA strands at 260 nm were then monitored using an Agilent Technologies Cary 60 UV−vis spectrophotometer. Their absorbances increased during melting because single-stranded nucleic acids are hyperchromic at 260 nm relative to the duplex.

Absorbances were normalized between zero and one to produce alpha curves, where $\alpha = 1$ corresponds to a sample containing only double helices, while $\alpha = 0$ contains only single strands. The melting point was defined at $\alpha = 0.5$. The equilibrium constant was then derived according to a two-state assumption for the transition using the equation for self-complementary helices.[19]

$$K = \frac{\alpha}{2C_T(1-\alpha)^2} \tag{1}$$

where C_T is the total strand concentration. The natural logarithm of the equilibrium constant was then plotted against the reciprocal temperature to derive the standard enthalpies and entropies of the transition according to the linear form of the van't Hoff equation. This assumes that the entropy and enthalpy are temperature-independent.

$$\ln K = -\frac{\Delta H^\circ}{RT} + \frac{\Delta S^\circ}{R} \tag{2}$$

The free energy of binding of the strands was then calculated according to $\Delta G = \Delta H - T\Delta S$.

Molecular Dynamics. Simulated structures of PNA homoduplexes were generated using the proto-Nucleic Acid Builder .[20] A set of initial helical parameters[21−23] were obtained from published crystal or NMR structures. A weighted Monte Carlo conformational search was conducted to generate chemical structures based on these initial parameters and on thresholds such as bonding energies and interatomic distances set by the user. The structures generated by this were imported into the Gromacs program using the July 2021 edition of the CHARMM36m forcefield,[24,25] and the N- and C-termini were acetylated and amidated, respectively.

Simulated homoduplexes were energy minimized in vacuum over 50,000 simulation steps using a steepest-descent algorithm. Structures were then solvated in explicit waters in dodecahedral boxes using the simple point charge (SPC) water model spc216.[26] The distance between any atom of the solute and the edge of the box was set to a minimum of 15 Å. Periodic boundary conditions[27] were enforced to imitate a bulk solvent, and solvated structures were energy minimized using steepest-descent minimization over 50,000 simulation steps.

For each PNA homoduplex, three replicates were obtained from this point onward. Each replicate was heated from 30 to 298 K over 500 ps, and Newton's equations of motion were integrated using a leapfrog integrator.[28,29] Temperature was weakly coupled to an external bath using a velocity rescaling[30] algorithm. Harmonic restraints of 24 kcal mol^{-1} Å^{-1} were applied to non-hydrogen atoms, whereas hydrogen atoms were constrained using the LINCS[31] algorithm for the suppression of high-frequency hydrogen oscillations such that 2 fs timesteps could be used. During this and for all future simulations, the van der Waals interactions were handled using a switched cut-off scheme, with switching to zero occurring from 10 to 12 Å. The electrostatic nonbonded interactions were treated using a particle mesh Ewald (PME)[32,33] algorithm with quartic interpolation and a grid spacing of 1 Å. The short-ranged component of the PME was computed to 10 Å, and long-ranged components were handled using the fast Fourier transform library FFTW.[34]

After heating, position restraints were halved every 2 ns, with the equations of motion and temperature coupling thereon

https://doi.org/10.1021/acs.jpcb.2c05547
J. Phys. Chem. B 2022, 126, 9528−9538

handled by a stochastic integrator, which incorporated friction and noise terms.[35] After 10 ns of simulation, the non-hydrogen restraints were removed entirely, and structures were equilibrated for 10 ns in the canonical ensemble. They were then equilibrated for a further 10 ns in the isobaric−isothermal ensemble with pressure coupling at 1 bar handled by a Berendsen barostat.[36] 100 ns production runs were then carried out for each of the three replicas of each sequence using Parrinello−Rahman pressure coupling.[37]

Trajectory Analysis. Trajectories were analyzed by gmx_mmpbsa, which enables MM-GBSA estimations of enthalpy, entropy, and free energy to be calculated.[38] The free energy of a molecule computed using MM-GBSA[39−41] is given by the average molecular mechanics potential energy E_{MM}, the solvation energy G_{solv}, and the entropic term TS. The difference in free energies of two single PNA strands and the PNA homoduplex is then given by the difference of these.

$$\Delta G = \Delta\langle E_{MM}\rangle + \Delta\langle G_{solv}\rangle - T\Delta S \quad (3)$$

E_{MM} is the sum of the bonded and nonbonded interactions of the molecule in vacuum, obtained from snapshots of the simulation trajectory, and accounts for the bond angle and dihedral energies between bonded atoms and the van der Waals and electrostatic interactions between nonbonded atoms.

$$\Delta\langle E_{MM}\rangle = \Delta\langle E_{bonded}\rangle + \Delta\langle E_{vdW}\rangle + \Delta\langle E_{elec}\rangle \quad (4)$$

In the single-trajectory approach used in this project, the individual PNA strands are isolated from the trajectory of the homoduplex, meaning that $\Delta\langle E_{bonded}\rangle = 0$. The single-trajectory approach assumes that the conformation of a PNA strand in the duplex is similar to a free PNA strand in solution but is computationally less expensive. The validity of the approach was tested by simulating three single-stranded PNAs, for a total of nine replicates, and calculating their enthalpies using the same generalized Born method. These enthalpies were then compared to the enthalpies of the single strands as evaluated by the single-trajectory approach, that is, when taken from the trajectory of the duplex. The difference in this conformation enthalpy[42,43] was limited (~1%) in all three cases (Table S2).

The solvation energy difference $\Delta\langle G_{solv}\rangle$ from eq 3 is calculated using a continuum solvent model, which approximates the explicit water solvent as a homogeneous medium with a dielectric constant of 78.5 at 298 K. It is composed of polar and nonpolar components. The polar term accounts for the electrostatic interactions between solute and solvent. The nonpolar component accounts for the work of formation of a cavity the size of the solute in the solvent and the short-ranged van der Waals interactions between the solute and the solvent.[44]

$$G_{solv} = G_{polar} + G_{nonpolar} \quad (5a)$$

$$G_{polar} \approx G_{GB} = -k\sum_{ij}\frac{q_i q_j}{f_{GB}} \quad (5b)$$

$$G_{nonpolar} \approx G_{SA} = \gamma SA + b \quad (5c)$$

where q_i and q_j are atomic charges and k, in the generalized Born model, is a function of the dielectric constants of the solvent and of free space, where it is unity. f_{GB} is a function of interatomic distances and Born radii of the ith and jth atoms, wherein the Born radius is the spherically averaged distance between an atom of the solute and the solvent. G_{SA} is a linear function of the solute's solvent-accessible surface area (SA).[45]

The final contribution to the free energy change according to eq 4 is then $T\Delta S$, which was evaluated in this study using a QH approximation.[46]

Using the MM-GBSA approach, the binding energies for PNA homoduplexes were evaluated. Snapshots from each production trajectory were stripped of explicit water, and the first 20 ns were discarded. The remaining 80 ns were analyzed using MM-GBSA. The version of the generalized Born model, *igb* = 5,[47] was used for the analysis, and enthalpy change was taken to be $\Delta\langle E_{MM}\rangle + \Delta\langle G_{solv}\rangle$ according to $\Delta G = \Delta H - T\Delta S$.

A standard state correction was applied to the free energy differences to correct for the box volumes and thereby standardize the concentrations.

$$\Delta G^{\circ} = \Delta G - RT\ \ln\left(\frac{V_{com}}{V^{\circ}}\right) \quad (6)$$

where $V^{\circ} = 1661$ Å^3 and V_{com} is the volume of the simulation box.[48]

The production trajectories with explicit waters were then analyzed using gmx_rms and gmx_hbond, which were used to compute the root-mean-square atomic deviations (rmsds) (Figure S1) over time and the number of Watson−Crick hydrogen bonds between the base pairs over time. Hydrogen bonds were defined as bonds between the atom triplets involved in Watson−Crick bonding with distances between donor and acceptor of less than 0.325 nm and angles of hydrogen-donor−acceptor of less than 30°.

To confirm that run durations were sufficient to obtain converged thermodynamic data, relative enthalpies, entropies, and free energies were plotted as functions of sequence length (Figures S2−S4). To determine whether simulations were equilibrated and repeatable, rmsd histogram distributions for the replicas of each sequence were overlapped (Figures S5 and S6).

Modeling Nucleic Acid Binding Energies. A nucleic acid duplex can be thought of as a series of overlapping base pair stacks (Figure 3). A stack is two base pairs in sequence,

$$\begin{array}{c} 5'-\mathrm{A}\,\mathrm{A}\,\mathrm{T}\,\mathrm{A}\,\mathrm{C}\,\mathrm{G}\,\mathrm{G}\,\mathrm{C}\,\mathrm{A}\,\mathrm{G}\,\mathrm{A}-3' \\ 3'-\mathrm{T}\,\mathrm{T}\,\mathrm{A}\,\mathrm{T}\,\mathrm{G}\,\mathrm{C}\,\mathrm{C}\,\mathrm{G}\,\mathrm{T}\,\mathrm{C}\,\mathrm{T}-5' \\ \downarrow \\ \frac{\mathrm{AA}}{\mathrm{TT}} + \frac{\mathrm{AT}}{\mathrm{TA}} + \frac{\mathrm{TA}}{\mathrm{AT}} + \frac{\mathrm{AC}}{\mathrm{TG}} + \frac{\mathrm{CG}}{\mathrm{GC}} + \frac{\mathrm{GG}}{\mathrm{CC}} + \frac{\mathrm{GC}}{\mathrm{CG}} + \frac{\mathrm{CA}}{\mathrm{GT}} + \frac{\mathrm{AG}}{\mathrm{TC}} + \frac{\mathrm{GA}}{\mathrm{CT}} \end{array}$$

Figure 3. An oligonucleotide duplex broken down into a series of overlapping stacks, as indicated by the asterisks. Stacks are written below the duplex, as indicated by the arrow.

and there are 10 unique stacks for homoduplexes. Overlapping means that the first pair of the following stack is the second pair of the preceding stack. This is known as the nearest-neighbor model. Given that the stabilities of nucleic acid duplexes are impacted both by their base-pairing and base-stacking energies, the nearest-neighbor method is theoretically rigorous in that it accounts for both.[49−51]

The energies and entropies of binding are given by summing over the energies and entropies of these stacks. In addition, the helix initiation energy or entropy and the terminal guanine−cytosine (GC) energy or entropy are added. The binding

https://doi.org/10.1021/acs.jpcb.2c05547
J. Phys. Chem. B 2022, 126, 9528−9538

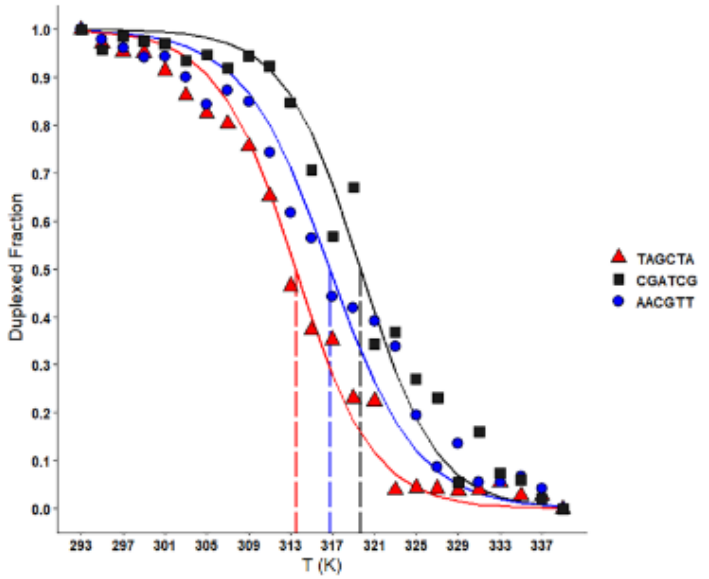


Figure 4. Normalized melting curves for the three PNA hexamers TAGCTA, CGATCG, and AACGTT, as measured by UV melting experiments. The duplexed fraction, α, is indicated on the y axis and the temperature in kelvin on the x axis. Melting points are shown by the dashed vertical lines. Points represent the means of independent triplicates and were obtained at 2 K intervals.

energy or entropy of a sequence is then a sum over the stacks, the helix initiation term, and the terminal GC term.

$$B\begin{cases}0 & n(\text{T. GC}) < 2\\ 1 & n(\text{T. GC}) = 2\end{cases} \tag{7a}$$

$$\Delta\chi = \Delta\chi_{\text{init}} + B\Delta\chi_{\text{T.GC}} + \sum_i j_i(\Delta\Delta\chi_i) \tag{7b}$$

where $\Delta\chi$ is a binding energy or entropy. $\Delta\chi_{\text{init}}$ is a helix initiation energy or entropy. $\Delta\chi_{\text{T.GC}}$ is the energy or entropy added so long as the number of terminal GC base pairs, n(T.GC), is two, as indicated by eq 7a. Put differently, $B\Delta\chi_{\text{T.GC}}$ evaluates to zero if there is a terminal adenine-thymine (AT) base pair. $\Delta\Delta\chi_i$ is the incremental change in binding energy or entropy given by the ith unique base pair with occurrence j_i.

For self-complementary helices, the binding entropy is decreased by 1.4 cal K^{-1} mol^{-1} to account for the C2 rotational symmetry of the double helix, which reduces the degrees of freedom upon binding.[51]

Equation 7b has the matrix form

$$\underbrace{\begin{pmatrix}\Delta\chi_1\\ \vdots\\ \Delta\chi_N\end{pmatrix}}_{\chi} = \underbrace{\begin{pmatrix}j_1^1 & \cdots & j_v^1\\ \vdots & & \vdots\\ j_1^N & \cdots & j_v^N\end{pmatrix}}_{S} \cdot \underbrace{\begin{pmatrix}\Delta\chi_1\\ \vdots\\ \Delta\chi_{v-2}\\ \Delta\chi_{\text{init}}\\ \Delta\chi_{\text{T.GC}}\end{pmatrix}}_{\varepsilon} \tag{8}$$

where χ is a $1 \times N$ column matrix containing either mean enthalpies, entropies, or free energies of hybridization of PNA homoduplexes. S is an $N \times v$ matrix containing the occurrences j of all stacked base pairs, terminal GC and initiation terms. ε is a $1 \times v$ column matrix containing the unknown energies or entropies of these terms. These unknowns are then evaluated by solving the linear least-squares problem, which minimizes the sum of the square of the residuals of χ with respect to S.

The problem is generalized by weighting each observable $\Delta\chi$ by its reciprocal standard error.

$$\sigma^{-1}\cdot\chi = \sigma^{-1}\cdot S\cdot\varepsilon \tag{9}$$

where σ^{-1} is an $N \times N$ diagonal matrix containing the reciprocal standard errors of each observable with non-diagonal elements of zero. The estimate ε obtained by solving the multiple linear regression problem thus contains the error-weighted least-squares estimates of the energies or entropies of the terms on the right-hand side of eq 7b.

Equation 9 can be used to decompose MM-GBSA estimates of binding enthalpies, entropies or free energies for PNA homoduplexes into sequence parameters.

RESULTS AND DISCUSSION

Thermal Melting Data. Melting points of the three PNA hexamers defined as $\alpha = 0.5$ were observed at 313.5 ± 0.4, 316.8 ± 0.4, and 319.7 ± 0.8 K for TAGCTA, AACGTT, and CGATCG, respectively (Figure 4).

Using eqs 1 and 2, the standard binding free energies, enthalpies and entropies were determined for the three hexamers. The standard errors of the measured enthalpies and entropies for all three sequences overlapped, whereas the free energy of hybridization for CGATCG was approximately 1 kcal mol^{-1} more negative. This suggested that CGATCG was stabilized relative to the other two homoduplexes as was corroborated by their melting temperatures.

The binding energies and entropies of the three sequences were compiled alongside the binding energies and entropies obtained from the literature. This produced a data set totaling

https://doi.org/10.1021/acs.jpcb.2c05547
J. Phys. Chem. B 2022, 126, 9528−9538

Table 1. Standard Free Energies, Enthalpies, and Entropies of Binding for the 10 PNA Homoduplexes from Thermal Melting Experiments[a]

sequence	$-\Delta G^{\circ}_{298}$(kcal mol^{-1})	$-\Delta H^{\circ}_{298}$(kcal mol^{-1})	$-\Delta S^{\circ}_{298}$(cal K^{-1} mol^{-1})
	(a) Experimental Thermal Melting Data		
CGATCG	13.64 ± 0.25	58.78 ± 4.09	151.42 ± 12.89
AACGTT	12.64 ± 0.39	52.05 ± 5.30	132.24 ± 16.45
TAGCTA	12.34 ± 0.29	55.08 ± 3.85	143.41 ± 11.95
	(b) Literature Thermal Melting Data		
GTAGATCACT[b,c,d,e,f,g]	18.99 ± 0.87	86.92 ± 5.00	227.97 ± 14.48
TGTTACGACT[h]	21.08 ± 1.00	92.60 ± 5.70	240.00 ± 16.11
AGGTAACCAG[d]	18.76 ± 0.60	83.30 ± 2.80	216.60 ± 7.40
AGTGAAGCAG[d]	19.02 ± 0.82	82.15 ± 4.56	211.85 ± 12.86
TGATCTAC[f]	13.10 ± 0.00	60.90 ± 0.00	224.70 ± 0.00
GTAGATCACTGT[f]	21.40 ± 0.00	97.10 ± 0.00	253.80 ± 0.00
GTAGATCACTGTCAC[f]	26.40 ± 0.00	117.10 ± 0.00	304.30 ± 0.00

[a]Standard errors are as provided by the study author or are the standard errors from reported energies or entropies of multiple publications if a sequence is reported multiple times. Standard errors of zero indicate that they were not standard errors but provided by the study author. [b]Tomac, S.; Sarkar, M.; Ratilainen, T.; Wittung, P.; Nielsen, P.; Nordén, B.; Gräslund, A. Ionic Effects on the Stability and Conformation of Peptide Nucleic Acid Complexes. *J. Am. Chem. Soc.* **1996**,*118*, 5544−5552. [c]Sen, A.; Nielsen, P. On the stability of peptide nucleic acid duplexes in the presence of organic solvents. *Nuc. Acid. Res.* **2007**,*35*, 3367−3374. [d]Sen, A.; Nielsen, P. Unique Properties of Purine/Pyrimidine Asymmetric PNA-DNA Duplexes: Differential Stabilization of PNA-DNA Duplexes by Purines in the PNA Strand. *Biophys. J.* **2006**,*90*, 1329−1337. [e]Ratilainen, T.; Holmén, A.; Tuite, E.; Haaima, G.; Christensen, L.; Nielsen, P.; Nordén, B. Hybridization of Peptide Nucleic Acid. *Biochemistry.* **1998**,*37*, 12331−12342. [f]Totsingan, F. Synthesis and Applications of PNA and Modified PNA in Nanobiotechnology. Ph.D. Thesis, University of Parma, Parma, Italy, **2007**. [g]Sforza, S.; Haaima, G.; Marchelli, R.; Nielsen, P. Chiral Peptide Nucleic Acids (PNAs): Helix Handedness and DNA Recognition. *Eur. J. Org. Chem.* **1999**, 197−204. [h]Jasiński, M.; Miszkiewicz, J.; Feig, M.; Trylska, J. Thermal Stability of Peptide Nucleic Acid Complexes. *J. Phys. Chem. B.* **2019**,*123*, 8168−8177.

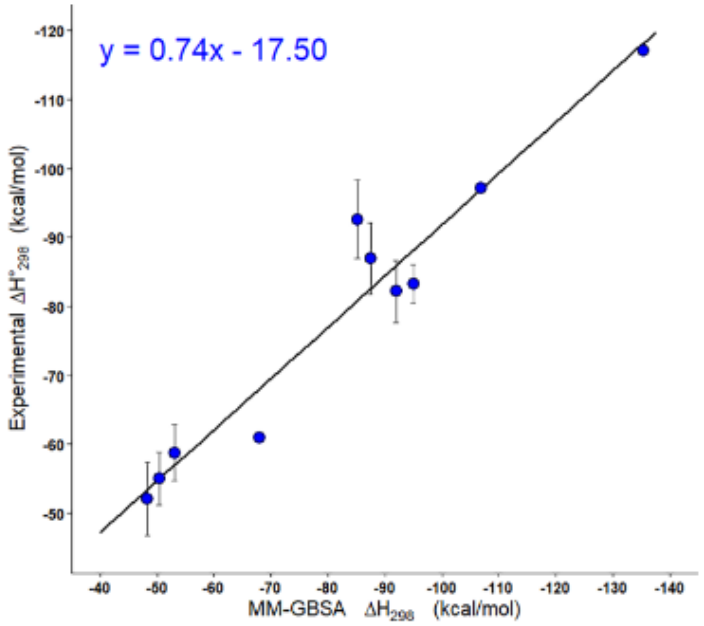


Figure 5. Correlation analysis for the enthalpy of binding of PNA homoduplexes from thermal melting experiments and from MM-GBSA analysis.

10 sequences, 7 of which were obtained from the literature,[2,3,14−18] against which computationally derived energies and entropies could be benchmarked (Table 1).

GBSA Benchmarking. The entropies, enthalpies, and free energies of binding for the 10 PNA homoduplexes were obtained using MM-GBSA. Means and standard errors were obtained from triplicates. Linear regression of the enthalpies of binding from thermal melting experiments against MM-GBSA analysis produced an R^2 of 0.93, indicating that they were well correlated. A high R^2 would be anticipated given that there were few sequences of the same length, meaning a positive correlation requires only that the simulation and reality both report a higher enthalpy of binding for longer lengths, which is what was observed. The mean absolute difference, taken as the error of the prediction, between the enthalpy of binding from thermal melting and from MM-GBSA experiments was 7.87

https://doi.org/10.1021/acs.jpcb.2c05547
J. Phys. Chem. B 2022, 126, 9528−9538

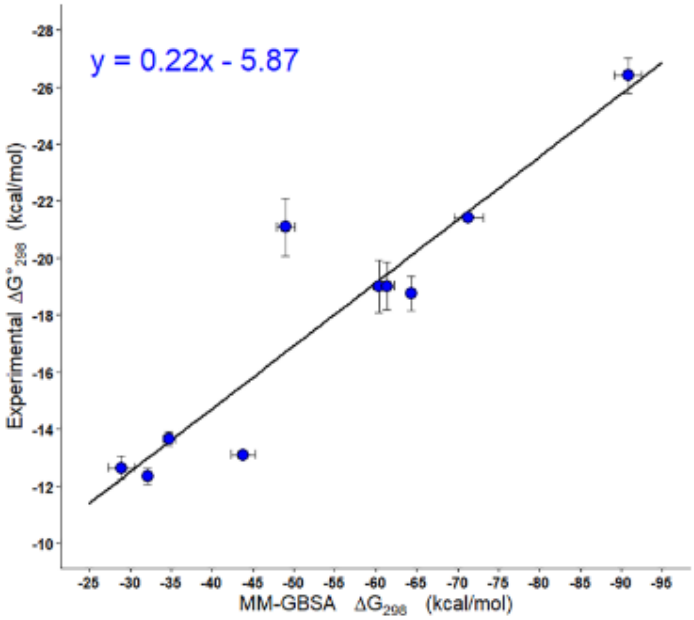


Figure 6. Correlation analysis for the Gibbs free energy of binding of PNA homoduplexes from thermal melting experiments and MM-GBSA.

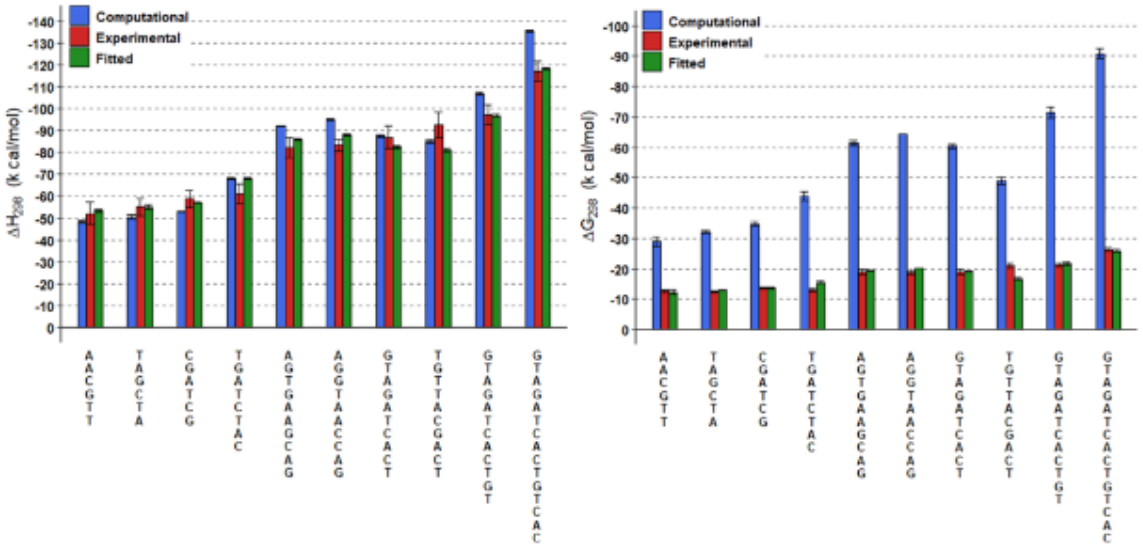


Figure 7. Enthalpies, on the left, and free energies, on the right, for the 10 PNA homoduplexes of the benchmarking set. Data from thermal melting experiments is in red. Data from MM-GBSA is in blue. Data after fitting according to $y = ax + b$ is in green. Error bars indicate the standard errors. Standard errors for computational and fitted data are obtained from triplicate simulations.

kcal mol^{-1}, and the mean relative error was 10.01%. The equation of the line between the thermal melting and MM-GBSA enthalpies (Figure 5) yielded a slope coefficient $a = 0.74$ and a shift coefficient $b = -17.50$. The large shift coefficient indicated a deviation between the thermal melting and MM-GBSA analysis. This was likely a result of a combination both of the assumptions made during simulation and of the small size of the data set, with only 10 sequences with thermal melting data being analyzed, causing the shift to be largely determined by only a few points.

Treatment of the binding enthalpies according to the equation of the line $\Delta H^{\circ}_{exp} = 0.74(\Delta H_{com}) - 17.50$ reduced the relative and absolute errors of the estimates to 5.25% and 4.12 kcal mol^{-1}, respectively.[2,3,14−18]

In contrast to the binding enthalpy, binding entropy and free energy were worse estimated, though they remained highly correlated with R^2 of 0.84 and 0.86, respectively. The mean

https://doi.org/10.1021/acs.jpcb.2c05547
J. Phys. Chem. B 2022, 126, 9528−9538

absolute and relative errors were 31.49 kcal mol^{-1} and 50.14% for $T\Delta S$ and 35.95 kcal mol^{-1} and 102.68% for the Gibbs free energy of binding. These errors are significantly higher than the error for the enthalpy and indicate that the configurational entropy as estimated by QH analysis does not accurately evaluate the entropy of binding from thermal melting experiments.

Similarly to enthalpy, treatment of free energy according to the equation of the line, $\Delta G^{\circ}_{exp} = 0.22(\Delta G^{\circ}_{com}) - 5.87$ (Figure 6) reduced the relative and absolute errors to 6.53% and 1.16 kcal mol^{-1}, respectively. This demonstrates that significant correction can be achieved via linear regression, enabling the prediction of experimental binding energies from computer simulations even when there is quantitative disagreement (Figure 7).

Nearest-Neighbor Modeling. To decompose binding energies into sequence parameters according to eqs 8 or 9, the coefficients in $\boldsymbol{\epsilon}$ must converge. For that to be the case, the number of binding energies in $\boldsymbol{\chi}$ must be large. The 10 PNA homoduplexes in Table 1 are not sufficient and so an additional 39 PNA homoduplexes were simulated (Table S3) and estimates of their binding enthalpies, entropies, and free energies were determined using MM-GBSA. A total of 49 PNA homoduplexes were used to produce the nearest-neighbor model. As a result, it is assumed that the errors described in the previous section are representative of the average error of the remaining 39 homoduplexes, which lack experimental data.

Benchmarking the subset of 10 sequences gave relative errors of 10.01 and 102.68% for the enthalpy and free energy of binding, respectively. A correction reduced these to 5.25 and 6.53%. This shows that thermal melting data can be accurately predicted using MM-GBSA without correction for enthalpy, though predictions of free energy require correction.

It is possible to develop a nearest-neighbor model with both corrected and uncorrected data. However, corrections according to the equations of the line, $\Delta\chi_{exp} = a\Delta\chi_{com} + b$, in Figures 5 and 6, impact the sequence parameters on the right-hand side of eq 7b in two ways

$$(\Delta\Delta\chi_i)_C = a(\Delta\Delta\chi_i) \tag{10a}$$

$$(\Delta\chi_{init})_C = a(\Delta\chi_{init}) + b \tag{10b}$$

where the subscript C indicates a term derived from solving the multiple regression problem using binding energies or entropies, which have been corrected according to $\Delta\chi_{exp} = a\Delta\chi_{com} + b$. The impact of the correction on the nearest-neighbor terms is not trivial. First, by scaling the stacked base pair energies according to the slope coefficient a (eq 10a), the equality $\Delta\Delta G = \Delta\Delta H - T\Delta\Delta S$ only holds when $a_1\Delta\Delta G = a_2\Delta\Delta H - a_3T\Delta\Delta S$, where a_i is the slope coefficient for the ith state variable such as the enthalpy. This is unlikely to be the case since the source of error differs between enthalpy and entropy: for enthalpy, it arises from the evaluation of potential energy, whereas for entropy, it arises from the evaluation of the fluctuations in atomic coordinates. In addition to this, the shift coefficient b may be large compared to the actual helix initiation energy or entropy. The corrected helix initiation value on the left-hand side of eq 10b thus may largely be determined by a metric of computational error as opposed to any physically relevant quantity.

So long as these considerations are taken into account, the model can be highly predictive of real-world measurements. This was the case for previous research on MM-GBSA estimates of DNA and RNA.[13]

In this research, corrected MM-GBSA data were not used for fitting the nearest-neighbor model so as to maximize the likelihood that sequence parameters were meaningful. Since the binding entropy, and therefore free energy, quantitatively disagreed with thermal melting data, a nearest-neighbor model was generated solely for the binding enthalpy. This was possible since, for the binding enthalpy, the experimental error was approximately equal to a typical experimental error of 10%.[13]

MM-GBSA enthalpies for all 49 sequences (Table S3) were used to solve the generalized multiple linear regression problem in eqs 5c−7a. Nearest-neighbor stacking, initiation, and terminal GC enthalpies were thus determined (Table 2).

Table 2. Nearest-Neighbor Binding Enthalpies for PNA Homoduplexes[a]

	$-\Delta H_{298}$ (kcal mol^{-1})	$-\langle\Delta H_{298}\rangle$ (kcal mol^{-1})
AA	8.65 ± 0.16	8.76 ± 0.31
AT	8.28 ± 0.32	
TA	9.34 ± 0.32	
AG	10.64 ± 0.27	9.72 ± 0.46
GA	8.91 ± 0.35	
AC	8.90 ± 0.19	
CA	10.31 ± 0.17	
GG	12.32 ± 0.19	10.91 ± 1.12
GC	8.69 ± 0.30	
CG	11.71 ± 0.34	
	$-\Delta H_{298}$ (kcal mol^{-1})	
T.GG	1.62 ± 0.35	
Init	2.27 ± 0.46	

[a]± indicates the standard error of the mean.

All parameters from Table 2 were significant to $p < 10^{-5}$ in solving the multiple linear regression problem. Each stack was categorized according to the number of canonical Watson−Crick hydrogen bonds between them. These were six for GG, GC, and CG; five for AG, GA, AC, and CA; and four for AA, AT, and TA. On average, the stacking enthalpy became more negative as the number of Watson−Crick hydrogen bonds increased. This was in line with expectations since the hydrogen bonds stabilize the duplex.

The stacks GG and CG were more stabilizing relative to GC despite having the same number of hydrogen bonds. This difference could arise from the base stacking interaction, since base pairing only partly accounts for the energies of interaction between the strands. However, it may also reflect the sequence selection used in this study since bias can be introduced by particular stacks often coinciding or by a few sequences accounting for a large amount of the total occurrences of a stack.

Atypical for nucleic acid binding enthalpies is the negative helix initiation enthalpy. This enthalpy is typically positive for DNA and RNA duplexes.[48,49] The explanation for it being stabilizing in PNA is due to PNA neutral backbone. In DNA and RNA, negative backbones repel, introducing a barrier to the association of the two strands. For PNA, however, this net repulsion does not occur, and so it is feasible that helix initiation is more favorable which would explain the negative helix initiation enthalpy. Similarly to other nucleic acid nearest-

https://doi.org/10.1021/acs.jpcb.2c05547
J. Phys. Chem. B 2022, 126, 9528−9538

Table 3. Experimental Binding Enthalpies ΔH°_{exp}, Nearest-Neighbor Predicted Binding Enthalpies ΔH°_{NN}, and Absolute $|-\Delta H^{\circ}_{exp}+\Delta H^{\circ}_{NN}|$ and Relative $|-\Delta H^{\circ}_{exp}+\Delta H^{\circ}_{NN}|/\Delta H^{\circ}_{exp}$ Differences for 10 Homoduplexes

sequence	$-\Delta H^{\circ}_{exp}$ (kcal mol^{-1})	$-\Delta H^{\circ}_{NN}$ (kcal mol^{-1})	$\lvert-\Delta H^{\circ}_{exp}+\Delta H^{\circ}_{NN}\rvert$ (kcal mol^{-1})	difference (%)
CGATCG	58.78 ± 4.09	53.40	5.38	9.15
AACGTT	52.05 ± 5.30	49.07	2.98	5.73
TAGCTA	55.08 ± 3.85	50.91	4.17	7.57
GTAGATCACT	86.92 ± 5.00	87.09	0.17	0.20
TGTTACGACT	92.60 ± 5.70	85.24	7.36	7.95
AGGTAACCAG	83.30 ± 2.80	92.28	8.98	10.78
AGTGAAGCAG	82.15 ± 4.56	89.95	7.80	9.49
TGATCTAC	60.90 ± 0.00	67.55	6.65	10.92
GTAGATCACTGT	97.10 ± 0.00	106.30	9.20	9.47
GTAGATCACTGTCAC	117.10 ± 0.00	136.03	18.93	16.17

neighbor models, the terminal GC term is negative, indicating that the absence of terminal AT base pairs stabilizes the double helix.

The parameters in Table 2 can be used to estimate the binding enthalpies of PNA duplexes according to eq 7b (Table 3). Predicted binding enthalpies had a mean relative error of 8.74% when compared to thermal melting binding enthalpies. This was a decrease from the 10.01% error from the direct comparison of MM-GBSA to thermal melting data for the subset of 10 sequences. Since the nearest-neighbor model was based on a total of 49 sequences, the decrease in error showed that the accuracy of prediction increased by accounting for unrelated sequences.

This indicated that binding enthalpy could be readily predicted as a function of sequence parameters and that MM-GBSA could derive these sequence parameters. It is consequently possible that the nearest-neighbor model developed herein could aid in the design and structural understanding of PNA homoduplexes.

Hydrogen Bond Analysis. Since the MM-GBSA nearest-neighbor predictions agree well with the available experimental data, it is possible to interpret them as having physical meaning. The first indication of this being the case is that, when the stacks are grouped according to the number of Watson−Crick hydrogen bonds involved, the binding enthalpy becomes, on average, more stable as the number of bonds increases.

Additional interpretations of physical meaning can be obtained through dynamics studies. The terminal GC term, T.GC, was investigated using hydrogen bond analysis. Watson−Crick hydrogen bonds of terminal and 3rd-position, classed as internal base pairs, were monitored. The number of Watson−Crick hydrogen bonds was normalized and these were expressed as percentages. These percentages were averaged over the simulation durations and for simulations involving sequences of the same length (Table 4).

The percentage of unbroken bonds, from Table 4, is independent of sequence length and always, on average, higher for internal than terminal base pairs. Terminal AT hydrogen bonds were broken more often than terminal GC hydrogen bonds, though both spent most of their time in the bound state. Additionally, the standard deviation in the percentage of unbroken bonds was greater, most often by around 5-fold but in one case over 15-fold, for terminal AT than GC base pairs. This indicated that the terminal AT hydrogen bonding behavior was characterized by fluctuations between bound and melted states.

Table 4. Mean Normalized Number of Hydrogen Bonds, between Zero and One, Expressed as a Percentage of Terminal and Internal AT and GC Base Pairs of Varying Sequence Lengths[a]

	% H-bonded	
length	AT	GC
	terminal	
6	70.4 ± 27.1	83.3 ± 4.3
8	74.4 ± 12.0	88.4 ± 2.9
10	84.7 ± 13.2	89.1 ± 2.1
12	79.3 ± 10.6	83.2 ± 7.1
14	80.3 ± 7.8	84.1 ± 6.5
16	68.4 ± 29.4	88.4 ± 1.7
18	67.6 ± 20.6	82.9 ± 3.9
	internal	
6	90.2 ± 2.4	95.7 ± 0.3
8	90.9 ± 0.5	94.0 ± 1.5
10	91.7 ± 0.9	95.3 ± 0.5
12	89.9 ± 1.6	96.0 ± 0.9
14	91.6 ± 0.7	96.0 ± 0.7
16	91.1 ± 0.4	95.0 ± 0.9
18	89.6 ± 2.4	95.8 ± 0.4

[a]Errors indicate standard deviations.

To determine the nature of these fluctuations, the duration of the melting events was plotted against the occurrence, per nanosecond, of a melting event of that duration being initiated (Figure 8). A melting event was defined as an uninterrupted period of time with at least one hydrogen bond broken.

From Figure 8, it was observed that, for both AT and GC terminal base pairs, melting events of a duration less than one nanosecond were similarly common. By contrast, melting events longer than this were more common for terminal AT than GC pairs. This indicated that the large deviations in the number of unbroken bonds for the terminal AT pairs were likely a consequence of long, persistent melting events as opposed to short-lived ones. Since long melting events of this manner may destabilize the duplex and decrease its binding enthalpy, this may partly explain why the T.GC enthalpy, which is present only in the absence of AT termini, is negative.

In addition, the total melting time, obtained by aggregating all individual melting events during a run, was obtained. Total melting times for terminal (Figure S2) and internal (Figure S3) AT and GC base pairs were then plotted against the occurrence of a run with at least that much melting time out of all runs.

https://doi.org/10.1021/acs.jpcb.2c05547
J. Phys. Chem. B 2022, 126, 9528−9538

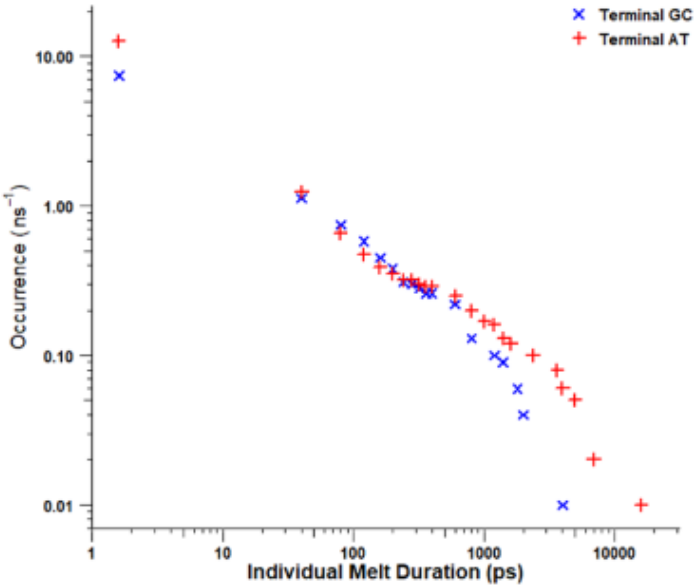


Figure 8. Occurrences per nanosecond of an individual melting event of at least a duration of picoseconds initiating for terminal GC base pairs, in blue, and terminal AT base pairs, in red.

Consequently, the behavior observed in the hydrogen bonding analysis can corroborate the predictions of the nearest-neighbor model. This supports the hypothesis that an MM-GBSA nearest neighbor model contains meaningful quantities that relate energy with structure.

■ CONCLUSIONS

Parameters for calculating the binding enthalpies of PNA duplexes as a function of their sequence have been developed. The difference between these predictions and PNA duplex binding enthalpies from thermal melting data was 8.74%. This was a reduction from an error of 10.01% from the direct comparison between MM-GBSA and thermal melting binding enthalpies. This suggested that the accuracy of the estimates improved when unrelated sequences were considered.

These results show that MM-GBSA can accurately decompose PNA binding enthalpies into parameters with structural meaning. This was demonstrated for the terminal GC term using a hydrogen bonding analysis, which showed that AT termini instabilities are characterized by persistent melting events.

Notably, the helix initiation enthalpy, which is positive for nearest-neighbor models of DNA and RNA,[48,49] was negative for PNA. A plausible reason for this is that the backbone of PNA is neutrally charged, and hence association does not have to overcome the like-charge repulsion that occurs between the phosphoribose backbones of DNA or RNA. Despite its approximate nature, MM-GBSA was able to resolve binding enthalpies with accuracy comparable to experimental error. This research demonstrates the usefulness of MM-GBSA for predicting thermodynamic quantities in cases of limited experimental data, as would be expected for studies into novel nucleic acids.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpcb.2c05547.

Binding energies and entropies for all models; convergence of MM-GBSA outputs; rmsd traces of example models and replicas; and rmsd histograms of all replicas (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Jack Goodman − *University of the West of England, Bristol BS16 1QY, U.K.*; orcid.org/0000-0003-4677-723X; Email: jack2.goodman@live.uwe.ac.uk

Authors

David Attwood − *University of the West of England, Bristol BS16 1QY, U.K.*

Janice Kiely − *University of the West of England, Bristol BS16 1QY, U.K.*

Pablo Coladas Mato − *GKN Aerospace, Bristol BS34 6FB, U.K.*; orcid.org/0000-0003-4347-6204

Richard Luxton − *University of the West of England, Bristol BS16 1QY, U.K.*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpcb.2c05547

Author Contributions

J.G. designed and conducted the research. J.G. and D.A. wrote the manuscript. The study was conceived by R.L. and J.K. All authors provided advice, reviewed the manuscript, and assisted in the direction of the research.

Notes

The authors declare no competing financial interest.

https://doi.org/10.1021/acs.jpcb.2c05547
J. Phys. Chem. B 2022, 126, 9528−9538

■ ACKNOWLEDGMENTS

This work was supported through joint funding from the University of the West of England and GKN Aerospace.

■ REFERENCES

(1) Wittung, P.; Nielsen, P. E.; Buchardt, O.; Egholm, M.; Nordén, B. DNA-like double helix formed by peptide nucleic acid. *Nature* **1994**, *368*, 561−563.
(2) Tomac, S.; Sarkar, M.; Ratilainen, T.; Wittung, P.; Nielsen, P.; Nordén, B.; Gräslund, A. Ionic Effects on the Stability and Conformation of Peptide Nucleic Acid Complexes. *J. Am. Chem. Soc.* **1996**, *118*, 5544−5552.
(3) Jasiński, M.; Miszkiewicz, J.; Feig, M.; Trylska, J. Thermal Stability of Peptide Nucleic Acid Complexes. *J. Phys. Chem. B* **2019**, *123*, 8168−8177.
(4) Nielsen, P. E. Peptide Nucleic Acid. A Molecule with Two Identities. *Acc. Chem. Res.* **1999**, *32*, 624−630.
(5) Eriksson, M.; Nielsen, P. E. PNA-nucleic acid complexes. Structure, stability and dynamics. *Q. Rev. Biophys.* **1996**, *29*, 369−394.
(6) Nielsen, P. E.; Egholm, M.; Berg, R. H.; Buchardt, O. Sequence-Selective Recognition of DNA by Strand Displacement with a Thymine-Substituted Polyamide. *Science* **1991**, *254*, 1497−1500.
(7) Shakeel, S.; Karim, S.; Ali, A. Peptide nucleic acid (PNA) - a review. *J. Chem. Technol. Biotechnol.* **2006**, *81*, 892−899.
(8) Gupta, A.; Mishra, A.; Puri, N. Peptide nucleic acids: Advanced tools for biomedical applications. *J. Biotechnol.* **2017**, *259*, 148−159.
(9) Ray, A.; Nordén, B. Peptide nucleic acid (PNA): its medical and biotechnical applications and promise for the future. *FASEB J.* **2000**, *14*, 1041−1060.
(10) Pellestor, F.; Paulasova, P. The peptide nucleic acids (PNAs), powerful tools for molecular genetics and cytogenetics. *Eur. J. Hum. Genet.* **2004**, *12*, 694−700.
(11) Moccia, M.; Adamo, M.; Saviano, M. Insights on chiral, backbone modified peptide nucleic acids: Properties and biological activity. *Artif. DNA PNA XNA* **2014**, *5*, No. e1107176.
(12) Jasiński, M.; Feig, M.; Trylska, J. Improved force fields for peptide nucleic acids with optimized backbone torsion parameters. *J. Chem. Theory Comput.* **2018**, *14*, 3603−3620.
(13) Golyshev, V.; Pyshnyi, D.; Lomzov, A. Calculation of Energy for RNA/RNA and DNA/RNA Duplex Formation by Molecular Dynamics Simulation. *Mol. Biol.* **2021**, *55*, 927−940.
(14) Sforza, S.; Haaima, G.; Marchelli, R.; Nielsen, P. Chiral Peptide Nucleic Acids (PNAs): Helix Handedness and DNA Recognition. *Eur. J. Org. Chem.* **1999**, *1999*, 197−204.
(15) Ratilainen, T.; Holmén, A.; Tuite, E.; Haaima, G.; Christensen, L.; Nielsen, P.; Nordén, B. Hybridization of Peptide Nucleic Acid. *Biochemistry* **1998**, *37*, 12331−12342.
(16) Sen, A.; Nielsen, P. On the stability of peptide nucleic acid duplexes in the presence of organic solvents. *Nucleic Acids Res.* **2007**, *35*, 3367−3374.
(17) Totsingan, F.Synthesis and Applications of PNA and Modified PNA in Nanobiotechnology. Ph.D. Thesis, University of Parma, Italy, 2007.
(18) Sen, A.; Nielsen, P. Unique Properties of Purine/Pyrimidine Asymmetric PNA-DNA Duplexes: Differential Stabilization of PNA-DNA Duplexes by Purines in the PNA Strand. *Biophys. J.* **2006**, *90*, 1329−1337.
(19) Marky, L.; Breslauer, K. Calculating Thermodynamic Data for Transitions of any Molecularity from Equilibrium Melting Curves. *Biopolymers* **1987**, *26*, 1601−1620.
(20) Alenaizan, A.; Barnett, J.; Hud, N.; Sherrill, D.; Petrov, A. The proto-Nucleic Acid Builder: a software tool for constructing nucleic acid analogs. *Nucleic Acids Res.* **2020**, *49*, 79−89.
(21) Yeh, J.; Pohl, E.; Truan, D.; He, W.; Sheldrick, G.; Du, S.; Achim, C. The Crystal Structure of Non-Modified and Bipyridine-Modified PNA Duplexes. *Chem.—Eur. J.* **2010**, *16*, 11867−11875.
(22) Kiliszek, A.; Banaszak, K.; Dauter, D.; Rypniewski, R. The first crystal structures of RNA-PNA duplexes and a PNA-PNA duplex containing mismatches-toward anti-sense therapy against TREDs. *Nucleic Acids Res.* **2015**, *44*, 1937−1943.
(23) He, W.; Hatcher, E.; Balaeff, A.; Beratan, D.; Gil, R.; Madrid, M.; Achim, C. Solution Structure of a Peptide Nucleic Acid Duplex from NMR Data: Features and Limitations. *J. Am. Chem. Soc.* **2008**, *130*, 13264−13273.
(24) van der Spoel, D.; Lindahl, E.; Hess, B.; Groenhoe, G.; Mark, A.; Berendsen, H. GROMACS: Fast, Flexible and Free. *J. Comput. Chem.* **2005**, *26*, 1701−1718.
(25) Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; de Groot, B.; Grubmüller, H.; MacKerell, A., Jr. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat. Methods* **2017**, *14*, 71−73.
(26) Berendsen, H. J. C.; Postma, J. P. M.; Gunsteren, W. F.; Hermans, J. Interaction Models for Water in Relation to Protein Hydration. *Intermolecular Forces* **1981**, *14*, 331.
(27) Bekker, H.; Dijkstra, E.; Renardus, M.; Berendsen, H. An Efficient, Box Shape Independent Non-Bonded Force and Virial Algorithm for Molecular Dynamics. *Mol. Simul.* **1995**, *14*, 137−151.
(28) Hockney, R. W.; Goel, S. P.; Eastwood, J. W. Quiet High Resolution Computer Models of a Plasma. *J. Comput. Phys.* **1974**, *14*, 148−158.
(29) Berendsen, H. J. C.; van Gunsteren, W. F.Practical algorithms for dynamics simulations. *Molecular-Dynamics Simulation of Statistical-Mechanical Systems*; Enrico Fermi Summer School, 1986; pp 43−65.
(30) Bussi, G.; Donadio, D.; Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **2007**, *126*, 014101.
(31) Hess, B.; Bekker, H.; Berendsen, H.; Fraaije, J. LINCS: A Linear Constraint Solver for Molecular Simulations. *J. Comput. Chem.* **1997**, *18*, 1463−1472.
(32) Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald: AnN·log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98*, 10089−10092.
(33) Essmann, U.; Perera, L.; Berkowitz, M.; Darden, T.; Lee, H.; Pedersen, L. A smooth particle mesh Ewald method. *J. Chem. Phys.* **1995**, *103*, 8577−8593.
(34) Frigo, M.; Johnson, S. The Design and Implementation of FFTW3. *Proc. IEEE* **2005**, *93*, 216−231.
(35) Goga, N.; Rzepiela, A.; de Vries, A.; Marrink, S.; Berendsen, H. Efficient Algorithms for Langevin and DPD Dynamics. *J. Chem. Theory Comput.* **2012**, *8*, 3637−3649.
(36) Berendsen, H.; Postma, J.; van Gunsteren, A.; DiNola, J.; Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **1984**, *81*, 3684−3690.
(37) Parrinello, M.; Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.* **1981**, *52*, 7182−7190.
(38) Valdés-Tresanco, M. S.; Valdés-Tresanco, M. E.; Valiente, P.; Moreno, Moreno E. gmx_MMPBSA: A New Tool to Perform End-State Free Energy Calculations with GROMACS. *J. Chem. Theory Comput.* **2021**, *17*, 6281−6291.
(39) Still, C.; Tempczyk, A.; Hawley, R.; Hendrickson, T. Semianalytical Treatment of Solvation for Molecular Mechanics and Dynamics. *J. Am. Chem. Soc.* **1990**, *112*, 6127−6129.
(40) Onufriev, A.; Case, D. Generalized Born Implicit Solvent Models for Biomolecules. *Annu. Rev. Biophys.* **2019**, *48*, 275−296.
(41) Bashford, D.; Case, D. Generalized Born Models of Macromolecular Solvation Effects. *Annu. Rev. Phys. Chem.* **2000**, *51*, 129−152.
(42) Jawad, B.; Poudel, L.; Podgornik, R.; Ching, W. Thermodynamic Dissection of the Intercalation Binding Process of Doxorubicin to dsDNA with Implications of Ionic and Solvent Effects. *J. Phys. Chem. B* **2020**, *124*, 7803−7818.
(43) Jawad, B.; Poudel, L.; Podgornik, R.; Steinmetz, N.; Ching, W. Molecular mechanism and binding free energy of doxorubicin intercalation in DNA. *Phys. Chem. Chem. Phys.* **2019**, *21*, 3877−3893.
(44) Gonçalves, P.; Stassen, H. Calculation of the free energy of solvation from molecular dynamics simulations. *Pure Appl. Chem.* **2004**, *76*, 231−240.

https://doi.org/10.1021/acs.jpcb.2c05547
J. Phys. Chem. B 2022, 126, 9528−9538

(45) Otsuka, K.; Jinno, K.; Morikawa, A. Active and Selective Catalysts for the Synthesis of C_2H_4 and C_2H_6 via Oxidative and Coupling of Methane. *J. Catal.* **1986**, *100*, 353−359.
(46) Janežič, D.; Venable, R. M.; Brooks, B. R. Harmonic analysis of large systems. III. Comparison with molecular dynamics. *J. Comput. Chem.* **1995**, *16*, 1554−1566.
(47) Onufriev, A.; Bashford, D.; Case, D. Exploring Protein Native States and Large-Scale Conformational Changes With a Modified Generalized Born Model. *Proteins* **2004**, *55*, 383−394.
(48) General, I. A Note on the Standard State's Binding Free Energy. *J. Chem. Theory Comput.* **2010**, *6*, 2520−2524.
(49) SantaLucia, J., Jr.; Allawi, H.; Seneviratne, P. Improved Nearest-Neighbor Parameters for Predicting DNA Duplex Stability. *Biochemistry* **1996**, *35*, 3555−3562.
(50) Xia, T.; SantaLucia, J., Jr.; Burkard, M.; Kierzek, R.; Schroeder, S.; Jiao, X.; Cox, C.; Turner, D. Thermodynamic Parameters for an Expanded Nearest-Neighbor Model for Formation of RNA Duplexes with Watson−Crick Base Pairs. *Biochemistry* **1998**, *37*, 14719−14735.
(51) SantaLucia, J., Jr. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Biochemistry* **1998**, *95*, 1460−1465.

■ NOTE ADDED AFTER ASAP PUBLICATION

Due to a production error, the version of this paper that was published ASAP November 14, 2022, contained an error in Table 2. This error was corrected and the paper was reposted November 15, 2022.

https://doi.org/10.1021/acs.jpcb.2c05547
J. Phys. Chem. B 2022, 126, 9528−9538