UNIVERSITY COLLEGE LONDON

Nucleic Acids at the Mineral Interface - An Origins of Life Study

by

Jacob B. Swadling

A thesis submitted in partial fulfillment for the
degree of Doctor of Philosophy

in the
Faculty of Mathematical & Physical Sciences (MAPS)
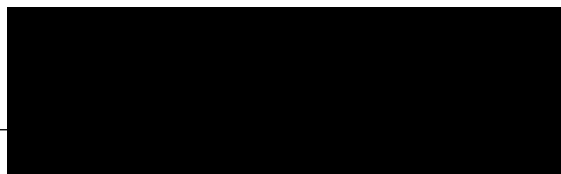&
Department of Chemistry

November 2013

Declaration of Authorship

I, Jacob B. Swadling, declare that this thesis titled, 'Nucleic Acids at the Mineral Interface - An Origins of Life Study' and the work presented in it is my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.

- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.

- Where I have consulted the published work of others, this is always clearly attributed.

- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.

- I have acknowledged all main sources of help.

- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date: 06/11/2013

UNIVERSITY COLLEGE LONDON

Abstract

Faculty of Mathematical & Physical Sciences (MAPS)

&

Department of Chemistry

Doctor of Philosophy

by Jacob B. Swadling

This thesis details work which employs classical simulation techniques to investigate the interactions of nucleic acid molecules with various charged clay mineral environments. There is a focus on the structure and stability of nucleic acids at mineral interfaces in order to understand how geological settings aided in fostering the first biomolecules at the time of the origins of life on Earth. A comparison of three nucleic acids, DNA, RNA and PNA, shows a difference in preferential stability in bulk water over the corresponding nucleic acid in a mineral environment. The comparative study showed that the prevailing geochemistry preferentially favoured DNA over potentially competing genetic candidates, such as RNA and PNA. This gives us a unique insight into how there may have been a transition from a proto-DNA world (such as the RNA world) to the current DNA/protein world. The structure and arrangement of single-stranded RNA on both cationic and anionic charged surfaces showed marked differences. Both cationic and anionic surfaces successfully adsorb on charged RNA oligonucleotides but show significant differences in the adsorbed structure. Cationic surfaces are successful in mediating the collapse of the RNA sequence from an elongated linear polymer into one that is capable of exhibiting catalytic function. The anionic surface elongates the RNA polymer and exposes the information carrying base sequence to the aqueous region allowing fidelity in templating information and replicating sequences. Studies of single-stranded RNA were extended to model a large biologically relevant RNA ribozyme using replica exchange sampling methods. The results elucidated the structure and arrangement of the catalytic centre of the ribozyme. The results in this thesis show that mineral mediated origins of life differ considerably from an aqueous one that is more commonly associated with the origins of life.

Acknowledgements

Contents

List of Figures

List of Tables

Abbreviations

LUCA	**L**ast Universal **C**ommon **A**nscester
RNA	**R**ibo**n**ucleic **A**cid
DNA	**D**eoxyribo**n**ucleic **A**cid
PNA	**P**eptide **N**ucleic **A**cid
UV	**U**ltra **V**iolet
PXRD	**P**owder **X-r**ay **D**iffraction
LDH	**L**ayered **D**ouble **H**ydroxide
XRD	**X-r**ay **D**iffraction
TGA	**T**hermo**g**ravimetric **A**nalysis
siRNA	**s**mall **i**nterfering **RNA**
mRNA	**m**essenger **RNA**
TIP	**T**ransferable **I**nteraction **P**otential
SPC	**S**ingle **P**oint **C**harge
SPC/E	**SPC/E**xtended
AMBER	**A**ssisted **M**odel **B**uilding with **E**nergy **R**efinement
CHARMM	**CH**emistry at **H**ar**v**ard **M**olecular **M**echanics
GROMOS	**Gro**ningen **M**olecular **S**imulation
ESP	**E**lectro**s**tatic **P**otential
DFT	**D**ensity **F**unctional **T**heory
MD	**M**olecular **D**ynamics
REMD	**R**eplica Exchange **MD**
TAD	**T**emperature **A**ccelerated **D**ynamics
XSEDE	**E**xtreme **S**cience and **E**ngineering **D**iscovery **E**nvironment
PRACE	**P**artner**s**hip for **A**dvanced **C**omputing in **E**urope
NMR NOESY	**N**uclear **M**agnetic **R**esonance, **N**uclear **O**verhauser **E**ffect **S**pectroscop**y**

CPU	**C**entral **P**rocessing **U**nit
GPU	**G**raphical **P**rocessing **U**nit
STFC	**S**cience and **T**echnology **F**acilities **C**ouncil
TNA	**T**hreose **N**ucleic **A**cid
QM	**Q**uantum **M**echanics
SIPF	**S**alt **I**nduced **P**eptide **F**ormation
LAMMPS	**L**arge-scale **A**tomic/**M**olecular **M**assively **P**arallel **S**imulator
NGS	**N**ational **G**rid **S**ervice
HECToR	**H**igh-**E**nd **C**omputing **T**erascale **R**esource
RDF	**R**adial **D**istribution **F**unction
VMD	**V**isual **M**olecular **D**ynamics
AHE	**A**pplication **H**osting **E**nvironment
AFT	**A**tomic **F**orce **M**icroscopy
RMSD	**R**oot **M**ean **S**quare **D**eviation
PCA	**P**rincipal **C**omponent **A**nalysis
MSD	**M**ean **S**quare **D**isplacement
MMSTB	**M**ultiscale **M**odeling **T**ools in **S**tructural **B**iology
trSAXS	**t**ime **r**esolved **S**mall **A**ngle **X**-ray **S**cattering
PDB	**P**rotein **D**ata**b**ase

Physical Constants

Constant Name	Symbol	=	Constant Value (with units)
Planck's Constant	h	=	1.582611×10^{-37} kcals
Boltzmann Constant	k_B	=	3.2976268×10^{21} kcal K^{-1}
Gas Constant	R	=	1.9858775×10^{-3} kcalK^{-1} mol^{-1}

Symbols

symbol	name	unit
a	distance	m
P	power	W (Js^{-1})
r_G	radius of gyration	Å
r_{RMSD}	root mean-square deviation	Å
$<r^2>$	mean square displacement	Å^2
ΔG	Gibb's free energy	kcal mol^{-1}

List of Publications

- Jacob B. Swadling, Peter V. Coveney, and H. Christopher Greenwell. Clay minerals mediate folding and regioselective interactions of RNA: A large-scale atomistic simulation study. *Journal of the American Chemical Society*, 132(39):13750–13764, 2010

- Jacob B. Swadling, Peter V. Coveney, and H. Christopher Greenwell. Computer simulation of clay mineral-biomolecule interactions. *Mineralogical Magazine*, 75 (1):1972, 2011. Conference Goldschmidt 2011, Prague, JUN 14-19, 2011

- Jacob B. Swadling, Peter V. Coveney, and H. Christopher Greenwell. Stability of free and mineral-protected nucleic acids: Implications for the RNA world. *Geochimica et Cosmochimica Acta*, 83:360–378, 2012

- Peter V. Coveney, Jacob B. Swadling, Jonathan A. D. Wattis, and H. Christopher Greenwell. Theory, modelling and simulation in origins of life studies. *Chemical Society Reviews*, 41(16):5430–5446, 2012

- Jacob B. Swadling, James L. Suter, H. Christopher Greenwell, and Peter V. Coveney. Influence of surface chemistry and charge on mineral-RNA interactions. *Langmuir*, 29(5):1573–1583, 2013

Chapter 1

Introduction

The subject of the origins of life is one that induces a great deal of interest in academia as well as in the public imagination. Given the vast amount of time which separates us from these events, the subject relies heavily on theoretical approaches in its understanding. The study of the chemical origins of life not only relies on a knowledge of biomolecular structure and function but the geochemical aspects which are inherent in prebiotic chemistry and important in mediating the chemical origins of life. This thesis starts with a review of current theories in origins of life studies, with particular emphasis on prebiotic chemistry where the chemical species and the fundamental phenomena defining life arose. Chapter 3 introduces the intricate structure of nucleic acids and the chemical composition of clay minerals which are adopted in the latter research chapters. Chapter 4 introduces the molecular simulation techniques which we have applied to the biomolecular/inorganic systems.

The question "How and where did life arise on Earth?" is one which has always attracted widespread attention, and increasingly so in recent years. Origins of life studies represent an exciting and highly multidisciplinary research field that incorporate contributions from many areas of research. Origins of life studies are an effort to understand the transition from chemistry to biology. Attempts to define life are irrelevant to scientific efforts; drawing an arbitrary line along the progression as *the point* at which life began would be difficult, and more importantly would not further our understanding of the transitions involved [6].

The uniformity of biochemistry in all living organisms indicates that life descended from a single last universal common ancestor (LUCA) [7]. It is possible, in principal, to compile a complete account of the origins of life if the stages in which LUCA evolved from abiotic components on early Earth was known. Stages which have been the focus of much attention in the origins of life community are concerned with finding what were

the sources of the organic molecules that made up the first self-replicating molecules, where was the source and how did biological organisation evolve from an abiotic source of organic molecules? [8].

Modern origins (of life) theories can be traced back to Charles Darwin's "warm little pond", often referred to as a "prebiotic soup". The warm little pond falls short due to a lack of any geological evidence and any apparent source of free energy. It is known that for the self-condensation required to form the first biopolymer two things are needed: energy and a sufficient concentration of precursor reactants; for RNA this would be an ample supply of ribonucleotides. Moreover, all life processes are inherently far from equilibrium whereas the prebiotic soup as normally conceived is in thermodynamic equilibrium; as a result of any such equilibrium, nucleotides would be reluctant to react further. To counter this a number of approaches where physical and/or chemical gradients impose far from equilibrium conditions have been hypothesised and tested including the use of minerals to concentrate monomers [9], salt induced peptide formation reactions which mimic tidal pools [10], and the concentration of monomers in freezing processes [11].

It is important when considering the formation of RNA strands during the origins of life not to ignore the nonlinear nature of RNA dynamical self-assembly. Too often the argument that the probability of synthesising a gram of the 'one' (or few) particular self reproducing sequences by a random assembly process would need more mass of substance than exists on Earth, has been echoed [12]. It is naïve to think that RNA sequences in a soup form by random synthesis (*i.e.* as if at equilibrium). It does establish that not all sequences have been made. Nature does not perform a global search for optimality over all possible chemical species. There is a large element of chance in determining which sequences have been produced. Life is indeed driven by a set of chemical processes taking place far from equilibrium [13–15]. To maintain these processes, all organisms are open systems; their complexity is founded on feedback involving autocatalytic and crosscatalytic molecules that assist reactions without being destroyed in the process. One metabolic or regulatory pathway may produce a molecule that accelerates other pathways (cross-catalysis) which, through a vast amount of interlinked chemistry, may end up indirectly catalysing the original pathway.

The shortcomings associated with the warm little pond idea were somewhat addressed by the seminal work of Miller and Urey. They passed an electric discharge through a mixture of methane, ammonia and hydrogen, producing amino acids, amongst other products. This reaction demonstrated that an energy input could generate biochemicals. Miller continued this research, showing that the major synthetic route to amino acids is through the Strecker reaction [16], although now there is evidence to suggest that

the early Earth did not have a strongly reducing atmosphere, making the Miller-Urey experiments less pertinent than originally thought [17]. Haldane proposed, sometime before the experiments of Miller that ultra-violet (UV) radiation provided the energy to convert methane, ammonia and water into the first organic compounds in the early ocean [18].

It was not until the 1970s that an alternative hypothesis was thought that provided a possible solution to the atmospheric conditions of the early Earth: The discovery of deep submarine hydrothermal vents and their subsequent implication as a possible geochemically plausible prebiotic setting. "Black smokers" were the first type of hydrothermal vent discovered. They arise in part from magma interacting with seawater at ocean spreading zones. These vents are far from equilibrium with the surrounding seawater [19]. Because of their extremely high temperatures, low pH, short lifetimes and a lack of three-dimensional compartmentalisation that would conceivably lead to irretrievable dilution of proto-biomolecules in the early ocean [20]. For these reasons, black smokers are not ideal settings for the origins of life.

A second type of hydrothermal vent was discovered at the turn of the millennium [21]. Unlike black smokers, these "alkaline vents" are not volcanic but form through serpentinisation. Serpentisation is the oxidation of olivine and consequent reduction of water to hydrogen [22]. The process creates temperatures of around 150-200°C and strongly alkaline fluids of pH 9-11 rich in hydrogen. Russell *et al.* suggested the role of these alkaline vent systems in prebiotic scenarios in 1993, before their eventual discovery at the "Lost City" of the mid-Atlantic [23]. Russel and co-workers hypothesised alkaline vents are potential candidates for producing life owing to their relatively benign temperature and pressure, apparent ability to incubate proto-metabolism through proton and redox potential gradients within emerging alkaline fluids, and the delicate internal porous structure of FeS minerals acting as an early protocells.

Possibly the most widely adopted theory in origins of life studies is the "RNA World" hypothesis, which states that at some point prior to the present deoxyribonucleic acid (DNA)/protein world, ribonucleic acid (RNA) provided the molecular basis for catalysis and replication [24]. The theory is attractive because it requires the synthesis of only one type of biomolecule which solves many of the problems of complexity. Evidence to support the RNA World hypothesis came from the work of Cech and Altman who won the 1989 Nobel prize in chemistry for their independent discovery of a group of catalytic RNA molecules known as ribozymes [25]; and further from the discovery, by Stietz and Moore, of the structure of the ribosome, a cellular component which makes proteins [26] that was shown to be a ribozyme. More recently, Sutherland and co-workers provided an abiotic synthetic route to nucleotides, under near prebiotic conditions [27].

In essence, there are many questions which remain unanswered in our quest to discover the origins of life on Earth (and of course elsewhere in the Universe) but theory and modelling are making a significant impact on our understanding of the stages in the transition from chemistry to biology. I believe that the research detailed within this thesis goes some way to answering various hitherto unanswered questions associated with stages in the origins of life. And it is hoped that these answers will go on to inspire new experiments and to enhance the reputation of modelling methods within many sections of the experimental prebiotic chemistry community.

f

Chapter 2

Theory, Modelling and Simulation in Origins of Life Studies

Origins of life studies represent an exciting and highly multidisciplinary research field. Contributions made by theory, modelling and simulation to addressing fundamental issues in the domain and the advances these approaches have helped to make in the field are discussed in detail within this thesis chapter. Theoretical approaches will continue to make a major impact at the "systems chemistry" level based on the analysis of the remarkable properties of nonlinear catalytic chemical reaction networks, which arise due to the auto-catalytic and cross-catalytic nature of so many of the putative processes associated with self-replication and self-reproduction. Here we review a description *inter alia* of nonlinear kinetic models of RNA replication within a primordial Darwinian soup, and state-of-the-art computationally-based molecular modelling techniques that are currently being deployed to investigate various related scenarios relevant to the origins of life.

2.1   Theoretical nonlinear chemical kinetics

For a complex system of interlocking chemical reactions, the second law of thermodynamics is a good place to start to guide us in setting parameters to limit the search for the origin of life? It defines a set of thermodynamic potentials whose extrema specify the equilibrium states of matter. In particular, under the usual conditions pertaining in chemical reactions (fixed temperature and pressure, or fixed temperature and volume), the Gibbs and Helmholtz free energies serve as potentials whose global minima completely define the equilibrium state [28].

If the system is prevented from attaining thermodynamic equilibrium through the imposition of external constraints (*e.g.* temperature or concentration gradients, *etc.*), but is able to get sufficiently close to equilibrium (technically, the thermodynamic forces and fluxes are linear in the displacement from their vanishing equilibrium values), then the appropriate thermodynamic potential is the internal entropy production of the system [29]. That is, its extrema (usually they are minima, but in some cases maxima) define the stationary states of the system.

However, for systems far from equilibrium (that is, for which the thermodynamic forces and fluxes are no longer linearly dependent on the distance from equilibrium), there is no thermodynamic potential that guides the evolution of the system. Although this result was established more than 35 years ago, remarkably little attention has been paid to it [30, 31]. Beyond the overriding importance of irreversibility (that is, entropy is always increasing) chemical thermodynamics in the conventional sense has very little to say about the global behaviour of systems far from equilibrium, such as living states of matter and those involved at the origins of life. Instead, a more quantitative insight can be gained from the study of nonlinear dynamics, the formal mathematical domain within which the theoretical basis of complex reaction kinetics resides, all compatible with the second law.

Enormous progress has been made in the understanding of the behaviour of nonlinear dynamical systems over the past forty years or so, much of it traceable to Turings seminal paper on the chemical basis of morphogenesis [32]. Chemical reactions far from equilibrium provide an example because many rate processes (including autocatalytic and cross catalytic ones), based on the law of mass action, are nonlinear. Qualitatively, what makes nonlinear systems so fascinating is that they exhibit "unexpected" properties from the standpoint of conventional linear, equilibrium, or steady-state theory. Nonlinear systems are much more reminiscent of the world we inhabit: they can exhibit multiple states for the same equations and parameters (such as multi-stability and hysteresis), various forms of spatial, temporal and spatio-temporal organisation (often referred to as dissipative structures or "self organisation"), chaos (in the technical sense) and its corollary, sensitive dependence on initial conditions [30]. It is through the application of the mathematics of nonlinear dynamical systems that a better understanding of the origins of life can be gained.

2.1.1 Recent assessment of the RNA world hypotheses

Although the RNA world is in various ways conceptually attractive, no one has yet produced experimental proof of its viability. As a result, various theoretical issues

concerned with the scenario are often subject to analysis.

Szostak *et al.* highlight the potential and the aims of the field of synthetic life and the RNA world in particular [33]. A central challenge is the discovery of a RNA replicase. This problem is also noted by Ma and Wu, who suggest that self replicating RNA molecules must be the forerunner of the RNA world, the information content coming in later [34]. Even in the presence of a bountiful supply of substrate, the assembly of RNA strands is inefficient; hence Szostak *et al.* suggest that activated nucleotides are probably required. There are no naturally-occurring ribozymes which can catalyse the necessary chemical reactions. A replicase has to function both as a ribozyme template for replication, which suggests that it has a naturally open structure, and to act as an active polymerase, suggesting that it has a stable folded structure.

These apparently contradictory statements can be reconciled by assuming that the environment undergoes some form of regular oscillatory behaviour; for example the tidal cycling resulting in temperature cycling or periodic fluctuations in the concentrations of monomers. The early impact which created the Moon also caused the Earth to undergo rapid rotation, with a periodicity in the range 2–6 hours [35, 36]. This caused significant tidal activity in coastal regions, the flooding and drying producing oscillations in salinity and temperature permitting the interstrand interactions between DNA chains to alternately promote association and dissociation. This would allow copying of the templated information in strands. Since there is only a relatively short period of a few hundred million years between the end of the late heavy bombardment and the origin of life, this theory brings with it the concomitant need to explain the origin of life over a rather short time scale of around 300,000 years.

Fernando *et al.* support the idea that systems with templating but no enzymatic catalysis undergo chain lengthening, which causes a reduction in the total concentration of chains and the overall reaction to slow down [37]. They construct a computer-based model of this process and find that tidal cycling significantly increases the rate of elongation.

Another model of replication was studied by Fernando and di Paolo, namely the Chemoton [38]. A definition of life was proposed by Ganti as involving metabolism, regulation, control, being stable and having an informational subsystem [39]. The Chemoton has autocatalysis, a membrane, and an informational subsystem. This advocates the idea that templated replication can occur without the need for enzymatic catalysis. Their model combines replication of cell-like structures, double-stranded polymers and metabolism. They find only rare elongation of the polymer and oscillations in the monomer concentration throughout the cell-cycle.

In addition to the replication of the information-carrying RNA, Szostak *et al.* also note that in order to replicate successfully, cells need to be able to grow and divide their cell membranes [33]. These membranes also need to satisfy key properties, namely that small molecules should be able to diffuse across on a reasonable timescale, yet large macromolecules should be generally confined to their interiors. Furthermore the chemical environment inside the membrane should be simultaneously beneficial to membrane growth and RNA replication.

The various potential mechanisms underlying the RNA world hypothesis are also discussed by Joyce [12]. In his view the main challenge in the RNA world hypothesis is overcoming the clutter associated with prebiotic chemistry. He notes that although replication is currently carried out in a residue-by-residue fashion, this need not be the case in the prebiotic environment. As long as the production rate exceeds the decay rate, a polymer or cell will proliferate. Furthermore, even had it existed, the RNA world was not necessarily the starting point for life, as there could have been simpler genetic systems prior to it, for example TNA or PNA (threose nucleic acid or peptide nucleic acid; see Figure 2.1), which then transformed into the RNA world. Alternatively, there could have been other replicating mechanisms allowing templating, for example involving inorganic clays [40]. Joyce considers more of the details of potential pathways for RNA and protein synthesis catalysed by RNA. Again the issue of compartmentalisation is raised; amongst the advantages of cell-like structures in which such reactions could occur are the maintenance of sufficient concentrations of product species and the suitability of the environment for reactions. Joyce notes that these compartments need not be the phospholipid bilayers common to modern cells; they could be airborn aerosols, rock pores, β-sheets of other polypeptides, or terpenoids [12]. A further unanswered question for the RNA world is how the transition to the DNA-makes-RNA-makes-protein biochemistry occurred.

The error threshold for the self-replication of a chemical species which acts as both an enzyme and a substrate is analysed by Obermeyer and Frey [41] via a model system of equations of the form

$$\frac{\mathrm{d}X_i}{\mathrm{d}t} = \sum_j M_{i,j} R_j X_j - X_i \sum_j R_j X_j, \qquad (2.1)$$

where $M_{i,j}$ is a mutation probability, $X_j(t)$ are the concentrations of chemical species j and $R_j = A_j + \sum_i B_{j,i} X_i$ is the replication rate of species j. They note that the potential for nonenzymatic replication is questionable. They find that only weak specificity is required in order for the distribution of species to be localised about the master sequence.
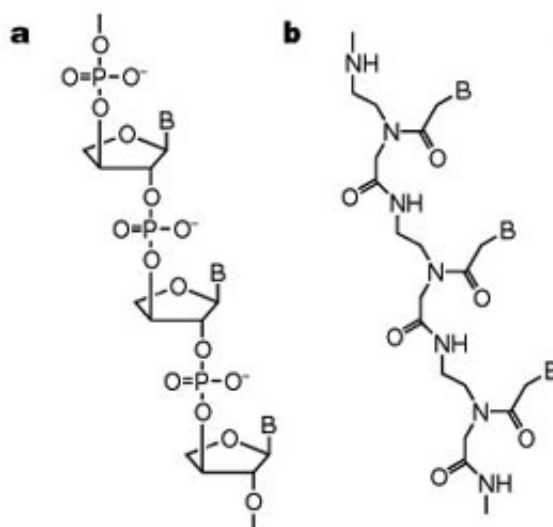
Figure 2.1: Chemical structure of a) threose nucleic acid; b) peptide nucleic acid, where B is a nucleotide base.

Stronger specificity constraints provide stronger localisation and so longer sequences can be replicated, which also furnishes more tolerance against mutations.

2.1.2 The origin of the RNA world

Wattis and Coveney constructed and analysed a microscopic kinetic model for the emergence of long chains of RNA from monomeric β-D-ribonucleotide precursors in prebiotic circumstances [42]. Wattis and Coveney's theory starts out from the general chemical assumption that catalytic replication can lead to a large population of long chains. The models incorporate the possibility of (i) direct chain growth (ii) template assisted synthesis and (iii) catalysis by RNA replicase ribozyme, all with varying degrees of efficiency. However, all chemical processes are "open"; they do not assume the existence of closed hypercycles which sustain a population of long chains, rather it is the feasibility of the initial emergence of a self-sustaining set of RNA chains from monomeric nucleotides that is of concern. Detailed microscopic kinetic models lead to kinetic equations which are generalisations of the Becker-Döring system for the stepwise growth of clusters or polymer chains; they lie within a general theoretical framework which Coveney and Wattis have successfully applied to a wide range of complex chemical problems [43].

The most detailed model Coveney and Wattis considered contains Becker-Döring aggregation terms, together with a general Smoluchowski fragmentation term to model the competing hydrolysis of RNA polymers [43]. This last process is a key one that enables all chemical ingredients to be recycled and leads to the massive amplification of some

self-replicating sequences over all others. The main reactions that such growing chains undergo are (i) the basic Becker-Döring rate processes controlling chain growth

$$C_r^\gamma + N_i \underset{}{\overset{slow}{\rightleftharpoons}} C_{r+1}^{\gamma + N_i}. \tag{2.2}$$

Here, the four nucleotide bases are denoted by N_i (with $i = 1$, 2, 3 or 4) and oligomeric ribonucleotide sequences by C_r^γ where r signifies the number of bases in the sequence and γ denotes the particular order in which they occur. (ii) Template-based chain synthesis (a form of catalysis mediated by Watson-Crick base pairing of ribonucleotides on complementary chains)

$$C_r^\gamma + N_i + C_s^\theta \underset{}{\overset{fast}{\rightleftharpoons}} C_{r+1}^{\gamma + N_i} + C_s^\theta. \tag{2.3}$$
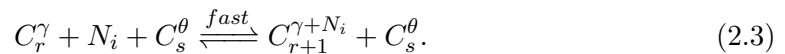
(iii) Hydrolysis, whereby a long chain is split into two shorter chains. Chemically this corresponds to the process

$$C_{r+s}^{\gamma + \theta} \longrightarrow C_r^\gamma + C_s^\theta. \tag{2.4}$$

This has the form of a general fragmentation process as modelled by the Smoluchowski equations, a mechanism which increases the number of chains but reduces the average chain length. (iv) Enzymatic replication (replicase ribozymal activity), where a third chain aids the growth of a chain which is already in close contact with another chain acting as a template is modelled by

$$C_r^\gamma + N_i + C_{r+k}^{\gamma* + \theta*} + C_s^\xi \rightleftharpoons C_{r+1}^{\gamma + N_i} + C_{r+k}^{\gamma* + \theta*} + C_s^\xi. \tag{2.5}$$
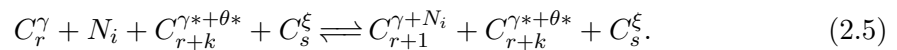
Here the combination of γ with N_i is a subsequence of the chain $\gamma + \theta$, while C_s^ξ plays the part of a replicase ribozyme. Needless to say, some of these replicases will have much higher efficiency than (most) of the others.

Wattis and Coveney demonstrated that it is possible to realise the selection of certain self-replicating RNA polymer chains in a reasonable amount of time starting from plausible assumptions about the chemistry and initial conditions that could have prevailed within a putative prebiotic soup comprised of β-D-ribonucleotide monomers [42, 43].

2.1.3 Modelling of elements of the RNA-world

All the models developed by Coveney and Wattis are based on deterministic nonlinear dynamics, being derived from macroscopic laws of mass-action. For situations that may frequently pertain at the cellular level, it is by no means clear that this "law" is always valid. With small numbers of molecules involved, more inherently probabilistic models have an important role to play. Thus, Hanel *et al.* consider the kinetics of stochastic minimally nonlinear models, of the form $\dot{x}_i = \sum_j A_{i,j} x_j + J_i + \nu_i$, where the connectivity matrix A has the form of an Erdos-Renyi network [44]. Whilst choosing A to be a random matrix, they impose the constraint $x > 0$, which makes the system nonlinear, hence the Lyapunov exponents may be positive, negative or zero. If all exponents are negative, the system converges to a steady-state or equilibrium solution; if the largest exponent is strictly positive then the system is chaotic and exhibits strong instabilities, sensitive dependence on initial conditions and other 'random' behaviour not conducive to the persistence of living systems. However, if the largest exponent is zero, then the system is at criticality, where a greater range of states can be explored and nonlinear effects are dominant, aiding the control of behaviour; this state is of great interest in the modelling of living systems, wherein the types of behaviour exhibited include oscillations. The average degree of the network $\langle k \rangle = L/N$, where L is the number of links, and N is the number of nodes. There is a considerable range of $\langle k \rangle$ where the largest eigenvalue is close to zero, indicating applicability to living systems.

Lehman proposes that even in the early RNA world self-replication occurred via RNA strands breaking into shorter strands which replicated the daughter strands, recombining to form multiple copies of the original RNA molecule [45]. The reasons for this are many: recombination is an energy-neutral process, so would have been common in prebiotic conditions; an RNA replicase ribozyme would have to be long, typically in excess of 100 nucleotides. This is too long to replicate accurately via catalysis, whereas shorter RNA strands could easily replicate. One consequence of this hypothesis is a lengthening of RNA in the population over time; this is entirely consistent with the development from simple prebiotic life to more advanced function.

As mentioned earlier, one problem in analysing models of the RNA world is the vast (combinatorially large) number of possible species. In general there is not enough mass for all to be explored, that is, for even one molecule of each to be created, so it is not a question of which has the most stable structure theoretically, but which *of those that have been created* is best at replicating. A simpler problem, advocated by many authors instead, is to consider a reduced system of species, in which there are only two species competing for a substrate. A good example of the two species are the left- and right-handed chiral structures found in many living systems. Chemically, physically and

thermodynamically they have the same stabilities, yet competition and other nonlinear effects cause one to be eliminated and the other to become dominant.

2.1.4   Compartmentalisation: self-replicating micelles and vesicles

Experimental scenarios leading to self-reproducing vesicles have been successfully developed by Luisi's group [46]. Walde *et al.* describe the conditions under which vesicles formed by caprylic acid and oleic acid in water undergo autopoietic self-reproduction. The increase in their population number is due to a reaction which takes place within the spherical boundary of the vesicles themselves. This happens by letting caprylic or oleic anhydride hydrolyze at alkaline pH. The initial increase of the concentration of acid/carboxylate is slow, but the presence of vesicles above a critical concentration brings about a rapid second phase leading to more vesicles being formed in an overall autocatalytic process. The process of autopoietic self-reproduction of caprylic acid and oleic acid vesicles was also studied as a function of temperature by Walde *et al.*. They showed that by increasing temperature (up to 70°C), the domain of rapid of vesicle formation becomes steeper while the long initial slow phase is significantly shortened.

Models of self-replicating vesicles have been proposed and analysed by Coveney and Wattis [47, 48]. They constructed a kinetic model based on a novel generalisation of the Becker-Döring cluster equations which describe the stepwise growth and fragmentation of vesicular structures. Their nonlinear kinetic model is complex and involves many microscopic processes; however, by means of a systematic contraction of the complete set of kinetic equations to the macroscopic limit, they showed that the model correctly captures the experimentally observed behaviour of a long slow induction phase, a rapid autocatalytic phase followed by slow convergence to equilibrium. This model was generalised further by Bolton and Wattis to account for the size templating effect observed by Lonchin *et al.* and Berclaz *et al.* When the systems studied there were initiated with no vesicles, a broad range of vesicle sizes developed slowly; whereas in a system with an initial distribution of vesicles of one particular size (essentially monodisperse), vesicles of the same size are produced more quickly [46, 49, 50].

2.1.5   Accumulation of nucleotides in hydrothermal pore systems

The RNA world requires high concentrations of small prebiotic molecules, whereas geochemical extrapolations suggest the presence of a dilute prebiotic ocean with concentrations comparable with modern day values [51]. Braun *et al.* proposed a potential solution to this discrepancy in concentrations using joint experiment and simulation [52, 53]. Combined solutions of the Navier-Stokes equations, molecular diffusion and heat transfer

in two dimensions using a finite element solver in order to simulate molecular transport in elongated hydrothermal pore systems influenced by a thermal gradient. They found extreme accumulation of molecules in a wide variety of plugged pores. The authors state that the mechanism is able to provide highly concentrated single nucleotides suitable for operations of an RNA world at the origins of life [52]. Accumulation is driven by the thermal gradient across a pore. The fluid is shuttled by thermal convection along the pore, whereas the molecules drift across the pore driven by thermal diffusion. Baaske *et al.* showed that millimetre-sized pores accumulate even single nucleotides more than 108-fold into micrometer-sized regions than would be the case without the thermal diffusion. An enhanced concentration of molecules is found in the bulk water region near the closed end of the pore. Because the accumulation depends exponentially on the pore length and temperature difference, it is robust with respect to changes in cleft geometry and molecular dimensions. Baaske *et al.*'s findings suggest that, for life to evolve, complicated active membrane transport is not required initially. Instead, interlinked mineral pores in thermal gradients can provide a high concentration starting point for the molecular evolution of life. A detailed overview of the simulation techniques used in the various simulations, which are described in the following sections, is given in the previous thesis chapter.

2.1.6 Free Energy

Free energy is the portion of energy that is available to perform thermodynamic work; *i.e.*, work mediated by thermal energy. Free energy is subject to irreversible loss in the course of such work. The free energy is an indicator of the thermodynamic stability of a system and it governs the direction of spontaneous change. The activated process barrier, the change in free energy, can give us the kinetics of a system. The Gibbs free energy is given by the equation:

$$\Delta G = \Delta H - T\Delta S, \tag{2.6}$$

where H is the enthalpy ($H = U + pV$, where p is the pressure and V is the volume), T is the absolute temperature and S is the final entropy.

The Helmholtz free energy is given by:

$$\Delta F = \Delta U - T\Delta S, \tag{2.7}$$

where U is the internal energy of the system.

In thermodynamics the chemical potential is a form of potential energy that can be absorbed or released during a chemical reaction. The chemical potential of a species in the mixture can be defined as the slope of the free energy of the system with respect to a change in the number of moles of just that species. Thus, it is the partial derivative of the free energy with respect to the amount of the species, all other species' concentrations in the mixture remaining constant, and at constant temperature. When pressure is constant, chemical potential is the partial molar Gibbs free energy. At chemical equilibrium or in phase equilibrium the total sum of chemical potentials is zero, as the free energy is at a minimum.

$$\mu_a = (\frac{\partial F}{\partial N_a})_{NVT} = (\frac{\partial G}{\partial N_a})_{NPT}, \tag{2.8}$$

A partition function describes the statistical properties of a system in thermodynamic equilibrium. They are functions of temperature and other parameters, such as the volume enclosing a gas. Most of the thermodynamic properties of the system, such as the total energy, free energy, entropy, and pressure, can be expressed in terms of the partition function or its derivatives.

$$Q(N, V, T) = \frac{1}{N!} \frac{1}{h^{3N}} \int d\mathbf{p}^N \exp[-p^2/2mkT] \int d\mathbf{r}^N \exp[-\beta U(\mathbf{r}^N)], \tag{2.9}$$

$$Q(N, V, T) = \frac{1}{\Lambda^{3N} N!} \int d\mathbf{r}^N \exp[-\beta U(\mathbf{r}^N)], \tag{2.10}$$

where $\beta = 1/kT$ and $\Lambda = (h^2/2\pi mkT)^{1/2}$.

$$F = -kT\ln Q, S = kT\frac{\Delta \ln Q}{\Delta T} + k\ln Q, U = kT^2\frac{\Delta \ln Q}{\Delta T}, \tag{2.11}$$

The challenge associated with calculating free energies is that a direct integration of the partition function (sample the entire phase space) is not feasible. There are five broad classes of methods which we can employ to calculate free energies. i) direct calculation of the partition function, *e.g.*, free energy of solids using lattice dynamics. ii) free energy perturbation, *e.g.*, umbrella sampling and expanded ensembles. iii) From the density of states using the multi-canonical partition function. iv) Using potential of mean force (free energy barriers) calculations. v) Thermodynamic integration.

2.2 Molecular modelling

Previous sections have been concerned with studying the range of chemical kinetic models, based on the macroscopic law of mass action. Microscopic, molecular theory also has much to say on origins of life. Origins of life studies have hitherto rarely used computer simulation based molecular modelling techniques to understand the possible chemical pathways to the formation of the first biomolecules. Simulation methods provide unlimited molecular information about systems that are difficult to characterise experimentally because of a lack of measurement resolution and/or the extreme temperature and pressure conditions associated with these origins of life systems. This section considers the computational approaches used and the scientific insights gained from these molecular simulation techniques in terms of origins of life studies.

2.2.1 Use of quantum mechanics simulations in origins of life research

Quantum mechanical (QM) treatments offer methods of simulating small numbers of atoms at high levels of accuracy. QM studies are also known as *ab initio*, or first principles, methods because, unlike molecular mechanics, they start by solving the ground state electronic structure of the system of interest and, other than the electronic configuration and atomic mass of the atoms present are essentially entirely, parameter free. These methods offer unique insight into bond forming mechanisms, or detailed structural and spectroscopic aspects that depend on small differences in local electronic structure. As such, the primary places that QM is used in origins studies are within probing catalysis (mainly mineral surface mediated) and chiral selectivity by minerals. In order to make QM calculations tractable, and applicable to extended structures such as minerals, the periodic boundary conditions are often imposed in all three spatial directions so that a small simulation cell is mathematically repeated in all space. By far the most employed QM technique is the density functional theory (DFT) approach.

2.2.1.1 Reactivity at mineral surfaces

In contrast to other research areas within origins of life studies, much of the focus of those involved in QM simulations has been in understanding the process of mineral catalysed peptide bond formation with numerous groups electing to simulate this road to a protein world as compared to those that favour the RNA world approach. One reason for this bias is that the peptide forming reaction involves a relatively simple mechanism, with small reactants, and hence is computationally tractable when compared to many of the reactions involved in nucleic acid synthesis. A comprehensive review of peptide

formation on minerals, including both experimental and molecular modelling studies, is presented by Lambert [54], who identifies a range of simulation studies performed between 1988 and 2007 on silicate, titania, clay and pyrites surfaces.

The peptide forming reaction requires a dehydration and condensation mechanism. It can be readily appreciated that this would be highly disfavoured for amino acids in dilute solution. As such, mineral surfaces [9] and concentrated salt solutions (salt induced peptide formation, SIPF [55]) have been postulated and experimentally shown to promote peptide bond formation, especially where these are coupled to wet/dry cycles as is plausible in a putative tidal lagoon on the Hadean Earth. It is noteworthy that the whole field of research into SIPF arose from Monte Carlo computer simulations of the dehydration of sodium chloride solutions, which showed an unsaturated inner hydration shell of sodium ions at above 3M concentration [56].

Mineral induced peptide formation also has a long research track record. Lahav and co-workers experimented on various ways to activate amino acids from the mid-1970's onward [57, 58] and this was then investigated computationally by Loew *et al.* who used QM (non-DFT) methods to look at the effect of these activating molecules (H_3PO_4, H_2SO_4) and included clusters of atoms to represent clay edges ($Al(OH)_4^-$ and $Si(OH)_4$) [59].

Observing the dearth of mechanistic detail in many of the experimental reports on clay mineral and alumina catalysed peptide bond formation, Aquino *et al.* looked at amide bond formation using DFT, choosing to include explicit and implicit water molecules, though approximating the amino acids in terms of the simpler species acetic acid (CH_3CO_2H) and methylamine (CH_3NH_2) as the reactants [60]. The authors extended the earlier studies of Loew *et al.* by exploiting increasing computing power to simulate a variety of Lewis and Brønsted acid sites, including $Al(OH)_3$, $[Al(H_2O)_5]^{3+}$, H^+, H_3O^+-H_2O, and H_2O and hence calculate activation energies with commensurate thermodynamic outputs. A key conclusion was the importance of Lewis acid sites and the effect of pH, which the authors saw as a challenge for future advances in the field.

Taking forward the seminal experimental work on SIPF of Rode *et al.* [61–64], which showed that Cu^{2+} plays a key role in promoting abiotic peptide bond formation in wetting and drying cycles, Rimola *et al.* [65] used DFT simulation methods to determine that the solvated Cu^{2+} cation with explicit water molecules reduces the activation barrier for peptide bond formation for di-glycine from 55 kcal.mol^{-1} to 20 kcal.mol^{-1} and therefore favoured the forward dehydration reaction for peptide formation.

2.2.1.2 Reactivity of amino acids in aqueous solution

In addition to the burgeoning literature on prebiotic peptide bond formation at a variety of mineral surfaces, Nair *et al.* employed first principles molecular dynamics to show that activating agents such as carbonyl sulphide (COS) in tandem with nonequilibrium high pressure and temperature conditions just under the critical point for water could also mediate peptide bond formation without needing to invoke additional mineral surfaces, though this does not fully address the concentration problem discussed [66]. Experimental evidence for such processes had already been offered by Leman *et al.* in the case of COS activation [67], and Huber and Wächtershäuser for hydrothermal conditions [30]. The authors identified key constraints in this reaction, including that: i) extreme conditions stabilise the required neutral forms of the reactants and as such enhance potential reactivity under neutral pH conditions; ii) high temperatures ensure increased thermal reactivity, and finally; iii) high pressure water changes the selectivity to concerted rather than stepwise reaction pathways.

2.2.1.3 Chiral selectivity in minerals

A key area where DFT can add insight is that of understanding the specific binding interactions of chiral proto-biomolecules and biomolecules at chiral surfaces in an effort to elucidate where initial asymmetric imbalances may have seeded kinetic resolution processes. One of the earlier studies using DFT, by Yu *et al.* [68], probed the structures of the enantiomers of a model di-peptide interacting with the interlayer of nontronite, an iron rich clay. The dipeptide of alanine (Ala) was found to exhibit different structures and different binding energies dependent on the enantiomer. The L-Ala-mineral system was shown to be 6 kcal.mol^{-1} more stable than the corresponding D-Ala system, with a readily identifiable conformational structural difference between the two enantiomer-mineral systems. The authors report the compatibility of repeating motifs of the clay structures with structural features of the di-peptides. Such motifs have also been identified by Thyveetil as important in other layered mineral-biomolecule systems [69].

2.2.2 Catalytic nature of mineral surfaces

Mathew and Luthey-Schulten used molecular dynamics, as described in the previous chapter, to investigate a proposed origins of life scenario involving clay montmorillonite and its catalytic role in forming oligonucleotides from activated nucleotides [70]. Their

simulations provide atomic detail of reactant conformation prior to polynucleotide formation, furnishing insight into reported experimental observations by Huang *et al.* [71]. The simulations clarify the catalytic role of metal ions, demonstrate that reactions leading to correct linkages take place primarily in the interlayer, and explain the observed sequence selectivity in the elongation of the chain. Mathew *et al.* went further to compare reaction probabilities involving L- and D-chiral forms of the reactants and found enhancement of homochiral over heterochiral products when catalysed by montmorillonite. The simulations confirmed the synthesis of oligonucleotides should proceed in the $3'$-$5'$ direction when intercalated, just as in template directed synthesis in the RNA polymerase. Interestingly, in terms of origins of life studies, Mathew and Luthey-Schulten's simulations reveal increased regioselectivity for $3'$-$5'$ over competing $2'$-$5'$ linkage formation, as well as an overall increased catalytic effect when the reaction takes place in the interlayer.

2.2.3   Stability of free and mineral-protected nucleic acids

Although deep ocean hydrothermal vents have generated particular interest as a possible source of the first life forms, there remains the question of how biopolymers such as RNA could have remained intact at the elevated temperatures and pressures around these vents. One possible explanation is that clay-like particles may have acted as structures which supported and protected nucleic acids once formed. Experimentally, it has been shown that alkanes are formed when methanol reacts with smectites such as montmorillonite under conditions similar to those at hydrothermal seafloor vents. Although layered double hydroxides (LDHs) are not as naturally widespread in present times, evidence suggests that during early ages of the Earth, called the Archean era, minerals such as green rust, may have been much more common due to the lack of oxygen in the atmosphere [72, 73].

Due to the nanoscale dimensions of the interlayer region of layered double hydroxides (LDHs) (see Figure 2.2), the exact conformation of the intercalated DNA is difficult to elucidate experimentally. Thyveetil *et al.* used molecular dynamics techniques performed on supercomputing grids to carry out large scale simulations of double stranded, linear and plasmid DNA up to 480 base pairs in length intercalated within a magnesium-aluminium LDH [69].

Thyveetil *et al.*'s models were found to be in agreement with experimental observations, according to which hydration is a crucial factor in determining the structural stability of DNA. Phosphate backbone groups were shown to align with aluminium lattice positions. At elevated temperatures and pressures, relevant to origins of life studies, some of which

maintain that the earliest life forms originated around deep ocean hydrothermal vents, the structural stability of LDH intercalated DNA is substantially enhanced as compared to DNA in bulk water. Thyveetil *et al.*'s simulations use high temperatures and pressures to observe how the structure and stability of DNA is altered under these conditions when intercalated. The increasing temperature and pressure of the simulations confirmed the DNA is stabilised once intercalated, since the number of Watson-Crick hydrogen bonds rapidly degrades for DNA in bulk water under similar conditions.

2.2.4   The role of host layer flexibility in DNA guest intercalation

Another set of simulations performed by Thyveetil *et al.* used the same molecular dynamics techniques to investigate how layered double hydroxides can form staged intermediate structures [72]. These staged intermediates have been observed experimentally, although the mechanism of their formation has not been determined. Thyveetil *et al.* showed that LDHs are flexible enough to corrugate around bulky intercalants such as DNA. The simulations explore three possible intermediate structures that may form during intercalation of DNA into Mg_2Al–LDH and how the models differ energetically. The results showed that when DNA strands are stacked directly on top of each other, the LDH system has a higher potential energy than when they are stacked in a staggered or interstratified manner. The simulations showed that, on average, greater diffusion coefficients arise for DNA strands in the Daumas-Hérold configuration compared to a Rüdorff model and Stage-1 structure - Figure 2.3 shows a schematic of the various staging structures described.

Peristaltic modes were shown to be more prominent in the Daumas-Hérold structure compared to the Rüdorff and Stage-1 structures and support a mechanism by means of which bulky intercalated molecules such as DNA rapidly diffuse within interlayers.

2.3   Conclusions

This Chapter has reviewed chemical kinetic and molecular modelling approaches that are now throwing very considerable light on numerous challenging issues associated with the origin of life on Earth (and possibly elsewhere in the Universe). The methods available are powerful and wide ranging. They span a host of length and time scales, from the quantum mechanical description of electron dynamics, through the atomistic and molecular levels which are described most often by classical (Newtonian) mechanics, to more mesoscopic and macroscopic levels which represent the collective kinetic behaviour
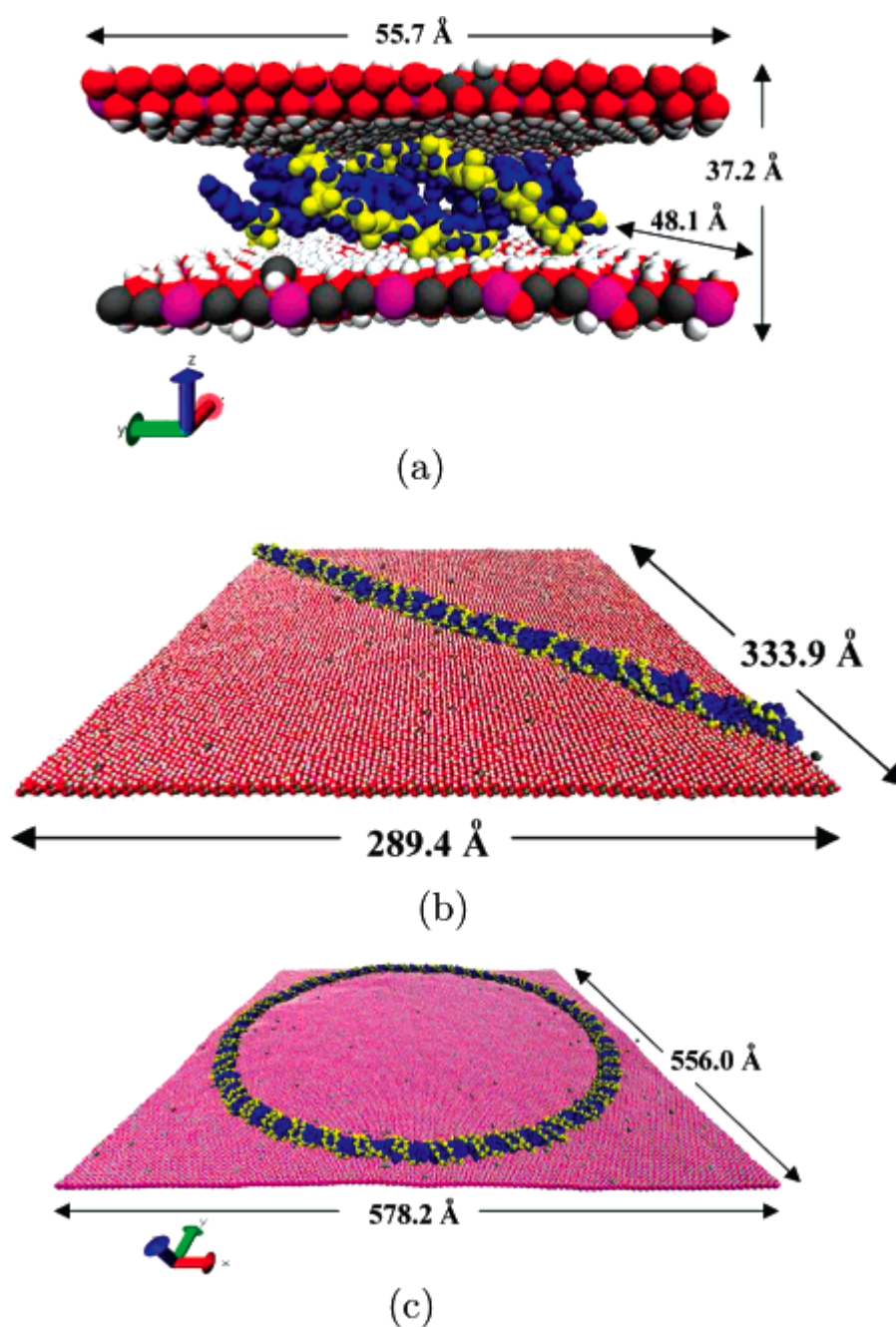
Figure 2.2: Visualizations of molecular dynamics simulation snapshot of (a) 12 base-pair DNA intercalated into an LDH; (b) 108bp DNA intercalated into a LDH; (c) 480bp plasmid intercalated into a LDH under ambient conditions (300 K and 1 atm). Magnesium, aluminium, oxygen and hydrogen atoms in the LDH sheets are represented as grey, pink, red and white spheres, respectively. The DNA strand has been coloured yellow to represent the phosphate backbone and blue for the sugar groups and base pairs. Water molecules have not been displayed and only one LDH sheet is visualised for (b) and (c). Reproduced from [72] with permission.
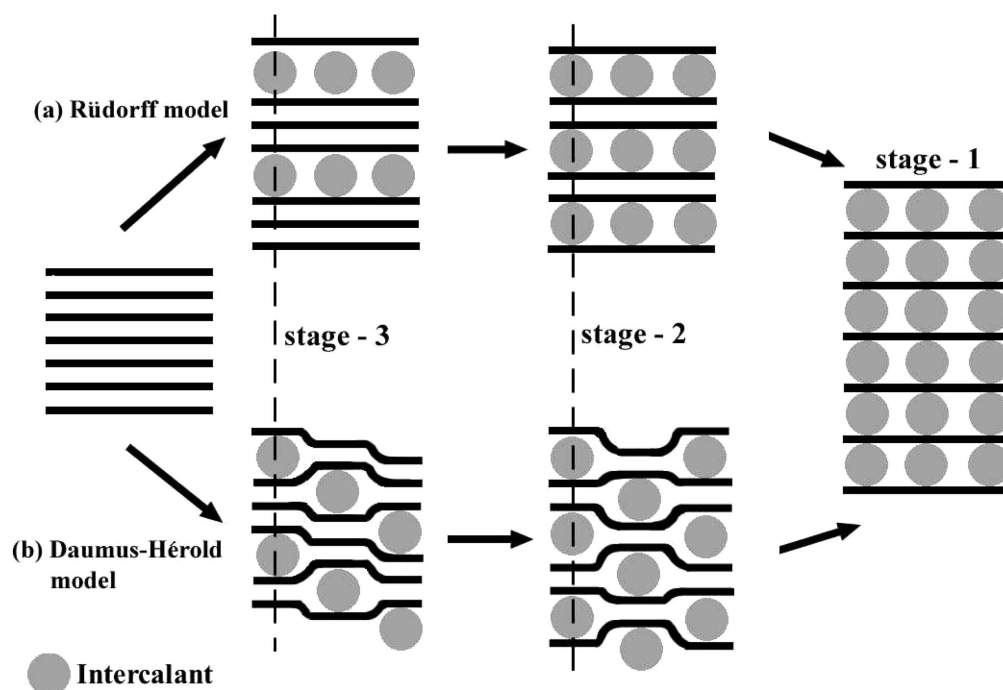
Figure 2.3: Schematic representation of the possible pathways by means of which layered materials could intercalate ionic species. LDHs are believed to form stage-2 intermediates through a Rüdorff model (a). More flexible materials such as graphite are believed to follow a Daumas-Hérold pathway (b). Thyveetil *et al.*'s simulations show that Mg_2Al LDHs are able to distort due to the size difference between intercalated DNA and chloride ions, supporting a Daumas-Hérold pathway for intercalation of large biomolecules. The dashed lines indicate similarly staged areas in the two pathways. Reproduced from [72] with permission.

of much larger assemblies of reacting and self-reproducing molecules. The research contained in this thesis should help the subject of origins of life gain respectability from within the scientific establishment and lead to stronger interactions between these theoretical approaches and related experimental research which can only serve to underpin further advances in our understanding of the events which took place at the dawn of life on Earth.

The intrinsic value of modelling approaches cannot be overestimated. Research by numerous experimental groups has shown that many reactions of interest to prebiotic chemists involve reactions between disordered pairs, or groups, of reactant molecules at ordered mineral catalytic surfaces. Such reactions are not amenable to direct chemical or physical analysis, not least as some occur within nanopore or micropore environments deep within the catalytic mineral host. In this thesis, simulations using computational chemistry methodology add insight into, amongst others, energetics at clay surfaces; the properties of RNA within clay interlayers; and stability and folding kinetics of nucleic acids at clay surfaces.

The following chapter details the complex inorganic mineral, and bimolecular models employed to investigate the various origins of life scenarios in the latter research chapters.

Chapter 3

Clay Mineral and Nucleic Acid Structure

This chapter is concerned with outlining the structure of the cationic and anionic clay minerals that are employed in our simulations in the following chapters. For both clay types the two-dimensional sheet-like structure of the inorganic clay framework is discussed. Following the definition of the clay structures, details of the arrangement of charge-balancing ions and associated water molecules within the interlayer region and at the exposed aqueous mineral surface are given.

Following the classification of cationic and anionic clays in 3.1, the known structure and conformation of nucleic acids in bulk water and intercalated within a mineral host is addressed.

3.1   Structure of anionic and cationic clays

The clay minerals which are studied here are types of abundant, naturally occurring materials formed by the weathering and decomposition of igneous rock and volcanic ash [74]. Clays belong to a wider class of compounds known as layered materials which are sometimes defined as "crystalline material wherein the atoms in the layers are cross-linked by chemical bonds, while the atoms of adjacent layers interact by physical forces" [75]. They usually comprise of the μm colloidal fraction of soil, sediment and rock. Both clay layers and the interlayer space between layers have widths in the nanometer range.

By convention, types of clay are named by the charge of the interlayer species. For example, naturally occurring cationic clay minerals have aluminosilicate sheets that posses a negative charge, that dictates the interlayer species carry a positive charge

(cationic). When the interlayer guest species is negatively charged (anionic) the mixed metal hydroxide sheet is positively charged and is referred to as an anionic clay mineral.

It is uncommon that sufficiently large crystals of clay platelets are obtained for full structural determination using single-crystal X-ray diffraction methods because of the lack of any significant long range order. The distance between clay layers, the bulk structure, can be inferred from the d-spacings gained from powder X-ray diffraction (PXRD) methods, but generally clays are characterised by the lack of any significant long range order.

It is possible to make the distinction between mono-layer and bilayer arrangements, and to infer the orientation of anisotropic interlayer guests from the interlayer spacing, determined by PXRD. The lateral arrangement of interlayer guest species however, cannot be easily obtained using these methods. For this reason simulation is well suited to the application of clay science for the resolution and accuracy available.

Neutron diffraction has been successfully used to ascertain the interlayer spacing of clay minerals. It has also been demonstrated as a method for determining the positions and self-diffusion dynamics of interlayer species that have been isotopically labeled [76, 77].

3.1.1 Layered double hydroxides

Layered double hydroxides (LDHs) are a family of compounds which have layers consisting of divalent and trivalent cations octahedrally coordinated with hydroxide ions (see Figure 3.1). LDHs are structurally similar to brucite whose general formula is $[Mg(OH)_2]$. The brucite sheets are neutral in charge but held together by a weak hydrogen bonding network which manifests in the material being soft and easy to cleave apart. The difference between brucite and LDHs is that the regular divalent cation have been isomorphically substituted with trivalent cations. This produces an overall positive charge which is compensated by the anionic interlayer guest species. The general chemical formula of the main group of LDHs is:
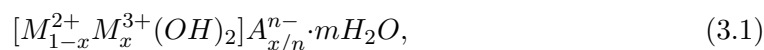
$$[M_{1-x}^{2+}M_x^{3+}(OH)_2]A_{x/n}^{n-}\cdot mH_2O, \tag{3.1}$$

where M^{2+} and M^{3+} are divalent and trivalent cations respectively, and A is an anion of valence n. The various possibilities of divalent and trivalent cation identities and ratios give rise to a wide variety of compositions and stoichiometry. M^{2+} can represent Mg, Mn, Fe, Co, Ni, Cu and Zn, whereas M^{3+} can represent Al, Mn, Fe, Co, Ni, Cr and Ga [78].
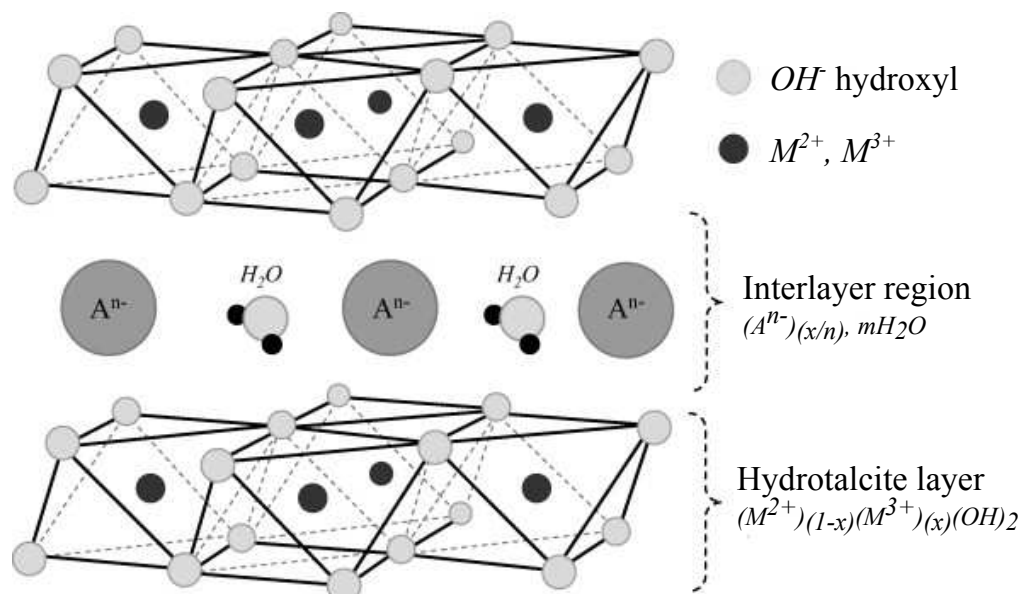
Figure 3.1: Schematic of a simple LDH. Two cation layers are shown with one interlayer region occupied by anionic species and water molecules.

It is not yet fully understood how the isomorphic substitution of divalent for trivalent metal ions arrange within the LDH layers. It is thought that the repulsion between ions could result in an ordered arrangement to minimise the energy. This is particularly true when the MgAl ratio is low. Experiments using atomic force microscopy show LDHs exhibit short range order [79]. Long range order within LDHs has only been observed in Li/Al and Mg/Ga LDHs [80, 81]. Short range ordering was found to minimise the cation/cation repulsion in anhydrous inorganic LDHs although the first principles approach used suffers from finite size effects [82].

A characteristic property of anionic clays is their ability to possess a variety of different stacking conformations. A layered compound which differs in stacking sequence can be called a polytype. Two principal forms, or polytypes, of LDHs are known: the two-layer repeat hexagonal structure, mannaseite, and the three-layer repeat rhombohedral structure, hydrotalcite. It is after the latter naturally occurring mineral that the Mg-Al LDHs are generically referred to as being hydrotalcite-like. Further polytypes appear in this class of minerals depending on the way in which successive layers stack [83]. Hydroxyl groups above the mid-plane are labeled A, and below labelled C. Cations occupy octahedral positions, C. When the hydroxyl groups are stacked directly on top of each other, this is called a prismatic arrangement (P-type). If the layers are offset by a lattice translation of (1/3, 2/3) or (2/3, 1/3) then this forms an octahedral arrangement (O-type). Naturally occurring hydrotalcite is an Al/Mg-LDH which has a 3R rhombohedral unit cell, consisting of three layers with the hydroxyl groups aligned in a sequence (BC-CA-AB-BC-...).

Different polytypes are often found intergrown in mineral samples. The manasseite forms the core with hydrotalcite forming the outer parts of the grain [84]. This indicates that manasseite is produced at higher pressures than hydrotalcite. Calle *et al.* [84] suggested that the reason for this is that $3R_1$ polytype can be constructed much more easily from a stack of identical brucite layers. The translation needed to produce the $3R_1$ polytype is *(2a/3+b/3)*, whereas for a $2H_1$ polytype the same translation is needed followed by a rotation of 60 degrees. It has also been shown that temperature has an effect on stacking order. Hines and Solin demonstrated a transition from the 3R at 25^oC to 2H at 150^oC for a Zn/Al-LDH [85].

The structure of the AlMg LDH has been investigated using a number of different experimental approaches. Properties such as cation bond lengths, the stacking of layers and cationic ordering have all been determined by Bellotto *et al.* using X-ray diffraction (XRD) methods [81]. To perform accurate XRD analysis samples of LDHs must be highly crystalline and have virtually no stacking faults. If they do it is difficult to carry out Rietveld analysis on the structure. Stacking faults arise from the random orientation of successively stacked layers and they prevent the ideal packing of the clay sheets. In particular, $(10l)$ reflections have a characteristic "shark fin" shape which is most commonly associated with basal plane slippage. Rietveld analysis of LDHs have often used the rhombohedral space group $R3m$. Bellotto *et al.* also tried to refine the structure with a $P3$ space group to test for the presence of cationic ordering. There was no improvement in the statistical indices, compared to the $R3m$ space group, indicating a lack of ordering.

The rigidity and bending modulus of clay platelets are difficult to measure experimentally because of their small dimensions. Thorpe presented a way to measure the rigidity of clay platelets which contain mixed intercalated ions [86]. The basal spacing surrounding large ions will decrease smoothly away from the intercalate site as the LDH layer corrugates around the ions. This area is known as the "catchment area" and is given the symbol p which corresponds to the number of lattice sites in this region. Solin *et al.* altered the composition of Ni/Al LDHs intercalated with CO_3^{2-} and found the resulting basal spacings using XRD [87]. The number of lattice sites in the catchment area was found to be $p \sim 5$. This showed that LDHs are less rigid than smectites but more rigid than graphite.

X-ray diffraction patterns are able to resolve the location and the bonding of interlayer atoms, but they are insufficient at elucidating the subtle structural information such as hydrogen bonding networks and inner sphere complexation. It is difficult to obtain fine grained crystals of LDHs and as such it is difficult to obtain this data from XRD experiments. The Rietveld technique is often used to refine the structural details like

atomic coordinates and site dependencies, but only if the ideal structure is known. The Rietveld refinement uses the observed powder X-ray diffraction pattern and compares it to the calculated pattern derived from a space group.

Another feature of LDH minerals that make them difficult to study experimentally is the lack of order in the arrangement of water molecules around the mineral interface. The nature of water at the mineral interface is determined by the charge density which also dictates whether the surface is hydrophobic or hydrophilic. Water associated with the LDH mineral can be divided up into three types: external surface, inter-particle pore and intercalated water. Constantino and Pinnavaia investigated the behaviour of water in MgAl LDHs as a function of charge density [88]. They showed the water content as a function of temperature using thermogravimetric analysis (TGA) revealing that interfacial pore water was readily removed at around 60°C. The results showed a step function between 110 and 250°C; evidence of the loss of interlayer water.

The arrangement of organic molecules within the LDH interlayer region is dependant on the structure of those molecules as well as the water content within the interlayer gallery. Early studies of the arrangement of terephthalate ions intercalated within an MgAl LDH used combined simulation and experiment to show that the water content can greatly effect the basal spacing distance and consequently how the intercalated molecules arrange themselves [89].

Newman *et al.* used molecular dynamics techniques to investigate the nonlinear variations in basal spacing as a function of hydration for LDH compounds [90]. They showed that at higher levels of hydrations the terephthalate ions arrange perpendicular to the mineral surface which gave a basal spacing of 14 Å whereas less hydrated systems forced the ions parallel to the surface giving a spacing of 8.4 Å.

The ratio between Mg^{2+} and Al^{3+} ions is an important feature of LDH compounds. Different ratios lead to a change in charge densities. Basal spacings have been shown to be lower for LDHs with a low MgAl ratios, which can be attributed to the greater energy of repulsion between layers.

3.1.2   Smectite clays

Smectite clays are a type of 2:1 layered compound which are built of an octahedral layer sandwiched between two tetrahedral layers, denoted by the terminology T-O-T. The central/inner layer consists of divalent or trivalent cations octahedrally coordinated with hydroxyl groups. The exposed outer layers consists of silicon atoms tetrahedrally

coordinated to oxygen. Smectites consist of negatively charged, crystalline aluminosilicate sheets. The montmorillonite type of smectite has a structure comprising stacks of pyrophyllite-like layers, $[Al_2Si_4O_{10}(OH)_2]$, each of which consists of an octahedral alumina sheet sandwiched between two tetrahedral silica sheets. The negative charge arises from partial substitution of metal ions in typically both the octahedral and tetrahedral sheets.
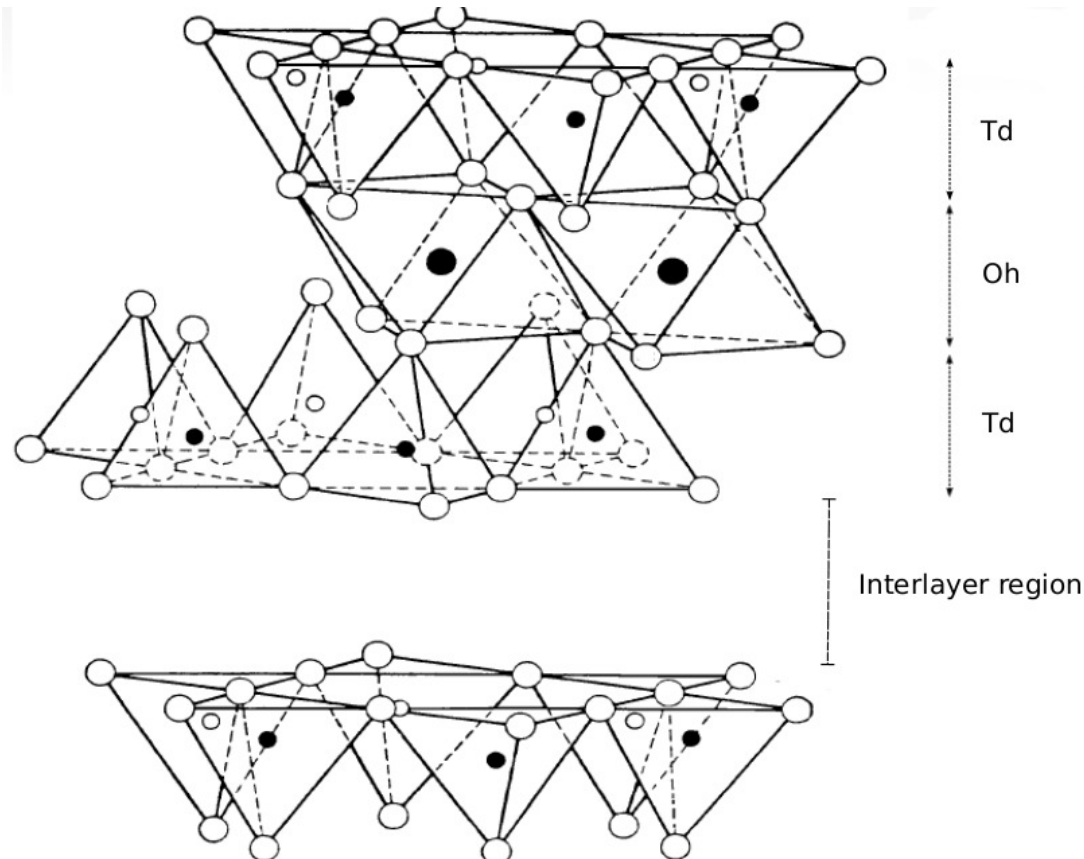
Figure 3.2: Schematic of a simple smectite clay showing a layer composed of an octahedral, Oh, alumina sheet between two tetrahedral silica sheets.

$$[Mg_xAl_{2-x}(OH)_2(Si_4O_{10})](C^{n+})_{x/n} \cdot nH_2O, \qquad (3.2)$$
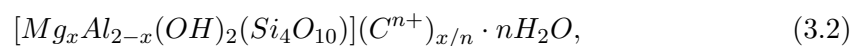
where C is a cation of valence n, commonly Na^+, Ca^{2+} and Mg^{2+}. Kaolinites have no partial substitution of metal cations in the octahedral sheet and as such are charge neutral. Smectites have variable basal spacing as a function of hydration and nature of the intercalated species. Simulation and experiment have both recreated hysteresis hydration curves of smectite clays. The function shows discrete steps that correspond to the adsorption of one, two and three layers of water to the mineral surface. Experimentally there are two types of water adsorption: intralayer swelling caused by hydration of the exchangeable intercalated cations and the insertion of water between clay platelets

[91]. The size and the charge of the saturating cations govern the adsorption of water into the interlayer. Na^+ clays are more likely to swell than K^+ clays and less likely than Li^+, as they have the largest hydration energy which gives them a larger tendency to swell [92].

Naturally occurring smectites tend to have inorganic species such as Li^+, Na^+ and K^+, adsorbed in the interlayer region. These can be exchanged for more complex charged species such as alkyl ammonium cations. There are two specific region within the smectite structure that are hydrophobic and hydrophilic. The structure and arrangement of cations and water within the interlayer region of smectites has been studied by various experimental techniques: neutron diffraction [93], XRD [91] and infrared spectroscopy [94].

Due to various properties such as good biocompatibility and ion exchange capability, smectite clays like montmorillonite have gained considerable interest in pharmaceutical applications. The strong association between the mineral and species such as metal oxides and silicates, as well as biological molecules like nucleic acids, proteins and sugars. Adsorption pathways of biomolecules and organic species at the smectite interface has gained much attention due to these applications [95, 96].

Experiments were carried out on montmorillonite interacting with the enzymes: lysozyme, lactoglobulin, chymotrypsin, pepsin and haemoglobin [97]. The theoretically calculated basal spacings were in agreement with those found experimentally, with the exception of pepsin which is thought to unfold whilst intercalated. Haemoglobin has been shown to be immobilised when intercalated within a Na^+-montmorillonite. This allows direct measurement of electrochemical properties of the haemoglobin in the native conformation.

The mechanism by which nucleic acids intercalate and adsorb on smectite surfaces is not fully understood and it is one of the main scientific questions that is addressed in this thesis. Pyrimidine and purine nucleotides are readily intercalated within montmorillonite despite both carrying a negative charge but some intercalate more readily than others, for example the purine, adenine adsorbs more readily than other nucleotides [98]. Several studies have shown that pyrimidine nucleotides are less likely to adsorb to the surface of montmorillonite than purines [99].

Winter *et al.* have shown that montmorillonite surfaces promote cooperative bonding whereby hydrogen bonds form between uracil purines which lay adjacent and parallel/flat to the mineral surface [100].

3.2 Structure of nucleic acids

Deoxyribonucleic acid (DNA) and ribonucleic acid (RNA) are linear polymers of nucleotide monomers. The length of these polymers can range from 20 nucleotides for small interfering RNA (siRNA) to 10^8 nucleotides for chromosomal DNA. In modern biology, DNA is the carrier of genetic information. DNA is passed on from parent to daughter cells and from progenitor to progeny. Messenger RNA (mRNA) is transcribed from DNA and used as a template for the synthesis of proteins, a process known as translation. Proteins carry out many essential tasks in the modern day cell including enzyme catalysis, motor activity and information relay.
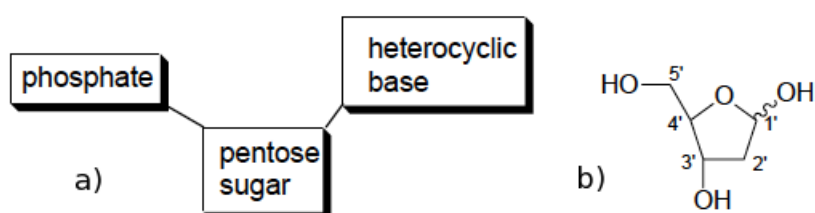
Figure 3.3: a) The basic constituent parts which make DNA and RNA: the phosphate backbone, pentose sugar and heterocyclic base. b) the chemical structure of the 2-deoxy-D-ribose sugar of DNA. RNA differs by an additional hydroxyl at the $2'$ position of the sugar.

3.2.1 Deoxyribonucleic acid

A simplistic view of DNA structure can be understood by viewing it as a constituent phosphate backbone, a 2-deoxy-D-ribose sugar and a heterocyclic base (see Figure 3.3), but this is now thought of as misleading. Sutherland *et al.* have demonstrated that activated pyrimidine ribonucleotides can be formed in a short sequence that bypasses free ribose and the nucleobases, and instead proceeds through arabinose amino-oxazoline and anhydronucleoside intermediates [101]. The heterocyclic base can be a purine (adenine or guanine) or a pyrimidine base (cytosine or thymine), as shown in Figure 3.4.

Throughout this thesis the terms 'nucleosides' and 'nucleotides' will be used. Nucleosides are deoxyribose sugars linked to a heterocyclic base, e.g. Deoxyadenosine. Nucleotides are phosphate esters of nucleosides, e.g. Deoxycytidine $3'$-phosphate which can be abbreviated to $3'$-dCMP, or dCp. The primary structure of DNA is a polymer strand formed by joining nucleotide from the $3'$ hydroxyl group to the $5'$ hydroxyl of the next nucleotide via a phosphate ester (see Figure 3.5). No $5'$-$5'$ or $3'$-$3'$ linkages are found in biological nucleic acids (except the mRNA cap). The primary structure of nucleic acids is only determined by the sequence of bases and is commonly written in the $5'$-$3'$
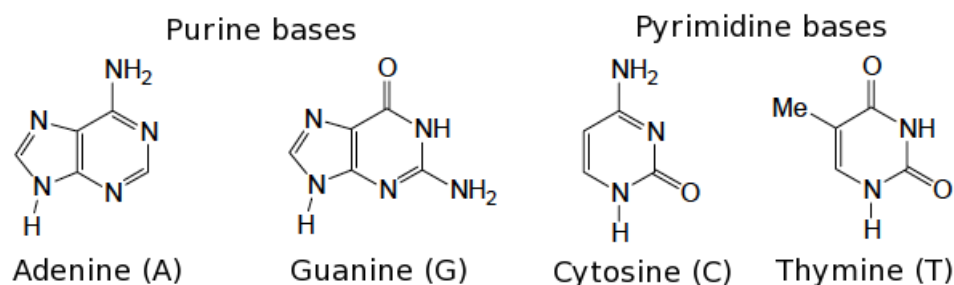
Figure 3.4: The four DNA bases: purine bases, adenine & guanine and the two pyrimidine bases, cytosine & thymine.

direction, e.g. the DNA shown in Figure 3.5 a) can be written like dpTpApT, or more simply as TAT.
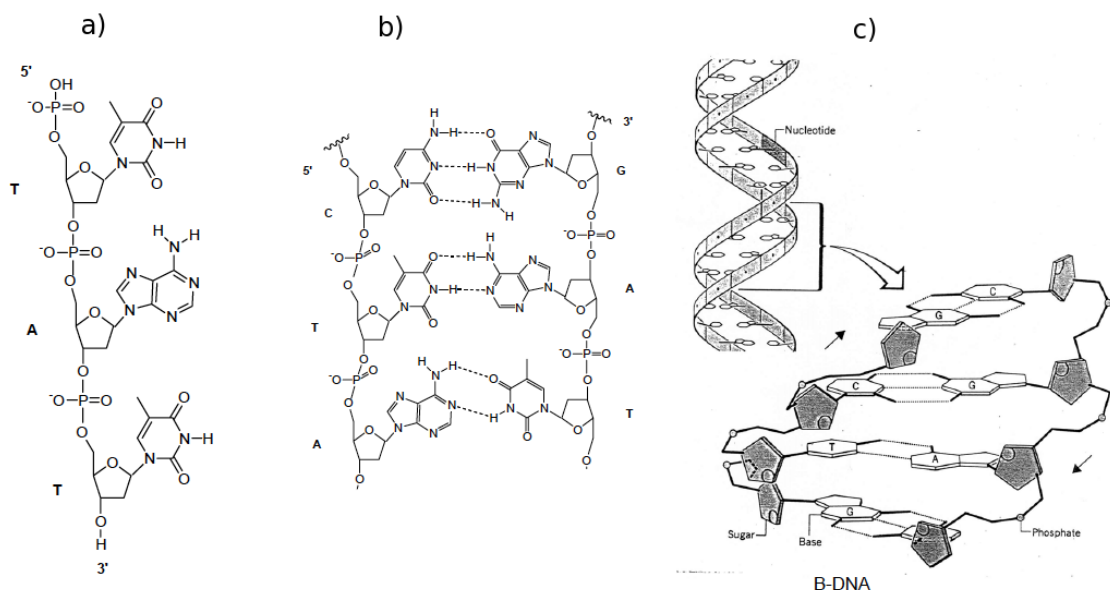
Figure 3.5: The secondary structure of DNA. a) A single strand of DNA of sequence adenine-thymine-adenine shown in the $5'$-$3'$ direction. b) A DNA duplex showing Watson-Crick hydrogen bonding between CG, TA and AT base pairs. c) The tertiary structure of DNA showing the orientation of bases relative to the phosphate backbone.

Hydrogen bonding between lone pairs of ring nitrogen or carbonyl oxygens with NH_2 or NH groups allow the formation of a duplex DNA helix. Figure 3.6 b) shows Watson-Crick hydrogen bonding between complementary bases which make up the secondary structure of DNA. The tertiary structure of DNA, shown in Figure 3.6 c), is made of two DNA strands which form an anti-parallel helix. The base pairs arrange perpendicular to the phosphate backbone. The minor and major grooves are a defining characteristic

of the tertiary DNA structure. The tertiary structure is determined by the conformational preferences of the nucleotides. Minor perturbations in these conformations lead to differently shaped DNA. DNA is most commonly found in the B-DNA form. B-DNA forms in high humidity and low salt concentrations. It is a right handed helical structure with 10 base pairs per turn with bases perpendicular to the helix axis (shown in Figure 3.5. A-DNA is right handed, like B-DNA, but there is an increase in the number of base pairs per turn that gives a more compact structure. Z-DNA has a left handed helical structure and is favoured for alternating guanine-cytosine sequences.
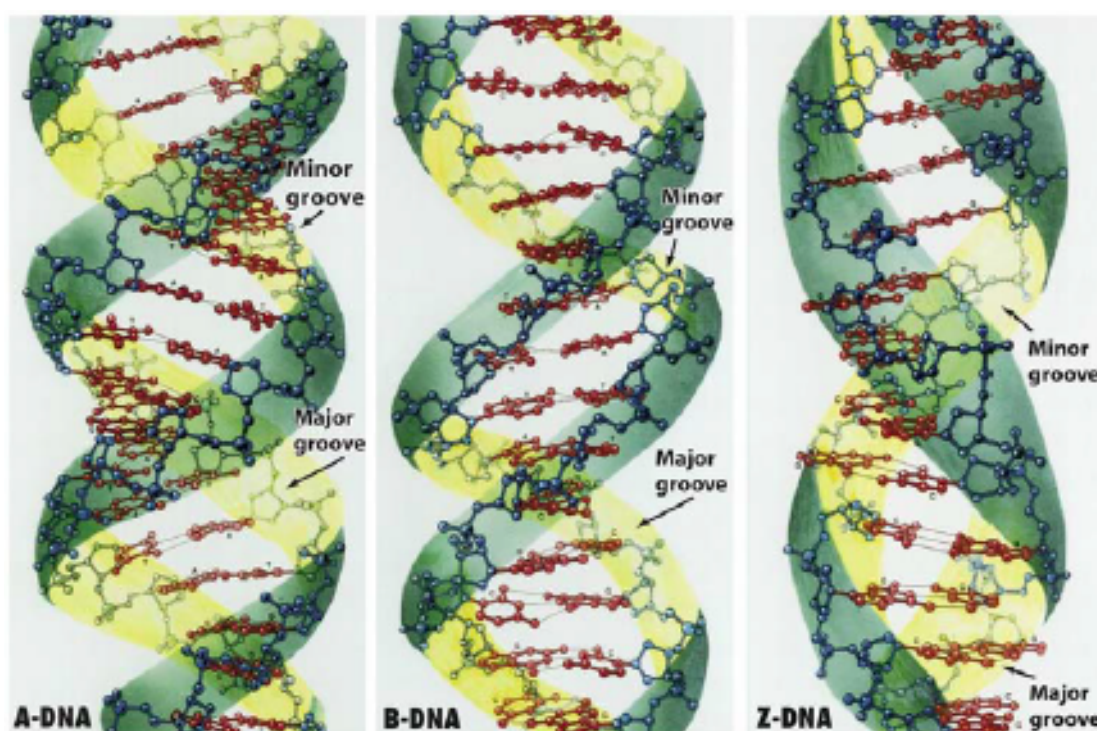
Figure 3.6: The tertiary structure of DNA showing A-, B- and Z-type DNA and their differences in major and minor groove width.

3.2.2   Ribonucleic acid

RNA differs from DNA in that it has an additional hydroxyl group at the $2'$ carbon position of the ribose. This difference produces a chemical instability in RNA that is absent in DNA. RNA is unstable to base cleavage at the $2'$ -OH. The RNA bases are adenine, guanine and cytosine which are the same as the DNA bases but RNA has a uracil base (see Figure 3.7) in place of the DNA base thymine.

Unlike DNA, RNA is usually single stranded, but it can form elements of secondary structure via hydrogen bonding networks with non adjacent bases.
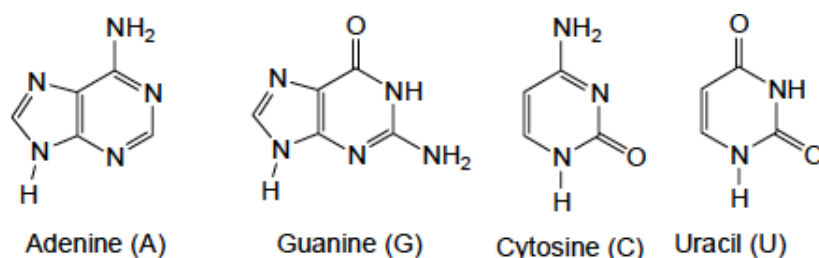
Figure 3.7: The four RNA bases, like those of DNA in Figure 3.4 but Thymine is substituted for uracil.
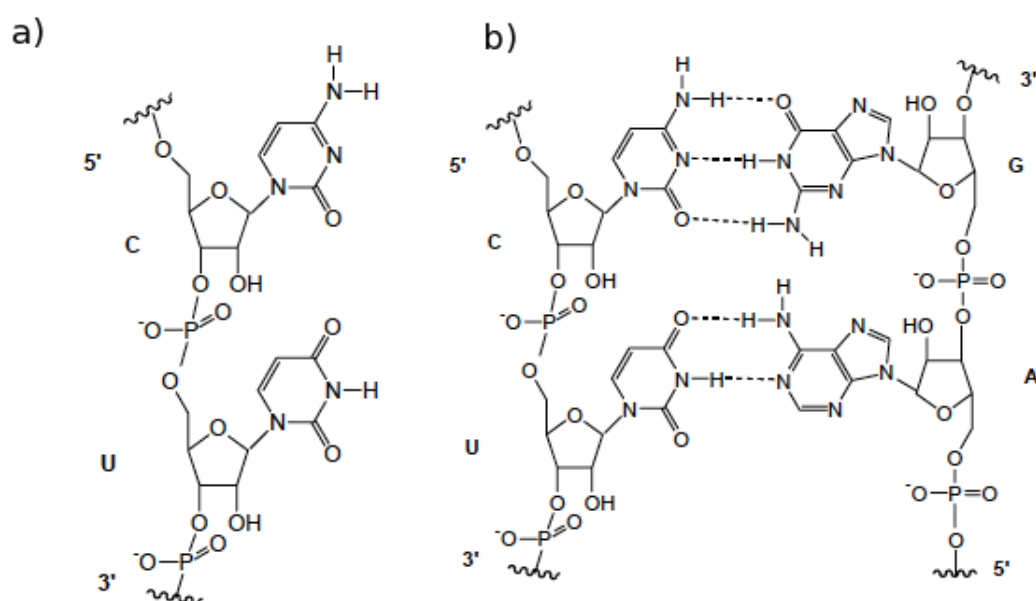
Figure 3.8: The secondary structure of RNA. a) a single strand of RNA of sequence CU. b) RNA duplex of sequence CU.

3.2.3 Peptide nucleic acid

Peptide nucleic acid is a synthetic nucleic acid first synthesised by Nielson *et al.* [102]. DNA and RNA have a charged phosphate backbone whereas PNA is composed of repeating N-(2-aminoethyl)-glycine units chemically linked via peptide bonds. Purine and pyrimidine bases (shown if Figures 3.7 and 3.4) are attached to the PNA backbone via methyl and carbonyl groups (see Figure 3.9).

It has been shown that the PNA:DNA binding is more energetically favourable than in a DNA:DNA duplex due to the decreased electrostatic repulsion [103]. The secondary structure of a PNA:DNA duplex is shown in Figure 3.9 b).
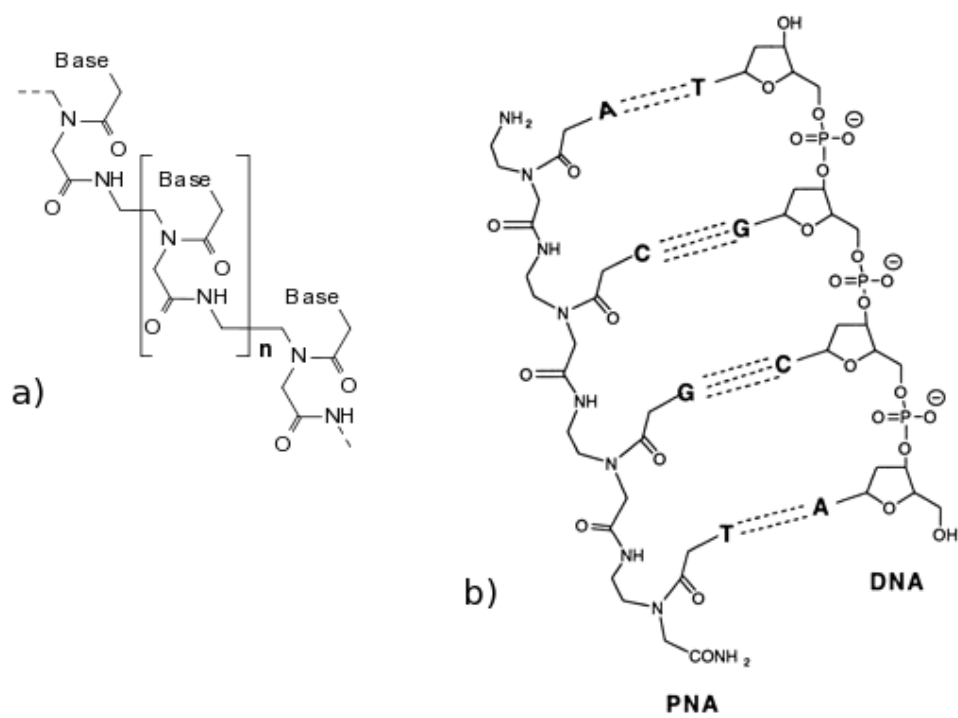
Figure 3.9: a) A single strand of PNA with the monomer in square brackets. b) PNA:DNA duplex strand.

Chapter 4

Biomolecular & Inorganic Simulation Techniques

The focus of this chapter is on the physio-chemical theories which underpin molecular simulation techniques. The development and selection of potentials used to describe the molecular interactions and detail specific inorganic and biomolecular force fields which we employ to run molecular dynamics simulations are discussed. The chapter concludes by presenting the substantial computational resources needed to run modern day molecular simulations.

4.1 Potential parameterisation

Potentials are algorithms that generate energy surfaces. In this respect, it is always possible to tackle ground state problems in terms of potentials. Because potentials are related to models of system behaviour, they often contain a number of physical approximations. They can be generic to a class of compounds or molecules or they can be specific to a material or molecule. Simulation potentials can be divided into three categories: i) Model potentials are primarily used for testing theories; ii) Valence force fields are intramolecular potentials for covalent systems. They are often used systems where there is no interest in bond formation/breaking; iii) Central force potentials are ionic; the pair interaction depends only on distances between atoms or ions *e.g.* rare gases and ionic solids.

Valence force fields feature potential functions that describe the degrees of freedom of the molecule. These components define the geometry: bonds, bond angles, torsions and non-bonded contacts. One functional form for such an empirical force field is:

$$U(\mathbf{r}^N) = \text{bonds} + \text{angles} + \text{torsions} + \text{non-bonded} \tag{4.1}$$

$$\text{bonds} = \sum_{bonds} \frac{k_i}{2}(l_i - l_{i,0})^2$$

$$\text{angles} = \sum_{angles} \frac{k_i}{2}(\theta_i - \theta_{i,0})^2$$

$$\text{torsions} = \sum_{torsions} \frac{V_n}{2}(1 - \cos(n\omega - \gamma))$$

$$\text{non-bonded} = \sum_{i=1}^{N}\sum_{j=i+1}^{N}(4\varepsilon_{ij}[(\frac{\sigma_{ij}}{r_{ij}})^{12} - (\frac{\sigma_{ij}}{r_{ij}})^6] + \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}}),$$

where $U(\mathbf{r}^N)$ denotes the potential energy as a function of the collective position $\mathbf{r}$ of N atoms. The various contributions are schematically represented in Figure 4.1. The first term in Equation 4.1 models the interaction between pairs of bonded atoms using a harmonic potential that gives the increase in energy as the bond length l_i deviates from the reference value $l_{i,0}$. The second term is a summation over all valance angles using a harmonic potential. The third is a torsional term which describes how the energy changes as the bonds rotate. The fourth term in Equation 4.1 is the non-bonded term which describes the interaction between all pairs of atoms (i and j) that are in different molecules or in the same molecule but separated by at least three bonds. It is common in force fields to model the non-bonding term using a Coulomb potential for the electrostatic interactions and a Lennard-Jones potential for the van der Waals interactions.

4.1.1 Water models

There is a large number of water models for (bio)molecular simulation. These models are classified by three types: Rigid models with a fixed geometry, flexible models that include vibrational degrees of freedom and polarisable models which account explicitly for polarisation. The most widely used water models for biomolecular simulation are of the rigid type. Two of the most popular models are the TIP (Transferable Interaction Potential) [104] and SPC (Single Point Charge) [105]. These have been parameterised to reproduce the properties of liquid water at ambient temperature and pressure conditions. All SPC water models have three interaction sites centred on the atomic nuclei. The model consists of positive partial charges on the hydrogen atoms and negative on the oxygen atom with a non-zero Lennard-Jones parameter on the oxygen only. The extended SPC models (SPC/E) corrects for the interaction energy from the self-polarisation of water. The
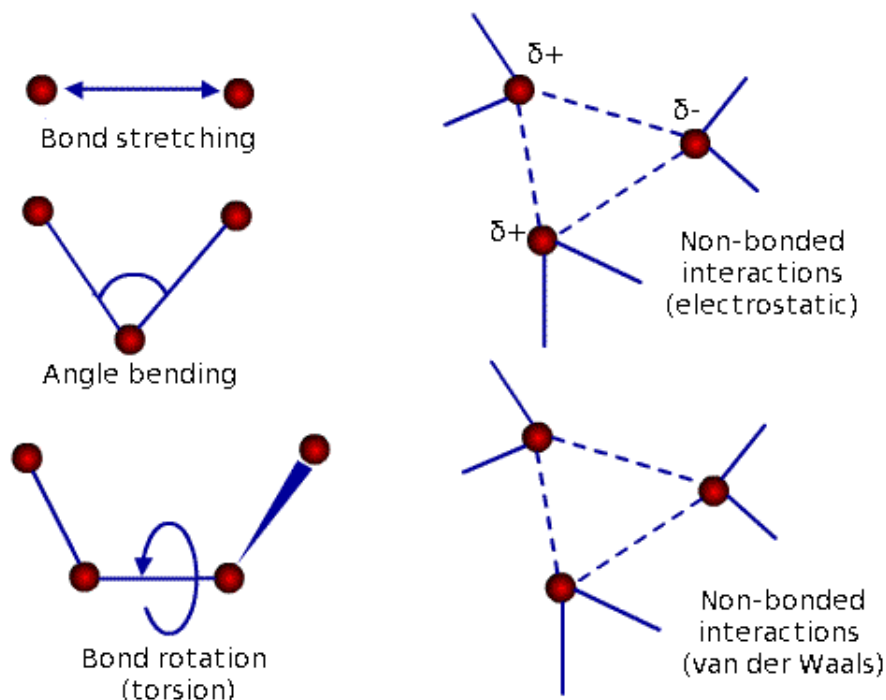
Figure 4.1: Schematic representation of the four key contributions to a molecular mechanics force field: bond stretching, angle bending, torsional terms and non-bonding interactions.

SPC/E model accurately models the bulk behaviour of water. The SPC water model reproduces many properties of bulk water, including radial distribution functions, diffusion coefficients, rotational relaxation times, dielectric permiativity, Debye relaxation time, heat capacity, excess Helmholtz energy and shear viscosity.

4.1.2 Biomolecular potentials

The AMBER (Assisted Model Building with Energy Refinement) biomolecular force field has been in constant development since the 1980s under the leadership of Peter Kollman at the University of California, San Francisco. The version termed 'ff99' is the most widely adopted for nucleic acid moieties [106]. The ff99 version is an all-atom force field with fixed partial charges centred on atoms. ff02 was developed as a polarisable variant of the ff99 force field [107]. ff02 has polarisable dipoles at the atoms that can be calculated iteratively at each step, or propagated with the atomic positions as additional dynamical variables.

An alternative to the AMBER force field set is the CHARMM (Chemistry at HARvard Molecular Mechanics) force fields developed by Martin Karplus at Harvard University. The param force fields (of which param27 is the most recent) have fixed partial charges

centred on atoms [108]. The charges are based on *ab initio* dimer energies and geometries. The CHARMM force field set has a fluctuating charge force field that is polarisable and allows the redistribution of atomic charges to yield equivalent electronegativity [109]. The CHARMM Drude force field is currently in development. It is based on the Drude model; *i.e.* introduction of 'massless' virtual sites (Drude particles) carrying partial electric charge and attached to individual atoms via a harmonic spring.

The final biomolecular force field set we will discuss is the GROMOS (GROningen MOlecular Simulation) set developed under the leadership of Wilfred van Gunsteren at ETH Zurich. The GROMOS force fields are normally used in conjunction with the GROMACS molecular dynamics code. The major version is '53A6' which is a united-atom force field (implicit aliphatic hydrogen), with fixed partial charges centred on atoms [110].

The cited force fields all have strengths and weaknesses. It would be difficult to demonstrate that one of them is better or worse that the others in all circumstances. There are a number of recent force field comparisons, mostly showing that there is still room for improvement in all of them [111–113].

4.1.3 Inorganic potentials

Due to the complexities of the structure and composition of clay and other hydrated minerals, and the inherent uncertainties of the experimental methods, it is important to apply theoretical molecular models for a fundamental atomic-level understanding, interpretation and prediction of these phenomena. The ClayFF force field was developed by Cygan and co-workers at Sandia National Laboratories to address these concerns. ClayFF is a general force field suitable for the simulation of hydrated and multicomponent mineral systems and their interfaces with aqueous solutions [114]. The interatomic potentials were derived from parameterisations incorporating structural and spectroscopic data for a wide variety of simple hydrated compounds. ClayFF adopts the SPC model to describe the water and hydroxyl behaviour (see Section 4.1.1). The metal-oxygen interactions are described by a Lennard-Jones function and a Coulombic term with partial charges derived by Mulliken and electrostatic potential (ESP) analysis of density functional theory (DFT) results. The force field was validated using experimental and spectroscopic data on bulk structure, relaxed surface structure and intercalation processes. The interatomic interactions in the ClayFF force field are treated as non-bonded which allows it to be applied to a wide range phases and accounts for the energy and momentum transfer between the fluid phase and solid.

4.1.4 Mixing rules

Ab initio, also known as first principals, calculations can calculate the energies and geometries of a system to a great deal of accuracy. However, these methods are limited by the size and length scales available, currently only tens of atoms for picoseconds.

Investigation of the conformation and dynamics for complex biomolecules and mineral surface structures in explicit aqueous solvent is only really feasible using classical simulation techniques that are dependant on the force field potentials used to describe the interactions.

A common method for developing a force field starts with choosing an appropriate functional form for the type of interaction to be modelled then fitting the free parameters to a number of observable properties or *ab initio* calculations [96]. Force fields that describe the interaction between biological and inorganic systems present additional challenges to those used in systems that fall into a one category. Force fields typed for use with biomolecules traditionally use covalent intramolecular interactions with intermolecular interactions modelled with van der Waals and Coulomb descriptions between partial charges. Force field models for the majority of ionic crystal, mineral systems are based on the Born model for solids, with the interactions between the ions being dominated by long-range electrostatic interactions between ions with formal charges, and short-range potentials describing the van der Waals interactions and the Pauli repulsion [114]. Because each model uses a different charge model, there is no one way to derive the cross-terms needed for a full description of the interactions. A possible method is to use the organic potential parameters for the ions of the inorganic crystal, along with the Lorentz-Berthelot mixing rules (See reference 126 in [96]) to generate the potentials required. The Lorentz rule [115] is given by:

$$\sigma_{ij} = \frac{\sigma_{ii} + \sigma_{jj}}{2}, \tag{4.2}$$

where sigma, σ, is the distance at which the intermolecular potential between the two atoms is zero. The Lorentz rule is only valid for the hard sphere model. The Berthelot rule [116] is given by:

$$\epsilon_{ij} = \sqrt{\epsilon_{ii}\epsilon_{jj}}, \tag{4.3}$$

where epsilon, ϵ, is the well depth of the Lennard-Jones potential function.

However, this approach often leads to high binding energies and short equilibrium interatomic distances [117–120]. This is a clear example of the transferability problem where the crystalline environment has a strong effect on interactions. A second problem encountered for interfaces is the lack of experimental data which can be used to fit potentials. Relevant experimental data, such as immersion and desorption enthalpies, are available for a limited number of systems, but there is insufficient data for fitting purposes.

The description of divalent cations associated with large polarisation and charge transfer still poses a challenge in molecular dynamics. Šponer *et al.* show the Mg^{2+}-N7(guanine) interaction energy term is 210 kcal/mol, while the force field accounts for only 150 kcal/mol [121]. Non-additivity in the first ligand shell of Mg^{2+} is 70 kcal/mol, mainly due to inter-ligand polarisation repulsion [122]. In reality, the first-shell water molecules are heavily polarised by the ion and their H-bonding properties are very different from those of bulk water molecules [123]. The force fields can be biased towards direct (inner-shell) binding of Mg^{2+} to solute. It is suggested that the sampling of divalent ions is entirely insufficient in affordable simulations.

4.1.5 Chosen models

The Amber parm99 forcefield was used to describe the bonding interactions and to provide the partial charges for RNA and DNA molecules [124] in Chapters 7 through to 8. The parmbsc0 refinement was used to provide an improved description of the α/γ concerted rotations within the RNA and DNA [125]. The Amber forcefield has been used extensively to simulate nucleic acids and reproduces well the structure and dynamics of nucleic acid moieties, including Watson-Crick base pairing [125–127]. We parameterised the PNA (peptide nucleic acid) in Chapter 5 using the forcefield developed by Shields *et al.* [128] which was found to reproduce PNA structural properties with good accuracy.

The ClayFF forcefield was used to provide parameters for atoms belonging to the inorganic LDH and montmorillonite in Chapters 7 to 8 [114]. ClayFF produces good agreement with experiment for layered double hydroxides in terms of lattice parameters, water diffusion coefficients and far-infrared spectra [129, 130]. Recently a reparameterisation of the ClayFF forcefield was performed using electrostatic potentials derived from plane-wave density functional theory calculations to address certain thermodynamic issues [131]. Because we are not concerned with clay "edges", as we are only examining the LDH basal surface, these thermodynamic issues do not arise in our work and Cygan *et al.*'s original ClayFF forcefield parameterisation more than suffices [114].

Both parm99 and ClayFF use a harmonic potential for bond terms, while ClayFF is an ionic forcefield with no angle or dihedral terms, making the two forcefields simple to combine. Lorentz-Berthelot mixing rules were used to supply the missing Lennard-Jones parameters. These mixing rules estimate intermolecular potential parameters of the Lennard-Jones potential using an arithmetic average for the collision diameter and a geometric average for the well depth; they work well in situations in which the dominant interactions are electrostatic, as is the case here [69, 72, 96, 132, 133].

Water molecules were described using the flexible single-point charge (SPC/E) model [105]. The SPC/E water model reproduces the bulk structure of water accurately but there are limitations of using non-polarizable water models compared with the computationally more expensive polarizable models. A review of various water models and their associated advantages and disadvantages can be found in Chapter 4.1.1 and in the following references [134, 135]. A cutoff of 10 Å was imposed on the Lennard-Jones interactions. Periodic boundary conditions were imposed in all three spatial directions. Coulombic interactions were computed using Ewald summation and the particle-particle, particle-mesh method with a precision value of 0.001 and a grid order of 4 [136].

Because the ClayFF and Amber forcefields are parameterized at ambient temperatures and pressures, the forcefields are likely to be less reliable at the higher temperatures and pressures. Nevertheless, the higher temperature and pressure simulations within Chapter 5 provide important qualitative comparisons between the different nucleic acid models and furnish insight into the effects of temperature and pressure on these models, as previously reported [137]. A discussion on the reliability of nucleic acid forcefields under various conditions (as well as at long timescales) is given by [126].

4.2 Molecular dynamics techniques

Molecular Dynamics is the solution of the classical equations of motion for atoms and molecules in order to obtain the time evolution of the system. The trajectories of the molecular system are obtained by solving the differential equations embodied in Newton's second law:

$$\frac{d^2 x_i}{dt^2} = \frac{F_{x_i}}{m_i} \tag{4.4}$$

where the motion of a particle of mass m_i along one coordinate (x_i) has the force F_{x_i} in that direction. Finite difference techniques are used to generate molecular dynamics trajectories with continuous potential models. The idea is that the integration is broken

down into small stages, each separated in time by a fixed time step, δt. The total force on each particle in the configuration at time t is calculated as the vector sum of its interactions with other particles. From the force we can determine the accelerations of the particles which can then be combined with the positions and velocities at a time to calculate the positions and velocities at time $t + \delta t$. The force is assumed to be constant during the time step. The force on the particle in their new positions are then determined, leading to new positions and velocities at time $t + 2\delta t$, and so on.

There are many algorithms for integrating the equations of motion using finite difference methods for molecular dynamics simulations. All algorithms assume that the positions and dynamic properties can be approximated as Taylor series expansions:

$$\mathbf{r}(t + \delta t) = \mathbf{r}(t) + \delta t \mathbf{v}(t) + \frac{1}{2}\delta t^2 \mathbf{a}(t) + \frac{1}{6}\delta t^3 \mathbf{b}(t) + \frac{1}{24}\delta t^4 \mathbf{c}(t), \tag{4.5}$$

$$\mathbf{v}(t + \delta t) = \mathbf{v}(t) + \delta t \mathbf{a}(t) + \frac{1}{2}\delta t^2 \mathbf{b}(t) + \frac{1}{6}\delta t^3 \mathbf{c}(t),$$

$$\mathbf{a}(t + \delta t) = \mathbf{a}(t) + \delta t \mathbf{b}(t) + \frac{1}{2}\delta t^2 \mathbf{c}(t),$$

$$\mathbf{b}(t + \delta t) = \mathbf{b}(t) + \delta t \mathbf{c}(t),$$

where $\mathbf{v}$ is the velocity (the first derivative of the positions with respect to time), $\mathbf{a}$ is the acceleration (second derivative), $\mathbf{b}$ the third derivative, and so on. The Verlet algorithm is widely used to integrate the equations of motion [138]. The Verlet algorithm uses the positions and accelerations at time t, and the position from the previous step, $\mathbf{r}(t - \delta t)$, to calculate the new positions at $t + \delta t$:

$$\mathbf{r}(t + \delta t) = \mathbf{r}(t) + \delta t \mathbf{v}(t) + \frac{1}{2}\delta t^2 \mathbf{a}(t) + ..., \tag{4.6}$$

$$\mathbf{r}(t - \delta t) = \mathbf{r}(t) - \delta t \mathbf{v}(t) - \frac{1}{2}\delta t^2 \mathbf{a}(t) - ...,$$

the addition of the equations above gives:

$$\mathbf{r}(t + \delta t) = 2\mathbf{r}(t) - \mathbf{r}(t - \delta t) + \delta t^2 \mathbf{a}(t), \tag{4.7}$$

The velocities can be calculated by dividing the difference in positions at two times, on the time step or alternatively at the half-step:

$$\mathbf{v}(t) = [\mathbf{r}(t + \delta t) - \mathbf{r}(t - \delta t)]/2\delta t, \mathbf{v}(t + \frac{1}{2}\delta t) = [\mathbf{r}(t + \delta t) - \mathbf{r}(t)]/\delta t, \tag{4.8}$$

The Verlet leapfrog algorithm is a development of the standard Verlet method that first calculates the velocities $\mathbf{v}(t + \frac{1}{2}t)$ from the velocities at time $t - \frac{1}{2}\delta t$ and the accelerations at time t. The positions are then deduced from the velocities just calculated together with the positions at time $\mathbf{r}(t)$ using this equation:

$$\mathbf{r}(t + \delta t) = [\mathbf{r}(t) + \delta t\mathbf{v}(t + \frac{1}{2}\delta t), \mathbf{v}(t + \frac{1}{2}\delta t) = \mathbf{v}(t - \frac{1}{2}\delta t) + \delta t\mathbf{a}(t), \qquad (4.9)$$

The velocities thus 'leap frog' over the positions. The leap frog approach is advantageous because it explicitly includes the velocity and does not require the calculation of the differences of large numbers. However, the positions and velocities are not synchronised. This means that it is not possible to calculate the kinetic energy contribution to the total energy at the same time as the positions are defined.

The Large-scale Atomistic/Molecular Massively Parallel Simulator (LAMMPS) [136] was used to perform all the simulations reported in this thesis. LAMMPS was used because of its highly scalable nature.

4.2.1 SHAKE

The integration time step in MD simulations is limited by the highest frequency motion in the system. This is usually the hydroxyl bond vibration which occurs on the timescale of approximately 50 fs. Treating bonds as a classical harmonic oscillator is not necessarily a better approximation to the quantum-mechanical nature of the bond than using a constraint. The SHAKE algorithm constrains the highest frequency motions [139]. As a result, the time step can be increased and the range of frequencies is reduced, leading to faster redistribution of energy between modes.

The iterative SHAKE method treats each constraint independently. The method may fail if the atomic displacement is very large (e.g. due to a too long time step or a spurious collision).

4.2.2 Improved sampling methods

Direct molecular dynamics can in principle simulate diffusion, but for many processes the timescales are too long. Direct MD simulations have time scales on the order of nanoseconds. Real time scales for processes such as protein folding are often of the order of microseconds or more. The motivation for using the following so called improved sampling methods is the enhanced configurational sampling they deliver. This section

mentions a number of different sampling techniques but will focus mainly on the replica exchange molecular dynamics (REMD) method.

4.2.2.1 Transition state theory

If the barrier between different states is large, the system spends most of the time exploring a single energy basin. Transitions to neighbouring basins are very rapid compared to this (the correlation time is much shorter than the residence time). The evolution of such a system is a series of discrete hops. In this limit, it is possible to calculate hopping rates without calculating trajectories. Transition state theory assumes a saddle hypersurface such that a system, once it has crossed it, cannot return [140]. Transition state theory gives us the Arrhenius expression:

$$k = \nu_0 \exp(\frac{E_a}{k_b T}), \tag{4.10}$$

where the rate constant, k, is given by the exponential of the activation energy E_a multiplied by the prefactor ν_0. Vineyard's formulation of the theory gives an expression for the prefactor:

$$\nu_0 = \frac{\Pi_{i=1}^{N} \nu_i}{\Pi_{j=1}^{N-1} \nu_j}, \tag{4.11}$$

the frequencies are evaluated at the ground state and primed in the chosen saddle hyperplane (the plane of maximum flux between the two basins) then this expression gives the upper bound to the hopping rate and the value of E_a the correct activation energy for the process. This is why the search methods for saddle points are worthwhile. It is possible to obtain correction terms (taking into account the true dynamics of the system). When such terms are used, it is possible to use any convenient hyperplane to obtain the hoping rate (although in practice it is sensible to get as close as possible to the true saddle hyperplane).

4.2.2.2 Accelerated dynamics

Parallel replica dynamics methods replicate a basin across many processors [141]. The system is run until one replica shows a transition then all replicas are stopped and the simulation clock is advanced by the total time summed over all processors. The transition replica is replicated over all processors and simulation is continued. The method assumes that infrequent events obey first order kinetics: the probability distribution of the waiting

time before the next jump (beyond the correlation time) is $p(t) = k \exp(-kt)$ where k is the hopping rate. Parallel replica dynamics simulations begin with a dephasing stage where the momenta is randomised in order to eliminate correlations between processors. The parallel replica procedure does not assume transition state theory and so can account for detailed correlated events (by letting the processor where an event occurred continue for the correlation time. If replicas of the system are run on M processors, then the escape time probability for the whole system (as a function of time on processor 1) is $p(t_{total}) = Mk \exp(-Mkt_1)$ and the total simulation time is $t_{total} = Mt_1$. For any probability distribution, $(1/M)p(t/M) = p(t)dt$; so $p(t_{total}) = k \exp(-kt_{total})$. Thus, the total system of M processors, using the total simulation time, gives the same probability distribution of escape times as the single processor.

Bias potential hyperdynamics methods add a bias potential [142]. Provided the potential goes to zero at the saddle plane, the dynamics of the biased system has the same relative escape probabilities as the real system, but has the advantage of running faster. It is also possible to calculate the true simulation time from the biased simulation time. The method assumes that system obeys transition state theory both on the original surface and on the biased surface. The probability of observing a given sequence of events is the same on the two surfaces. After n steps the times on the biased surface and the original surface are related by:

$$t_{hyper} = \sum_{j=1}^{n} t_{MD} \exp(\Delta V[r(t_j)]/kT), \tag{4.12}$$

Any form of bias potential that obeys the rules is possible. Complex methods of constructing potentials tend to be numerically stable; simpler models are more robust but not efficient. The construction of good bias potentials is an active area of current research.

Temperature accelerated dynamics (TAD) use high temperature simulation to construct a faithful representation of a low temperature trajectory by extrapolating the individual processes at high temperature back to low temperature and assuming that the escape probabilities from a basin are governed by first order kinetics [143]. This method assumes first order kinetics and harmonic transition state theory for the extrapolation to low temperature. Simulation is run at high temperature in a single basin. When a transition is found, the activation energy for the process is obtained using a nudged elastic band method and the simulation continued in the original basin. This generates a list of processes, activation energies and escape times. The TAD calculation is run as a high temperature MD. The MD stage is used only to explore a single basin. When the system tries to cross the dividing surface between states the velocities are reversed, reflecting

the system back into the initial basin. This is called basin constrained dynamics. The simplest method for detecting a transition is to minimise the structure produced when the system evolves beyond an exclusion zone and test whether it is different from the starting point.

4.2.3 Replica exchange molecular dynamics

Replica exchange molecular dynamics (REMD), sometimes called parallel tempering, is a type of coupled ensemble high temperature molecular dynamics used to enhance the configurational sampling [144]. The method uses a system of N atoms with M non-interacting copies (replicas) in the canonical ensemble at M different temperatures, T_M. There is a one-to-one correspondence between replicas and temperatures: the label i ($i = 1, ..., M$) for replicas is a permutation of the label m ($m = 1, ..., M$) for temperatures and *vice versa*:

$$\begin{cases} i = i(m) \equiv f(m) \\ m = m(i) \equiv f^{-1}(i) \end{cases} \tag{4.13}$$

where $f(m)$ is a permutation function of m and $f^{-1}(i)$ is its inverse. The microstate X of the generalised ensemble is given by;

$$X = (x_{m(1)}^{[1]}, ..., x_{m(M)}^{[M]}) = (x_{1,...,x_M^{[i(M)]}}^{[i(1)]}) \tag{4.14}$$

$$x_m^{[i]} \equiv (r^{[i]}, p^{[i]})_m$$

$$r \equiv (r_1, ..., r_N), p \equiv (p_1, ..., p_N)$$

for non-interacting replicas, the weight factor of the microstate X in the generalised ensemble is given by the product of the Boltzmann factors:

$$W(X) = \exp\{-\sum_{i=1}^{M} \beta_{m(i)} H(r^{[i]}, p^{[i]})\} = \exp\{-\sum_{m=1}^{M} \beta_m H(r^{[i(m)]}, p^{[i(m)]})\}, \tag{4.15}$$

$$\beta_m = (k_B T_m)^{-1}; H(r^{[i]}, p^{[i]} = V(r^{[i]}) + K(p^{[i]})$$

we exchange replicas i at T_m and j and T_n, or we can exchange temperatures instead of replicas. Convergence towards equilibrium distribution is ensured by imposing detailed balance on the transition probability $P(X \rightarrow X')$. A detailed balance condition can be satisfied by using the Metropolis criterion for the transition probabilities.

Each replica in the generalised canonical ensemble is simulated for a given number of MD steps. A pair of replicas at neighbouring temperatures are exchanged with the probability $P(x_m^{[i]}|x_{m+1}^{[j]})$.

4.3 Analysis of simulations

The structure and arrangement of the various models were analysed using methods reported in our previous studies [69, 72, 132, 145–147], including radial distribution functions (RDF) and atomic density profiles. The radius of gyration, r_G, estimates the molecule's effective size during the simulation:

$$r_G = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\mathbf{r}_i - <\mathbf{r}>)^2}, \tag{4.16}$$

where $<\mathbf{r}>$ is the mean position of all N atoms.

The end-to-end distances represent the average distance between the first and last segments of a linear polymer. The end-to-end distance of nucleic acid strands in this study were calculated as the linear distance between phosphorus atoms at the exposed $3'$ and $5'$ ends of the nucleotide and averaged over simulation time. Changes in end-to-end distance provide an indication of the extent of folding and unfolding the molecule has undergone over time.

Root mean square deviations (RMSD) from the crystal structures were calculated by removing the translational and rotational degrees of freedom through a least-squares fit followed by rotation to the original reference frame of the molecule, using the relation:

$$r_{RMSD}(\mathbf{r}, \mathbf{r}_0) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\mathbf{r}_i - \mathbf{r}_{i,0})^2}, \tag{4.17}$$

where $\mathbf{r}$ is the current position of atom i, and $\mathbf{r}_{i,0}$ is its initial position.

Detailed structural information regarding the Watson-Crick hydrogen bonds in nucleotide models was gleaned using the 3DNA software analysis tool [148] by least-squares comparison with crystal structures of model nucleic acids. In order to identify a Watson-Crick base pair, the following geometric criteria must be met: the distance between the origins of two bases must be less than 15 Å, the vertical separation must be less than 2.5 Å, the normal vectors of the base-pairs must be rotated less than $65.0°$ with respect to each other and there must be at least one pair of nitrogen/oxygen base atoms that

are within 4.05 Å of each other. The interlayer structure was also analysed using radial distribution functions as described in our previous work [137, 149]. The radial distribution function describes how the atomic density varies as a function of the distance from a central atom.

In order to compare the mobility of the nucleic acid strands within mineral interlayer galleries, the self-diffusion coefficients of all nucleic acid duplex strands were calculated and averaged over the number of duplexes in each model. The diffusion coefficient is calculated from the slope of the mean square displacement (δr^2) *vs.* time graph. The MSD was computed from the centre of mass of each molecule in order to include only translational motion in the calculation. Graphs were constructed using multiple time origins in order to improve the statistics of the calculation.

The MSD was calculated using:

$$< r^2(t) >= \frac{1}{N} \sum_{i=1}^{N} (\mathbf{r}_i - \mathbf{r}_{i,0})^2, \tag{4.18}$$

where $\mathbf{r}$ is the current position of atom i, and $\mathbf{r}_{i,0}$ is its initial position. The diffusion was subsequently estimated from the diffusion coefficient using:

$$< r^2(t) >= 6Dt + C, \tag{4.19}$$

where C and D are constants; D is the diffusion coefficient and defines the rate of diffusion of the molecule, and t is time. The diffusion coefficient can be found from the gradient of the mean-squared displacement as a function of time.

As the molecules are constrained to move within the interlayer planes the diffusion is confined to the xy-plane. Therefore, we compute values for the two dimensional diffusion coefficient, which is given by $4D_{xy} = \frac{d}{dt}\delta r^2$ where $\delta r^2 = (\delta x^2 + \delta y^2)/2$. The slope of the δr^2 *vs.* time plot exhibits a linear region of slope D, and the error on the least square fit estimates the error of the diffusion coefficient.

Visualisation is an essential tool for examining and analyzing the structure and behavior of molecular systems. The systems were visualised using the AtomEye atomistic configuration viewer [150] and the Visual Molecular Dynamics program (VMD) [151]. AtomEye and VMD are both freely available visualisation software packages. AtomEye treats atoms as spheres and bonds as cylinders. These images are rendered as primitive objects rather than composites of polygons. AtomEye was employed for its fast rendering qualities. VMD was used for its excellent representation of nucleic acid structures.

4.4   Scientific, parallel and grid computing

Simulating biological and inorganic systems using MD methods is computationally very expensive due to the sheer number of calculations the classical MD code has to iterate through. In order to get the most out of modern day MD codes in terms of simulation time and number of atoms simulated, it is necessary to employ the large scale supercomputing resources that are available to the scientific community. It is the parallelisation of modern MD codes which lend them so well to supercomputers. The near-linear scaling relationship between simulation time and the number of processors used shown by some MD codes means that large atomic/molecular systems can be simulated using MD in realistic wall-clock times.

Grid supercomputing, defined as "distributed computing performed transparently across multiple administrative domains" [152], is one of the more appealing methods in modern high performance computing. It aims to provide a framework for scientists to run simulations on machines ranging in size from small local clusters to massive international supercomputers. Just as the electrical grid seamlessly delivers electrical power to our homes, the dream behind the original computing grid was to deliver uniform methods of transparent access to supercomputing resources. There are a number of grid projects which aim to unite the resources to this end. These include The US XSEDE [1] grid and PRACE [2].

XSEDE (Extreme Science and Engineering Discovery Environment) in particular supports 16 supercomputers and high-end visualisation and data analysis resources across the US. Currently the largest resource on XSEDE is Stampede at the Texas Advanced Computing Center. Stampede is a Dell PowerEdge C8220 Cluster with Intel Xeon Phi coprocessors. The 102,400 processors are capable of 2.66 PetaFlops (2.66×10^{15} floating-point operations per second). Machines such as Stampede allows the simulate fully atomistic biological and inorganic systems on the scale of millions of atoms for microseconds, as well as performing computationally expensive sampling and ensemble simulations.

There has been a trend in recent times to employ GPUs (Graphics Processing Unit) for parallel computing because of their relatively low cost. The UK's most powerful GPU/CPU machine resides at the Center for Innovation in High Performance Computing at the STFC Rutherford Appleton Laboratory [3]. The Emerald system is an 84-node cluster equipped with 372 NVIDIA Tesla M2090 GPUs, delivering more than 114 teraflops of performance.

---

[1] https://www.xsede.org/
[2] http://www.prace-ri.eu/
[3] http://www.einfrastructuresouth.ac.uk/cfi/emerald/

Chapter 5

Stability of Free and Mineral-Protected Nucleic Acids

This chapter investigates the structural stability of three different nucleic acids intercalated within a magnesium aluminium layered double hydroxide (LDH) mineral, at varying degrees of hydration, and free in aqueous solution using molecular dynamics techniques. The nucleotides investigated are ribose nucleic acid (RNA), deoxyribose nucleic acid (DNA), and peptide nucleic acid (PNA), all in duplex form. This study is used to gain insight into which candidate genetic material arising at the time of the origin of life on Earth may have been preferentially favoured by the prevailing geochemistry, in particular when interacting with anionic clays.

5.1   Introduction

Identifying the nature of the chemical processes that led to the origin of life on Earth poses a major intellectual challenge to modern science. The complete absence of a "fossil record" for the type of proto-biomolecules present in this era means that there is no way any hypothesis can be proven to be correct. However, it is possible to make reasonable proposals and then experimentally investigate the processes that may have led to the emergence of the first biomolecules, including the first genetic material and the first living organisms [153, 154].

The RNA World view is one hypothesis that has attracted widespread support amongst those investigating the origin of life [155]. The RNA World hypothesis asserts that the first life forms on Earth were based on RNA molecules, which were able to self-replicate,

preceding the appearance of proteins and subsequent protein-based (bio)chemistry [155–157]. The essence of the RNA World hypothesis is that the transfer of genetic information relied solely upon the self-replication of RNA, for which Watson-Crick base-pairing provided the key mechanism [158]. The hypothesis concurs with the doctrine of biopoiesis, which asserts that life originated from simple organic molecules already present on Earth [159]. However, a gap exists in the process of biopoesis in that it relies on there being a pre-biological route to an initial source of nucleotides for the formation of RNA. Though there is evidence showing that RNA could have catalysed most of the steps involved in the synthesis of nucleotides [160], and possibly the coupling of redox reactions to the synthesis of phosphodiesters and peptides, no complete abiotic synthetic pathways for all of the nucleotides have hitherto been reported. Recent work has suggested potentially plausible prebiotic routes to RNA synthesis, including from activated pyrimidine ribonucleotides [27, 161], whilst layered minerals have also been shown to be catalytically active in RNA elongation [162]. Notwithstanding these problems, it is conjectured that the transition to the DNA genetic apparatus came at a later time, possibly via an RNA-protein intermediate stage [163, 164].

Although it is conceivable that the first forms of living organism were comprised of RNA, oligomers of RNA are highly sensitive to hydrolysis, making it challenging to explain how such species may have survived long enough to create more complex forms of proto-life. Moreover, it has been proposed that RNA's arrival on the scene may have been preceded by simpler organic genetic materials. One such molecule is peptide nucleic acid (PNA) which differs from RNA and DNA in its backbone structure; PNA has an uncharged protein backbone unlike the charged phosphorus backbone of RNA and DNA [165] (see Figure 5.1). Unlike RNA and DNA, PNA has not been observed in Nature; PNA was originally designed using computer modelling and subsequently synthesised in a laboratory.

Experiments have shown that PNA is able to displace complementary oligonucleotide strands from existing DNA duplexes, indicating that PNA·DNA hybrids are more stable than duplex DNA. Other enhanced structural features include faster rates of reformation of the duplex upon denaturation and comparatively higher melting temperatures [102]. These properties and its similarities with DNA have led to PNA being proposed as a possible proto-RNA [166] in origin of life scenarios [167, 168]. But which nucleic acid arose first and how? To answer these questions we must consider the conditions present on the early prebiotic Earth [169].

Assuming that the initial prebiotic synthesis of proto-biomolecules took place in aqueous solution and arose within the oceans of the Archean Earth (3.5Ga ago), it is likely that the chemical and temperature gradients at early hydrothermal vents may have provided
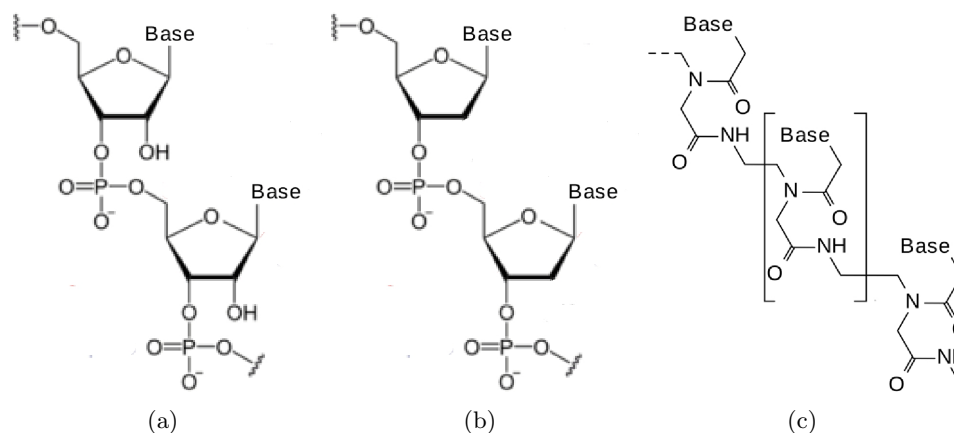
Figure 5.1: The structural formulae of (a) ribonucleic acid (RNA), (b) deoxyribonucleic acid (DNA) and (c) peptide nucleic acid (PNA). PNA is an artificially synthesized nucleic acid. PNA has an uncharged backbone which is composed of repeating N-(2-aminoethyl)-glycine units linked by peptide bonds.

the necessary driving forces for chemical reactions to occur [170]. The production of simple organic molecules has been observed in present day vents [171] and replicated in laboratory studies [172]. For organic macromolecules to have been generated, small abiotic substances had to be concentrated to the extent that reactive centres came within sufficient proximity for further chemistry to have occurred, while the products would have needed to be protected from degradation. One way to realise these conditions is at the surfaces and within the internal pores of minerals present at such vent systems [173]. A recent theory, by Hansma, suggests that layered clay minerals may have acted as the first proto-cells [174].

Layered double hydroxides (LDHs) are a class of layered mineral, sometimes known as anionic clays (the sheets are positively charged and take up anions), which have been the focus of origins of life studies for a number of years [175–178]. It is thought that in the reducing conditions present on the early Earth, green rust (Fe^{2+}/Fe^{3+}) LDHs may have been prevalent [73]. LDH minerals have structures that are conceptually derived from the layered mineral brucite ($Mg(OH)_2$), which has divalent Mg^{2+} metal cations octahedrally coordinated with hydroxide ions. In LDHs, the isomorphic substitution of trivalent in place of divalent cations causes a net positive charge which is counterbalanced by the introduction of anions within the interlayer region. Anions of particular interest to organic molecular evolution include Cl^-, SO_4^{2-}, CO_3^{2-}, PO_4^{3-} and linear oligophosphates [177]. In the LDH interlayer, anions can be readily exchanged with other anionic species so that carbonates, cyanide, oligomeric phosphates and amino acid guests may all be intercalated into the LDH host structure.

Glyco-aldehyde phosphate (GAP) has been postulated as a candidate for a primitive

building block *en route* to nucleic acids [179]. These molecules become highly concentrated through intercalation into LDHs from a dilute external solution [175]. Experimental results by Pitsch et al. [175] show that LDHs are able to adsorb GAP anions from concentrations of a few μM to yield a $\sim$10M confined environment within the LDH interlayer. Subsequent condensation reactions of GAP ions in LDHs give a high yield of sugar phosphates [178].

Owing to interest in LDH-DNA compounds for drug delivery [180–185] and molecular code systems [186], several experimental studies have been undertaken reporting intercalation of DNA into LDHs. Experimental studies in these areas have provided some understanding of DNA-LDH systems; powder X-ray diffraction (PXRD) has demonstrated the changes in interlayer spacing in these materials [187], while circular dichroism and infrared spectroscopy have been used to study the intercalated nucleic acids [188]. Although all forms of polyanionic DNA have been found to intercalate into LDHs [187], to the best of our knowledge no evidence of RNA or PNA intercalation into LDHs yet exists in the literature. Greenwell and Coveney further pointed out that, conceptually, LDHs might represent idealised information storage and transfer systems and that the separation distance of intralayer charge sites is remarkably similar to the distance between phosphate groups in nucleic acid structures [189].

Origins of life studies have hitherto rarely used computer simulation techniques to understand the possible chemical pathways to the formation of the first biomolecules. However, computational methods provide powerful molecular level insights into the structure and properties of cationic and anionic clay based systems which are difficult to characterise experimentally [190]. As a result, interest has begun to grow in the use of computational modelling to investigate systems of possible relevance to the origins of life [70, 132, 137, 191]. Furthermore, computer simulation can rapidly access environmental conditions difficult to attain by experiment.

This study employs large-scale molecular dynamics simulation techniques, similar to those previously performed [132, 137, 191], to compare the relative structural stability of DNA, RNA and PNA all within layered double hydroxides and free in bulk aqueous solution. Double stranded RNA was used to investigate the potential for prebiotic LDH-RNA systems to protect early forms of these nucleotides from extreme environmental conditions. The results of double stranded DNA and PNA structures were compared to RNA thus providing insight into the comparative properties of these nucleic acids in mineral and bulk aqueous environments.

5.2 Methods

This section discusses the approach employed to simulate the LDH-nucleic acid-water models. The choice of forcefield needed to describe the interaction between the nucleic acids and LDH is given in Chapter 4. The models simulated contain up to around a quarter of a million atoms that can be classed as large-scale molecular dynamics. A range of powerful supercomputing resources were invoked to perform the simulations in as short a wallclock time as possible.

The layered double hydroxide which forms the basis for the mineral models employed in this study has unit formula $[Mg_2Al(OH)_6]\cdot nH_2O.Cl$ and is the same as that used in recent studies [137, 149, 191]. Chloride ions are present to counter the positive LDH charge. The LDH models were replicated from a unit cell with dimensions 16.34 Å×18.82 Å×25.34 Å, which was obtained by the refinement of powder X-ray diffraction data on hydrotalcite using Rietveld methods [81]. In this work, models containing either 3 or 6 interlayer regions were simulated.

The sizes and sequences of the nucleic acid molecules used in this study are given in Table 5.1. Three different forms of double stranded nucleic acid were constructed, as shown in Figure 7.6. The first structure was that of a dodecamer duplex of A-DNA assembled using the Nucleic Acid Builder, which is part of AmberTools [127], with base-pair sequence d(5′-CTTTTGCAAAAG-3′) [192]. The DNA sequence was chosen as it has been well characterised in past molecular dynamics simulations [137, 192]. The second structure was a dodecamer of A-RNA also built using the Nucleic Acid Builder with corresponding base-pair sequence d(5′-CUUUUGCAAAAG-3′) (the RNA sequence corresponds to the DNA sequence with thymine groups substituted by uracil). All phosphate groups have an unprotonated hydroxyl group and carry a -1 negative charge, as a pH of ≥ 7 is assumed where all phosphate groups are known to be ionised. The third structure was a dodecamer of PNA, in the P-form, with the same base-pair sequence as the DNA dodecamer. The PNA strand's peptide backbone was terminated with N-terminus and C-terminus end-groups, as reported by Shields et al. [128] and Soliva et al. [193] in their MD studies of PNA:DNA and PNA:RNA molecules.

Nucleic acids RNA and DNA can adopt two different right-handed double helix structures, A and B. The A-form has a shorter more compact helical structure which has an increase in the number of base-pairs per rotation over the B-form. The A-form of DNA was chosen to allow a direct comparison with the A-form used for RNA. The A-form of

(a)   (b)   (c)
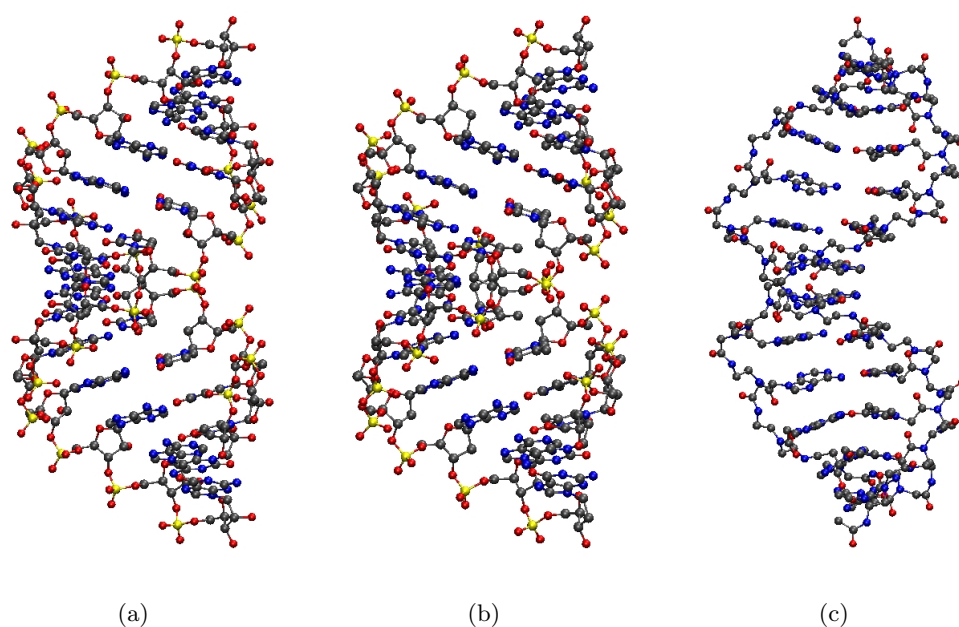
Figure 5.2: A schematic three dimensional ball-and-stick representation of (a) A-RNA in Models **I, IV** and **VII**, (b) A-DNA in Models **II, V** and **VIII** and (c) PNA in Models **III, VI** and **IX**. All models are described in more detail in Table 5.1. Carbon, nitrogen, hydrogen, oxygen and phosphorus atoms are represented as grey, blue, red and yellow spheres respectively. The structure of RNA differs from the structure of DNA by the additional hydroxyl group at the C2 position on the nucleotide ribose.

DNA is known to undergo a transition to the B-form (or something closely approximating it) in bulk water. The A-form is known to be the preferred structure for DNA under conditions of high salinity and low hydration. The P-form of PNA, which is the name given to its native structure in aqueous solution (shown in 7.6), appears structurally different to the A-form of DNA and RNA in Figure 7.6 because of the difference in composition of the peptide backbone in PNA from the phosphate backbone in DNA and RNA. The P-form can be characterised by the wide helix (28 Å) with almost twice the pitch (18 base-pairs per turn) of an A- or B-form helix (10-11 base-pairs per turn) [194]. In order to make comparisons between the nucleic acids in water and when intercalated, simulations of the three types of nucleic acid were also performed solvated in water (Models **VII-IX**, see Table 5.1); the nucleic acids within these bulk water simulations were charge-balanced by sodium ions.

Many studies have shown that water plays a vital role in the structure of nucleic acids in aqueous solution; the role of water also governs the structural stability of nucleic acids once intercalated into layered double hydroxides [137, 182, 187]. The first part of our study therefore aims to investigate the impact of hydration on Models **I, II** and **III**, which consist of a nucleic acid intercalated within a LDH as shown in Figure 5.3, by varying their water content. Models of varying degrees of hydration were created

by changing the water content only in the interlayer containing the nucleic acid, while keeping the other interlayer at two water molecules per unit formula $Mg_2Al(OH)_6$. Previous studies of MgAl-LDH produced simulated hydration curves in general agreement with experimental data using the ClayFF forcefield when intercalated with chloride ions [129, 149], and with DNA [137]. In order to create hydration curves for the RNA, DNA and PNA models, sixteen different LDH-nucleic acid models were built with an increasing number of water molecules, from 2 water molecules per Al atom up to 16 water molecules per Al atom. In cases where nucleic acids are included in adjacent interlayers, the molecules were offset with respect to each other. Our previous simulations of DNA-LDH have shown that the model adopts the lowest potential energy when the nucleic acids in adjacent interlayers are not stacked directly on top of each other [137, 191].
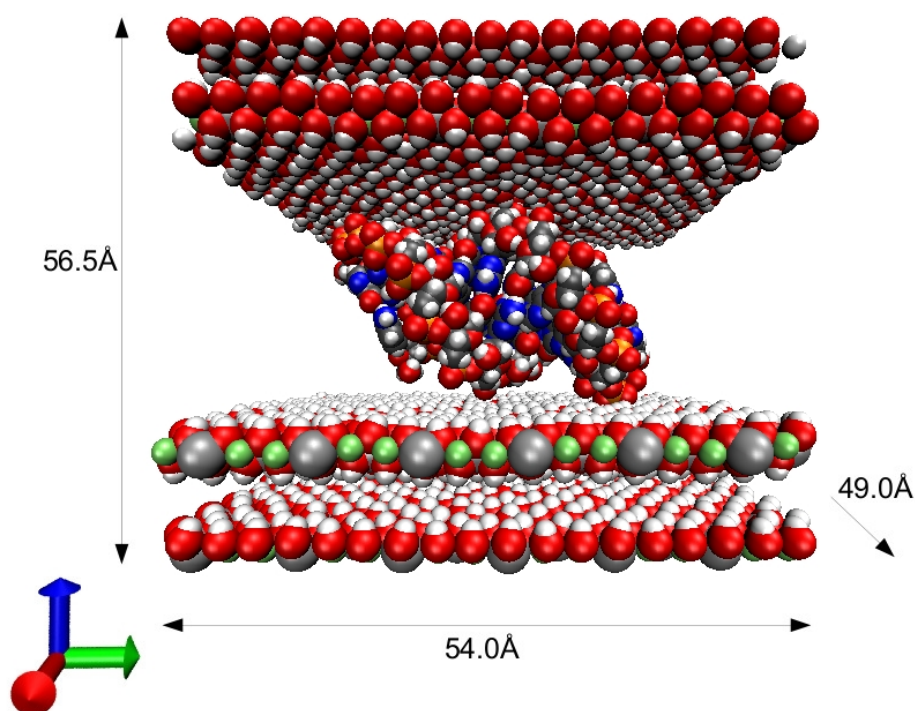
Figure 5.3: Initial structure of the LDH-RNA Model **I** at the start of the simulation. Model **I** consists of three Mg_2Al LDH sheets with a 12 base pair RNA duplex placed within one of the interlayers. For clarity, water molecules and chloride ions are not displayed. Magnesium, aluminium, oxygen, hydrogen, nitrogen, carbon and phosphorus are displayed as green, silver, red, white, blue, grey and orange spheres respectively. The x, y and z axes are displayed as red, green and blue arrows respectively.

Small LDH models suffer from finite size effects which cause the suppression of thermal undulations. The collective thermal motion of atoms in the LDH sheets causes undulatory modes. The largest wavelength of thermal undulations was found to be 40 Å [149]. The lateral dimensions of the largest models used in this Chapter were more than twice

this distance. Model **IV** contained an RNA dodecamer duplex while Model **VI** contained a PNA duplex of similar length; the hydration level which yielded basal spacings comparable with experimental findings for analogous DNA-LDH models are 21.1 - 23.9 Å. Experimental results were taken from the variation in the *d*-spacing as a function of the number of water molecules calculated from the thermogravimetric analyses recorded in a static air atmosphere with a heating rate of 278 K/min, in the temperature range of 298-1073 K. [182, 187]. Larger models based on this experimental hydration state were created containing six LDH interlayers for the RNA, DNA and PNA duplex models (Model **IV-VI**, Figure 5.4), with four RNA/DNA/PNA duplices in each interlayer, in order to simulate the nucleic acid-LDH models at higher temperatures and pressures. Models **IV-IX** were simulated at five different conditions: 300 K at 1 atm, 350 K at 1 atm, 400 K at 50 atm, 450 K at 100 atm and 500 K at 100 atm, in order to study the stability of nucleic acids intercalated and free in bulk water under different conditions associated with possible origins of life environments [137]. The various temperature and pressure conditions were selected to span mild and extreme conditions associated with hydrothermal vent models and mineral mediated origins [169].

The models were energy-minimized using the steepest descent method before carrying out the MD simulations using the NPT ensemble (which means the number of particles (N), the pressure (P) and the temperature (T) are kept at a constant value), using techniques discussed in our previous studies [132, 137, 149] and in Chapter 4. After equilibration was deemed to have been established, production runs were performed for 30 ns to provide enough statistics to capture diffusion data for the nucleic acid molecules [149]. The models were deemed to have reached equilibrium before 500 ps by monitoring the model potential energy and the cell parameters. The data show that all models are at equilibrium (or at least in long lasting steady states) as adjudged by the figures of merit, including potential energy and cell parameters, and that any further simulation beyond 30 ns is not necessary. A Nosé-Hoover thermostat/barostat was used to regulate the temperature and pressure of each simulation.

5.3 Results

As reported in our previous work on DNA-LDH models [137], the nucleic acids studied here do not exhibit large changes in position and conformation within the interlayer over the duration of the simulation; conversely the LDH sheets themselves show evidence of their flexibility around bulky intercalants [137, 191]. Visualisation of the final structures of the larger models (Models **IV, V** and **VI**) at ambient conditions are shown in Figure 5.5.
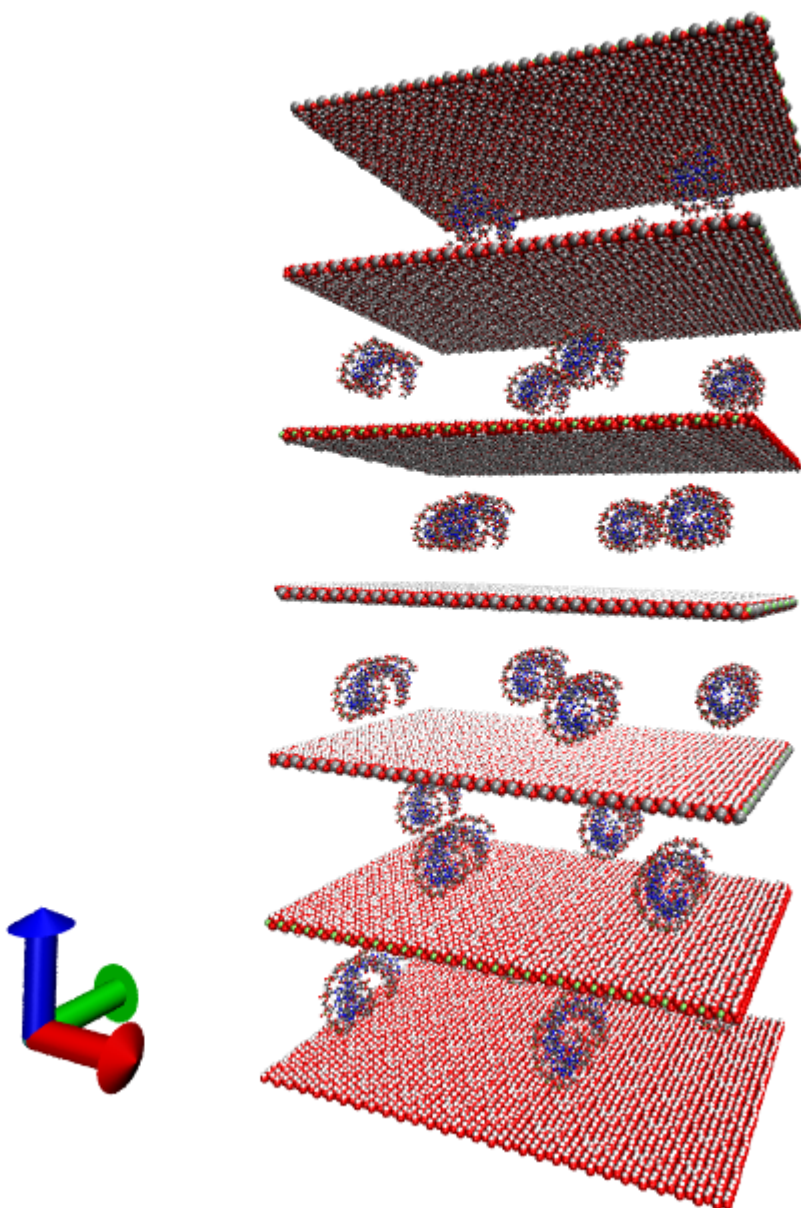
Figure 5.4: Initial structure of the large LDH-RNA Model **IV** at the start of the simulation. The model consists of 6 Mg$_2$Al interlayers with four 12 base pair RNA double strands per interlayer. See Table 5.1 for more details of sequence and structure. For clarity, water molecules and chloride ions are not displayed. Colour scheme is the same as that used in Figure 5.3.

Section 5.3.2 describes in detail the hydration behaviour of LDH models containing RNA, DNA and PNA, and compares these properties to our previous results for similar sized DNA-LDH models [137]. In Section 5.3.3 the behaviour of RNA, DNA and PNA intercalated in larger LDH structures, which eliminate finite size effects and capture emergent thermal undulations within the LDH sheets [149] is reported. Previous experimental work and simulations of DNA, RNA and PNA in bulk aqueous solution were used to validate the nucleic acid models reported.

5.3.1 Validation of models for nucleic acids in bulk water

In order to provide confidence in the nucleic acid models, various parameters describing Watson-Crick base-pair geometry were calculated for each nucleic acid system (models **VII**, **VIII** and **IX**) and compared with previously reported parameters found experimentally as well as those derived from simulation (see Table 5.2). Parameters obtained from the nucleic acids in our systems are in good agreement with values obtained from both experiment and simulation and therefore provide confidence in properties reported for both free nucleic acids and nucleic acids within clay galleries.

As reported in previous MD simulations of DNA in bulk water using the parm99 force-field [201, 202], DNA undergoes a spontaneous conformational change from the A-form to the B-form in simulation times of less than 3 ns. Simulation of DNA in bulk water (model **VIII**) displays the transition from the A to the B form of DNA, as shown in Figure 5.6 and as a movie in Appendix A. Simulation over longer timescales of 30 ns did not show any back-transition to the A-form.

5.3.2 Effect of hydration on nucleic acid-LDH models

The swelling/hydration curves were produced by simulating Models **I-III** with an increasing amount of water molecules per nucleic acid containing interlayer, beginning with a dehydrated model and adding two water molecules per LDH unit formula ($[Mg_2Al(OH)_2]$) up to sixteen water molecules per LDH unit. The variable n refers to the number of water molecules per unit formula. Each hydrated model was simulated for 3 ns once equilibrium was deemed to have been established.

Computing the hydration curves (clay layer separation as a function of water content) gives insight into how the nucleic acid structures change with varying amounts of water. Figure 5.7 shows the direct comparison between the change in basal spacing, RMSD from the crystal structure and disruption of Watson-Crick hydrogen bonded base-pairs, with increasing water content for DNA, RNA and PNA intercalated in LDH. Some similarity

exists in basal spacings between DNA and PNA, whereas on average RNA has generally lower basal spacings, especially when the number of water molecules per unit formula ranges between $n = 6 - 15$.

At low hydration states, below 10 water molecules per Al atom, significant structural deformation can be seen as evidenced by the relatively high RMSD and correspondingly low percentage of Watson-Crick hydrogen bonds remaining intact, but as the water content of the interlayer reaches similar values to bulk water, at 16 water atoms per Al, the nucleic acids return to their bulk-water solvated double helix structures (see Figure 5.8). Plots of RMSD *vs.* hydration in Figure 5.7 show that the structures of both RNA and PNA become increasingly similar to that of the same nucleic acid in bulk water as the interlayer water content increases. However, the RMSD values of RNA and PNA are significantly greater than DNA after $n = 8$. In addition, the RMSD between the intercalated and reference structures does not decay monotomically, but rather exhibits numerous local minima and maxima. These "steps" are also observed in Figure 5.7 (a) for the basal spacing and are attributed to hydrogen bonding networks formed between water and the clay surfaces, which only allow expansion of the interlayer once distinct hydration layers are fully filled [129, 149, 203, 204]. This "step" behaviour is prominently seen in Figure 5.7, where the percentage of intact Watson-Crick hydrogen bonds is reported, and is particularly obvious for PNA. From the results for RMSD values, it can be inferred that intercalated PNA strands show lower structural stability, in terms of number of Watson-Crick base-pairs, at higher hydration states ($n > 8$) compared to RNA and DNA.

Presently, there are no experimental results which describe the basal spacing of RNA-LDH or PNA-LDH models as a function of water content; instead, the results from our DNA-LDH models are used to give a rough estimate as to which hydration state to use [181, 187, 205], when building larger simulation models. The experimental value of the basal spacing under ambient conditions for DNA-LDH models is reported to be between 21.1 - 23.9 Å [182, 187], which corresponds to a water content, n, between 10 - 13 water molecules per unit formula for RNA-LDH models and 8 - 11 for PNA-LDH. As RNA-LDH models have lower basal spacings compared to PNA-LDH and DNA-LDH models (Figure 5.7), these LDH-nucleic acid models were built with $n = 10$ water molecules per unit formula $[\mathrm{Mg_2Al(OH)_6}]{\cdot}n\mathrm{H_2O}$. As the PNA double strand is less easily compressed, the number of water molecules per unit formula chosen was $n = 9$, to ensure values are comparable with experimental basal spacings for DNA-LDH models. The following sections describe the results obtained for larger LDH models (Models **IV, V** and **VI**), each in turn containing a different nucleic acid moiety, at different temperatures and pressures.

5.3.3   RNA, DNA and PNA intercalants

Model **IV** was used to study the properties of RNA strands intercalated within a six layer LDH supercell, with four double strands per interlayer. Under ambient conditions (300 K and 1 atm) Table 5.3 shows that, on average, base-pairing is substantially less than for RNA in bulk water, indicating that the environment within the LDH interlayers causes the Watson-Crick hydrogen bonding to be significantly disrupted. Table 5.3 provides structural information for a DNA-LDH model; the intercalated DNA strands retain ~60% of all base-pairing up to 450 K and 100 atm, while that for RNA strands decreases monotonically from a maximum of 25% with increasing temperatures and pressures. RNA within bulk water appears to have better retention of Watson-Crick base-pairing than DNA and PNA in bulk water, but the RMSD values are greater above 400 K and 50 atm than for the intercalated RNA, indicating that the LDH sheets constrain the structure of intercalated RNA molecules at elevated temperature and pressure, albeit at the expense of Watson-Crick pairing.

Figure 5.9(a) shows the RMSD of RNA intercalated within the LDH, as a function of simulation time, relative to the A-form and B-form crystal structures of RNA. The RNA is found to deviate only by a few Ångstroms from the A-form crystal structure at the highest hydration level investigated. At all other lower levels of hydration the RNA structure deviates much more (5-6 Å) from the A-form crystal structure, and even more so from the B-form.

Figure 5.9(b) shows the RMSD of DNA intercalated within the LDH, as a function of simulation time, relative to the A-form and B-form crystal structures of RNA. The findings indicate that the DNA in a fully hydrated LDH clay adopts a structure intermediate between those of the crystalline A- and B-forms. At lower levels of hydration, the confinement and the clay environment lead to a greater distortion of the nucleic acid; the equilibrated structures are closer to the A-form than the B-form.

Figure 5.9(c) shows the RMSD of PNA whilst intercalated within an LDH, as a function of simulation time, from the crystal structure of PNA, which is refered to as the P-form, at various levels of hydrations. The intercalated PNA most resembles the initial P-form when simulated with 16 water molecules per unit formula (*i.e.* at higher hydration levels). At lower hydration levels the structure is considerably distorted from the P-form.

The radii of gyration for all three duplex nucleotides intercalated in a LDH and in bulk water at ambient conditions, shown in Figure 5.10, demonstrate that PNA has the largest relative size both when free in bulk water and intercalated within the LDH. There is a change in the relative order of size between the three different nucleic acids at ambient conditions when intercalated within the LDH compared to the nucleic acids

in bulk water; DNA has the smallest relative size when intercalated (Model **V**) whilst RNA has the smallest size when free in bulk aqueous solution (Model **VII**). At elevated temperatures and pressures RNA, DNA and PNA within the LDH (Models **IV-VI**) take on the same order of size as at ambient conditions, whilst the nucleic acids free in bulk aqueous solution at elevated temperatures and pressures have a reversed order of size compared to the order at ambient conditions. At 450 K and 100 atm PNA has the smallest size in bulk water relative to RNA and DNA, while they all show increased fluctuations in size in bulk water compared to that under ambient conditions, indicating that the extreme conditions alter the size and structure of the nucleic acid in bulk water. The relative duplex nucleotide size shown by the radius of gyration at elevated temperatures and pressures further indicates that the LDH protects the structure and conformation of the nucleic acid from changes caused by extreme environments.

The RDF plots in Figure 5.3.3 provide insight into the overall structure of the RNA-LDH model and how it varies with increasing temperature. Under ambient conditions, the RDF plots show that, as expected, LDH surface hydrogen atoms are closest to phosphorus atoms in negatively charged phosphate groups [137]. Although the peak intensity of the RDF curves for aluminium and magnesium ions are very similar, the intensities corresponding to the closest surfaces, $r = 5.3$ Å for Al^{3+} ions and $r = 5.4$ Å for Mg^{2+} ions, manifest a slight preference for proximity to aluminium compared with magnesium ions. At higher temperatures and pressures, this behaviour in the RDF plots is enhanced, suggesting greater interaction of the phosphorus and LDH atoms. Conversely, peak intensities for RDF plots of water surrounding phosphorus atoms diminish with increasing temperature and pressure.

RDF plots in Figure 5.12, centred on the phosphorous atoms within the nucleic acids, except for PNA which is centred on the sp^2 hybridised amide carbonyl carbon atom in the PNA backbone (see Figure 5.1) show the peak intensities for the RDF plots between phosphate group P atoms and selected LDH atoms increase with increasing temperature, suggesting that the phosphate groups in PNA have stronger interactions with the LDH surface at higher temperatures and pressures. By contrast, the peak intensities in the RDF plots for phosphate group P atoms and water O and H atoms decrease with increasing temperatures and pressures. This indicates that hydrogen bonding between phosphate groups and water molecules is disrupted with increasing temperature. This behaviour is also seen in Model **IV**. However, the principal component analysis (PCA) plots shown in Figure 5.13 reveal a much more well defined structure and show little difference to the behaviour obtained for PNA in bulk water (see Figure 5.14). This figure shows that the amplitude of motion along the first eigenvector is ~ 8 orders of magnitude larger for PNA in bulk water compared to PNA intercalated within LDH, reflecting the heavily constrained motion of PNA inside LDH sheets.

Figure 5.15 shows that the main contribution to the dynamics for intercalated RNA and DNA arises from the motion of the phosphate groups furthest away from the LDH. These phosphate groups oscillate towards/away from the central axis, tensioned by opposing forces of attraction towards the LDH surface and hydrogen bonds between intra-strand base-pairs. The main contribution to the dynamics of the PNA strand differs from DNA/RNA, as the PNA has an uncharged peptide backbone which does not interact so strongly with the LDH surface. The peptide backbone oscillates on both sides towards/away from the central axis.

As seen in previous studies [137, 146, 149, 191], thermal undulations of LDH sheets are evident in large scale molecular dynamics simulations. These are found to significantly affect the motion of intercalated DNA strands [137, 191] due to strong electrostatic interactions between the intercalant and the LDH sheet. PCA provides clear evidence that RNA strands are also greatly affected by the thermal motion of LDH sheets. Figure 5.16 and 5.17 compares PCA results for intercalated RNA and RNA within bulk water. Under ambient conditions, RNA intercalated within LDH as well as within bulk water has a unimodal probability distribution along the first eigenvector, suggesting that a well defined equilibrium structure has been reached. Figures 5.16 and 5.17 shows the subspace spanned by the first two principal components (the first principal component is shown in Figure 5.15) for Models **IV** and **VII** at ambient temperatures and pressures. At higher temperatures and pressures, it becomes increasingly evident that projections along the first two eigenvector pairs are coupled, which is most likely due to thermal motion in the LDH sheets increasingly influencing the motion of intercalated RNA [137]. The main contribution to the dynamics at this hydration level is seen at the ends of the nucleic-acid molecules.

PNA intercalants are composed of an uncharged PNA protein backbone. In turn, this causes the dynamics of PNA to be different from duplex strands of RNA and DNA each of which possesses two charged complementary strands. At temperatures and pressures above ambient conditions, the number of retained Watson-Crick hydrogen bonds is larger than those found in intercalated 12 base-pair RNA, suggesting stronger base-pair interactions within the PNA strand. Basal spacing values for PNA-LDH are on the whole larger than RNA-LDH in Model **IV**, suggesting that the 12 base-pair RNA strand has less resilience to the LDH sheet dynamics and distorts more readily. The RMSD values for the PNA-LDH Model **VI** in Table 5.3 are larger than those for RNA-LDH in Model **IV** indicating that, overall, the PNA double strand deviates more from its starting structure at all temperatures and pressures compared to RNA, whilst retaining a higher proportion of Watson-Crick bonding. By contrast, the structural parameters of intercalated DNA with the same base-pair sequence, as detailed in Table 5.3, shows that the DNA double helix retains its own integrity more than the PNA duplex, relative

to their individual starting structures. Table 5.3 shows that the LDH confers a different relative preferential stability on the three nucleic-acid duplexes compared to their homologues in bulk water. The relative trend in stability in terms of intact Watson-Crick base-pairs when intercalated is PNA<RNA<DNA. The reverse trend is observed for nucleic acids free in aqueous solution.

Although the structure of RNA is perturbed whilst intercalated, the RMSD and PCA data indicate that overall structure retention, at high temperatures and pressures, is similar to that exhibited at ambient conditions.

At ambient temperatures of 300 K PNA is found to have the smallest self-diffusion coefficient of the three intercalated nucleotides whilst RNA has the largest (shown in Table 5.4). The trend in diffusion coefficients is altered at elevated temperatures of 500 K where PNA has the highest and DNA the lowest self-diffusion coefficient.

5.4   Conclusions

A comparative study of the structure, properties and stability of selected double-stranded sequences of RNA, DNA and PNA has been undertaken In this chapter. The intention was to gain insight into which candidate genetic material arising at the time of the origin of life on Earth may have been preferentially favoured by the prevailing geochemistry, in particular when interacting with anionic clays.

Study of the root-mean-squared deviation relative to the starting structures together with principal component analysis of the double-stranded nucleic acids indicate that the intercalated RNA, DNA and PNA molecules all have significantly reduced motion relative to the nucleic acids in bulk water. However, the relative similarity between starting and end-point structures does not offer the best insight into structure retention of each double stranded nucleotide. A more important measure of retention of structural integrity from a genetic information-transfer perspective is the number of Watson-Crick bonds maintained.

The simulations demonstrate that hydration plays an important role in determining the structural stability of all three intercalated nucleic acids. This is in agreement with our previous findings on DNA moieties within LDH [137]. The results show that the three nucleic acids are affected differently by the degree of hydration of the LDH. While RNA continues to adopt a structure closer to the crystalline A-form at all levels of hydration, at lower hydration levels there is also a tendency for DNA to adopt structures closer to the A-form than the B-form found in bulk water. With regard to Watson-Crick hydrogen-bonding, the DNA duplexes retain a greater structural integrity as compared

to the intercalated RNA and PNA double strands which manifest significant degradation in base-pairing.

Having examined the properties of selected 12-mer duplexes of RNA, DNA and PNA in bulk water and when intercalated in the interlayer of LDH minerals, together with the hydration properties of the nucleic acid-LDH complexes, a conclusion can be reached that the properties of the DNA, RNA and PNA duplexes are strongly dependent on the state of hydration of the LDH interlayers. For the assembly and maintenance of the first double strand genetic system, a hydrated environment is necessary and, therefore, it is vital for there to be water within the LDH interlayers.

All of the nucleic acids retain some degree of structural integrity when intercalated in LDH minerals, even at elevated temperatures and pressures. This supports the notion that such anionic clays, possibly present within early Archean hydrothermal systems as Fe(II)/Fe(III) green rusts, may have played a significant role in concentrating and catalysing early prebiotic chemistry, as proposed by Arrhenius [73].

LDH-intercalated 12 base-pair double-stranded RNA oligomers were generally found to deform from their initial structure more readily compared to DNA and PNA. In addition, the 12 base-pair RNA and PNA strands are more easily deformed from their initial starting structure in bulk water compared to DNA; significantly more water is needed within LDH sheets to retain the double helix shape of *ds*-RNA and *ds*-PNA compared to the situation in bulk water, while there is a noticeable preference for intercalated DNA to retain an A-form, especially at lower levels of hydration.

To get a proper handle on the behaviour of these nucleotides within clays, a range of analysis tools needs to be brought to bear on the molecular dynamics trajectory data. Though the nucleic acid strands may lose their Watson-Crick base pairing in layered double hydroxides, the root-mean-squared deviation relative to the initial nucleic acid structure is invariably smaller for the intercalated nucleic acid than for the nucleic acids in bulk water. The intercalated species, stabilised through strong Coulombic interactions with the layered double hydroxides sheets, is unable to distort much from its initial coordinates; instead the Watson-Crick base-pairing becomes disrupted at increased temperature and pressure, assisted by the thermal motions of the layered double hydroxides sheets.

Of the three nucleic acids under investigation, DNA is observed to be the most stable when intercalated within an LDH, with respect to the percentage of Watson-Crick base-pairs. The opposite is found in bulk water where both PNA and RNA are more stable than DNA. The uncharged protein backbone of PNA has a detrimental effect on the overall stability of the polymer when intercalated as a duplex, as it experiences a greatly

reduced electrostatic interaction with the charged layered double hydroxides sheets. These results indicate that a mineral based origin of life may well have been rather different from the aqueous, bulk water based one more commonly considered in origins of life scenarios, DNA being the most stable genetic material within the simulations. The findings further suggest that a mineral mediated origin of life may have favoured DNA as the informational storage biomolecule over competing RNA and PNA, providing a route to modern biology from the RNA world.

Model	No. of atoms	Nucleic acid	No. of duplexes	Base sequences	Starting supercell dimensions (Å)	No. LDH sheets	Simulation time (ns)
I	12,650	RNA	1	d(5′-CUUUUGCAAAAG-3′)$_2$	49.0 × 56.5 × 56.5	3	25
II	12,650	DNA	1	d(5′-CTTTTGCAAAAG-3′)$_2$	49.0 × 56.5 × 56.5	3	25
III	12,650	PNA	1	d(5′-CTTTTGCAAAAG-3′)$_2$	49.0 × 56.5 × 56.5	3	25
IV	249,600	RNA	24	d(5′-CUUUUGCAAAAG-3′)$_2$	160.2 × 340.0× 129.9	6	30
V	249,600	DNA	24	d(5′-CTTTTGCAAAAG-3′)$_2$	160.2 × 340.0× 129.9	6	30
VI	250,800	PNA	24	d(5′-CTTTTGCAAAAG-3′)$_2$	160.2 × 340.0× 129.9	6	30
VII	218,863	RNA	24	d(5′-CUUUGCAAAAG-3′)$_2$	124.3 × 126.7 × 145.0	-	30
VIII	221,101	DNA	24	d(5′-CTTTTGCAAAAG-3′)$_2$	124.1 × 129.5 × 142.9	-	30
IX	230,159	PNA	24	d(5′-CTTTTGCAAAAG-3′)$_2$	125.5 × 128.0 × 148.9	-	30

Table 5.1: Simulation cell compositions and dimensions for the Models **I-VI** containing MgAl-LDH intercalated with 12 base-pair linear duplex RNA, DNA or PNA, together with chloride ions and water. The large clay Models **IV-VI** were built to remove any finite size effects, including the suppression of thermal undulations in the clay sheets. Models **VII**, **VIII** and **IX** are the nucleic acids solvated in water with counterbalancing sodium ions. Models **I-III** were studied at a range of water contents, while Models **IV-IX** were studied at a fixed level of hydration at a range of temperatures and pressures between 300 K, 1 atm and 500 K, 100 atm.
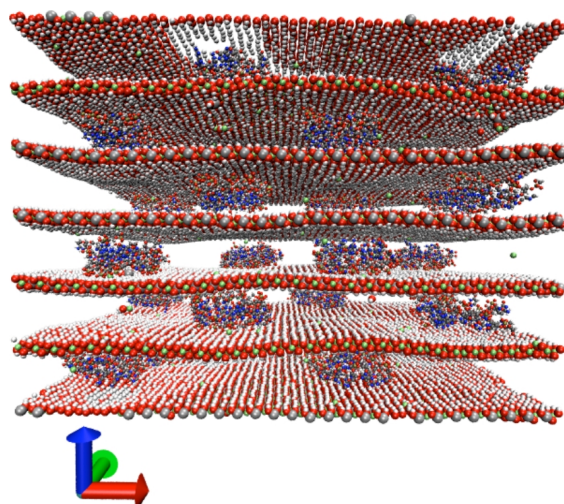
(a) **IV** RNA
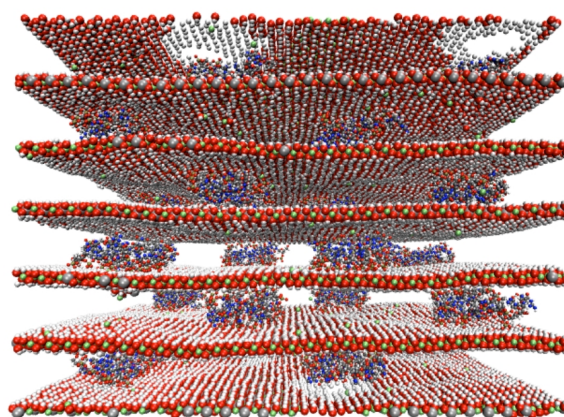
(b) **V** DNA
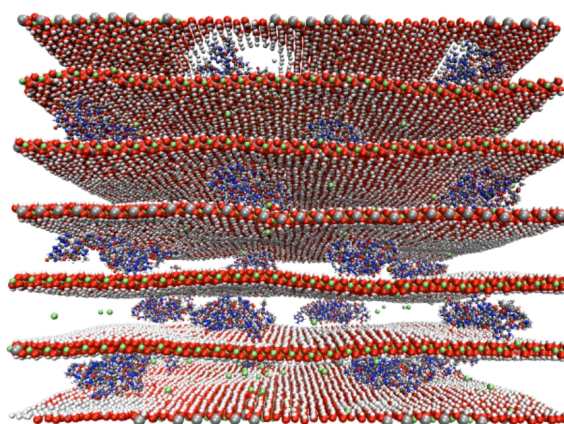
(c) **VI** PNA

Figure 5.5: Final structures of the large LDH-nucleic acid models: (a) Model **IV**, (b) Model **V** and (c) Model **VI** after the 30ns production phase of each simulation. The colouring scheme used is the same as that of Figure 5.3, with water and chloride ions not displayed to aid viewing. All nucleotide motion within the LDH sheets is significantly restricted compared to that in bulk water. In addition, visualisation reveals properties such as thermal undulations in the LDH sheet, as well as corrugation of the sheets around the nucleotides.

Property	Model **VII**	Experiment$^\alpha$	Simulation$^\beta$
Propeller (degrees)	-5.738	-4.688	-5.443
Roll (degrees)	3.103	6.881	2.800
Rise(Å)	2.991	3.349	2.600
Twist (degrees)	31.301	30.842	31.3
slide (Å)	2.391	1.990	2.3
Property	Model **VIII**	Experiment$^\gamma$	Simulation$^\delta$
Propeller (degrees)	-15.001	-14.824	-14.750
Roll (degrees)	2.166	2.278	1.955
Groove width (Å)	6.290	6.827	6.188
Rise(Å)	2.619	3.250	2.5
Twist (degrees)	33.102	32.948	35.8
slide (Å)	0.712	0.583	-1.0
Property	Model **IX**	Experiment$^\epsilon$	Simulation$^\zeta$
Propeller (degrees)	-10.876	-7.223	-
Roll (degrees)	2.665	3.772	0.588
Groove width (Å)	9.725	-	10.1
Rise(Å)	3.101	3.344	2.9

Table 5.2: Comparison of experimental and simulated parameters describing Watson-Crick base-pair geometry in different aqueous nucleic acid duplexes: RNA (model **VII**), DNA (model **VIII**) and PNA (model **IX**). α) Experimental values reported for RNA duplexes taken from the X-ray crystal structure at 2Å resolution [195]. β) Comparative simulation data for RNA duplexes taken from molecular dynamics using the AMBER forcefield [196]. γ) Experimental values reported for DNA duplexes from solution NMR [197]. δ) Simulation of DNA duplex using the AMBER forcefield [198]. ϵ) Experimental values for PNA from X-ray crystal structure at 1.70Å resolution [199]. ζ) Molecular dynamics study of a PNA duplex [200].
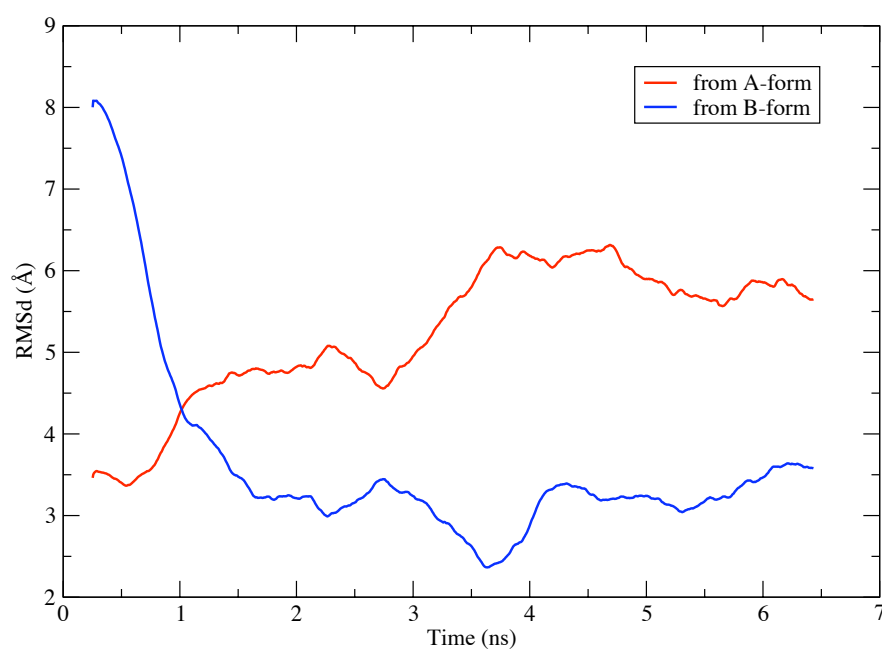
Figure 5.6: Root Mean-Squared deviation (RMSD) of the A-form of DNA in bulk water, with respect to time, from the crystal structure of the A-form (red) and of the B-form (blue). The two lines intersect at 1 ns; the interception shows that the DNA is deviates away from the A-form towards the B-form.
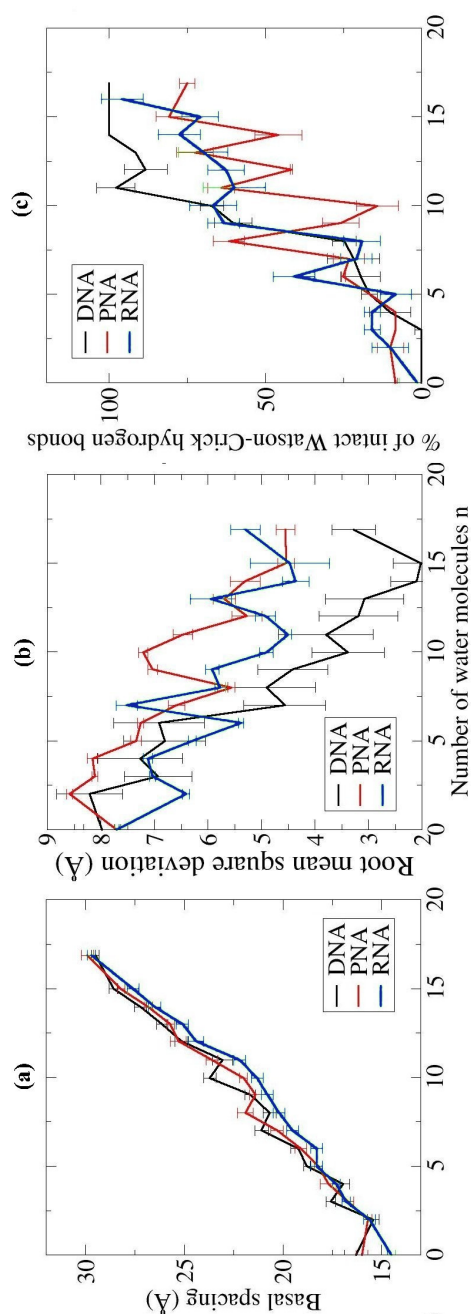
Figure 5.7: Comparison of the hydration properties of intercalated RNA duplex in Model **I**, DNA in Model **II** and PNA in Model **III**, at 300 K and 1 atm: (a) interlayer (basal) spacing, (b) root mean square deviation (RMSD) relative to the nucleic acid crystal structure and (c) percentage of Watson-Crick bonded base-pairs remaining. Error bars show the standard deviation for each value. From the change in basal spacing in (a), it appears that the intercalated RNA double strands are the most easily compressed, possibly owing to the different base-pair sequence resulting in weaker hydrogen-bonding between the strands compared to those of the DNA and PNA models. However, in general the intercalated PNA duplex strands have the largest RMSD values (b), indicating that the base-pairing in this model is significantly disrupted, with decreasing interlayer water, compared to the DNA and RNA duplexes, as confirmed by (c).

(a) RNA, n=14

(b) DNA, n=14

(c) PNA, n=14

(d) RNA, n=8

(e) DNA, n=8

(f) PNA, n=8

(g) RNA, n=0

(h) DNA, n=0

(i) PNA, n=0
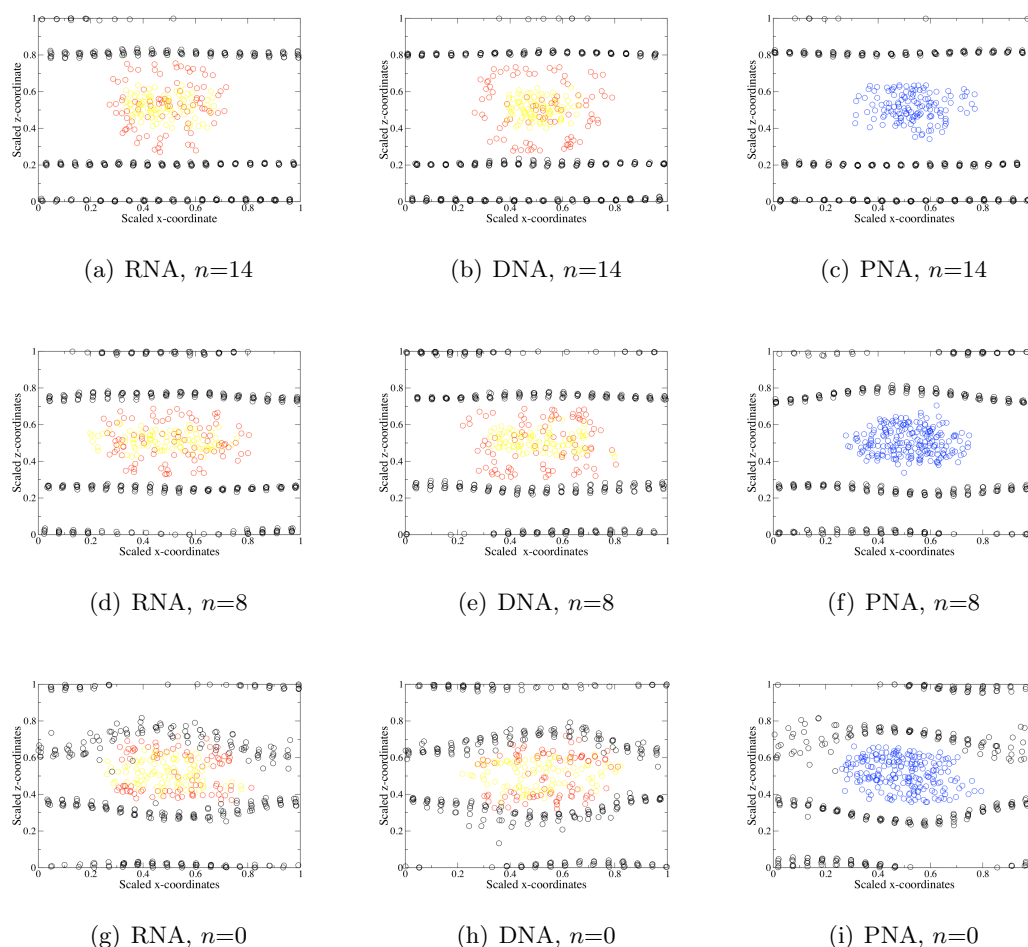
Figure 5.8: Cross section in the xz plane of Models **I – III** for various hydration states at 300 K and 1 atm, taken from the final snapshot of the simulation. Black circles correspond to positions of aluminium ions in the LDH sheets. Periodic boundaries cause some aluminium ions to reside at the top of the simulation cell. Red, yellow and blue circles correspond to phosphate groups, nitrogen atoms in the ribose sugar and the nitrogen atoms belonging to the PNA peptide backbone respectively. The variable n refers to the number of water molecules per unit formula $[Mg_2Al(OH)_6]\cdot nH_2O$. (a), (b) and (c) show RNA, DNA and PNA respectively, hydrated with n=16; (d), (e) and (f) show RNA, DNA and PNA respectively, hydrated with n=8 and (g), (h) and (i) show dehydrated RNA, DNA and PNA respectively. The cross section for the full hydrated PNA model, which is shown in subfigure (c), shows the PNA adopting a different conformation to the circular tube-like structure of RNA and DNA. This observed difference may be due to Cl^- ions assembling near to the LDH sheet surface.

Model	Temp. (K)	Press. (atm)	Mean RMSD (Å) per duplex	Mean % of intact Watson-Crick bonds per duplex	Max. basal spacing (Å)
IV	300K	1 atm	4.7 ± 0.9 Å	26.3 ± 1%	22.9 ± 0.03
(RNA-LDH)	350K	1 atm	6.2 ± 0.3 Å	21.2 ± 1%	22.7 ± 0.03
	400K	50 atm	6.7 ± 0.5 Å	17.1 ± 1%	22.8 ± 0.03
	450K	100 atm	7.0 ± 0.5 Å	13.7 ± 2%	24.0 ± 0.07
	500K	100 atm	8.1 ± 0.8 Å	5.1 ± 0.7%	27.2 ± 0.7
VII	300K	1 atm	2.5 ± 0.5 Å	89.7 ± 4%	-
(RNA in bulk	350K	1 atm	6.9 ± 1.5 Å	82.6 ± 7%	-
water)	400K	50 atm	9.3 ± 1.8 Å	68.2 ± 7%	-
	450K	100 atm	13.3± 0.5 Å	55.2 ± 5%	-
	500K	100 atm	11.2± 1.0 Å	14.7 ± 14%	-
V	300K	1 atm	4.3 ± 0.1 Å	62.1 ± 3%	19.2 ± 0.02
(DNA-LDH)	350K	1 atm	4.3 ± 0.1 Å	54.8 ± 6%	20.3 ± 0.02
	400K	50 atm	4.4 ± 0.1 Å	57.8 ± 4%	19.6 ± 0.02
	450K	100 atm	4.4 ± 0.1 Å	39.0 ± 2%	21.0 ± 0.02
	500K	100 atm	4.4 ± 0.1 Å	18.5 ± 8%	22.2 ± 0.03
VIII	300K	1 atm	4.8 ± 0.3 Å	82.0 ± 1.0 %	-
(DNA in bulk	350K	1 atm	5.5 ± 0.4 Å	74.1 ± 1.9 %	-
water)	400K	50 atm	4.8 ± 0.5 Å	40.5 ± 1.9 %	-
	450K	100 atm	5.7 ± 0.2 Å	25.1 ± 2.5 %	-
	500K	100 atm	4.9 ± 1.0 Å	14.7 ± 3.3 %	-
VI	300K	1 atm	6.9 ± 0.2 Å	23.3 ± 2.1%	23.0 ± 0.03
(PNA-LDH)	350K	1 atm	7.4 ± 0.2 Å	26.1 ± 2.2%	23.9 ± 0.3
	400K	50 atm	7.6 ± 0.2 Å	28.0 ± 2.1%	25.0 ± 0.04
	450K	100 atm	7.9 ± 0.2 Å	19.7 ± 1.5%	25.0 ± 0.05
	500K	100 atm	9.2 ± 0.3 Å	14.4 ± 3.9%	27.7 ± 0.10
IX	300K	1 atm	1.5 ± 0.4 Å	99.9 ± 0.4%	-
(PNA in bulk	350K	1 atm	6.3 ± 3.0 Å	99.9 ± 1.1%	-
water)	400K	50 atm	2.7 ± 0.5 Å	99.8 ± 1.2%	-
	450K	100 atm	9.9 ± 1.7 Å	49.2 ± 1.1%	-
	500K	100 atm	10.5± 1.3 Å	1 4.9 ± 4.3%	-

Table 5.3: Comparison of structural parameters for RNA, DNA and PNA duplexes intercalated in LDH and free in aqueous solution (Models **IV - IX**). The mean RMSD is calculated by comparing, from the initial structure, the coordinates of the nucleic acid over 30ns of MD with that of the original structure; the error is computed from the standard deviation. Of the three nucleic acids under investigation, DNA is observed to be the most stable when intercalated within an LDH, with respect to the percent of Watson-Crick base-pairs. The opposite is found in bulk water where both PNA and RNA are shown to be more stable than DNA. These trends suggest that clay minerals could have played a major role in mediating the transition from RNA to DNA as the main informational biomolecule.
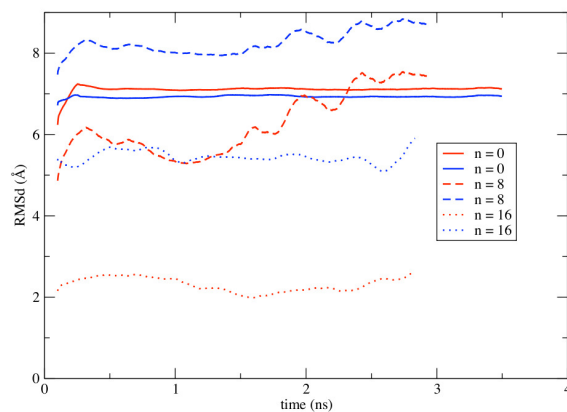
(a) model **I**, RNA
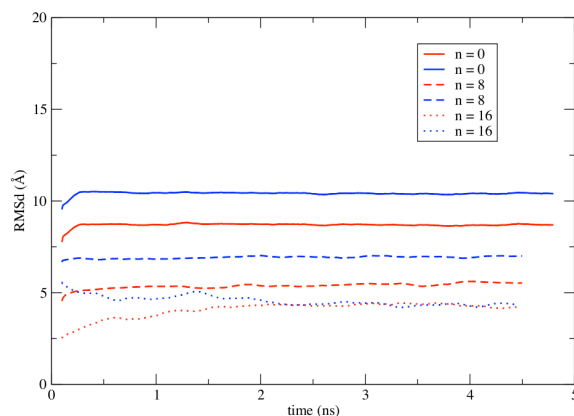
(b) model **II**, DNA
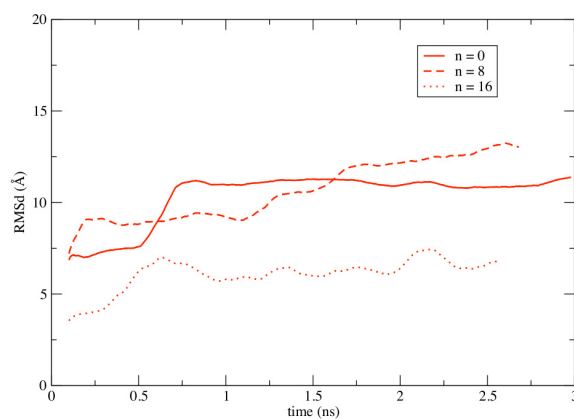
(c) model **III**, PNA

Figure 5.9: Root mean squared deviation (RMSD) of nucleic acids, RNA and DNA, intercalated within the interlayer of an LDH, from their respective crystal structures in the A-form (shown in Red) and in the B-form (shown in Blue), at various levels of hydration. Figure (c) shows the RMSD of PNA, intercalated within the interlayer of an LDH, from the initial P-form. The variable n refers to the number of water molecules per unit formula.
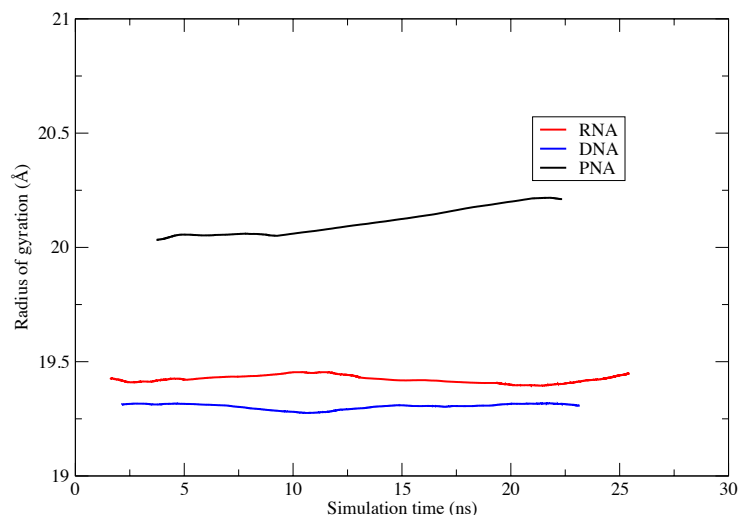
(a)
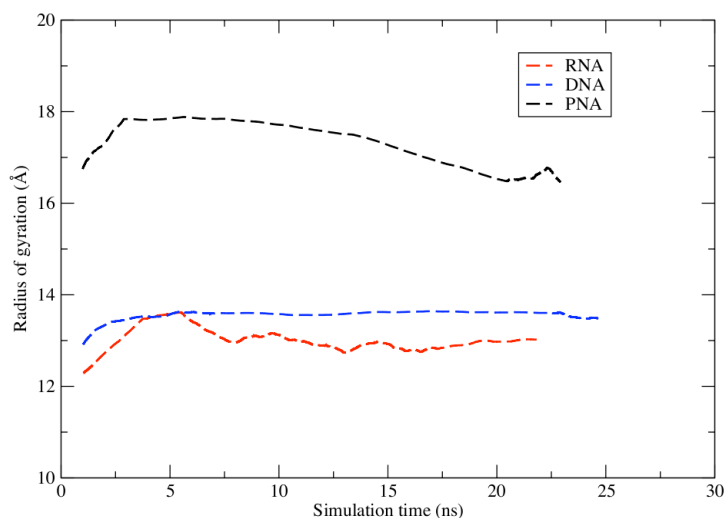
(b)

Figure 5.10: Radii of gyration of nucleic acids intercalated within an LDH for (a) Model **IV** (solid red), Model V (solid blue) and **VI** (solid black) at 300 K and 1 atm. Radii of gyration of nucleic acids in bulk water for (b) Model **VII** (dashed red), Model **VIII** (dashed blue) and **IX** (dashed black) at 300 K and 1 atm. Compared to RNA and DNA, the relative size of PNA is the largest both when free in bulk water and intercalated within the LDH. There is a change in the relative order of size between the three different nucleic acids at ambient conditions when intercalated within the LDH compared to the nucleic acids in bulk water; DNA has the smallest relative size when intercalated whilst RNA has the smallest relative size when free in bulk aqueous solution.
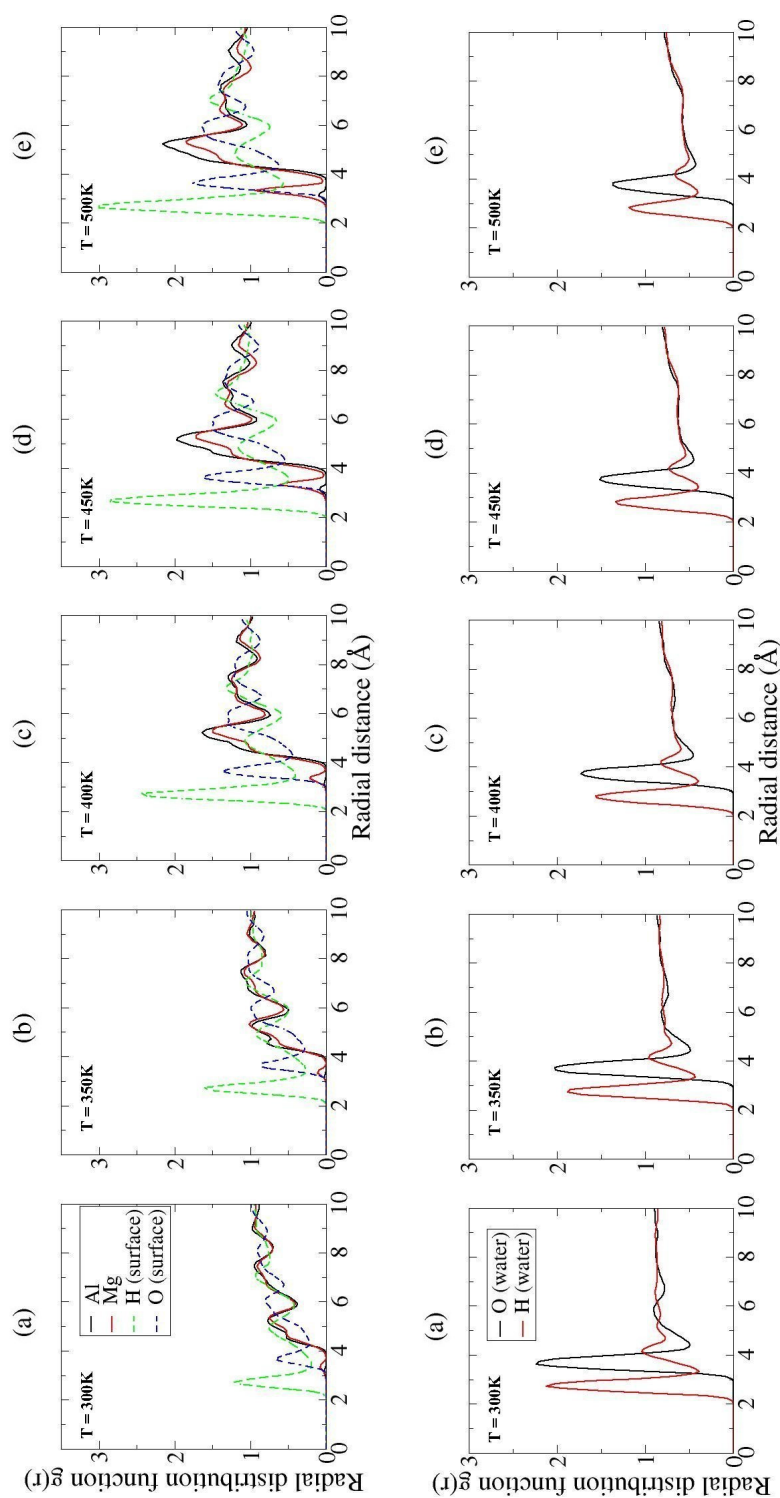
Figure 5.11: Average radial distribution functions for Model **IV**, which consists of RNA intercalated LDH, between phosphorus atoms in phosphate groups and LDH atoms (top row) and water molecules (bottom row), at conditions of (a) 300 K and 1 atm, (b) 350 K 1 atm, (c) 400 K 50 atm, (d) 450 K 100 atm and (e) 500 K and 100 atm. With increasing temperature, the peaks within the RDF for LDH atoms increase in amplitude, suggesting greater interaction of phosphate groups with the LDH surface. The RDF peak intensities decrease for water; presumably hydrogen bonding between phosphate groups and water molecules becomes increasingly disrupted at higher temperatures and pressures.
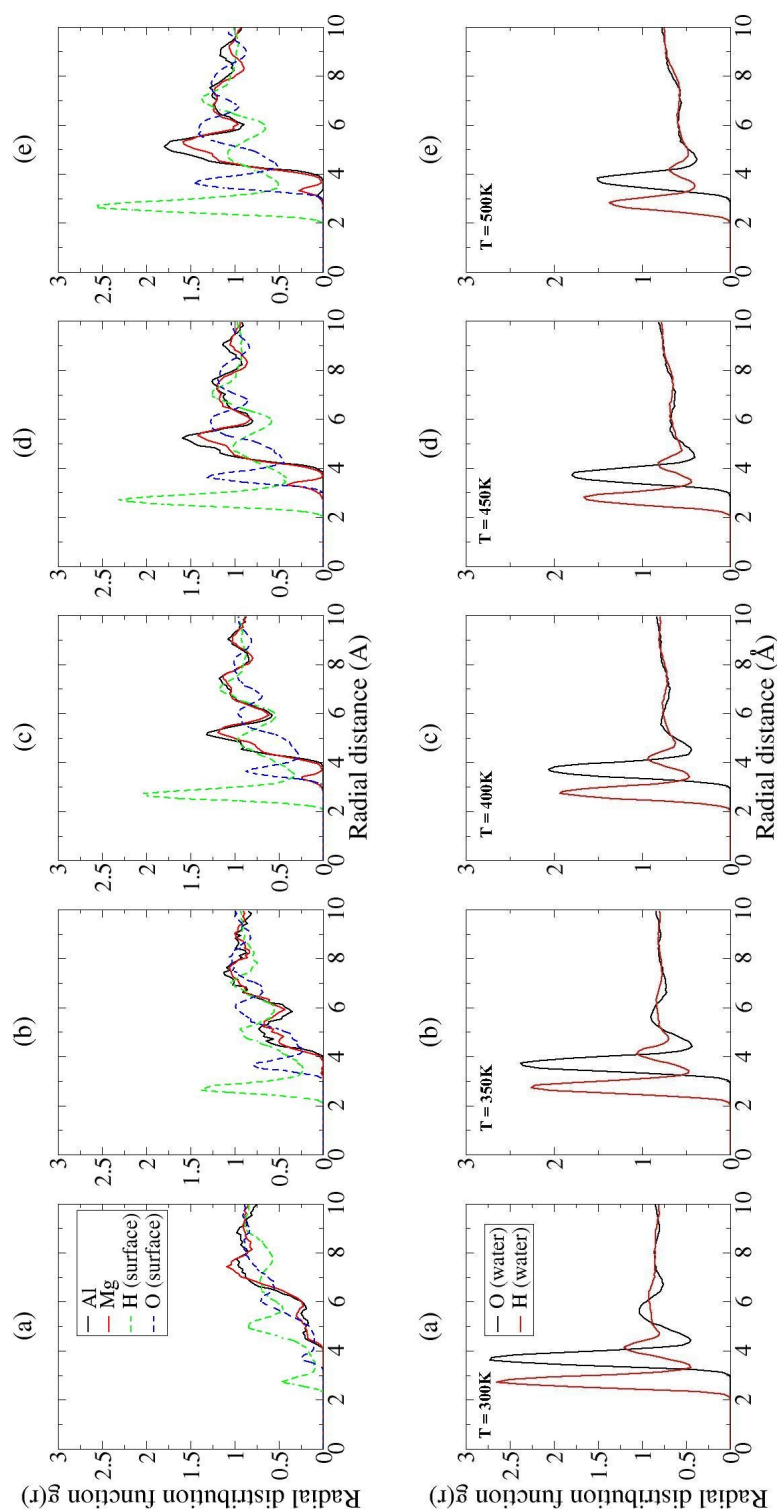
Figure 5.12: Average radial distribution functions for specified atoms (indicated in legend at top right of first panel in each row) in Model **VI**, which consists of PNA intercalated LDH, at (a) 300 K and 1 atm (b) 350 K and 1 atm (c) 400 K and 50 atm (d) 450 K and 100 atm (e) 500 K and 100 atm, calculated with respect to the phosphorus atoms in nucleic acid phosphate groups. With increasing temperature, the peaks within the RDF for LDH atoms increase, suggesting greater interaction of phosphate groups with the LDH surface. The RDF peak intensities decrease for water; presumably hydrogen bonding between phosphate groups and water molecules becomes increasingly disrupted at higher temperatures and pressures.
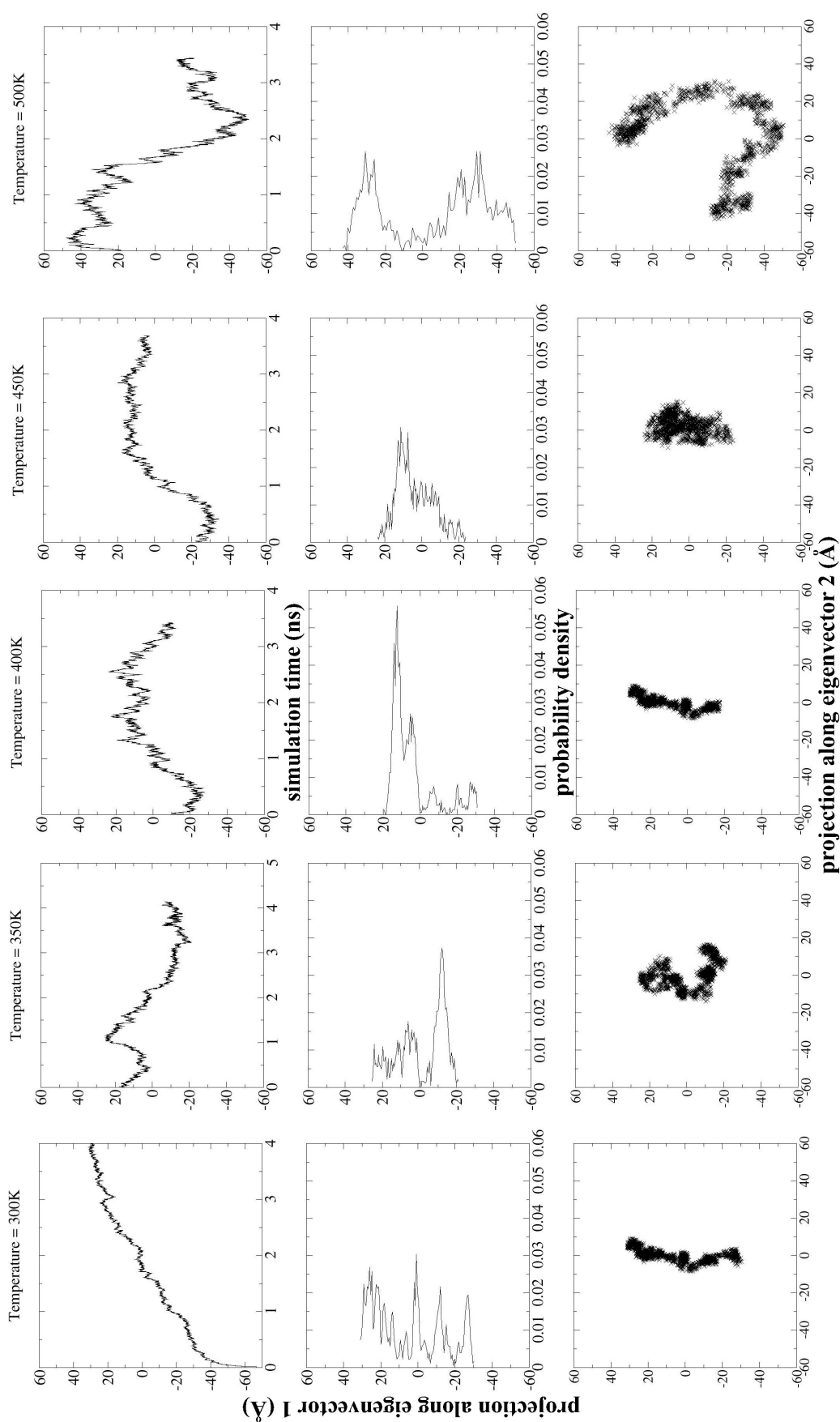
Figure 5.13: Comparison of principal component analysis for 12 base-pair PNA strands in LDH (Model **VI**) as a function of temperature. The layout is identical to that used in Figure 5.16. The first row displays the motion of heavy atoms along the first eigenvector, while the second row plots the probability density that the model occupies a particular configuration. The lowermost row shows the projections of the trajectory on planes defined by pairs of eigenvectors 1 and 2. Unlike the RNA in Models **IV** and **VII**, shown in Figure 5.16 and Figure 5.17, the PNA model has more unimodal probability distribution up to 400 K and 50 atm, suggesting that thermal undulations of the LDH sheets do not affect the intercalated PNA as much as RNA and DNA, presumably because the peptide chains only interact weakly with the LDH sheets through van der Waals forces.
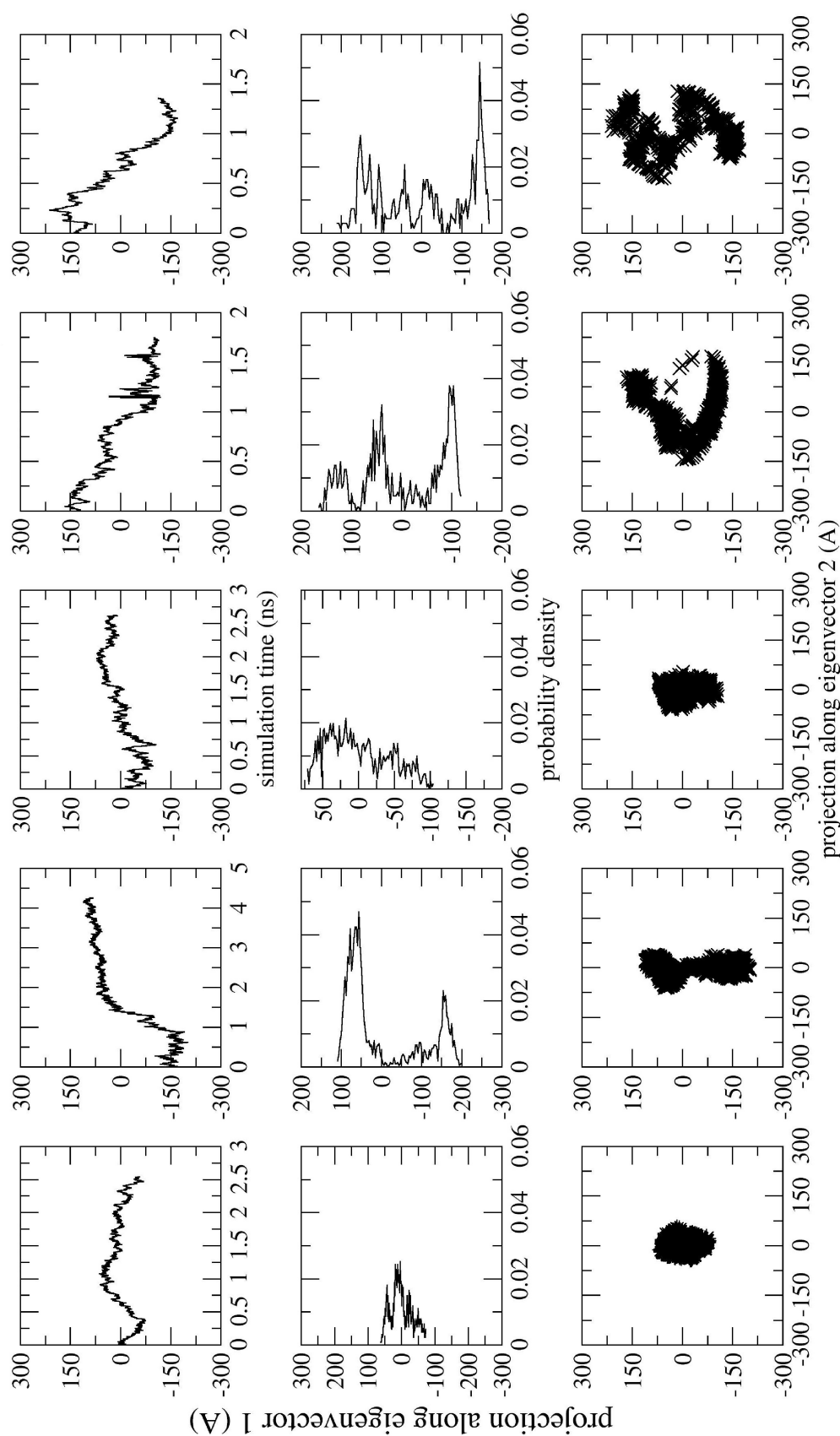
Figure 5.14: Comparison of principal component analysis for 12 base-pair PNA strands in bulk water (Model **IX**) as a function of temperature. The layout is identical to that used in Figure 5.17. The first row displays the motion of heavy atoms along the first eigenvector, while the second row plots the probability density that the model occupies a particular configuration. The lowermost row shows the projections of the trajectory on planes defined by pairs of eigenvectors 1 and 2.

(a) RNA-LDH　　　　(b) DNA-LDH　　　　(c) PNA-LDH

(d) RNA-Water　　　　(e) DNA-Water　　　　(f) PNA-Water
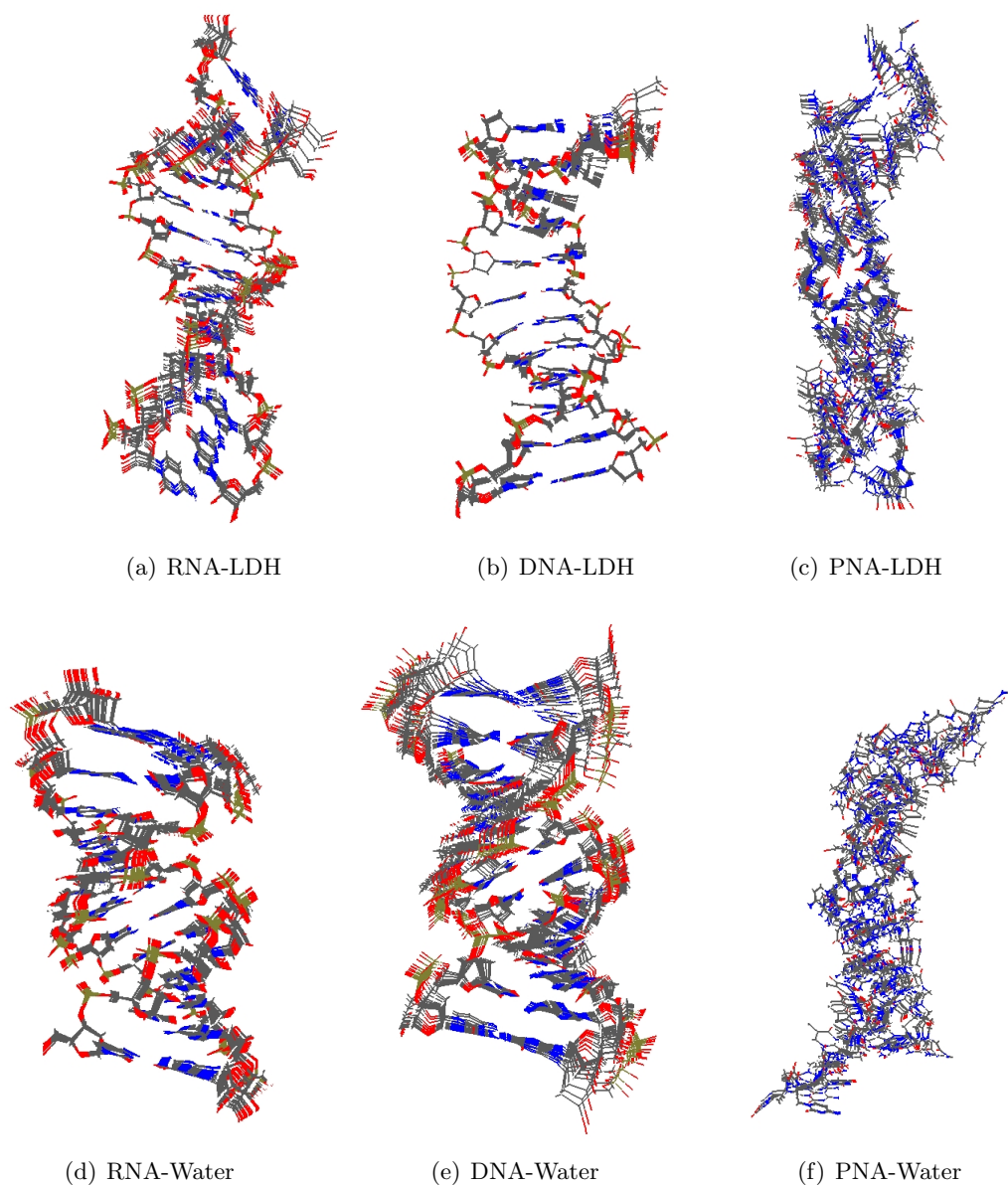
Figure 5.15: Superposition of configurations obtained by projecting the motion of all atoms onto the first eigenvector using principal component analysis for (a) Model **I**; (b) Model **II**; (c) Model **III**; (d) Model **VII**; (e) Model **VIII**; and (f) Model **IX**. The configurations are averaged over the last nanosecond of simulation for the studied models. The colour scheme is the same as that used in Figure 5.3. Intercalated nucleic acids show reduced motion compared to the corresponding bulk water models for the dominant mode of motion. DNA displays less motion, with respect to the first eigenvector, than all other models indicating that this is the most structurally stable of the models tested.
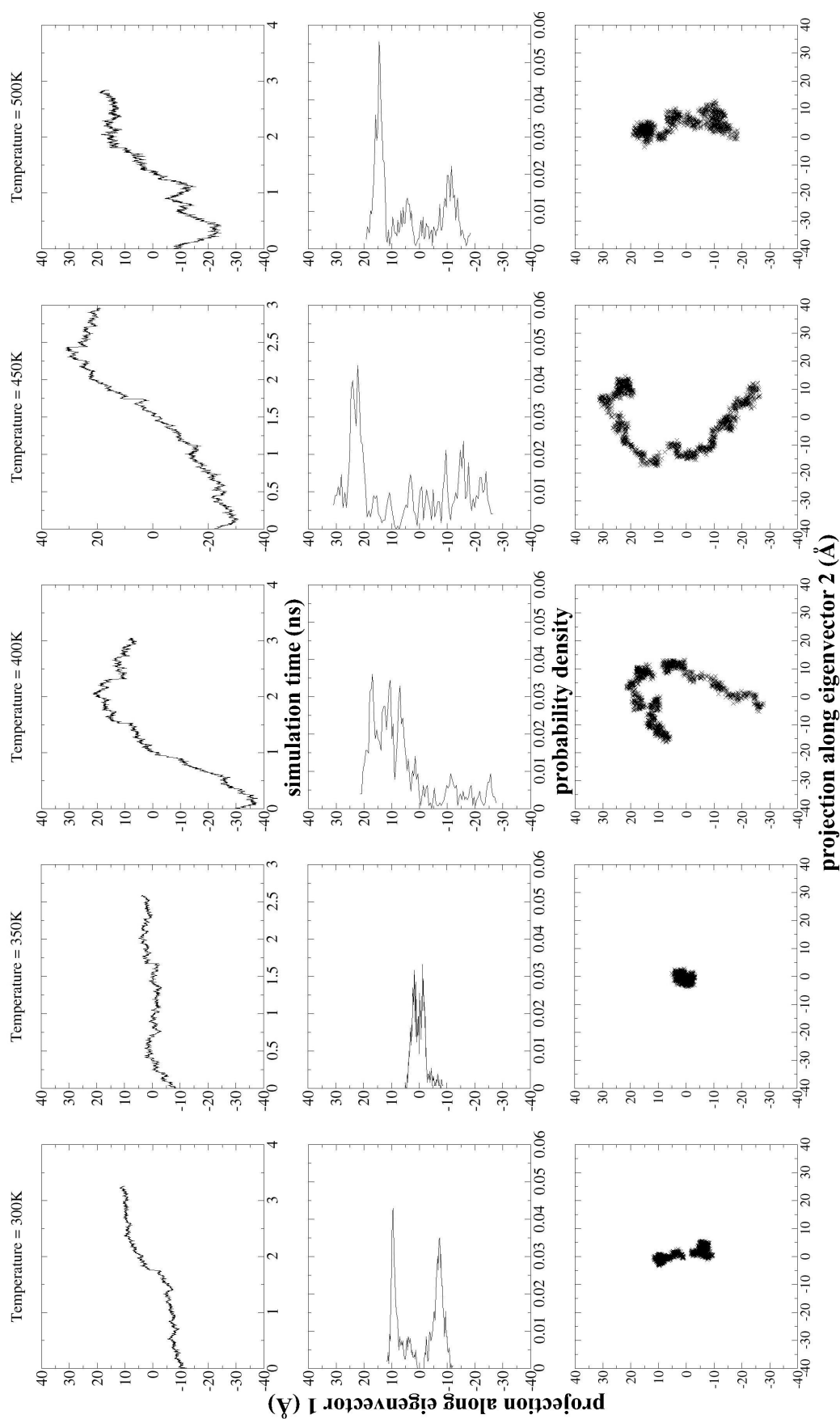
Figure 5.16: Comparison of principal component analysis for 12 base-pair RNA strands in an LDH corresponding to Model **IV** at the temperatures indicated. The first row displays the motion of heavy atoms along the first eigenvector, while the second row exhibits the probability that the model occupies a particular configuration. The lowermost row shows the projections of the trajectory on planes defined by pairs of eigenvectors 1 and 2. Comparison of the axes scales shows that RNA intercalated within LDH has almost 10 times less degrees of freedom than RNA in bulk water (See Figure 5.17). With increasing temperatures and pressures the strands have much larger amplitudes in their structural fluctuations; for example consider the effect of heavy atom motion along the first eigenvector. The plots of the subspace occupied by the RNA atoms also appear more coupled when the RNA strands are intercalated in LDH compared to those in bulk water. This is evidence that the strands are influenced by the movement of the LDH sheets, as previously shown [137].
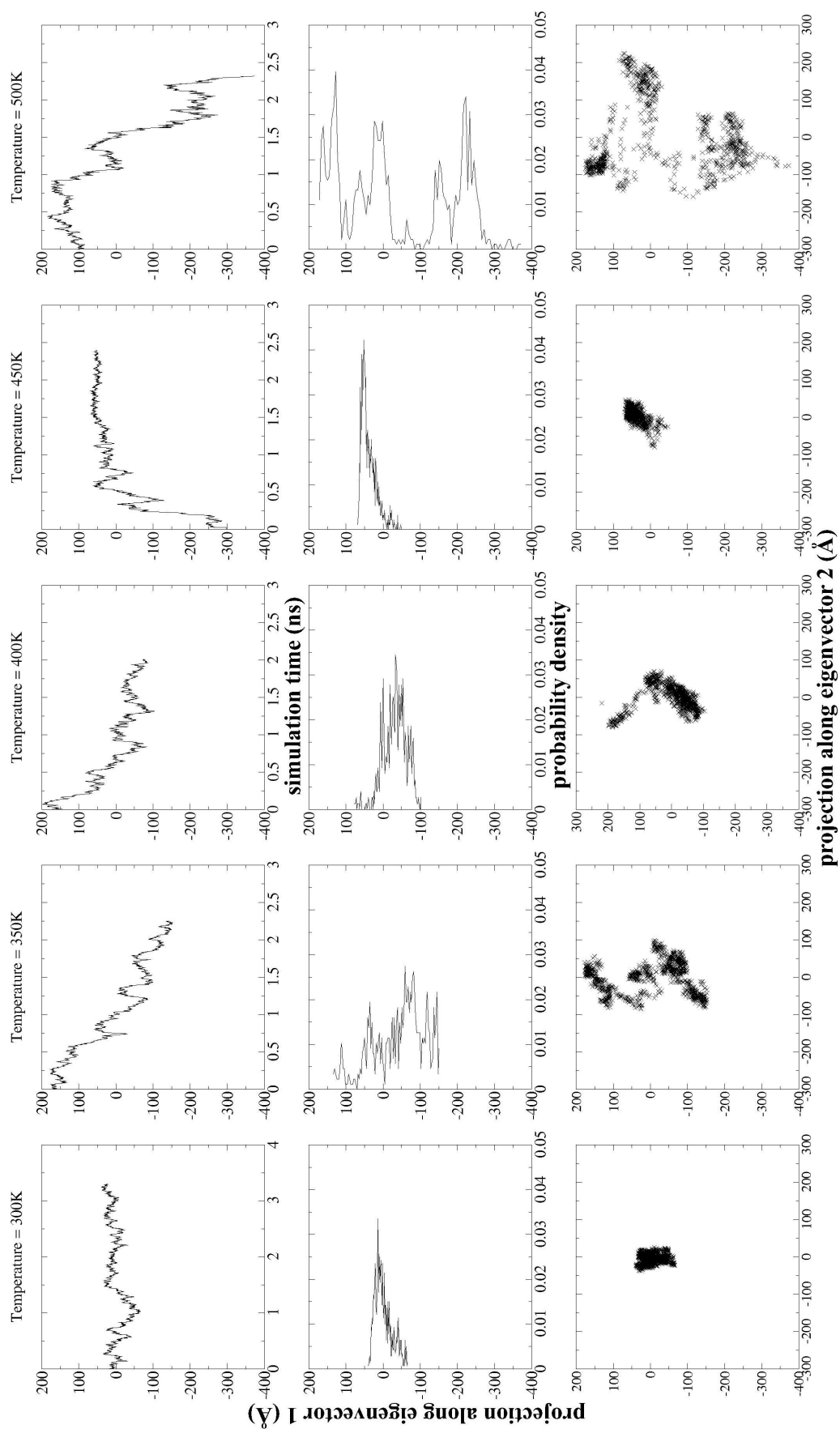
Figure 5.17: Comparison of principal component analysis for 12 base-pair RNA strands in bulk water corresponding to Model **VII** at the temperatures indicated.

Temperature	PNA Diffusion Coefficient ($\mathring{A}^2 ns^{-1}$)	RNA Diffusion Coefficient ($\mathring{A}^2 ns^{-1}$)	DNA Diffusion Coefficient ($\mathring{A}^2 ns^{-1}$)
300K	0.20 ± 0.000	0.33 ± 0.00	0.28 ± 0.01
350K	0.90 ± 0.001	0.48 ± 0.00	0.55 ± 0.02
400K	1.40 ± 0.002	0.83 ± 0.00	0.58 ± 0.03
450K	2.60 ± 0.002	1.45 ± 0.00	1.80 ± 0.09
500K	4.00 ± 0.004	2.34 ± 0.01	2.11 ± 0.12

Table 5.4: Diffusion coefficients for 12 base pair PNA oligomers in Model **VI**, RNA oligomers in Model **IV** and DNA oligomers in Model **V** calculated from the production phase of the simulations. Error bars are obtained from the least squares error fit on the gradient of the mean square displacement *vs.* time graph. As the temperature increases, the average diffusion within the basal plane increases.

Chapter 6

Clay Minerals Mediate Collapse and Regioselective Interactions of RNA

In this chapter molecular dynamics techniques are used to carry out large-scale simulations of various 25 mer sequences of ribonucleic acid (RNA), in bulk water and with aqueous montmorillonite clay over many tens of nanoseconds. The motivation of this was to elucidate the nucleic acid–clay interactions that have been observed in previous experimental studies, but have hitherto remained unexplained at a molecular level.

6.1   Introduction

Montmorillonite clay is a principal alteration product of volcanic ash, and it is relatively widespread within bentonite deposits in the Earths crust. It is postulated that similar processes would also have occurred on the early Earth [206], with interpetation of the oxygen isotopes of Hadean zicrons indicating the oceanic conditions needed for sodium montmorillonite at up to 4.4 billion years ago [207, 208]. Under laboratory conditions, montmorillonite has been shown to catalyse the formation of long ribonucleic acid (RNA) oligomers, along with a large number of other organic reactions [209]. Montmorillonite is a 2:1 clay consisting of an octahedral alumina (AlO_6) sheet sandwiched between two tetrahedral silica (SiO_4) sheets. Naturally occurring montmorillonite, such as Wyoming montmorillonite, contains partial isomorphic substitutions in both tetrahedral and octahedral sheets, which confer a net negative charge. Silicon (4+) ions in the tetrahedral sheets are commonly substituted with aluminum (3+) ions, and octahedrally coordinated aluminum (3+) by magnesium (2+), and iron (2+) ions. The negative charge

is balanced by various exchangeable cations, often sodium (Na^+) and calcium (Ca^{2+}) ions, which reside on the surfaces and within the interlayer region of the clay. Work performed by Ferris *et al.* has shown that of twenty-two chemically distinct varieties of natural montmorillonite, the Wyoming type has greatest catalytic activity [162].

The ability of montmorillonite to catalyse the formation of RNA oligomers has been demonstrated in a number of experiments by Ferris and co-workers [71]. In their laboratory studies Ferris *et al.* used artificial activated phosphorimidazolides of nucleosides together with homoionic Na^+ montmorillonite to form oligomers of 6-14 mers, in which montmorillonite was shown to enhance the rate constant for oligomer formation by 100-1000 times over the hydrolysis of the imidazole activating group [210, 211]. Ferris *et al.* then experimented with the use of a primer strand of 10 mers which produced 30-40 mers [212]. By changing the activating group from imidazole to 1-methyladenine, Ferris *et al.* formed single strands of up to 50 mers of poly-adenine and poly-uracil in a couple of days, without the need for a primer strand. RNA sequences of this length are sufficiently long to exhibit fidelity in replication as well as catalytic activity [71]. How such polymerisation could have occurred in the pre-biotic era remains to be fully understood.

As each montmorillonite clay sheet carries a net negative charge and each RNA strand has negatively charged phosphate groups, even the basic mechanism by which the RNA interacts with clays needs clarification. Franchi *et al.* have proposed that cations play a key role in the adsorption of the strands on clay surfaces [213]. The cations could potentially mediate the interaction by screening the negative charge using monovalent cations, or by creating a divalent cation bridge between the two layers of negative charge from the nucleic acid and mineral surface. These authors have demonstrated that the quantity of single stranded adenine and uracil RNA (poly[A] and poly[U]) adsorbed is greater in the presence of Ca^{2+} than Na^+ ions.

One of the leading theories concerning the origin of life is the RNA world hypothesis, based on the assumption that an RNA based world preceded the world we live in now, in which according to the central dogma of modern biology DNA makes RNA which in turn makes proteins [214]. The hypothesis is attractive in part because it only requires the formation of one type of biopolymer. The RNA world hypothesis has gained significant support from the work of Cech and Altman, who won the 1989 Nobel prize in chemistry for the discovery of a group of catalytic RNA molecules called ribozymes [25]. Steitz, along with many others, elucidated the structure of the ribosome, which is the molecular machine responsible for making proteins and amino acids [26, 215], showing that the ribosome is a ribozyme. More recent support has come from Sutherland *et al.* who synthesised purine based RNA [101]. Montmorillonite catalysis could potentially resolve some of the problems in RNA synthesis, by concentrating and polymerising the RNA

and producing stereo-enantiomers [216]. Much experimental work has been performed on the interaction of nucleic acids (both single and double stranded RNA and DNA) interacting with mica surfaces [217, 218]. Although the mica/nucleic acid system is not identical to that of montmorillonite/nucleotide systems, the mica surface can be thought of as an analogous system as both mica and montmorillonite have negatively charged surfaces. For many years montmorillonite has been associated in part with the RNA world hypothesis, and therefore the origins of life on Earth. Indeed, the idea that clay minerals played an important role in the early chemical processes leading up to the origin of life was proposed independently by Goldschmidt [219] and Bernal [220] as long ago as 1947.

Even today, however, the structure and conformation of RNA interacting with montmorillonite remains uncertain, as well as the mechanism of elongation of the polymer by addition of nucleic acid monomers. Research into origins of life has hitherto rarely used simulation techniques to understand the possible chemical pathways to the formation of biomolecules [70, 191, 221]. However, over the past ten to fifteen years, molecular simulation has played an increasingly important role in clay chemistry [137, 145], as well as in the study of RNA [222]. Computer simulations complement current experimental techniques by providing atomic-scale resolution of such systems. In particular, simulation provides insight into how molecules adsorb on and intercalate within clays.

This chapter explores the mechanism by which RNA adsorbs on an aqueous Wyoming montmorillonite mineral surface, using fully atomistic large scale classical molecular dynamics techiques. The interaction of various RNA molecules, of differing base sequence, with the mineral surface in the presence of Ca^{2+} or Na^+ charge balancing cations is investigated. How the differing counterions alter the RNA conformation, folding kinetics and structure is studied, and a comparison to its behavior on montmorillonite to that in aqueous solution is observed. The chapter looks at how RNA changes its structure and conformation when interacting with a montmorillonite surface and contrast the structure and conformation in bulk water.

Section 6.2 of this chapter describes techniques that we have employed to simulate the clay/nucleic acid systems reported here. Section 6.3 discusses the results obtained in these simulations, while Section 6.4 draws conclusions from this study.

6.2 Methods

This section discusses the techniques that have been employed to simulate the aqueous nucleic acids and clay systems in this paper. Fully atomistic, large scale molecular dynamics techniques have been used to simulate these fully flexible systems containing hundreds of thousands of atoms. A collection of three federated supercomputing grids coupled to local resources at UCL were utilized in order to perform these simulations, and their associated data analysis and transfer.

Model	RNA	Sequence	Cation	Number of Atoms	Clay Surface	Starting Cell Dimensions (Å³)	Simulation time (ns)
I	A	(5′-AAAAAAAAAAAAAAAAAAAAAAAAA-3′)	Na^+	96,025	Yes	82.74 × 143.28 × 180.62	33.866
Ia		(5′-AAAAAAAAAAAAAAAAAAAAAAAAA-3′)	Na^+	96,025	Yes	82.74 × 143.28 × 180.62	23.116
II		(5′-AAAAAAAAAAAAAAAAAAAAAAAAA-3′)	Ca^{2+}	95,821	Yes	82.74 × 143.28 × 180.62	34.008
III		(5′-AAAAAAAAAAAAAAAAAAAAAAAAA-3′)	Na^+	267,062	No	125.52 × 177.26 × 125.68	22.588
IV		(5′-AAAAAAAAAAAAAAAAAAAAAAAAA-3′)	Ca^{2+}	267,050	No	125.52 × 177.26 × 125.68	22.840
V	B	(5′-AAAGCAGAGUAAAGUCUCUGAGUAG-3′)	Na^+	96,011	Yes	82.74 × 143.28 × 180.62	25.240
VI		(5′-AAAGCAGAGUAAAGUCUCUGAGUAG-3′)	Ca^{2+}	95,807	Yes	82.74 × 143.28 × 180.62	31.340
VII		(5′-AAAGCAGAGUAAAGUCUCUGAGUAG-3′)	Na^+	261,084	No	125.52 × 177.26 × 125.68	20.786
VIII		(5′-AAAGCAGAGUAAAGUCUCUGAGUAG-3′)	Ca^{2+}	263,835	No	125.52 × 177.26 × 125.68	20.112
IX	C	(5′-AAAGUCUGGGCUAAGCCCACUGAUG-3′)	Na^+	96,011	Yes	82.74 × 143.28 × 180.62	25.886
X		(5′-AAAGUCUGGGCUAAGCCCACUGAUG-3′)	Ca^{2+}	96,011	Yes	82.74 × 143.28 × 180.62	31.484
XI		(5′-AAAGUCUGGGCUAAGCCCACUGAUG-3′)	Na^+	261,084	No	125.52 × 177.26 × 125.68	20.470
XII		(5′-AAAGUCUGGGCUAAGCCCACUGAUG-3′)	Ca^{2+}	261,084	No	125.52 × 177.26 × 125.68	20.866
XIII	dA	d(5′-AAAAAAAAAAAAAAAAAAAAAAAAA-3′)	Na^+	96,798	Yes	82.74 × 143.28 × 180.62	24.792

Table 6.1: Simulation cell compositions and dimensions for all the systems studied in this chapter. Three single-stranded RNA sequences were simulated, sequence **A** in models **I-IV**, sequence **B** in models **V-VIII** and sequence **C** in models **IX-XII**. Each sequence was simulated in the presence of the montmorillonite surface and in aqueous solution, and simulated with Na^+ ions or Ca^{2+} ions as a counter cation. Sequence **A** is a 25 mer poly-adenine RNA. Sequences **B** and **C** are both 25 mer strands with mixed base sequences taken from the hammerhead ribozyme which exhibits secondary structural features. A further model **XIII** was built consisting of a double strand of sequence **A**, denoted **dA**, on an aqueous montmorillonite surface. Model **Ia** is identical to model **I** except for the initial distribution of the charge balancing Na^+ ions on the surface of the clay, and was used to investigate the effects of changing the initial configuration of Na^+ ions on the ensuing structure and conformation of the RNA.

The montmorillonite used in this study was derived from an initial structure with the chemical formula $Na_3[Al_{14}Mg_2][Si_{31}Al]O_{80}(OH)_{16}$ shown in Figure 7.2, which is consistent with naturally occurring Wyoming montmorillonite. Partial substitutions were created in both tetrahedral and octahedral sheets, although there is no evidence that the negative charge arising from partial substitution of Al^{3+} ions for Mg^{2+} ions in the octahedral sheet are implicated in the bonding between RNA and montmorillonite. The montmorillonite structure was constructed using the Materials Studio (Accelrys) software. The montmorillonite structure was chosen as it is widely used in experiments [213, 223] and it has been shown to possess catalytic activity with respect to RNA oligomerisation [162].

The sizes and sequences of the RNA molecules used in this study are given in Table 6.1 and shown in Figure 7.2. All the polymers in models **I-XII** are 25 base groups in length and are single-stranded. All the RNA molecules, denoted **A, B, C** and **dA**, were assembled using the Nucleic Acid Builder which is part of AMBER Tools 1.2. Models **I** and **II** consist of a single-stranded poly-adenine RNA molecule, sequence **A**, aligned with the aqueous montmorillonite surface with differing cations neutralizing the negative charge from the partial substitutions in the clay and charged un-protonated phosphate groups. From the partial substitutions within the clay sheets, as shown in the clay chemical formula $Na_3[Al_{14}Mg_2][Si_{31}Al]O_{80}(OH)_{16}$, a net negative charge of 408 e^- was generated on the clay structure. The RNA strands used in this study all bear a net negative charge of 24 e^-. As such, a corresponding number of Na^+ or Ca^{2+} ions were added in order to neutralise the negative charge in all the systems. All phosphate groups were un-protonated and carrying a negative charge as we are assuming a pH of 7 where all phosphate groups are known to be ionised. This sequence was chosen to mimic the strand used in the experiments performed by Franchi *et al.* [213]. Models **III** and **IV** are the corresponding RNA in bulk aqueous solution (*i.e.* in the absence of montmorillonite); these are used to infer the effects of the clay surface on the structure, dynamics and conformation of the RNA molecule. In models **V** to **XII** two different single stranded RNA sequences were used (**B** and **C**). The base sequences were taken from two stem-loop secondary structural motifs in the type III hammerhead ribozyme [224]. The stem-loop motifs were chosen to observe the effects of the clay surface on the formation and stabilisation of secondary structures. All sequences were also simulated using different cations and in bulk aqueous solution. The two cations used were monovalent Na^+ ions and divalent Ca^{2+} ions, in order to investigate the effects of cation charge density on RNA conformation and on mediating surface interactions; they were selected to mimic those used in experiments [213]. Model **Ia** was constructed in order to test the effect of the initial cation distribution on the simulation; in this model the cations were distributed randomly around sequence **A** using a stochastic

algorithm. Cations in other models were entered manually into random positions within the simulation cell. Model **XIII** was simulated in order to provide a comparison with the double stranded DNA and RNA experiments preformed on a mica surface.

The sizes and sequences of the RNA molecules used in this study are given in Table 6.1 and shown in Figure 7.2. Table 6.1 uses standard notation for displaying RNA sequences, where the strand is displayed in the $5'$-$3'$ direction exhibiting individual nucleotide bases as A, U, C & G for adenine, uracil, cytosine and guanine respectively. All the RNA molecules were built using the Nucleic Acid Builder, which produces nucleic acid structures in a hierarchical fashion based on three successively applied steps: i) transformation of nucleotide bases to achieve desired helical and base-pairing configurations, ii) geometry optimisation which allows molecular structures to be built that satisfy sets of distance constraints and iii) potential energy minimisation. Models **I** and **II** consisted of a single-stranded poly-adenine RNA molecule aligned with the aqueous montmorillonite surface with differing cations neutralising the negative charge from the partial substitutions in the clay and charged un-protonated phosphate groups. This sequence was chosen to be identical to that used in the experiments performed by Franchi *et al.* [213]. Models **III** and **IV** are the corresponding RNA in bulk aqueous solution (*i.e.* in the absence of montmorillonite); these are used to infer the effects of the clay surface on the structure and conformation of the RNA molecules.

The two stem-loop motifs in models **V-VII** were chosen to study the effects of the clay surface on the formation and stabilisation of secondary structures. These sequences were also simulated using different cations and in bulk aqueous solution. Distances of 20 Å in the clay models and 50 Å in the aqueous models were used to minimise the interaction between RNA and its periodic images. Over 25,000 water molecules were placed above the surface of the montmorillonite clay within the periodic simulation cell, giving it a length of 180 Å, sufficient to decouple the clay layer from its periodic image in the *z*-direction. The initial configuration of the RNA and the clay surface is shown in Figure 6.2. Charge balancing cations were placed in the interlayer region of the clay, and also between the RNA molecule and the surface manually. In the bulk water models cations were placed at random locations around the molecule using the 'solvate' and 'addions' routines within AMBER Tools.

The models were minimised using the steepest-descent method. A thermalisation step was run, heating the system up from 50 K to 300 K over 0.01 ns. The whole system was allowed to evolve and equilibrate using the NPT ensemble. Equilibration was deemed to have been established by monitoring the potential energy of the system as a function of time, the radius of gyration of the RNA molecule and by looking for the cell dimensions to take on stable average values. Fourteen models were simulated in total. All simulations

were run on the geographically distributed computational resources. Clay models were run on 512 processing cores and the bulk water models on 1024 cores. The number of cores was selected based on the number of atoms of each model and the scaling properties of the LAMMPS code. Each model was simulated in production conditions for the specified number of nanoseconds following equilibration, shown in Table 6.1. Models **I** and **II** were simulated for 30 ns, where it was shown that equilibrium had been established for over 20 ns; as a consequence all other simulations were ran for at least 20 ns.

In order to perform the study reported here, very substantial computing resources were needed. To this end, three supercomputing grid infrastructures were utilized in the USA and Europe. These include computing nodes at Leeds and Rutherford Appleton Laboratories on the UK's National Grid Service (NGS) [1]; Bigben at Pittsburgh Supercomputing Center (PSC), Ranger at the Texas Advanced Computing Center (TACC) and Kraken at the National Institute for Computer Science (NICS), all on the US TeraGrid [2]. RZG at Rechenzentrum Garching on the Distributed European Infrastructure for Supercomputing Applications (DEISA) [3] and Intrepid at the Argonne Leadership Computing Facility, were also employed to run these simulations, as well as local machines at UCL, including the Centre for Computational Science's Linux cluster, Mavrino, and UCL Research Computing's Legion machine.

6.3 Results

This section discusses the various differences in the interaction of RNA with an aqueous clay surface in the presence of monovalent or divalent cations, and discuss how the different cations mediate the interaction between the clay and RNA. The observed collapse and formation of secondary structural motifs in the various RNA sequences is reported. The mechanism by which RNA tethers to the montmorillonite clay surface, and account for the observed increase in adsorption of RNA on the clay surface in the presence of divalent cations is also discussed.

6.3.1 Cations as mediators of the adsorption of RNA on clay surfaces

Model **Ia** was built in order to test the effects of the initial structure and distribution of the counterions on the evolution of the MD simulation. Model **Ia** is identical to model **I** in every aspect except for the initial location of the Na^+ counterions. A comparison of

---

[1] http://www.ngs.ac.uk/
[2] http://www.teragrid.org/
[3] http://www.deisa.eu/

the first 23 ns of simulation from models **I** and **Ia**, as well as average structures, showed that once equilibrium had been reached both models exhibited more or less identical RNA structure and conformation and counterion arrangement; this is supported by the similarities in the density profiles produced for each system. Insignificant differences in models **I** and **Ia** can be attributed to the random nature of MD simulations and did not effect the time averaged data of the two systems.

RNA-montmorillonite systems were simulated with two different choices of cation to balance the negative charge due to the unprotonated phosphate groups present in the RNA molecule and from the partial substitutions in the clay sheets. The RNA-clay models were simulated with Na^+ and Ca^{2+} ions for the RNA sequences **A**, **B** and **C**, as described in Table 6.1. Previous experimental studies carried out on nucleic acid-mineral systems [213, 217] have shown that the cations play a key role in mediating the interaction between RNA and clay surfaces. The role cations play in mediating the adsorption of nucleic acids on clay surfaces was reported by Franchi *et al.* using adsorption assay experiments [213]. Franchi and co-workers reported that divalent cations (Ca^{2+} and Mg^{2+}) are more efficient than monovalent cations (Na^+) in mediating the adsorption of nucleic acids on clay minerals including montmorillonite and kaolinite. This effect can be explained by the higher affinity of divalent cations for nucleic acid molecules and their greater ability to counter-balance the negative charges present in the nucleic acid-clay system. Visualisation of the average structure taken after 20 ns, shown in Figure 6.4, displays the extent of the differences caused by the different cations. The RNA molecule interacts very strongly with the surface when the divalent Ca^{2+} ions are present in models **II, V** and **X**, compared to the RNA strand in models **I, IV** and **IX** when monovalent ions are present. Despite the like-charge of the RNA and clay surface, the RNA polyanion is attracted to the surface when bridging divalent cations are available in models **II, V** and **X**. Thus, as in experiments, our simulations show significant differences in the interaction of RNA with the clay surfaces in the presence of Na^+ and Ca^{2+} cations.

Plots of atomic density for models **I** and **II**, displayed in Figure 6.5, show the differences in density of the phosphorus atoms near the clay surface when either Na^+ or Ca^{2+} ions are present. These plots clearly show that more phosphorus atoms in model **II** (Ca^{2+}) are concentrated near the clay surface than in model **I** (Na^+), indicating that RNA has a greater propensity to be closer to the surface in the presence of divalent cations. The mean-squared displacement plots in Figure 6.6 for models **I, II, V** and **VI** show greater diffusive molecular motion for RNA when Na^+ are present, as compared to the models containing Ca^{2+}. Thus, the mean-squared displacement of RNA as a function of simulation time for models **II** and **V** shows very little diffusive motion, characterized by the gradient of the mean squared displacement plots in Figure 6.6. Thus, this support the

earlier claim that divalent cations are more efficient than monovalent cations in mediating the adsorption of nucleic acids on clay minerals. Franchi's experiments showed no significant differences in the adsorption of polypurines and polypyrimidines on montmorillonite, even though previous work by Ferris *et al.* [226] reported a higher affinity of free purines and purine nucleotides than that of free pyrimidines and pyrimidine nucleotides.

Visualisation of the snapshots taken from MD trajectories of the entire system in models **II, V** and **X**, displayed in Figure 6.7, shows the RNA strand tethering to the surface through a single planar purine base of the biopolymers with sequences **A**, **B** and **dA**. Model **XIII** shows a Watson-Crick double-helical strand of RNA, sequence **dA** (Table 6.1), tethered to the montmorillonite surface through a planar purine base at the 5′ end of one of the strands, rather than midsection nucleotide base. Average structures of the RNA molecule, shown in Figure 6.8, reveal the extent to which the conformation of RNA has altered in the presence of the clay surface, compared to the elongated helical conformation of the strands in aqueous water. Experiments exhibit folding of single stranded RNAs in bulk water and RNAs on mica surfaces [218, 227].

Simulations show that the nucleotide base contributes to the mechanism by which RNA adsorbs onto the aqueous montmorillonite surface. This disagrees with the notion [213] that the sugar and base groups of the RNA molecule do not take part in the adsorption of RNA onto the surface and that the negatively charged phosphate groups are the best candidates for the interaction with clay surfaces mediated through cations, suggested by Franchi *et al.*. The visualisations from simulations in Figure 6.7 show the planar adenosine base lying flat on the montmorillonite surface at a tetrahedral Si^{4+} site, away from the Al^{3+} partial substitutions. The adenine ring interacts with the clay surface atoms via a combination of van der Waals and electrostatic forces.

Franchi and co-workers experiment indicates that double-stranded nucleic acids need higher cation concentrations than single-stranded ones to establish an interaction with the clay surface. One proposed mechanism by which nucleic acids interact with a montmorillonite surface is through a "cation bridge" where divalent cations locate between the mineral surface and the nucleic acid; the nucleic acid is then cooperatively bound to the surface through the divalent cations [213, 217]. Simulations do not exhibit the binding of RNA through a "cation bridge" mechanism, but rather indicate that the divalent cations neutralise the negative charge present on both the nucleic acid and the clay surface. The RNA then binds to the surface though a single nucleotide base essentially mediated by Lennard-Jones dispersion forces.

Pastré *et al.* used atomic force microscopy (AFM) to probe DNA adsorbed to a mica surface. Their experiments suggest that DNA attraction is due to the sharing of the DNA and mica counterions. The correlation between divalent counterions on both the

negatively charged DNA and the mica surface can generate a net attractive force whereas the correlation between monovalent counterions is relatively ineffective. These experimental observations regarding the counterion valency correlate with the simulations reported here in which Na^+ is ineffective in binding the single-stranded RNA in models **I, VI** and **X**. The degree to which the Na^+ ions are ineffective at attracting the RNA to the montmorillonite surface can be viewed qualitatively in the animation of model **VI** supplied in Appendix A.

6.3.2 RNA collapse and formation of secondary structural motifs

RNA molecules are known to fold into specific, intricate three-dimensional structures which are necessary to perform numerous biological functions [228]. Thus, understanding RNA folding is important in unravelling RNA's putative role in the early chemical processes which may have led to the origin of life on Earth. In addition to the models consisting of RNA on an aqueous montmorillonite surface, corresponding models of RNA in bulk water were built and simulated. These are models **III, IV** for sequence **A**, **VII, VIII** for sequence **B** and models **XI, XII** for sequence **C**; see Table 6.1. The bulk water simulations were designed to understand how the structure and conformation of different RNA sequences are modified in the presence of a mineral surface, by comparison with their corresponding behaviour in bulk aqueous solution. Visualisations taken from our simulations reveal significant changes in the conformation of RNA between the clay surface models **I** and **II** and the bulk water models **III** and **IV**, as well as notable differences in structure and conformation when different cations are present in models **I** and **II**.

The first experiments on the initial collapse of the RNA Tetrahymena ribozyme, induced by various counterions, were presented by Thirumalai and Heilman in 2001 [229, 230]. Experiments using time-resolved small-angle X-ray scattering (trSAXS) have revealed that, upon addition of divalent Mg^{2+} cations, RNA in bulk water rapidly compacts from an extended state to a globular state of nearly native dimensions, well before the formation of any stable tertiary contacts [228]. Gel electrophoresis experiments on the Tetrahymena ribozyme performed by Kucoli *et al.* show that the rate of formation of the catalytically active native state, from the initial collapsed state, increases with decreasing counterion charge [231]. It is intuitive that the transition to the active state would be slower compared to the initial collapse as it requires breaking of hydrogen bonding and re-folding. These collapsed RNA states can be considered analogous to the molten globule states that appear in early stages of protein folding which are stabilised by hydrophobic interactions. Our simulations show a collapse of the initial elongated RNA structures, for sequences **A**, **B** and **C** shown in Figure 7.2, into folded, more

compact, structures in the models in which RNA is interacting with a montmorillonite surface (models **I, II, V** and **VI**), over a time scale of 5 ns. The collapse of the initial structure can be seen in plots of radius of gyration of the RNA strand as a function of simulation time, shown in Figure 6.9. The radius of gyration is calculated here for the RNA molecule in each model. The radius of gyration plots in Figure 6.9 for models **I** and **II** show a decrease in the effective size of the molecule over the first 15 ns of simulation, from 20 Å to around 15 Å. The radius of gyration plots for models **V** and **VI** show a decrease in the effective size from 25 Å to around 15 Å. The reduction in size of the RNA single-strand depicted here is supported further by the observed collapse and compaction of RNA seen during the simulations. The 5 ns collapse observed in the simulations is analogous to the 1 micro-second fast collapse of the tetrahymena ribozyme reported experimentally using trSAXS [228]. The initial folding of the ribozyme happens over a longer timescale due to the size (400 mer) and complexity of the ribozyme compared with the smaller RNA sequences studied here. The ribozyme has complex junctions stabilised by hydrogen bonds and base stacking interactions that are compatible with the RNA's electrostatically extended conformation. Base stacking refers to the stabilising effect of parallel nucleotide base groups that interact through the π-bonding orbitals of their aromatic rings. The breaking of these initial interactions slows down the rate of the earliest global conformational changes. In the tetrahymena ribozyme experiments discussed here, the ribozyme's initial "extended" structure already contains preformed secondary structures; it is not a random elongated strand.

The collapse that is observed in models **I, II, V, VI, IX** and **X** containing clay surfaces is not observed in the corresponding bulk water simulations in this study. The single strand RNA sequences in models **III, IV, VII, VIII, XI** and **XII** retain their initial helical structure over the course of each simulation. The absence of folding of RNA in these bulk water simulations is characterised by the constant effective size of the RNA molecule for the duration of each simulation, which can be seen in plots of the radius of gyration in Figure 6.9.

The outcome of the initial collapse of RNA has been suggested to result from a trade-off between the electrostatic screening of negatively charged un-protonated phosphate groups *versus* the hydrophobic collapse of the strand to shield the hydrophobic base groups [228]. The collapse of the initial structures in the simulations leads to the formation of secondary structural motifs in models **I, II, V** and **VI**. The formation of the stem-loop secondary structure, shown in Figure 6.8, is stabilised by the stacking of adjacent and non-adjacent bases, and by the formation of Watson-Crick base pairs in models **V** and **VI**. The occurrence of Watson-Crick base-pairs and base stacking is known to stabilise the formation of optimal and sub-optimal folded secondary structures [222].

To analyse the principal modes of motion of RNA in all models, PCA was employed; the results are shown in Figure 6.8. In published MD simulations of RNA in bulk water the dominant modes of motion are a twisting of the strand around a central axis, junction bending and wedge bending [222]. Models **III, IV, VII, VIII, XI** and **XII**, which consist of RNA in bulk water, confirm these findings. The dominant modes of motion of RNA in bulk water can be seen in Figure 6.8. The clay-surface interacting RNA in models **I, II, V, VI, IX** and **X** have different modes of motion since it is the interaction with the clay surface which restricts molecular movement. The clay/RNA models show a strong dependence between the first two eigenvectors, suggesting that forces acting along these eigenvectors are coupled. The presence of the clay surface and the resultant tethering of the RNA significantly restricts the movement of the RNA molecule. The main contribution to the dynamics is seen at the ends of each of the RNA oligomer chains, analogous to previous simulations of DNA intercalated within a layered double hydroxide (LDH) matrix [149].

The compact folded structures observed in the simulations are analogous to the collapsed, folded and mis-folded hairpin loops observed during replica-exchange molecular dynamics performed by Sorin *et al.* [232], who reported the stabilisation of mis-folds by non-native interstrand nucleotide stacking. Sorin *et al.* employed large ensembles of simulations using the *Folding@home* distributed computing grid, thereby showing the competing pathways in hairpin folding and unfolding. Sorin *et al.* concluded that the conformational dynamics for folding small 12 mer RNA sequences happens over tens of milliseconds. The models in this chapter indicate that the 25 mer RNA, of sequences **A, B** and **C**, in the presence of a clay surface, enjoy an increased rate of folding due to the change in their modes of motion.

For sequences **B** and **C** the optimal, and lowest energy, folded structures were calculated using the *RNAfold* tool which is part of the Vienna RNA software suite [225]. The *RNAfold* tool reads in RNA sequences and calculates their minimum free energies using the minimum free energy algorithm of Zuker & Stiegler [233]. The optimal folded structures of sequences **B** and **C** are displayed schematically, and as a three dimensional structure, in Figure 6.3. The optimal folded structures and energies were used to compare against the folded structures of the RNA sequences in the simulations, shown in Figure 6.11. To obtain free energies for the folded structures that arise in the simulations, the *RNAinverse* program, which is also part of the Vienna suite, was used. The *RNAinverse* program searches for sequences that map into a predefined tertiary structure, thereby inverting the folding algorithm. Target structures and starting sequences are read in alternately. The optimal and sub-optimal structures and energies are tabulated in Table 6.2 where the secondary structures have been displayed using dot-bracket notation in which the characters "(" and ")" correspond to the 5′ and the 3′ base which

close the stemloop motif, while "." denotes an unpaired base. The RNA folding pathway can be discerned from the free energy landscape which may be visualised in terms of a tree diagram. The tree arranges local minima and their corresponding saddle points in a hierarchical fashion. Figure 6.12 shows the tree diagrams for both sequences **B** and **C**. The optimal fold, the lowest energy fold, which has the structure shown in Figure 6.3, is given the label 1 in the tree plots.

Model	Sequence	Secondary Structure	Type	Free energy (Kcal/mol)
	B	....(((((......)))))....	optimal	-6.15
V	**B**	..((...............))...	sub-optimal	2.51
VI	**B**	........(......).........	sub-optimal	2.50
	C	...(((((((((...))))))..))).	optimal	-7.40
IX	**C**	....	un-folded	0.00
X	**C**	...(.........)...........	sub-optimal	4.20

Table 6.2: RNA sequences and their corresponding optimal folded secondary structures for sequences **B** and **C** as well as the suboptimal structures in models **V, VI, IX** and **X**. Secondary structures are shown in the "dot-bracket" notation (explained in the main text), along with their calculated free energies.

The folding landscape in Figure 6.12 shows the compact structures for sequences **B** and **C** within the folding pathway to their optimal lowest energy fold in bulk aqueous solution. The structures taken from simulations are suboptimal structures which are all present along the predicted folding pathway. This indicates that these local minima are occupied at intermediate stages in the folding of the optimal secondary structure conformations. As folding is not observed after 20 ns of simulation in the bulk water models **VII, VIII, XI** and **XII**, but compact conformations of sub-optimal energies are observed along folding pathways for our systems containing RNA and a clay surface in models **V, VI, IX** and **X**, we infer that the clay surface has the effect of increasing the rate at which these RNA sequences compact and fold along pathways that might eventually reach their optimally folded conformations, associated with global free energy minima, even though these may not be found in a clay/RNA system due to the presence of many additional interactions.

Ion size determines the closest distance between ion and RNA and effects ion-ion distance and hence the strength of Coulombic and excluded volume correlation. Folding studies have been performed on the tetrahymena ribozyme looking into the effects of differing ions on folding kinetics [234]. These experiments show that there is a linear relationship between increase in folding stability and the increase in (divalent) ion charge density (i.e. charge/volume of the ion). This experimental observation supports the differences in folding rates shown between the simulation of models **I, V** and **IX** with models **II, VI** and **X**. The increase in charge density of Ca^{2+} over Na^+ leads to increased folding stability and drives the increase in rate observed in these simulations.

During the full 25 ns MD simulation of model **XIII** the double strand of RNA does not undergo any folding, but retains its double helical conformation; this indicates that the Watson-Crick base pairing is stable under these conditions, and no unzipping of one strand from its complementary strand is observed. Figure 6.7 shows the double-stranded RNA tethering to the montmorillonite surface through an adenine base group at the 5′ end of the tethered strand; this suggests that a tethered strand provides a stable template for growth of a complementary strand in the 5′ to 3′ direction.

6.4   Conclusions

In this present study simulations have elucidated a number of facets of nucleic acid–clay interaction and reactivity that have been observed in previous experimental studies, but have hitherto remained unexplained at a molecular level.

One prevailing theory in origins of life studies is that RNA in prebiotic conditions may have been confined and restricted in motion by mineral surfaces, much as it is confined within the ribosome of living cells today [215]. Due to the ubiquitous presence of clay minerals and their associated cations in nature, these aqueous environments may have provided prebiotic habitats for nascent nucleic acid polymers. Indeed, clays may thus have played a key role in the formation and preservation of these polymers and/or their precursors. Analysis of the dynamics of the RNA oligomers in all of the models we have studied reveals that the behaviour of RNA interacting with these surfaces is significantly different from that in bulk water.

In general, these simulations have shown that strong electrostatic forces act between the montmorillonite clay surface, the interacting RNA molecule and the aqueous monovalent or divalent cations. Divalent cations are more effective in mediating the interaction between the clay surface and RNA sequences, due to increased electrostatic screening. The electrostatic repulsion of the unprotonated phosphate groups within the RNA molecule from the negatively charged sites associated with aluminium ions on the clay surface is overcome through the presence of solvated divalent cations, allowing the RNA molecule close proximity with the clay surface. These simulations have identified a mechanism for clay-RNA interactions, whereby a planar purine nucleotide base interacts with the clay surface mainly through Lennard-Jones forces. The electrostatic attraction and tethering of the RNA strand to the clay surface causes it to be significantly pinned and hence restricted in movement compared to RNA in bulk water. Structural analysis of surface-interacting RNA molecules in all of the models we have studied confirms that the motion of RNA is indeed substantially restricted compared to RNA in bulk water. Tethering of single-sranded RNA to the montmorillonite surface is only observed

in models with Ca^{2+} counterions: equivalent models which have Na^+ counterions do not produce similar effects. This leads to the inference that divalent ions such as Ca^{2+} very likely play a key role in tethering, as reported experimentally [213]. The observed tethering of single-sranded RNA to the clay surface is of considerable relevance to pre-biotic chemistry: $5'$-$3'$ regioselective RNA polymerisation has been reported by Ferris and co-workers [71, 210, 211], but was hitherto difficult to explain mechanistically.

These simulations show that single strands of RNA fold at a considerably enhanced rate when a montmorillonite surface is present compared with the same sequences in bulk water. For both Na^+ and Ca^{2+} containing clay and single-sranded RNA models, the single-sranded RNA undergoes a fast collapse stage of folding into a more compact structure in less than 5 ns, which is not observed in bulk water with the same cations. The increased rate of folding at clay surfaces is of importance to the RNA world view, according to which early RNA polymers would have needed to produce well defined folded structures to support enzymatic activity in order to cleave self-replicating RNA strands, essential components in the first steps of molecular evolution [42]. Accelerated folding kinetics considerably increases the parameter space that may be explored by clay-RNA complexes in constructing catalytically active structures under conditions which may have been far from benign.

When interacting with the clay surface the motion of the end groups of the oligonu-cleotide dominate the motion of the molecule; as a result the single RNA strands collapse and fold into more compact structures, made possible by the increased concentration of cations screening the mutual electrostatic repulsion from phosphate groups in clay-containing systems with both Na^+ or Ca^{2+}. Simulations show that the compact, folded structures of these RNA sequences are stabilised by the stacking of adjacent and non-adjacent base groups as well as via hydrogen bonds between nucleotide groups, giving rise to secondary structural hairpin-loop motifs. Conceptually, once RNA chains are stabilised on clay surfaces, double-stranded RNA may be formed, for example as part of a template-directed replication process; folded into tertiary structures, such ribozymal RNA can then act as a catalyst [42] for self-replication.

Through these studies of mineral bound nucleic acids, the insight gained from many large-scale molecular simulations run for tens of nanoseconds adds considerably to the interpretation of experimental data from previous studies through provision of highly detailed molecular level insight. It is evident that, when considering the RNA world, minerals may substantially modify the chemistries that can occur: an RNA-mineral-water world may have differed considerably from a mineral-free RNA world.

(a)

(b)
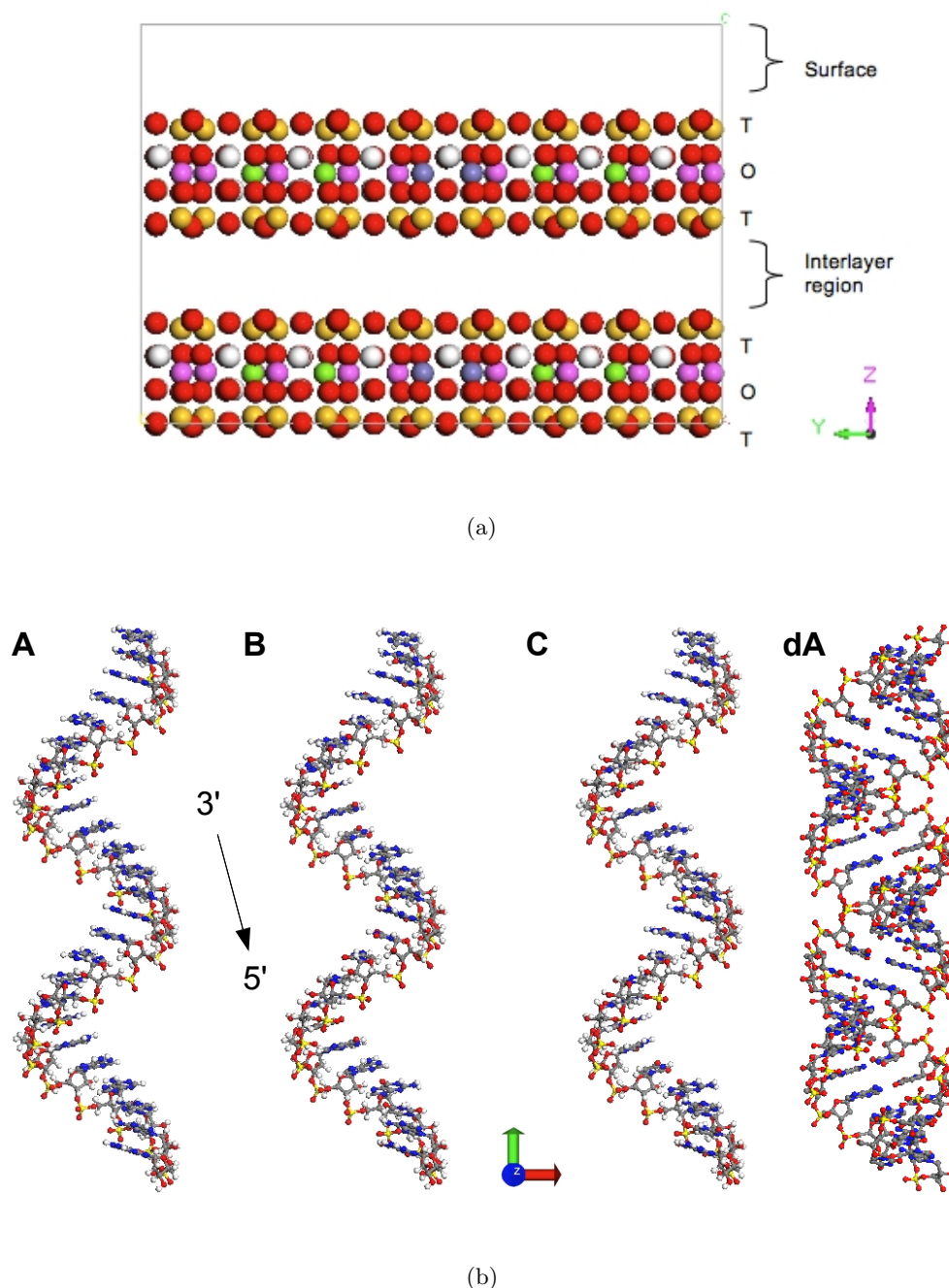
Figure 6.1: a) Initial structure of the unit cell of Wyoming type montmorillonite with charge balancing ions and water removed for clarity. Atoms are coloured as follows: Si (gold), O (red), H (white), Al (purple), Mg (green) and Fe (blue). The visualisation depicts the tetrahedral and octahedral layers of the clay, as well as interlayer regions and the exposed surface. (b) RNA sequences **A**, **B** & **C** show the three different 25 mer single strands of RNA and **dA** depicts the double stranded RNA used in this study. RNA sequence **A** is used in models **I-IV** and consists of 25 adenine base groups. RNA sequence **B** is used in models **V-VIII** and **C** in models **IX-XII**; both consist of different nucleotide base sequences described in more detail in Table 6.1. **dA** consists of a 25nt adenine sequence and complementary uracil strand. Colour scheme is as follows: O (red), H (white), N (blue), C (grey), and P (yellow). The visualisation shows the RNA strands in the $3' \rightarrow 5'$ direction, where the prime indicates the carbon atom on the ribose sugar component of the RNA monomer.
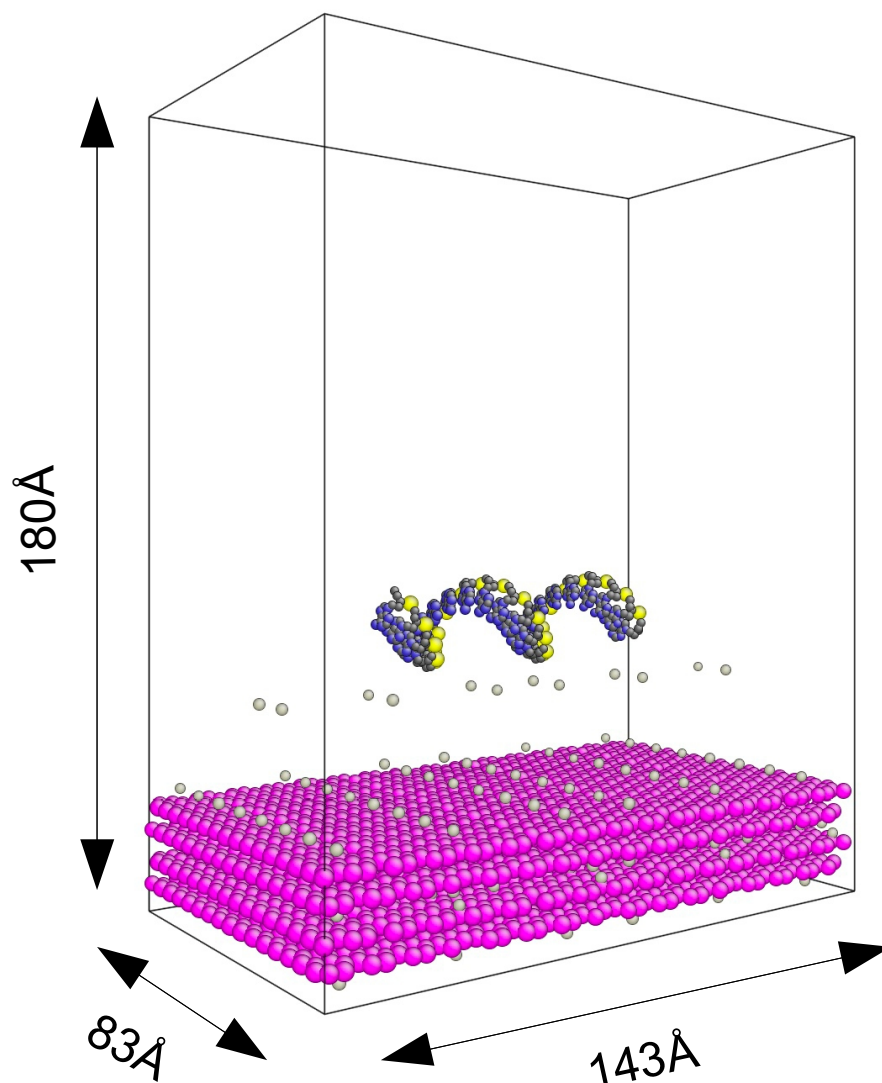
Figure 6.2: Starting structure of model **II** in Table 6.1, the silicon atoms within montmorillonite being shown as magenta spheres; remaining clay atoms have been hidden to aid viewing. Nitrogen, carbon and phosphorus atoms belonging to the RNA molecule are coloured the same as in Figure 7.2. Ca^{2+} ions are shown as grey spheres which were placed manually at random locations. Water molecules are not displayed. The helical structure of the RNA strand was created using the Nucleic Acid Builder.
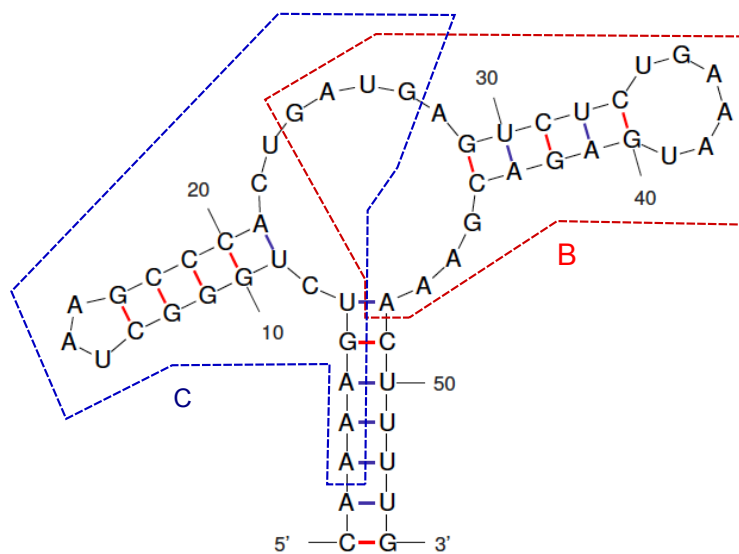
(a)
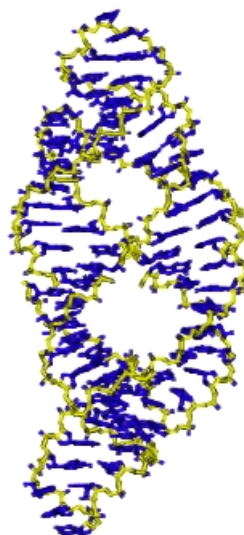
(b)

Figure 6.3: (a) Secondary structure of the hammerhead ribozyme (type III) and (b) the crystal structure of the RNA (type III) hammerhead ribozyme. Colour scheme as in Figure 7.2. Secondary structure schematic in Figure (a) was generated using the Vienna RNA software suite [225]. The tertiary structure of the ribozyme (b) was taken from the protein database file 1MME and visualised using VMD. Both Figures show the two stemloop motifs which were used in sequences **B** and **C**; sequence **B** consists of the nucleotides 24-48, and sequence **C** of nucleotides 3-27 shown in (a).

(a)

(b)

(c)

(d)
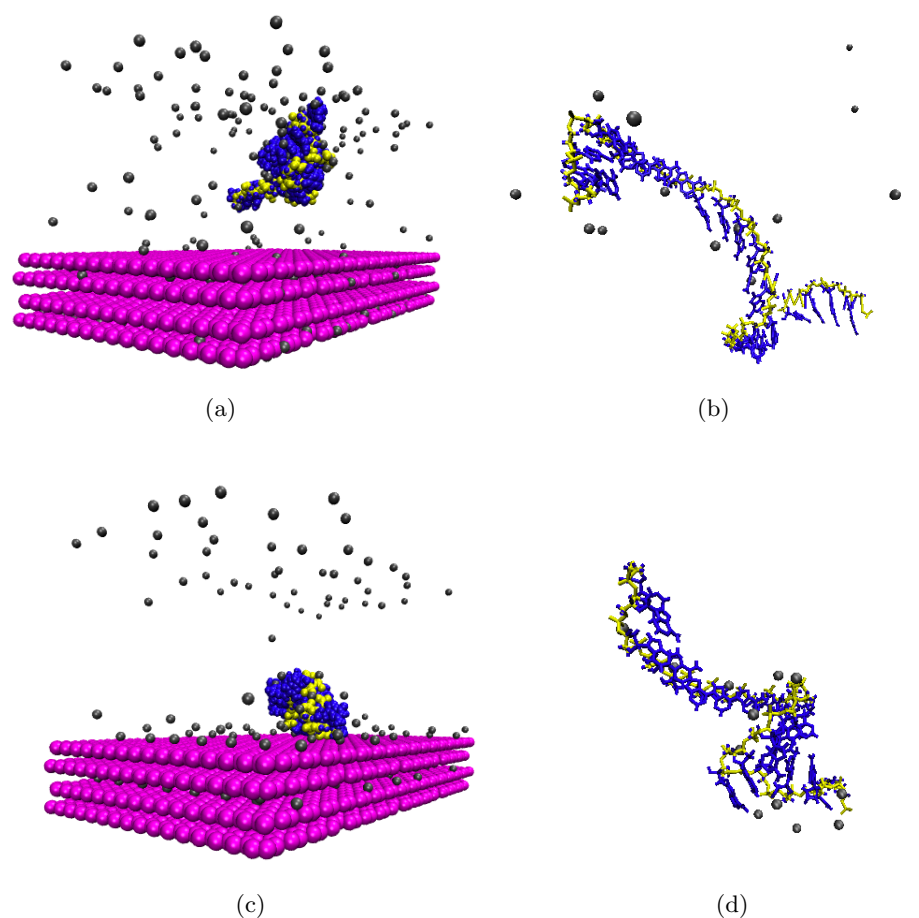
Figure 6.4: Average structures taken from simulations after 20 ns of simulation time. Atom colours are as described in Figure 6.2, water is not shown for clarity. The visualisations of the trajectories here reinforces the differences observed in simulations between the various models described. Visualisation in Figure (a) and (c) show the RNA strand adopting a more compact conformation in model **I** and **II** respectively, which are composed of an aqueous clay surface, compared to the extended structure of RNA in bulk water, observed in models **III** and **IV** in Figures (b) and (d) respectively. The visualisation displayed here also show the non-adjacent stacking of nucleotide bases in models **I** and **II**, Figures (a) and (c) respectively, which is not observed in the bulk water models. Colour scheme as in Figure 6.2.
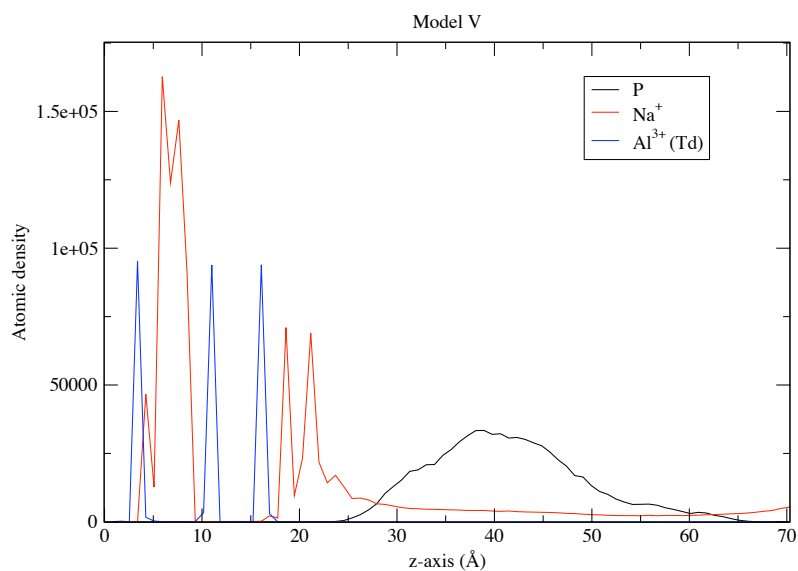
(a)
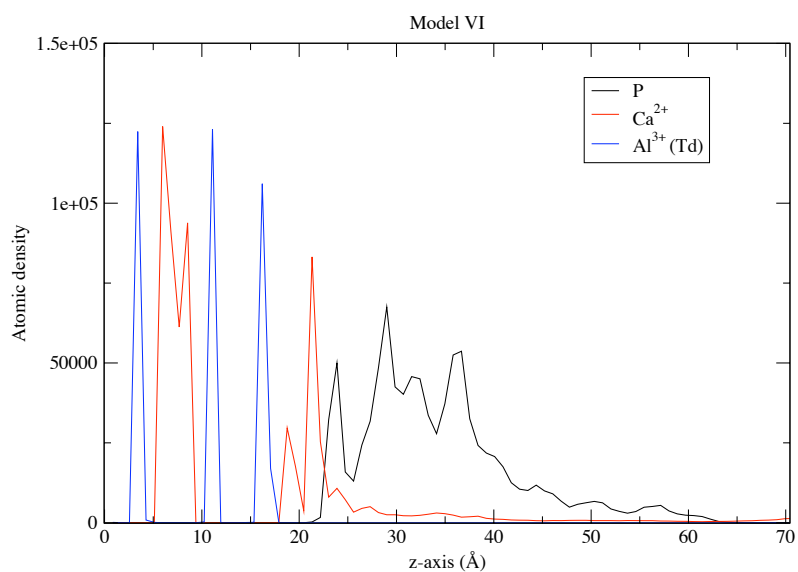
(b)

Figure 6.5: Atomic density profiles taken from 20 nanosecond simulations, of models **V** in Figure (a) and **VI** in Figure (b). The atomic density profile shows the time averaged atomic density of the specified atoms as a function of distance along the z-axis of the simulation cell. Atomic density refers to the number of atoms per unit volume. These profiles show the density of phosphorus atoms belonging to the RNA molecules, surface Al^{3+} ions and Na^+ or Ca^{2+} cations. The plots show the density of phosphorus atoms to be concentrated closer to the surface in model **VI** than in model **V** where the density is spread-out around 40 Å. The density of phosphorus atoms in model **VI** shows more well defined peaks indicating that these atoms occupy more regions relative to the clay surface than those in model **V**. These differences can be attributed to the increased electrostatic screening by divalent Ca^{2+} ions in model **VI** relative to that of monovalent Na^+ ions in model **V** in Figure (a).
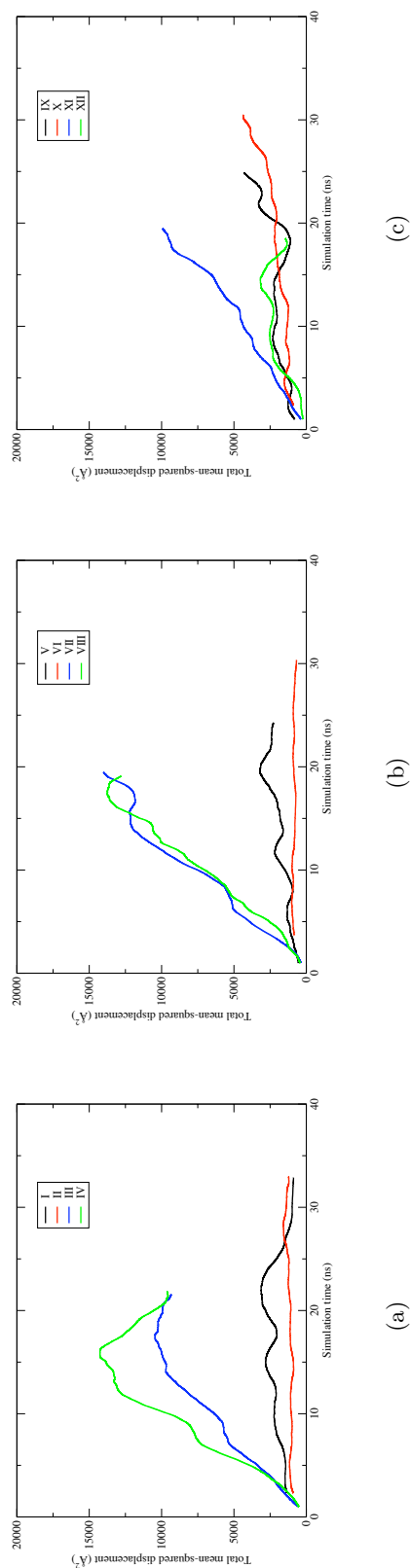
Figure 6.6: (a) Total mean-squared displacement of RNA for sequence **A**, systems **I-IV**, (b) sequence **B**, **V-VIII**, and (c) sequence **C**, **IX-XII**, all at ambient temperature and pressure (300 K, 1 atm), relative to the initial structure, as a function of simulation time. The linear portion of the MSD plot is proportional to the diffusion coefficient of the diffusing atoms. The plots show a significant difference in the diffusion of RNA in an aqueous solution (green & blue) as compared to RNA interacting with a clay surface (black & red). When interacting with the clay surface and calcium ions (systems **II**, **VI** & **X**) the RNA strand exhibits little diffusive molecular motion, characterised by the low gradient of the slope. This is due to the increased screening of negative charges from phosphate groups and surface sites by divalent calcium ions, allowing RNA to become attracted to the surface cooperatively, bringing the strand in close proximity to the surface and permitting it to tether through a nucleotide base. The RNA strand combined with the clay surface and sodium ions (system **I**, **V** & **IX**) displays more diffuse motion than with Ca^{2+} ions due to the reduced screening of negative charges by Na^+ ions. This is observed in the gradient of the MSD plots (black line). The MSD plots for RNA in bulk water, models **III**, **IV**, **VII** & **VIII**, show interesting step-function-like behaviour. The RNA-in-bulk-water models (blue and green curves) show a decrease in diffusivity between *ca* 5-10 ns for a duration of 2.5 ns, which then returns to its original behaviour.
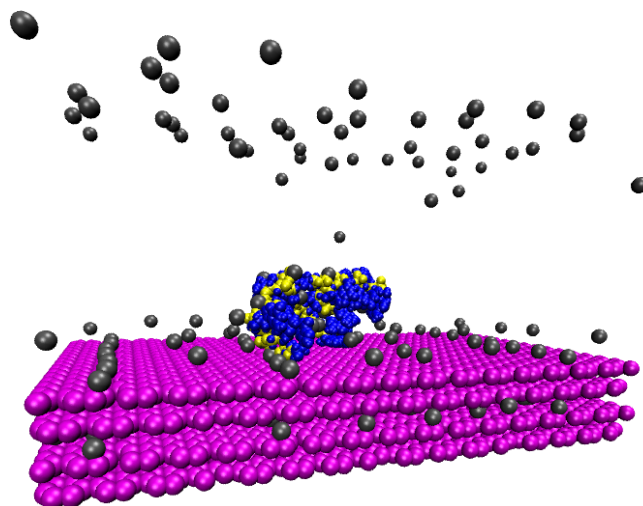
(a)
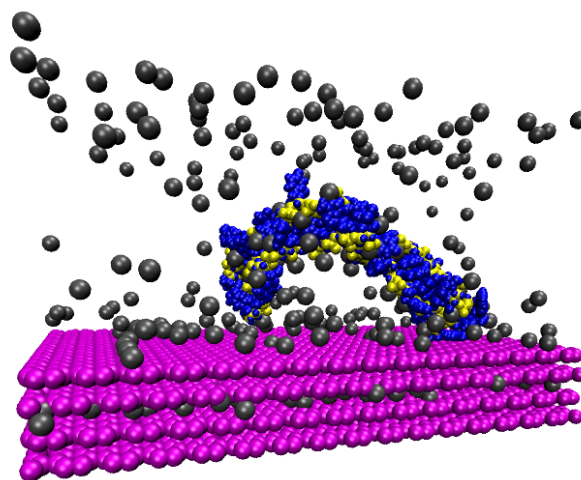
(b)

Figure 6.7: Simulation snapshots taken at 20 ns from simulations of (a) model **VI** and (b) model **XIII**, after 30 and 20 ns respectively. The colour scheme is the same as in Figure 6.4. These visualisations show the conformation of RNA and throw light on the mechanism by which RNA tethers to the surface. The visualisations in this Figure show the tethering of (a) the single strand, of sequence **B**, and (b) the double strand **dA** of RNA to the montmorillonite clay surface, through a planar purine ring, in this case adenosine, at the 5′ end of the biopolymer. The screening of the negative charges between the surface and phosphate groups by the cations allows the RNA strand to get close enough to the surface for it to tether through the nucleotide base and the surface via attractive dispersion forces. Water molecules are hidden to aid in viewing.
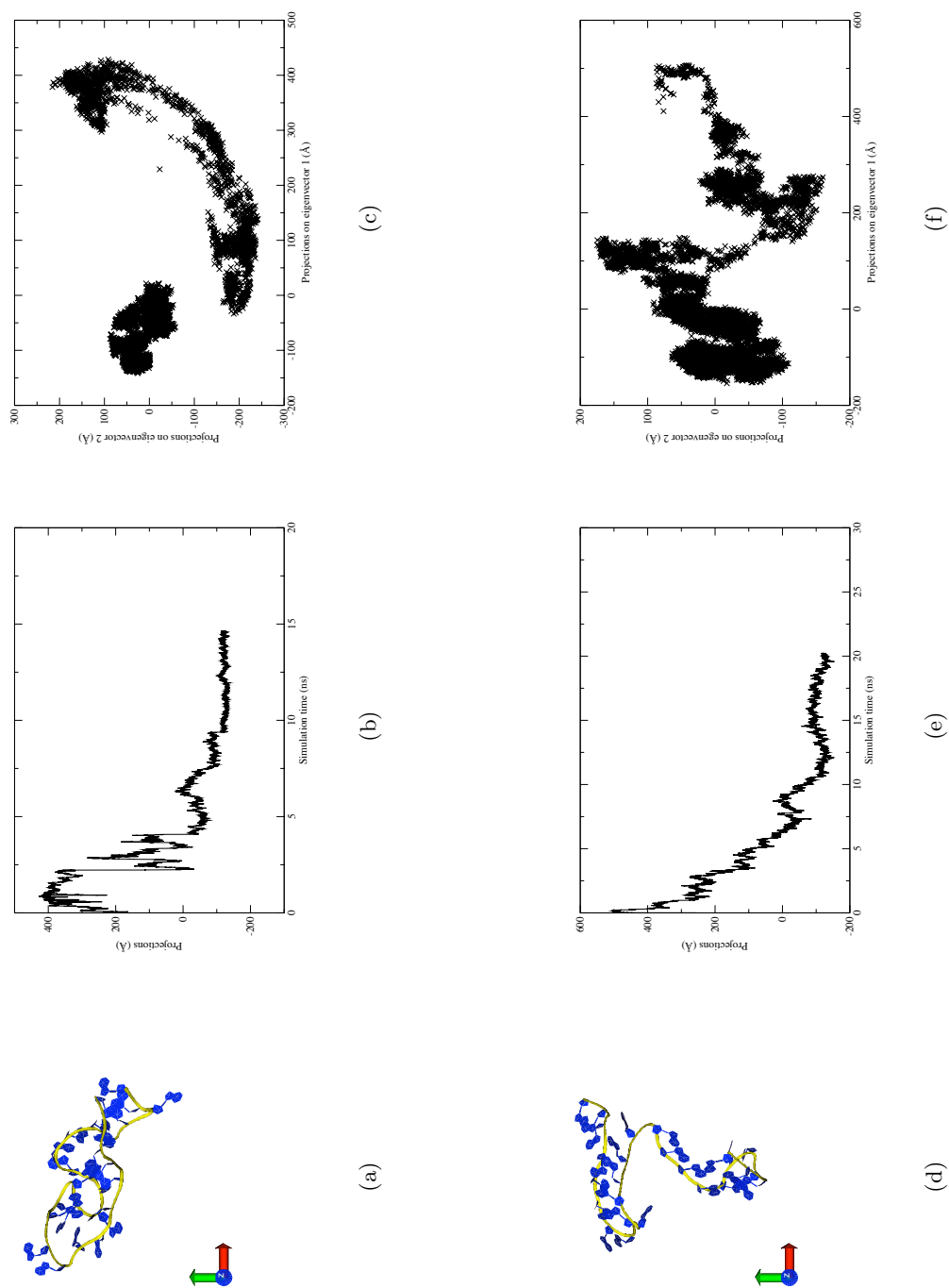
Figure 6.8: Principal component analysis of RNA interacting with a clay surface used to elucidate the principal modes of motion within systems **I-IV** of sequence **A**. (a) and (d) show the average structure of RNA in models **I** and **II** respectively, averaged over 20 ns of simulation. (b) and (e) show plots of projections against simulation time, while (c) and (f) show plots of projections of mode 1 against projections of mode 2. The motions along the first eigenvector obtained from coordinates of the RNA molecule and their planes of motion were defined by the first two eigenvectors of model **I** in (a), (b) and (c) and model **II** in (d), (e) and (f). The colour scheme is the same as Figure 7.2. Both models show a strong dependence on the first two eigenvectors, indicating that forces acting along these eigenvectors are coupled. The addition of the clay surface and the tethering of the RNA significantly restrict the movement of the RNA molecule. The main contribution to the RNA dynamics comes from the ends of the molecule.
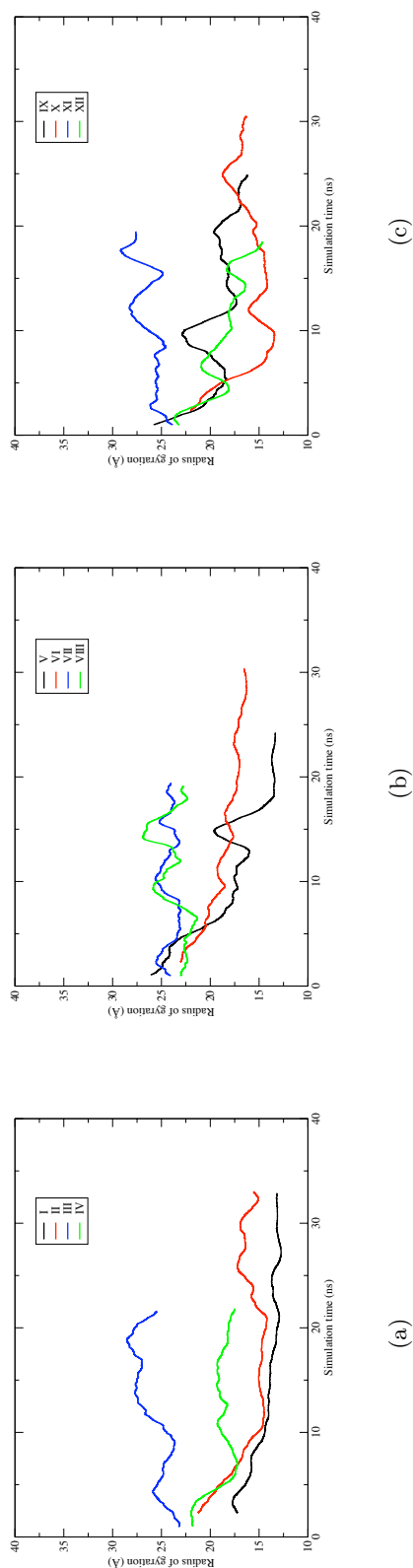
Figure 6.9: Radius of gyration, r_G, of RNA as a function of simulation time, relative to the initial structure for (a) models **I-IV**, (b) models **V-VIII** and (c) **IX-XII**. Lines in black and red indicate models with a clay surface, whilst blue and green lines refer to models in bulk water. The radius of gyration is a measure of the effective size of the RNA sequence. The radius of gyration is significantly smaller in the case of RNAs interacting with the montmorillonite surface, most likely due to the increased concentration of cations near the surface of the clay more effectively neutralising the charge of the phosphate groups, reducing the electrostatic penalty of having a compact structure and encouraging a folded conformation, stabilised by the "burying" of hydrophobic nucleotide bases and the formation of adjacent and non-adjacent base stacking, as well as Watson-Crick base pairing. The concentration of ions in the models of RNA in bulk aqueous solution (blue & green) is less than in the clay surface systems; as a result the RNA favours a conformation with a greater effective size, which corresponds to an elongated helical structure close to that in the initial models.
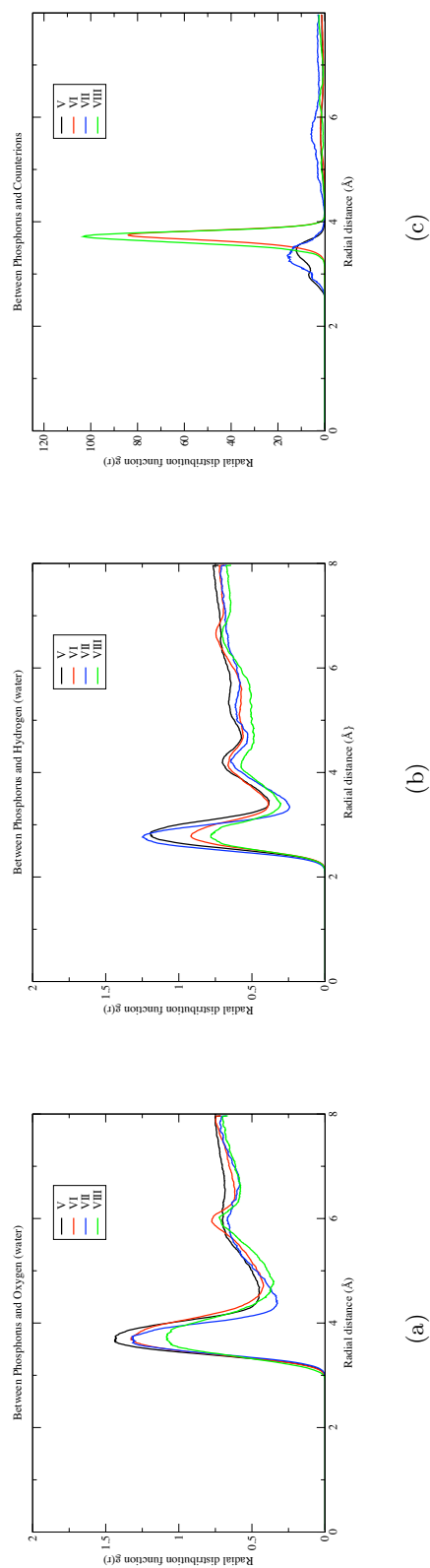
Figure 6.10: Average radial distribution functions for all the models which include sequence **C** with a clay surface and in bulk water (models **V**, **VI**, **VII** & **VIII**), calculated with respect to phosphorus atoms in phosphate groups of RNA molecules. Radial distribution functions show similar trends for coordination of water molecules around a central phosphorus atom. The radial distribution functions show (a) water oxygen atoms at 4 Å, (b) water hydrogen atoms localized at 2.8 Åand (c) counterions with respect to phosphorus, where the cations are Na^+ (models **V** & **VII**) and Ca^{2+} (models **VI** & **VIII**). (c) shows a significant difference in the radial distribution for the two different cations. That for Na^+ ions exhibits a "shoulder" on the main peak at 3.4 Å, indicating that Na^+ ions occupy two sites around phosphorus atoms at 3 Åand 3.4 Å. The Ca^{2+} case by contrast only has one site at which the ions are located.

(a)

(b)

(c)

(d)
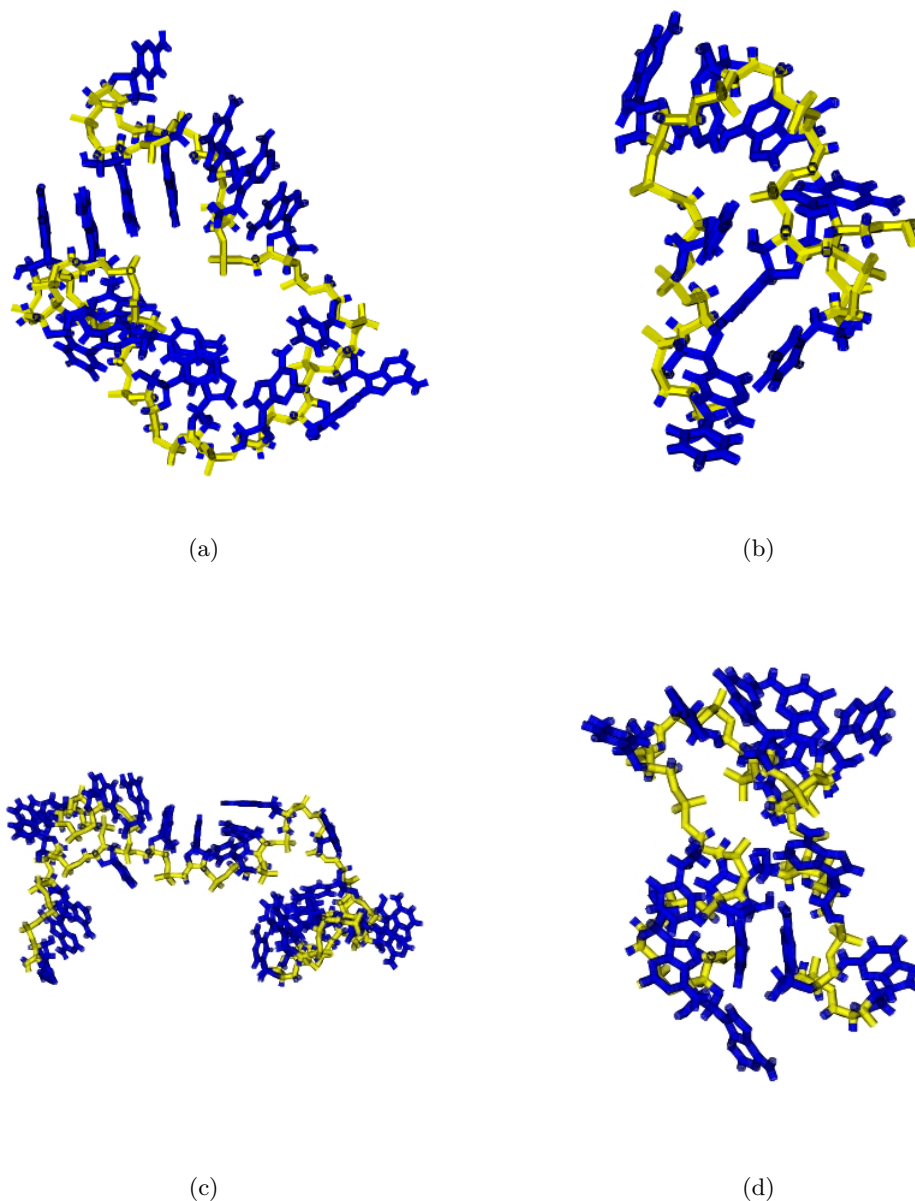
Figure 6.11: Folded sections of the average RNA structure taken from the entire production simulation for (a) model **V** , (b) **VI**, (c) **IX** and (d) **X**. Colour coding as in Figure 7.2. Images show the folded secondary structures of RNA for sequences **B** and **C** when interacting with a clay surface. Figures (a), (b) and (d) show the formation of hairpin loop type motifs stabilised by stacking of nucleotide bases, seen as a closed loop in the 3D structures. Figure (c) shows sequence **C** in model **IX** which has not yet formed a secondary structure motif, but exhibits a more compact conformation than the corresponding sequence in bulk water in model **XI**, in which the end segments have folded in onto the rest of the molecule.
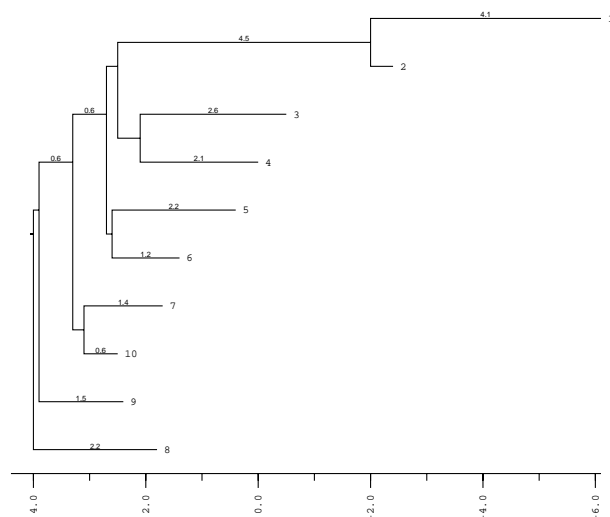
(a)
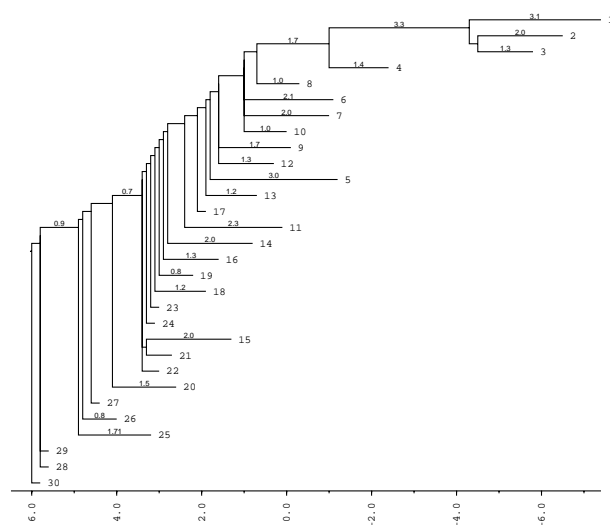
(b)

Figure 6.12: A representation of the folding free energy landscape of (a) sequence **B** and (b) sequence **C**. The tree arranges local minima and their corresponding saddle points in a hierarchical fashion. The optimal and lowest free energy fold is labeled as 1 in each tree. Folding landscapes were calculated using the Vienna RNA software suite [225]. Energies calculated for each RNA structure in Figure 6.11 correspond to sub-optimal folds and energies in the folding landscapes shown here. The energy values given in the horizontal axis are in units of Kcal/mol. The vertical axis gives shows the relative stability of the folded structure, the highest being the most stable. The free energies of the RNA sequences in models **V** and **VI** correspond to the suboptimal structures labelled 9 and 10 respectively in (a). Model **IX** corresponds to the suboptimal structure labelled 27 in (b).

Chapter 7

Influence of Surface Chemistry and Charge on LDH-RNA Interactions

This chapter details the results of large-scale molecular simulations, run over several tens of nanoseconds, of 25-mer sequences of single-stranded ribonucleic acid (RNA), in bulk water and at the surface of three hydrated positively charged MgAl layered double hydroxide (LDH) minerals. The three LDHs differ in surface charge density, through varying the number of isomorphic Al substitutions. We constructed these simulations in order to probe the interaction of various aqueous RNA sequences with LDH of differing charge density at the mineral surface.

7.1   Introduction

Clay and clay-like layered materials have gained considerable attention due to their ability to intercalate and adsorb guest molecules or to exchange charged species for externally present interlayer ions [235]. In principle, through the incorporation of charged species into and onto a layered host, novel solids may be engineered with desired physical and chemical properties [236].

LDHs have a structure similar to that of brucite, $Mg(OH)_2$, which consists of charge-neutral 2-dimensional layers of edge-sharing hydroxide octahedra. In the largest family of LDHs, isomorphic replacement of a fraction of divalent cations, for example Mg^{2+}, with trivalent ions such as Al^{3+}, occurs. This yields a positive charge on the layer, and

in turn necessitates the uptake of charged species into the interlayer region. This LDH family, which we study here, can be represented by the general formula:
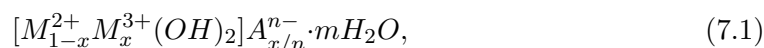
$$[M_{1-x}^{2+}M_x^{3+}(OH)_2]A_{x/n}^{n-}\cdot mH_2O, \tag{7.1}$$

where M^{2+} and M^{3+} are divalent and trivalent cations respectively, and A is an anion of valence n.

Layered double hydroxide-nucleic acid nanohybrids have proven to be a useful class of materials. Their application as efficient drug delivery vehicles for treatment of various neurodegenerative conditions [237] means that a detailed understanding of the interaction between biological molecules and the inorganic LDH interface is much sought after in the biotechnology/biomedicine domain.

Laboratory studies have shown that small interfering ribonucleic acids (siRNA) can be used in treating neurodegenerative conditions such as Parkinson's, Alzheimer's and Huntington's disease. Small interfering RNA are a potentially powerful class of bioactive drugs able to target and destroy messenger RNA (mRNA) [238]. Delivery of unprotected RNA based therapeutics is ineffective due to their inability to penetrate the plasma membrane and their susceptibility to degradation by nucleases. siRNAs play a variety of roles in biology; they interfere with specific gene expressions in RNA inference pathways and help shape the chromatin structure of the genome. However, these bioactive agents have yet to be successfully delivered to the required region in patients. For this reason there has been limited use of RNAs in treating these conditions [239, 240].

LDHs have the ability to exchange their interlayer anions with other charged species; this enables them to intercalate functionally bioactive anionic molecules, such as nucleic acids, with a very high loading capacity [241, 242]. It is thought that the positive surface charge of the LDH further allows an association with negatively charged plasma membranes, thus making the LDH particularly adept at transporting charged biomolecules across a cell membrane. LDHs have very low cytotoxicity and are highly biocompatible.

There has also been much interest in mineral surfaces, and in particular the interlayer region of layered minerals, in prebiotic or "origin of life" chemistry [243, 244]. We have recently discussed the possible role of LDHs in stabilising nucleic acids under potential early Earth conditions [133, 245].

Certain Fe^{2+}/Fe^{3+} LDHs may well have been prevalent on the early Earth and would undoubtedly have interacted with a range of prebiotic molecules and species associated with the putative RNA world. In addition to providing sanctuary for early RNA sequences against thermal or ultraviolet light degradation, such LDHs might have also

afforded effective surfaces for templated polymerisation of further RNA sequences as an early form of heredity [73].

Montmorillonite, by contrast, is a smectite clay. Smectites consist of negatively charged, crystalline aluminosilicate sheets. The montmorillonite type of smectite has a structure comprising stacks of pyrophyllite-like sheets, $[Al_2Si_4O_{10}(OH)_2]$, each of which consists of an octahedral alumina layer sandwiched between two tetrahedral silica layers. The negative charge arises from partial substitution of metal ions in typically both the octahedral and tetrahedral layers.

In Chapter 6 molecular dynamics is used to explain how the smectite, montmorillonite, catalyses the polymerisation of RNA and the mechanism of RNA adsorption at the montmorillonite surface [132]. Montmorillonite has the general formula:

$$[Mg_xAl_{2-x}(OH)_2(Si_4O_{10})](C^{n+})_{x/n} \cdot nH_2O, \tag{7.2}$$
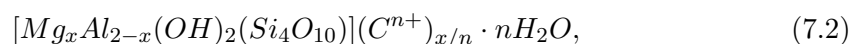
where C is a cation of valence n, commonly Na^+, Ca^{2+} and Mg^{2+}.

The ability of montmorillonite to catalyse the formation of RNA oligomers has been demonstrated in a number of experiments by Ferris and co-workers [71]. In their laboratory studies, Ferris *et al.* used artificial activated phosphorimidazolides of nucleosides together with homoionic Na^+ montmorillonite to form oligomers of 6-14 mers, in which montmorillonite was shown to enhance the rate constant for oligomer formation by 100-1000 times over the hydrolysis of the imidazole activating group [210, 211]. By changing the activating group from imidazole to 1-methyladenine, Ferris *et al.* formed single strands of up to 50 mers of polyadenine and polyuracil without the need for a primer strand [223]. RNA sequences of this length are sufficiently long to exhibit fidelity in replication as well as catalytic activity [71]. It is important to note here that the activated monomers that Ferris *et al.* used in these studies have yet to be synthesis under what are considered prebiotically plausible conditions.

A question arises when considering the interactions of nucleic acids with layered minerals as to whether the nucleic acid intercalates between the mineral layers, or binds to the edge sites. Many of the studies on RNA and layered minerals have used smectite clays, with a low charge density and more complex edge site chemistry [246]. In recent studies by Ferris & X-ray diffraction studies have shown the intercalation of the nucleic acid [247]. In the case of LDH-DNA systems, more relevant to the present study, the intercalation is unambiguous by X-ray diffraction (an interlayer expansion from 13.3 to 20.0 Å is observed) and at extremely high density [69]. Other studies have shown that simple phosphate containing biomolecules also intercalate into LDHs [248, 249].

Ertem and Ferris showed that the presence of $2'$-$5'$ linkages in these initial oligomers would not have significantly deterred the origin of the RNA world since oligo(C) oligomers containing less than 50% $3'$-$5'$ linkages are effective templates for the formation of the complementary RNAs [250, 251]. Gallori and co-workers have also shown how montmorillonite can protect RNA sequences such that they remain active as ribozymes after exposure to UV radiation [252].

A challenge that arises for intercalated RNA-LDH hybrids in origin of life scenarios, and one encountered more generally in the field of clay-polymer nanocomposites [190, 253], is that of releasing the oligomer/polymer formed within the mineral interlayer. Multiple attachment points result in a biomolecule that is difficult to displace. There are a number of pathways by which this may occur in our prebiotic setting. One understood and experimentally determined pathway for the release of DNA/RNA from the mineral surface is the acidic dissolution of the mineral layer leaving the nucleic acid free and exposed. Experiments by Tyner *et al.* showed that the weak base LDH nanocrystallites were completely dissolved upon exposure to a solution with pH 5.0 while kept almost unchanged at pH 7.2 [254]. Certainly, LDHs are broken up in order to buffer the pH in mammalian cells after transfection of the nucleic acid/LDH system. Release mediated by pH decrease is particularly relevant as, if the LDH minerals form near alkaline vent systems [52, 169], then, at some point in the Archean ocean, the alkaline source may of diverted to remove the mineral-RNA system to equilibrate with seawater which is thought to have had a pH of about 6 [255]. Two other potential routes exist: i) carbonate is very efficient at displacing anions from LDH systems through competitive anion exchange, though this has not been tried for polynucleic acids, to our knowledge, and ii) since we use Mg/Al LDH as a proxy for Fe^{II}/Fe^{III} LDH, as the oceans became toxic, it is possible that a transformation of the mixed valence LDH to non-layered iron oxides would also result in polymer release.

7.2 Methods

Computer simulation of clays has gained wide currency in recent years as it is capable of providing information and insight into these often amorphous and structurally disordered systems in a way that remains largely inaccessible experimentally [147, 245].

This section describes the large-scale computational methods and resources used to perform the aqueous-RNA and LDH-RNA simulations reported in this Chapter. The choice of interaction potentials needed to describe the clay-organic interactions is described in more detail in Chapter 4.

The sizes and sequences of RNA polyanions used in this study are given in Table 7.1. Table 7.1 uses standard notation for displaying RNA sequences, in which the strand is displayed in the 5′-3′ direction showing individual nucleotide bases as A, U, C & G for adenine, uracil, cytosine and guanine respectively. The RNA strands were chosen as they have been well characterised in previous molecular dynamics (MD) simulations [132]. Sequence **A** is a 25-mer polyadenine RNA. Sequences **B** and **C** are both 25-mer strands with mixed base sequences taken from the hammerhead ribozyme which exhibits secondary structural features [256]. All nucleic acids were built using the Nucleic Acid Builder, which is part of AmberTools 1.5. All phosphate groups were un-protonated and carry a negative charge as we assume a pH of 7 for which all phosphate groups are known to be ionized. Since the conditions on the early Archaean Earth remain unknown, Models were built replicate the conditions of previous simulations which work with a pH of 7.

The LDH structure used in this study was obtained by the refinement of powder X-ray diffraction (PXRD) data on hydrotalcite using Rietveld methods and has an initial structure with chemical formula $[Mg_2Al(OH)_6]Cl\cdot_2H_2O$ (see Figure 7.1) [81]. Two additional LDH surfaces were constructed by altering the isomorphic substitutions present in the LDH sheet to produce $[Mg_3Al(OH)_6]Cl_3\cdot2H_2O$ and $[Mg_6Al(OH)_6]Cl_9\cdot2H_2O$.

Water molecules were placed above the surface of the LDH within the periodic simulation cell, with a perpendicular length of 140 Å to the adjacent periodic image layer, sufficient to decouple the mineral layer from its periodic image in the z-direction. This represents either an external surface of the LDH. The LDH was charge balanced using the anionic nucleic acid and chloride ions. Chapter 6 investigates the effects of a cationic montmorillonite clay surface on the structure and stability of RNA [132]. As mentioned earlier, the cationic montmorillonite is a 2:1 clay consisting of an octahedral alumina sheet sandwiched between two tetrahedral silica sheets (see Figure 7.1). Partial substitutions in both sheets give rise to a net negative surface charge. The present chapter compares and contrasts the differing behaviour of RNA on cationic and anionic clays.

The models were minimized using the steepest-descent method. A thermalisation step was run, heating the system from 50 K to 300 K over 1 ps. The system was allowed to evolve and equilibrate using the NPT ensemble. Equilibration of models which are dominated by electrostatic interactions, like those reported in this chapter, can be difficult due to the time required to traverse the local minima associated with these interactions. Equilibration was deemed to have been established by monitoring the potential energy of the system as a function of time and by looking for the cell dimensions to take on
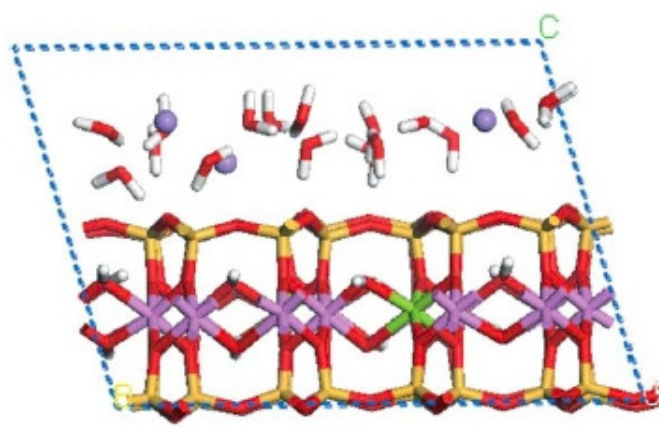
(a)
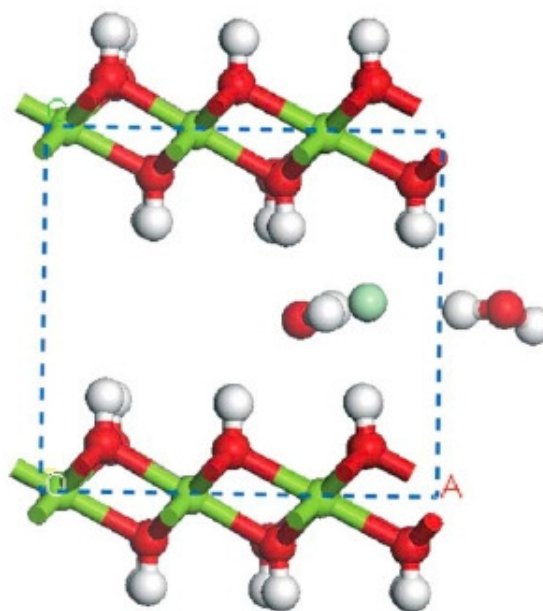
(b)

Figure 7.1: Structure of (a) the cationic clay mineral sodium montmorillonite interca-lated with water and sodium ions and (b) the anionic MgAl layered double hydroxide intercalated with water and chloride ions. The colour scheme is O (red), Si (orange), Cl (green), Na (brown), H (white), Mg (pink) and Al (light green). Blue dashed lines indicate periodic cell boundaries.

stable average values. Twelve models were simulated in total. Each model was simu-lated in production mode for ten nanoseconds following equilibration as shown in Table 7.1. This is sufficient to demonstrate equilibration in our studies [69, 146], it should be noted that the models may become trapped in energy minima on such molecular
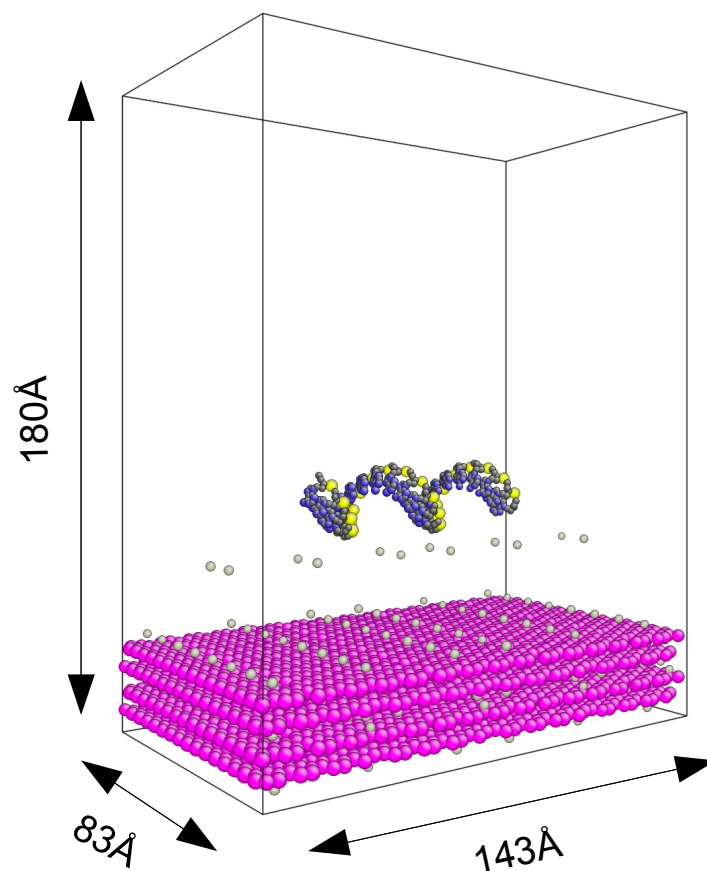
Figure 7.2: Starting structure of model **A**-Mg_2Al (see Table 7.1) which consists of an Mg_2Al LDH surface with RNA. Atom colours: oxygen, red; hydrogen, white; nitrogen, blue; phosphorus, orange; aluminium, silver; magnesium, light green; chlorine, dark green. Water molecules have been hidden to increase clarity. Periodic boundaries are imposed in xyz direction

dynamics timescales. Extended computational work, including much longer timescales and ensembles of many MD simulations (of relatively short duration) would enable us to probe further the general significance of our findings, but at far greater computational cost.

In order to perform the study reported here, very substantial computing resources were needed. To this end the XSEDE[1] supercomputing infrastructure was utilised in the USA which allowed access to Ranger at the Texas Advanced Computing Center (TACC) and Kraken at the National Institute for Computer Science (NICS). As well as machines on

---

[1]https://www.xsede.org/home

(a)   (b)   (c)
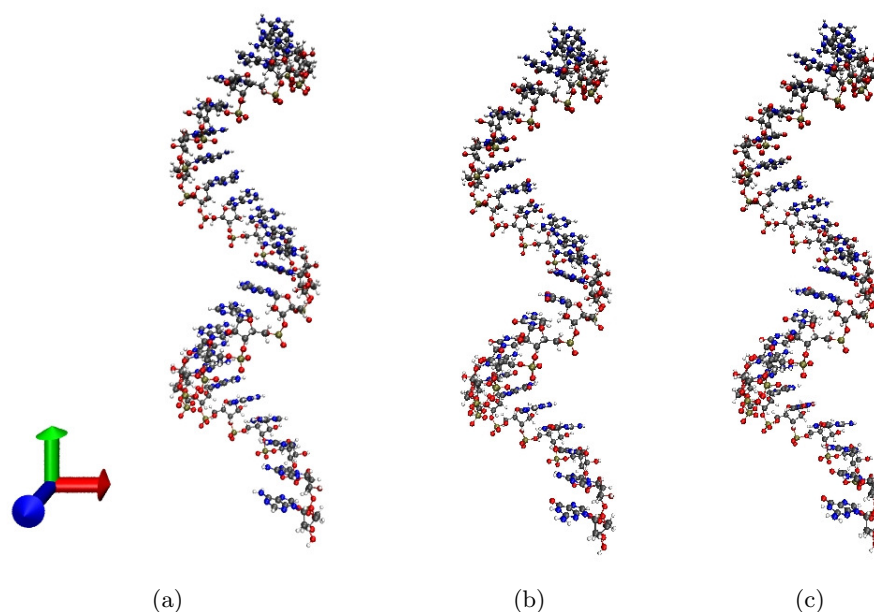
Figure 7.3: Starting structures of RNA sequences (a) **A**, (b) **B** and (c) **C**. Sequence **A** is a 25-mer polyadenine RNA. Sequence **B** and **C** are both 25-mer strands with mixed base sequences taken from the hammerhead ribozyme which exhibits secondary structural features [256] (See Table 7.1).

the XSEDE, local computers at UCL were invoked, including the Centre for Computational Science's Linux cluster, Mavrino, UCL Research Computing's Legion machine and the UK's national super computing resource HECToR.

7.3 Results

The results of the RNA/LDH surface and RNA/montmorillonite surface nanohybrid simulations are discussed in this section. For the former, by comparison to the RNA-aqueous system (Models **A**-Bulk, **B**-Bulk and **C**-Bulk), a consideration of i) the adsorption and interaction of RNA with the LDH mineral surface interface and look at how this affects the structure and stability of the polynucleotide molecules; ii) the structure of the water at the mineral interface and around the RNA; iii) the hydrogen bonding structure between the biomolecule and the mineral surface/water. This section subsequently examines how a change in the LDH surface charge density modifies the RNA/LDH interaction.

A comparison is made between the most similar LDH charge case with that of Ca^{2+}-montmorillonite-RNA, where the surface of the mineral is a silicate rather than hydroxide and in which electrostatic charge interactions occur via bridging cations to enable the anionic RNA to bind to the mineral.

7.3.1 Adsorption of RNA at the LDH mineral surface interface

The atomic density profile, shown in Figure 7.4, displays the density of atoms as a function of distance along the z-axis (perpendicular to the plane of the mineral sheet). The density profile clearly shows the two LDH layers from the well-defined aluminium and magnesium peaks at 5 Å and 25 Å. Water and charge balancing chloride ions occupy the region between the two well seperated LDH layers in the periodic box. The interlayer region can be seen between 10 and 20 Å in Figure 7.4. Water forms a mono-layer at the LDH surface which is shown by the two oxygen density maxima within the interlayer. Hydrogen atoms face away from the hydroxyl surface, producing a large hydrogen density peak at approximately 15 Å. At the exposed aqueous surface the formation of a bilayer of water is seen indicated by the large oxygen atom densities at around 35 and 40 Å. Beyond the 40 Å water layer a density profile characteristic of bulk water is shown, exhibiting no long range order. The atomic density profile also shows phosphorus atoms in close proximity to the exposed mineral surface, suggesting a strong electrostatic attraction between the charged phosphate groups which make up the backbone of the RNA strand and the oppositely charged mineral surface. The density profile implies that the RNA strand arranges such that the phosphate groups are orientated towards the surface while the hydrophobic base groups are exposed to the bulk water region.

Many of the properties suggested by the atomic density profile are reinforced by direct visualisation of the model. Figure 7.5 (a) shows the RNA strand, of sequence **A** in model **A**-Mg_2Al, adsorbing to the LDH surface and adopting a configuration in which the phosphate backbone faces the positively charged surface while the adenine base groups are orientated normal to the surface towards the bulk water region. The orientation of RNA with respect to the mineral surface is depicted as a schematic diagram in Figure 7.5 (b). The visualisation in Figure 7.5 (a) shows that the RNA strand has undergone a change in conformation: it has unwound from its helical-like starting structure, which model **A**-Bulk shows is conserved in bulk water over the 10 ns timescale of the simulation.

The structure and arrangement of RNA at the LDH mineral surface, observed in these simulations, give us an indication of the mechanism by which RNA based therapeutics are bound and protected when an LDH is used as a drug delivery vehicle [257].

Images of the three different RNA sequences in Figure 7.6 show the full extent to which the mineral surface has altered the RNA strands' structures. All three sequences, when interacting with a Mg$_2$Al LDH, show a deviation from the helical-like initial structure, shown in Figure 7.2, towards a more elongated configuration. Figure 7.6 shows all three sequences in this elongated structure with conserved linear regions. The free energy penalty arising from adopting this new linear type configuration is overcome
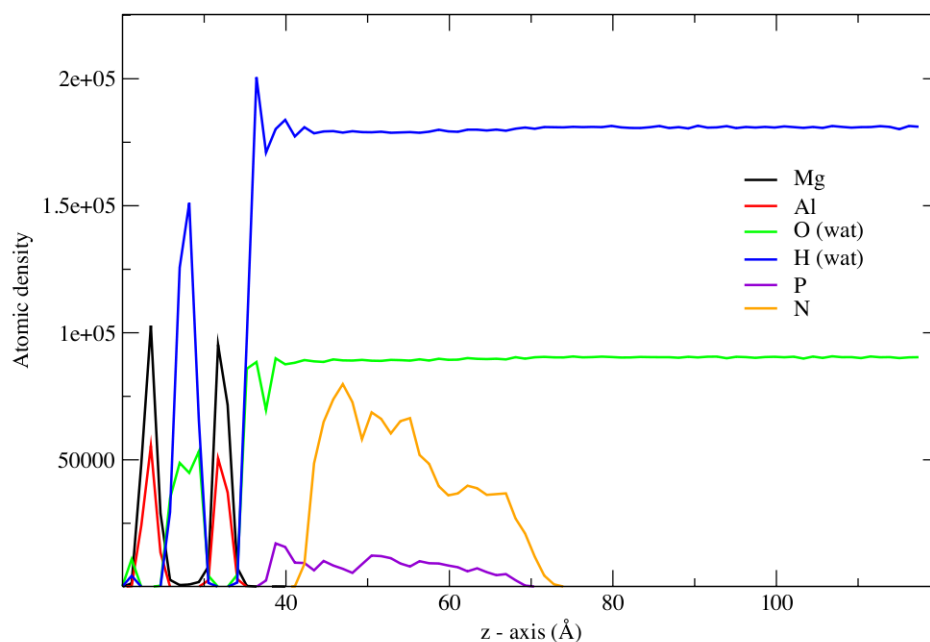
Figure 7.4: Atomic density profile of model **A**-Bulk produced using time-averaged data from the last 4 ns of molecular dynamics simulation. The atomic density plots for both nitrogen and phosphorus are scaled by 100 in the y-direction to provide better clarity. The profile shows strong, well-defined, aluminium and magnesium peaks at 24 Å and 32 Å which indicate the positions of the LDH sheets. The large oxygen (and hydrogen) peak at 36.5 Å indicates a monolayer of water at the LDH surface, and a double layer at 40 Å. The phosphorus peak at 39 Å and the nitrogen peak at 47 Å reveal that the RNA strand aligns in such a way that the phosphate groups face the mineral surface and the nitrogen-rich base groups are exposed to the aqueous region.

by the electrostatic attraction between the positively charged mineral surface and the negatively charged phosphate groups along the backbone of the RNA strand.

It is clear that the LDH mineral interface significantly alters the conformation of RNA when compared to the corresponding structure in bulk water, which is conserved over the duration of the simulation. RNA is concentrated on these clay surfaces, to which it is pinned by strong electrostatic forces (as shown in Figure 7.5), while facilitating templating with complementary nucleic acid bases. The exposed base groups aligned normal to the mineral surface are well positioned to form hydrogen bonds with incoming mononucleotides. Orgel *et al.* originally showed the oligomerisation of activated mononucleotides using polycytosine and polyuracil homopolymers as templates for these reactions [258]. Further studies are needed to fully understand if and how a complementary monomer or polymer of RNA would template to the adsorbed RNA.

(a)

(b)
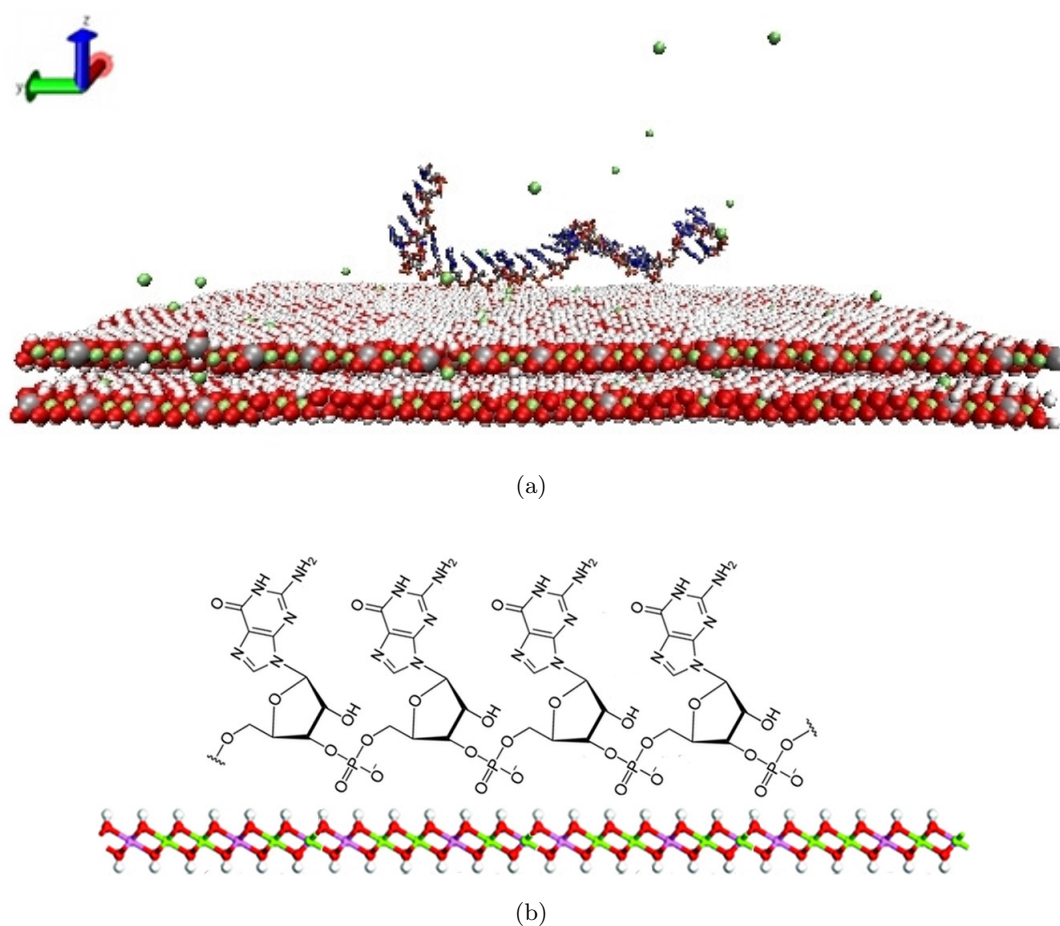
Figure 7.5: (a) Final snapshot of model **A**-Mg_2Al after 8 ns of production molecular dynamics simulation. Colour scheme is the same as in Figure 7.2. The visualisation shows the RNA interacting strongly with an Mg_2Al LDH surface through the phosphate groups, leaving the hydrophobic base groups exposed to the aqueous region. (b) Schematic diagram displaying the orientation that the RNA adopts when interacting with the LDH surface, shown from simulation in (a).

The radial distribution function with respect to phosphorus atoms, displayed in Figure 7.7, shows water coordinating the P atoms at 2.6 Å. The LDH surface hydroxyl groups are seen to reside at approximately 5 Å from the phosphorus groups and the LDH metal ions at 7 Å. The radial distribution function shows a slightly increased peak for aluminium over magnesium, presumably because Al^{3+} confers a net positive charge on the mineral; as a result the phosphate groups have a higher affinity for these ions.

7.3.2 Comparison with RNA-cationic clay

Results obtained from the simulation of LDH-RNA systems compared with a further analysis of montmorillonite-RNA systems simulated in Chapter 6 [132] is given here. The main difference, as far as the periodic simulation cells are concerned, is the replacement
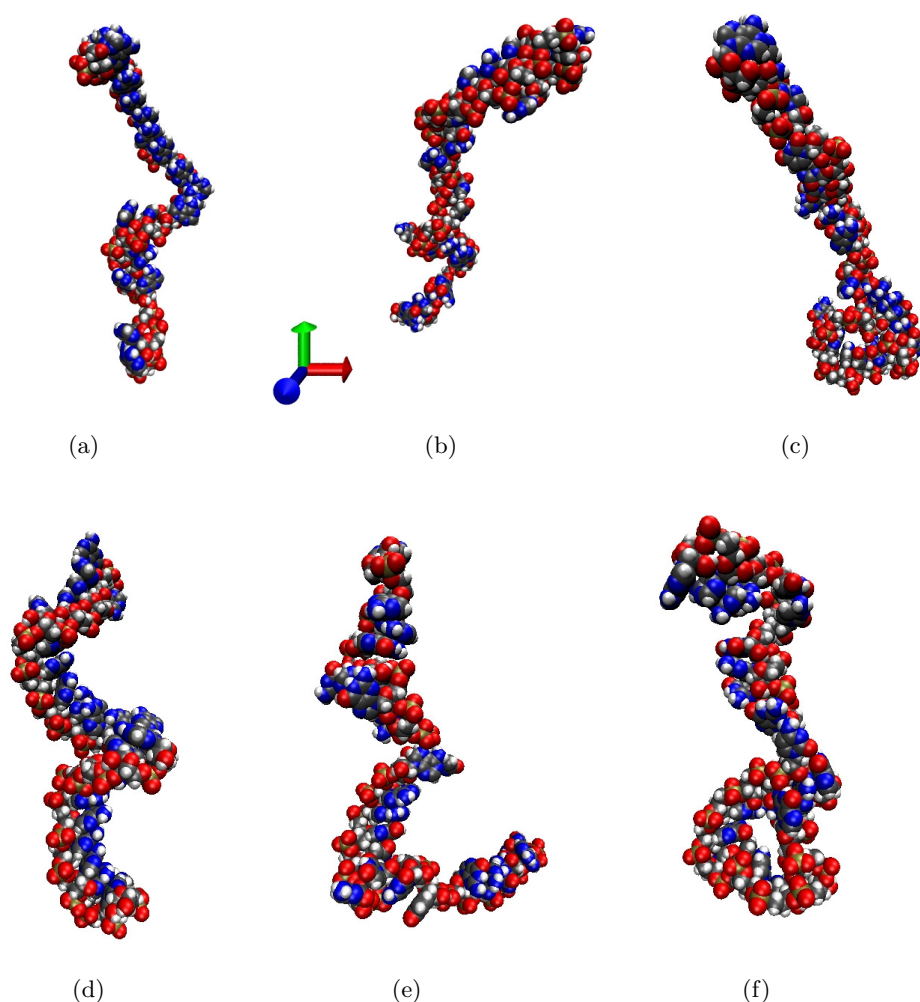
Figure 7.6: Snapshots of RNA taken after 5 ns of production molecular dynamics simulation for (a) sequence **A** (model **A-**Mg_2Al, (b) sequence **B** (model **B-**Mg_2Al) and (c) sequence **C** (model **C-**Mg_2Al), all of which are interacting with an Mg_2Al LDH (not shown). Models (d)-(f) are the corresponding sequences in bulk water, models **A**-Bulk, **B**-Bulk and **C**-Bulk respectively. Strands (a)-(c) show sequences with elongated portions over the corresponding conserved helical structures in bulk water (d)-(f) which better resemble the initial structure. Colour scheme as in Figure 7.1.

of the hydroxide layer by the aluminosilicate layer. Though montmorillonite layers are negatively charged, through having Ca^{2+} as charge balancing ions, the montmorillonite layers effectively interact as positively charged layers.

The results shown here for single-stranded RNA interacting with an exposed aqueous anionic surface are very different to those reported for the same single-stranded RNA sequences interacting with an aqueous montmorillonite cationic clay surface [132]. When interacting with the montmorillonite surface the RNA strands undergo a hydrophobic collapse into a globular-like structure where the RNA bases become buried within the charge of the phosphate backbone. The globular structure then forms characteristic
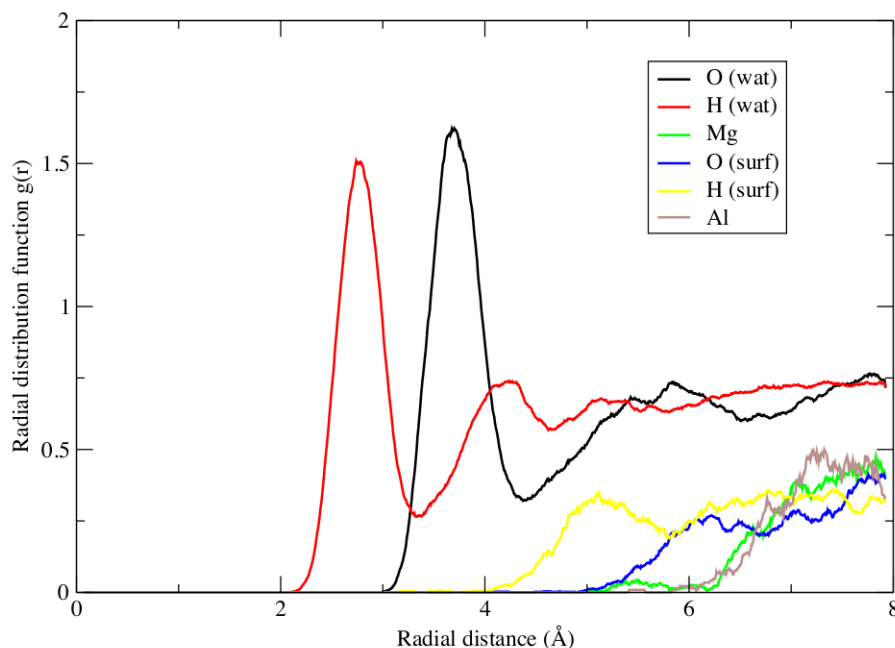
Figure 7.7: Radial distribution functions of atoms (shown in legend) with respect to phosphorus for the RNA sequence **A** within the Mg_2Al LDH in model **A**-Mg_2Al.

secondary structures, which are stabilised by hydrogen bonding and base-stacking. The extent to which the RNA folds at the cationic montmorillonite interface is shown in Figure 7.8. By contrast, RNA responds very differently at an LDH surface. Instead of folding, the elongation of the RNA molecule, due to the electrostatic penalty that would be paid in an anionic (chloride) environment is observed. The calcium ions associated with the montmorillonite clay effectively screen the phosphate charge, allowing the RNA to collapse.

While it appears that Ca^{2+}-montmorillonite and Mg_2Al LDHs both effectively adsorb RNA at the mineral surface, these minerals have contrasting effects on the conformation and arrangement of these biomolecules. While the Ca^{2+}-montmorillonite clay strongly promotes folding of the RNA structure the LDH prevents folding, possibly nullifying any enzymatic activity the RNA may possess. Instead, LDH arranges RNA in a way that may promote elongation reactions through a templating mechanism which makes LDH a prime candidate for amplifying the population of a known RNA sequence. It should be noted that the folding of 25-mer RNA sequences on clays as studied by us [132] does show folding to known secondary structural motifs.

(a)   (b)
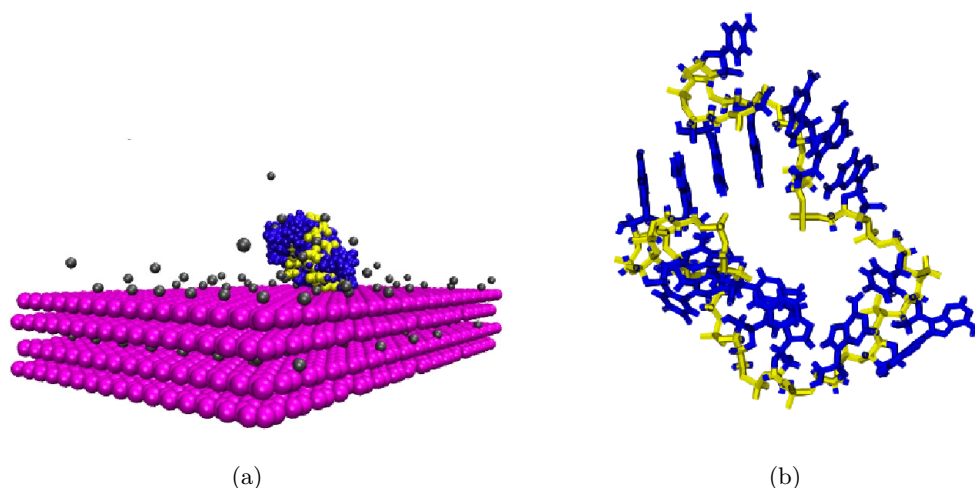
Figure 7.8: (a) Snapshot of RNA, of sequence **B** in Table 7.1, interacting with divalent calcium ions, water and a montmorillonite mineral surface after 30 ns of MD simulation. (b) Closeup of the folded RNA secondary structural motif in (a). Free energies indicate that (b) is a fold along the folding pathway to the native stemloop secondary structure. The folded RNA structure is taken from Chapter 6 [132].

Previous simulations have suggested how single-stranded RNA may adsorb to montmorillonite surfaces via a tethering mechanism in which the nucleotide base at the 3′ end of the RNA attaches to the surface with the rest of the strand aligned perpendicular to the surface, exposed to the aqueous region [132]. RNA interacts with the oppositely charged LDH surface in a different way: it is the charged phosphate group which is adsorbed to the surface, with the strand lying flat, parallel, to the surface. The interaction of RNA with anionic and cationic clays is analogous to the interaction of amino acids with these two types of clay minerals. Simulations by Newman *et al.* suggest that the anionic amino acid (S)-phenylalanine arranges in a similar manner to RNA with an Mg_3Al LDH; the amino acid interacts with the surface in such a way that the amino acid faces towards the aqueous region [259]. Newman *et al.*'s simulations also show that the cationic (S)-tyrosine amino acid arranges in an analogous way to RNA with the hydrophobic side-chain adjacent to the montmorillonite surface.

The following section describes the effect of changing the surface charge density of the LDH has on the adsorption and surface interactions of the different RNA sequences.

7.3.3 RNA - surface bonding

An important factor which controls the structural stability of the LDH is the development of a surface bonding network between the metal hydroxide LDH layers, water

molecules and intercalated organic molecules.

The electrostatic attraction between the charged phosphate groups in RNA and the charged metal sites in the LDH surface sheet allows the RNA to closely approach the LDH surface which in turn admits inner-sphere complexes to be established between the phosphate oxygen atoms and the metal hydroxide hydrogen atoms (see Figure 7.9). Inner sphere complexes form when ions bind directly to the surface without any bridging water molecules to mediate the adsorption. This is where the complexation is established, as opposed to RNA interacting with negatively charged montmorillonite via mediating cations.
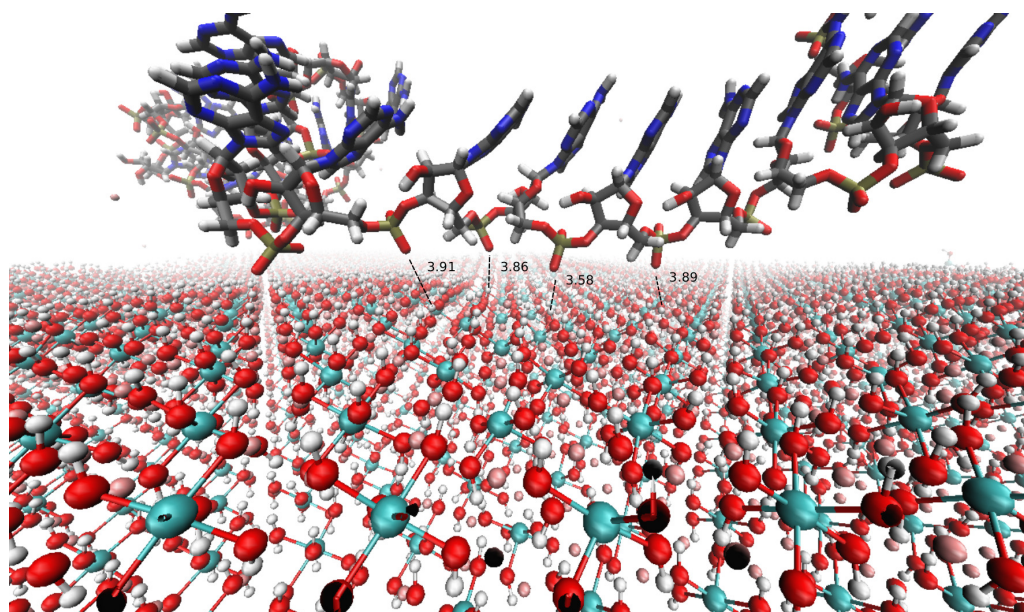
Figure 7.9: Simulation snapshot depicting the inner-sphere complexes taking place. The numerical values shown in black show the distance between interaction donors and acceptors in Ångstroms. Colour scheme: O(red), H(white), P(gold), C(grey) and N(dark blue).

The enhanced interaction gained through the inner-sphere complexation adds considerably to the adsorption of the RNA to the mineral surface. Due to the orientation of approach the RNA has to the LDH and the type of bonding that takes place the RNA is effectively pinned to the LDH by each phosphate group which fixes the RNA in a position which appears to facilitate the templating of a complementary strand through hydrogen bonding [260]. As mentioned earlier, further studies are needed to fully understand if and how a complementary strand would template to the adsorbed RNA.

7.3.4 Effect of varying surface charge density on adsorption of RNA

In Figure 7.10 the radial distribution function with respect to phosphorus for RNA sequence **A** interacting with Mg_3Al and Mg_6Al LDHs is displayed. These can be compared to that of the same RNA sequence for the Mg_2Al LDH in Figure 7.7 in order to ascertain how differing surface charge densities alter the RNA-LDH interaction. The radial distribution of water around the phosphorus atoms is the same for each of models **A-**Mg_2Al, **A-**Mg_3Al and **A-**Mg_6Al. The radial distribution of surface alumina ions around the phosphorus atoms differs as the surface charge density increases; a peak emerges at 3 Å indicating that the RNA has adsorbed onto the LDH surface and that phosphorus atoms are occupying space in close proximity to the alumina sites of the LDH sheet. The radial distribution shows an Al peak at 5 Å for both models **A-**Mg_3Al and **A-**Mg_6Al. The radial density for the LDH surface hydroxyl hydrogen atoms with respect to phosphorus show peaks at 5, 5.5 and 6 Å for Mg_2Al, Mg_3Al and Mg_6Al respectively. Thus, the greater the surface charge density the closer the RNA can approach. In this case Mg_2Al has the greatest charge density of the models tested, and exhibits the closest binding of RNA to the mineral surface. With this in mind, Mg_2Al would prove a better delivery vector for RNA based drugs through increased adsorption/protection over LDHs with less surface charge density e.g. Mg_3Al and Mg_6Al. The increased surface charge density would also lead to a higher loading capacity of RNAs and as a consequence endow the LDH drug vector with a more effective anion exchange capacity.

The radius of gyration, shown in Table 7.2, indicates how the size of the RNA molecule changes over the last 4 ns of simulation. The smallest averaged radius of gyration is found in RNA in bulk water. The helical starting structure of RNA, similar to that in Table 7.2, is conserved in the bulk water simulations and corresponds to the RNA with the smallest effective size compared to the RNAs in LDH models. The LDH adsorbed RNAs all show larger radii of gyration which confirm their elongated configurations when interacting with the surface. The RMSD shows that the RNA in bulk water deviates the least from the initial helical crystal structure due to the conservation of the initial structure from strong base-stacking interactions, shown in Figure 7.3, compared to the LDH-bound RNAs. The radii of gyration values do not show an increasing trend. A possible reason for the smaller radius of gyration values for RNA at the Mg_6Al LDH is that there is a decreased number of Cl^- counterions as compared to Mg_2Al and Mg_3Al LDHs.

The ionic environment in the vicinity of the surface of Mg_2Al LDH is greater than that of Mg_6Al LDH which means there is a greater electrostatic penalty for neighbouring phosphate groups. The difference in ionic environment caused by counter cations has been suggested as a possible reason for the difference in catalytic activity of various

montmorillonite clays by Joshi *et al.* [162]. This implies that Mg$_2$Al LDH may be a better surface for the adsorption of RNA compared to Mg$_3$Al and Mg$_6$Al LDH.

The end-to-end distance of each RNA was taken as an estimate of the linear distance between phosphorus atoms at each exposed end of the RNA sequence, as a function of time, and is displayed in Table 7.2. End-to-end distances provide an indication of the extent to which RNA is folding; shorter distances indicate that more folding has occurred. All RNA/LDH models exhibit an end-to-end distance which is comparable to that of their corresponding sequence in bulk water, except for models **B**-Bulk and **C**-Bulk. These two models show an increased RNA end-to-end distance over the same sequences in bulk water. This suggests that the RNA helical-like conformation is unwinding, giving a more elongated structure. Although the end-to-end distances are similar for most of the LDH bound RNAs, the RMSD shows that the structures of the RNA sequences in the LDH environment are far removed from their bulk water counterparts which are conserved over the 10 ns simulations (shown in Table 7.2). Figure 7.6 indicates that the differences in RMSD between RNAs arise from the LDH environment. Each RNA, when interacting with the LDH surface, forms elongated segments through the central stretches of its sequence. The exposed ends of the RNA sequence then tuck-in towards the rest of the molecule giving somewhat smaller than expected end-to-end distances for sequences which nevertheless exhibit elongated segments.

For a given LDH charge density, variations in the base pair sequences lead to different structures. This effect is observed in the varying structures shown in Figure 7.6 and shown for a given cationic clay in [132]. Put another way, the mineral surface interacts differently with each RNA oligomer as a function of its sequence (i.e. depending on the genetic information contained in the biomolecule), due at least in part to different base stacking interactions in the sequences studied. Extensive sampling across different sequences have not been performed. The relatively long time-scale of the simulations suggests this may be a genuine effect and one that merits further investigation in the future. Such an effect would have implications for the role of charged, layered minerals in prebiotic chemistry.

7.4   Conclusions

Molecular dynamics computer simulations have been used to probe the interaction of various aqueous RNA sequences with LDH of differing charge density at the mineral surface. The simulations provide detailed insight into the arrangement of water and counter-ions and suggest possible RNA-surface interactions. This information is particularly useful since it is not generally available from experiment for these systems due

to their poor crystallinity and disordered nature. The simulations demonstrate that RNA adopts significantly different conformations and interacts with LDH surfaces in very distinct ways than it does with montmorillonite surface.

The structure and arrangement of RNA at the LDH mineral surface, observed in these simulations, support the currently accepted mechanism for the LDH-mediated transfection of mammalian cells with nucleic acids. This mechanism holds that once the RNA is tightly bound/adsorbed through strong attractive electrostatic interactions between the surface and negatively charged RNA phosphate groups, the cationic LDH mineral is attracted to the cell surface where the RNA-LDH is taken-up via receptor-mediated endocytosis [182, 257]. Once inside the cell, the lower pH in the endosome causes the LDH to dissolve, thus delivering the nucleic acid to the cell.

According to our simulations, while cationic montmorillonite clays promote folding of the RNA structure, LDH prevent folding. The folded tertiary structure induces the catalytic capacity of the RNA ribozymes; unfolding inhibits catalytic activity. From simulation, the LDH arranges the RNA in a way which might promote a templating mechanism, making LDHs a potential candidate for amplifying the population of RNA sequences.

The findings confirm that the distance between LDH charge sites is equidistant to that of the phosphate groups of the RNA [189]; this permits strong docking of RNA onto the LDH surface. The insight gained from these many large-scale molecular simulations run for tens of nanoseconds adds considerably to the interpretation of experimental data through provision of highly detailed molecular level insight about the way RNA interacts with LDH mineral surfaces.

The interactions described here between RNA and LDH have relevance to origins of life studies, as the arrangement of single RNA strands in this manner might have made them amenable to Watson-Crick templating as part of an early hereditary process [244, 245].

Model	Sequence	RNA	Number of Atoms	Clay Surface	Starting Cell Dimensions (Å³)	Simulation time (ns)
A-Mg_2Al		(5′-AAAAAAAAAAAAAAAAAAAAAAAAA-3′)	284,706	Mg_2Al	82.74 × 143.28 × 180.62	10.0
A-Mg_3Al		(5′-AAAAAAAAAAAAAAAAAAAAAAAAA-3′)	284,607	Mg_3Al	82.74 × 143.28 × 180.62	10.0
A-Mg_6Al	**A**	(5′-AAAAAAAAAAAAAAAAAAAAAAAAA-3′)	284,472	Mg_6Al	125.52 × 177.26 × 125.68	10.0
A-Bulk		(5′-AAAAAAAAAAAAAAAAAAAAAAAAA-3′)	260,274	*None*	125.52 × 177.26 × 125.68	10.0
B-Mg_2Al		(5′-AAAGCAGAGAGUAAAGUCUCUGAGUAG-3′)	284,684	Mg_2Al	82.74 × 143.28 × 180.62	10.0
B-Mg_3Al		(5′-AAAGCAGAGAGUAAAGUCUCUGAGUAG-3′)	284,579	Mg_3Al	82.74 × 143.28 × 180.62	10.0
B-Mg_6Al	**B**	(5′-AAAGCAGAGAGUAAAGUCUCUGAGUAG-3′)	284,444	Mg_6Al	125.52 × 177.26 × 125.68	10.0
B-Bulk		(5′-AAAGCAGAGAGUAAAGUCUCUGAGUAG-3′)	161,655	*None*	125.52 × 177.26 × 125.68	10.0
C-Mg_2Al		(5′-AAAGUCUGGGCUAAGCCCACUGAUG-3′)	284,678	Mg_2Al	82.74 × 143.28 × 180.62	10.0
C- Mg_3Al		(5′-AAAGUCUGGGCUAAGCCCACUGAUG-3′)	284,573	Mg_3Al	82.74 × 143.28 × 180.62	10.0
C-Mg_6Al	**C**	(5′-AAAGUCUGGGCUAAGCCCACUGAUG-3′)	284,438	Mg_6Al	125.52 × 177.26 × 125.68	10.0
C-Bulk		(5′-AAAGUCUGGGCUAAGCCCACUGAUG-3′)	161,169	*None*	125.52 × 177.26 × 125.68	10.0

Table 7.1: Simulation cell compositions and dimensions for all the models studied in this chapter. Each sequence was simulated in the presence of the LDH surface and in bulk aqueous solution. Sequence **A** is a 25-mer polyadenine RNA. Sequence **B** and **C** are both 25-mer strands with mixed base sequences taken from the hammerhead ribozyme which exhibits characteristic secondary structural features.
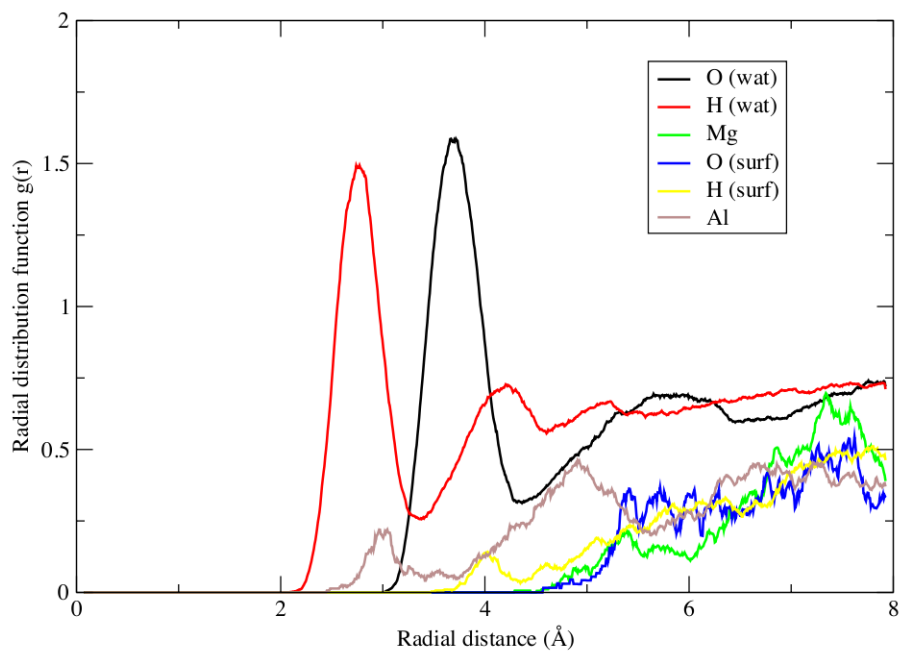
(a)
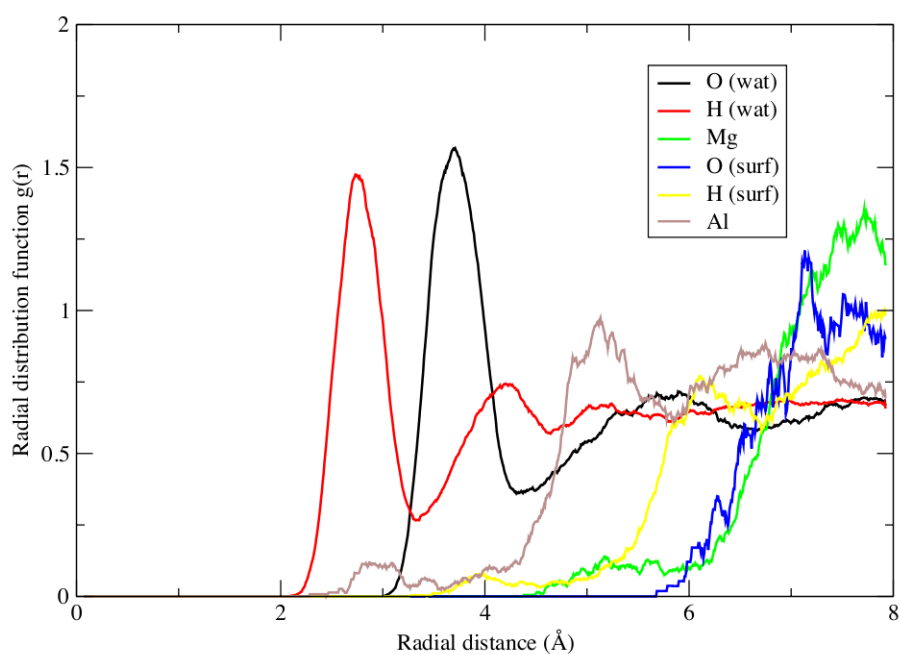
(b)

Figure 7.10: Radial distribution functions of atoms (shown in legend) with respect to phosphorus for the RNA sequence **A** with (a) an Mg$_3$Al LDH in model **A-**Mg_3Al and (b) an Mg$_6$Al LDH in model **A-**Mg_6Al.

Model	R_g (Å)	RMSD (Å)	end-to-end (Å)
A-Mg_2Al	23.914±1.142	6.638±0.601	62.952±5.390
A-Mg_3Al	28.028±1.214	4.788±0.462	75.510±2.949
A-Mg_6Al	25.888±0.516	5.948±0.536	73.420±3.958
A-Bulk	20.502±1.690	1.174±0.472	73.986±4.289
B-Mg_2Al	21.720±0.597	3.320±0.280	53.228±2.258
B-Mg_3Al	25.432±0.737	2.785±0.187	85.772±1.430
B-Mg_6Al	21.131±1.449	3.987±0.222	59.789±4.354
B-Bulk	20.335±1.165	1.182±0.922	68.447±5.217
C-Mg_2Al	25.370±0.862	3.029±0.241	65.421±3.659
C-Mg_3Al	25.231±0.955	7.491±0.698	68.064±3.330
C-Mg_6Al	26.235±1.804	6.493±0.698	81.705±3.563
C-Bulk	20.521±1.376	2.090±0.511	64.685±2.944

Table 7.2: Radius of gyration (R_g), root-mean-squared deviation (RMSD) and end-to-end distances of RNA calculated as the mean dependent variable from linear regression as a function of simulation time taken from the final 4 ns of simulation, displayed with the corresponding standard deviation. Radius of gyration shows that the RNA in the bulk water simulation (model **A**-Bulk) has the smallest size, with RNA on the clay models adopting an elongated structure. The RMSD show that the water model deviates the least from the initial crystal structure shown in Figure 7.3. End-to-end distances were calculated as the linear distance between the phosphorus atom at the 3′ and 5′ ends of the RNA strand. The end-to-end distances show no significant change in the length of the strand, indicating that no folding is taking place; on the other hand the RMSD clearly shows that the RNA has deviated from its crystal structure when interacting with the LDH surface.

Chapter 8

Folding and Catalytic Structure of the Hammerhead Ribozyme from Replica Exchange Molecular Dynamics

Enhanced sampling of replica exchange molecular dynamics is used in this chapter to study the complex folding pathway and the catalytically active structure of the full-length hammerhead ribozyme. This study furnishes molecular level resolution of the structure and mechanism of the hammerhead ribozyme and resolve hitherto open questions related to the catalytic activity of the ribozyme including the role of metal ions in mediating the reaction.

8.1 Introduction

Ribonucleic acid (RNA) was first suggested to be capable of exhibiting catalytic activity by Woese, Crick and Orgel in 1967 after the discovery that RNA could fold to form secondary structures in an analogous way to protein enzymes [155, 261, 262]. It was not until the 1980s that Cech and Altman independently discovered the first ribozymes, for which they won the 1989 Nobel Prize in Chemistry [263, 264].

The hammerhead ribozyme has long been regarded as a prototype for the understanding of the catalytic nature of RNA molecules because of its relatively small length and well characterised crystal structure. Notwithstanding the wealth of experimental work which has gone into the elucidation of the catalytic structure of the hammerhead ribozyme

[265, 266] the precise details of the catalytic mechanism remain unknown and a topic of much debate, as does the role of divalent the Mg^{2+} ions in the catalysis. The metal ions have not been observed crystallographically to bridge the adenine and scissile phosphate so as to promote acid/base catalysis.

The hammerhead ribozyme is derived from a small, self-cleaving genomic RNA discovered in satellites of various plant RNA virus genomes [267, 268]. It catalyses the RNA self-cleavage phosphodiester isomerisation reaction producing two separate RNA strands [269]. The hammerhead itself consists of two RNA strands. Experimentally, the strand which is to be cleaved can be supplied in excess while multiple turnover can be demonstrated and shown to obey Michaelis-Menten kinetics [270], perhaps the best known model for protein enzyme kinetics.

Ribozymes play an integral part in the popular 'RNA world hypothesis' which maintains that a world based entirely on RNA preceded the DNA/protein based one we now live in. The hypothesis holds that at some point in time the more stable nucleic acid deoxyribonucleic acid (DNA) took over the role of informational storage from RNA and that proteins, which are better catalysts than RNA, took over the enzymatic role from RNA. It has been suggested that many of the RNA biomolecules such as viruses still present on the Earth today are molecular fossils from this RNA world. The RNA world hypothesis is attractive because it only requires the prebiotic synthesis of one (bio)molecule rather than three in the case of modern 'DNA makes RNA makes proteins' biology.

Molecular simulation methods provide a wealth of information about ribozyme systems which can often be difficult to characterise experimentally because of the lack in measurement resolution and the complex nature of RNA structure. Up until now only relatively short RNA sequences have been studied using REMD methods to investigate the folding pathway and structure. Both groups of Wang and Paschek have reported REMD simulations of 8-mer RNA sequences in explicit water [271, 272]. Single uncoupled 12 ns MD simulations have been reported for the hammerhead ribozyme, by Lee *et al.* [273]. Single MD simulations tens of nanoseconds in length are relatively short for modern day standards and MD does not sample configurational space to the extent that REMD and metadynamics techniques can. The simulation of protein folding and structure is in many ways an analogous process to RNA ribozyme study. Pandé *et al.* have been leading the way in terms of distributed ensemble simulations of small proteins using the Folding@home network [274]. REMD simulations have been reported for the study of proteins of 8 residues using 20 replicas which is a comparable scale to the REMD studies being performed on RNAs [272].

Origins of life studies lends itself well to theoretical and simulation studies given the vast span of time which separates us. This is the first case of large-scale replica exchange sampling performed on petascale computers to investigate processes relating to origins of life studies, and at an unprecedented scale for ribozyme simulation.

8.2 Methods

This section describes the large-scale computational methods and resources used to perform the aqueous full-length hammerhead ribozyme folding simulations reported in this article, including the computational sampling algorithms used. This Chapter uses fully atomistic, large-scale replica exchange molecular dynamics techniques to simulate these flexible systems. Distributed high performance supercomputers, Kraken & JUGENE, coupled to local resources at University College London (UCL) were utilised in order to perform these simulations and their associated data analysis.

8.2.1 Model construction & potential parameterisation

The starting structure of the hammerhead ribozyme was modified from the 2.2 Å resolution crystal structure of the full-length catalytically active hammerhead ribozyme (PDB 2GOZ and NDB ID: UR0084), See Figure 8.1. The full-length hammerhead ribozyme consists of two RNA chains 43 and 20 nucleotides in length. The chains form three base-paired helix-like stems. 20 Å of SPC/E water and 50 charge balancing Na^+ ions were placed around the ribozyme using the *solvate* function in AmberTools. Crystal waters molecules were kept along with the divalent metal ion sites which were populated with 5 Mg^{2+} ions. The complete model consists of 54210 atoms and has unit cell dimensions of 83.6 x 73.8 x 95.4 Å. The RNA sequences are as follows, with each letter representing a base (A=adenine, U=uracil, G=guanine and C=cytosine):

Chain A: *GGAUGUACUACCAGCUGAUGAGUCCCAAAUAGGACGAAACGCC*
Chain B: *GGCGUCCUGGUAUCCAAUCC*

The bonding interactions and partial charges for the RNA ribozyme molecule were described using the Amber parm.rna.ff99 force field parameters [124]. The parmbsc0 refinement was used to provide an improved description of the α/γ concerted rotations within the RNA molecule [125]. The Amber forcefield has been used extensively to simulate nucleic acids and reproduces the structure and dynamics of nucleic acid moieties, including Watson-Crick base pairing [126, 132, 133].
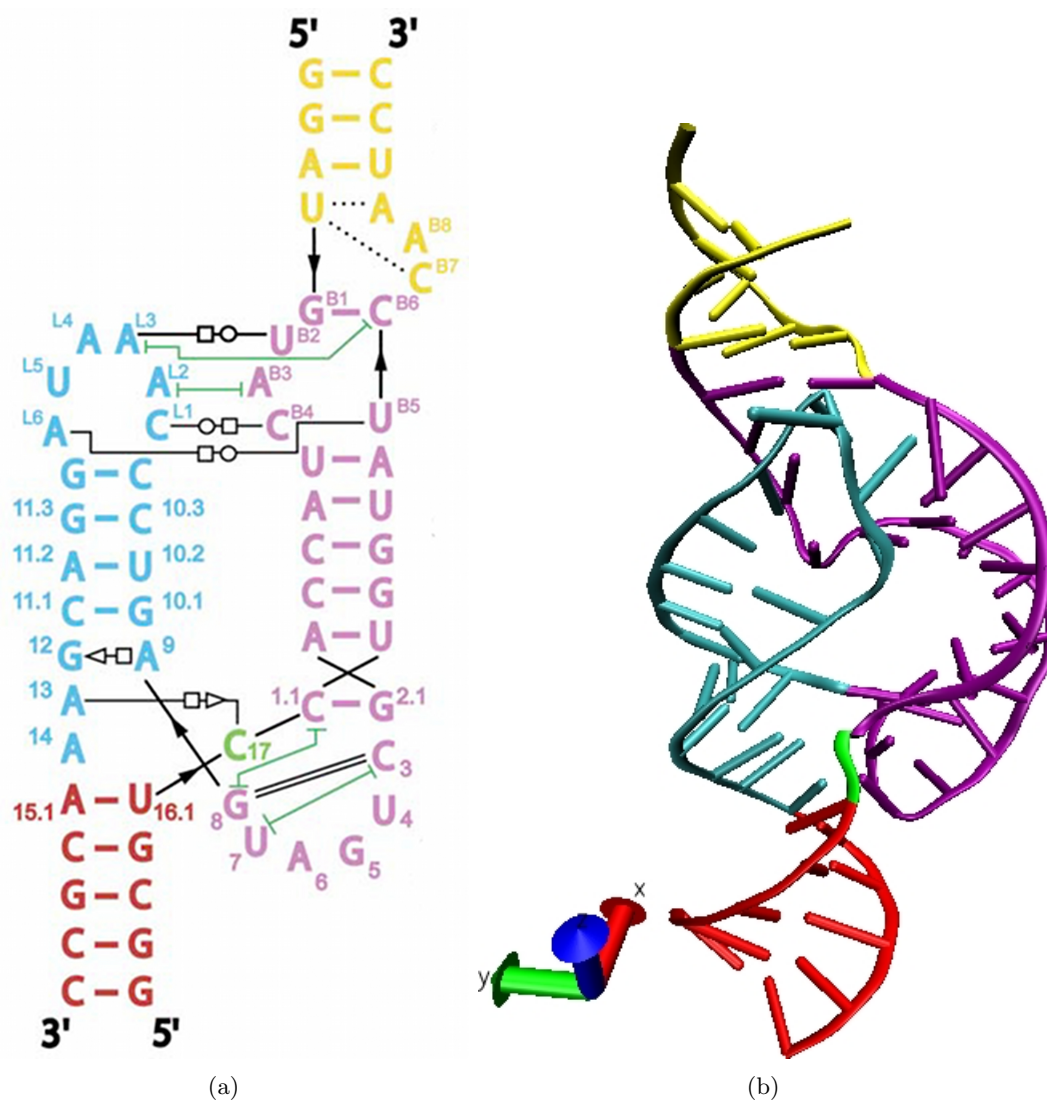
Figure 8.1: (a) Schematic representation of the secondary structure of the full-length hammerhead ribozyme and (b) the three-dimensional representation taken from the crystal structure. Annotations in (a) follow the scheme developed by Yang *et al.*[275]: canonical base-pairing is shown as coloured lines and hydrogen base pairing is shown as black lines. Open circle next to open square = Watson-Crick/Hoogsteen; open square next to open triangle = Hoogsteen/sugar edge; dashed line = H-bond; green line = non-adjacent base stacking. The thick black line shows the direction and continuity of the backbone. The colour scheme depicts Stem I in yellow and purple, Stem II in blue, Stem III in red and the scissile phosphate group in green [266].

The SHAKE algorithm was employed in order to constrain bonds and angles within hydroxyl groups and water [139] (see Section 4.2.1). This allowed us to use a timestep of 2 fs within the molecular dynamics algorithm.

8.2.2   Replica exchange molecular dynamics techniques

All simulations reported here were run using the Large Atomistic/Molecular Massively Parallel code (LAMMPS) developed by Plimpton *et al.* [136]. The code exhibits a near linear scaling relationship between the number of cores used and the speed-up in wall-clock time.

Before production simulations, the starting structure was energy minimised using the steepest descent method, then thermalised with molecular dynamics from 0 K, 0 atm to 300 K and 1 atm using the isothermal-isobaric ensemble.

The high level of conformational sampling in this study was achieved using the Replica Exchange Molecular Dynamics (REMD) algorithm [144]. The REMD method has been widely used in the computational study of protein folding [276], and also the analogous process of RNA folding [271, 272, 277]. REMD consists of M non-interacting copies (or, replicas) of the original system in the canonical ensemble at M different temperatures T_m ($m = 0, ..., M - 1$). The replicas are arranged so that there is always one replica at each temperature. The trajectory of each independent replica is computed using molecular dynamics (MD). After a certain number of MD steps, adjacent replicas (replicas i and $i + 1$) exchange temperatures according to the probability distribution:

$$ P_{ex} = \begin{cases} 1 & \text{for } U_{i+1} < U_i \\ exp[(\beta_i - \beta_{i+1})(U_i - U_{i+1})] & \text{for } U_{i+1} > U_i \end{cases} $$

β_i is equal to $1/k_B T$ where k_B is the Boltzmann constant; U_i is the instantaneous total potential energy of replica i. Typically the temperature spacing is selected so that $P_{ex} \approx 20 - 30$ % for all replicas [278]. REMD enhances the sampling of configurational space and provides a way of obtaining equilibrium properties of a system described by complicated energy landscapes [144].

100 replicas of the hammerhead ribozyme system were simulated, ranging in temperatures from 300.00 K to 570.94 K. An exponential temperature spacing was selected in order to achieve an exchange frequency of 23% which is essentially the same over all replicas, as shown in Figure 8.2.

Each replica was simulated for 50 ns which equates to 5 μs of fully atomistic molecular dynamics simulation for the entire ensemble of replicas. The equations of motion were integrated with a timestep of 2 fs and periodic boundaries were imposed in all three spatial directions. Electrostatic interactions were computed using the particle mesh Ewald method. A clustering process was applied to the large dataset of structures in order to group mutual similarity and reduce the complexity of the data, and to identify

a set of identifiable structurally/conformationally distinct clusters in terms of which the sampling can be understood. k-means clustering based on the cartesian coordinate RMSD between structures simulated from 300.00 to 307.90 K with a fixed radius of 4 Åwas applied. The k-means clustering was performed using the MMSTB (Multiscale Modeling Tools for Structural Biology) toolset [279].

The radius of gyration, r_G, estimates the molecule's effective size during the simulation:

$$r_G = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (\mathbf{r}_i - <\mathbf{r}>)^2}, \tag{8.1}$$

where $<\mathbf{r}>$ is the mean position of all N RNA atoms. Principal component analysis (PCA) was used to describe the main changes in RNA conformation with time. PCA is widely exploited to reduce the dimensionality of an MD trajectory and it identifies the dominant collective modes of motion of molecules. In a Cartesian coordinate system, the covariance matrix can be defined as:

$$\mathbf{C} = <(\mathbf{r}- <\mathbf{r}>)(\mathbf{r}- <\mathbf{r}>)^T>, \tag{8.2}$$

where $\mathbf{r}$ represents the atomic positions of the RNA molecule in a $3N$ dimensional configuration space, $<\mathbf{r}>$ is the mean position of atoms over all snapshots, and the superscript T denotes the matrix transpose. In PCA, the eigenvectors and corresponding eigenvalues of $\mathbf{C}$ are found by diagonalisation of the covariance matrix. The eigenvectors denote the orthogonal modes of motion and the eigenvectors with the largest eigenvalues dominate the dynamics of the system.

8.2.3 Convergence verification

In order to ascertain whether our REMD simulations have achieved reasonable convergence the difference between the free energies of states and the free energy of the lowest energy, native state for all temperatures was calculated and inspected the values as block averages over the last 10 ns of simulation time (see Figure 8.3). The free energy difference was calculated on the basis of the relative populations of states using the equation:

$$\Delta G = -RT\ln[(1 - P_n)/P_n], \tag{8.3}$$

where P_n is the fraction of native structures, R is the gas constant and T is the temperature. The number of replicas used, the high temperatures sampled, the convergence demonstrated and the long simulation times for models of this size provide confidence that all of the major states of the hammerhead ribozyme have been sampled and that we can make sufficient conclusions based on the structure and function of the ribozyme.

High performance supercomputing infrastructure

It is important not to overlook the considerable computational power required to run simulations of this type. To this end, supercomputing grid infrastructures were utilised in the UK and Europe. The primary computing source in this study came from the petascale IBM Blue Gene/P machine at Forschungszentrum Jülich in Germany called JUGENE and Kraken Cray XT5 system at the National Institute for Computational Science (NICS). We also employed local machines at UCL, including Mavrino, the Centre for Computational Science's Linux cluster, and UCL Research Computing's 5,500 core Legion machine.

The simulation was run on 76,800 cores. The total combined computing time expended in this work was in excess of 20 million CPU hours and resulted in more than 5 μs of production molecular dynamics. Simulations produced in excess of 6 terabytes of data which was transferred via the grid infrastructure to local storage at UCL for analysis.

Visualisation is an essential tool for examining and analysing the structure and behaviour of molecular systems. Systems were visualised using the Visual Molecular Dynamics program (VMD) for its fast rendering qualities and excellent representation of nucleic acid structures.

8.3   Results

The original hammerhead ribozyme crystal structure, which is refered to as the minimal hammerhead ribozyme [280], only resolved the partial crystal structure and produced apparent inconsistencies between mutational experiments [281] and the interpretation of thio substitutions [282, 283] within the crystallographic structural information. Many facets of the minimal structure did not agree with the experimentally determined catalytic mechanism. The rate of cleavage of the crystallised minimal structure was faster than that of the minimal hammerhead ribozyme in aqueous solution [284], which suggests that the global fold must correspond to the catalytically active form of the molecule as the aqueous minimal structure would have undergone a structural relaxation from the crystal minimal structure.

The crystal structure of the full-length hammerhead ribozyme was recently obtained at 2.2 Å resolution [269]. The full-length hammerhead ribozyme structure, which will be refered to from now on as the 'hammerhead ribozyme', is 50% bigger in size than the minimal hammerhead. The hammerhead ribozyme tertiary structure as described by the replica exchange molecular simulations will be discussed in the following section. Various likely structural conformations that the hammerhead ribozyme can have will be monitored. In the Section entitled 8.3.2 the conformation of the hammerhead ribozyme active site will be investigated and correlated to the tertiary structure in the Section entitled 8.3.1. Tertiary interactions are important in maintaining the active site conformation and will be discussed. Finally, the conformation of the active site is shown which points towards a well defined mechanism for the catalytic self-cleavage reaction and compare to other mooted reaction mechanisms that have been suggested in the literature.

8.3.1 Tertiary structure and folding of the hammerhead ribozyme

The clustering analysis performed uses the MMTSB algorithm to iteratively go through the structures and locate clusters of structures that are the same. From this it creates centroids describing each cluster and then gives an RMSD for each structure in the trajectory with respect to each identified cluster. The centroids themselves are not physically meaningful since they are effectively a mathematical construct based on the member of a cluster. However, the structure that has the lowest RMSD with respect to each centroid is meaningful and much easier to understand. In this way it is possible to look at the structure that is closest to the centroid representing each of the clusters found.

The clustering analysis grouped the hammerhead ribozyme structures into clusters where each structure has a RMSD within 4 Å from the centre of the cluster. The routine decomposed the replica exchange trajectories of replicas at ambient temperatures (300.00 - 307.90 K) into 30 distinct structural clusters. The histogram in Figure 8.4 shows the population of each cluster. Cluster FL01 has the largest population, being composed of 1753 like structures. The structures which belong to clusters that have a population under the threshold of 500 were deemed to be insignificant and were not considered in the majority of the ensuing analysis. With the threshold in place, 12 structures remain which are significantly populated.

In order to visualise the twelve most populated clusters, structure were selected which had the lowest RMSD value from the centre of each cluster. The hammerhead ribozyme structure with the lowest RMSD value from the centre of its cluster we refer to as the

'best member' of that cluster. The best member is more meaningful than the average structure of each cluster. The best member of each of the twelve most populated clusters is displayed in Figure 8.5, which visually demonstrates the variety of structures the hammerhead ribozyme is likely to adopt in conformational space - from compact, folded, structures to elongated ones. We will repeatedly refer to this figure in the remainder of the Chapter in order to help visualise the theoretical clusters.

To display the conformational landscape of the hammerhead ribozyme, showing the various structures the molecule is likely to adopt, a plot of each structure as a function of principal components 1 and 2 was made. Figure 8.6 depicts the conformational landscape and displays an excellent correlation between the first two principal components (which contribute the majority of the dynamics) and the clusters. When the conformational landscape is viewed in three dimensions (by plotting the first three principal components) I can be seen that each cluster 'island' is independent and separated by unoccupied conformational space. This indicates that there are only twelve significant structures that the hammerhead ribozyme is likely to adopt.

Figure 8.7 displays the dominant principal component projected onto the hammerhead ribozyme molecule. The first principal component contributes a significant majority of the dynamics to the overall motion of the molecule. The principal motion observed is similar to the junction bend motion observed in DNA and RNA duplexes [222]. The second principal component, which contributes significantly less than the first to the motion of the hammerhead ribozyme, is consistent with a spring type compression along the central axis of the molecule; this is also observed in DNA and RNA duplexes. Movies of the first two principal components are given as Supporting Information.

Figure 8.8 finds how the structural cluster islands relate to the hammerhead ribozyme crystal structure. The minima shown in Figure 8.8 shows the location of the structure which deviates the least from the crystal structure. The crystal structure-like conformation is a member of cluster FL03 (which can be seen in Figure 8.5). The exact structure lies 1.33 Å from the centre of cluster FL03. From this minimum the surface ascends out towards three distinct maxima which relate to clusters FL02, FL05 and FL08. The RMSD from the hammerhead ribozyme crystal structure was determined for each best cluster member and is reported in Table 8.1.

If this hammerhead ribozyme conformation, which is closest in terms of RMSD to the crystal structure, was the lowest energy native structure it would theoretically belong to the most populated structure as we can assume that the relative free energy is proportional to the population difference for a converged system [278]:

Table 8.1: Cluster number as shown in 8.6 with associated values of root mean-squared deviation (relative to the crystal structure and cluster centre) and radius of gyration for the 'best member' structure of each of the dominant clusters.

Cluster	RMS (from crystal)	R_g	e2e 1	e2e 2 in Å
FL01	8.89±0.25	18.59±0.23	44.66±1.74	46.63±1.36
FL02	10.34±0.25	17.46±0.16	29.65±2.41	27.34±1.51
FL03	6.10±0.15	18.48±0.14	48.75±1.13	50.60±1.04
FL04	8.33±0.26	18.43±0.22	43.79±5.01	49.99±1.40
FL05	11.56±0.21	17.89±0.12	51.29±2.13	44.60±1.34
FL06	7.91±0.31	19.13±0.16	51.97±4.10	53.52±2.33
FL07	10.61±0.14	16.99±0.11	34.26±1.80	38.00±1.24
FL08	10.07±0.21	20.14±0.21	61.61±1.63	63.52±2.13
FL09	10.00±0.24	18.53±0.30	57.70±1.89	30.18±2.16
FL10	10.04±0.19	18.76±0.12	56.12±1.98	38.69±0.76
FL11	10.55±0.10	16.56±0.07	40.36±0.50	31.81±0.57
FL12	10.74±0.59	17.10±0.28	24.78±1.30	37.08±4.66

$$ln\left(\frac{P(x,y)}{P_{max}}\right) = \frac{\Delta F}{k_\beta T} \qquad (8.4)$$

where P_{max} is the population of the highest populated cluster, $P(x,y)$ is the cluster population and $\Delta F/k_\beta T$ is the relative free energy. The relative free energy is shown in Figure 8.9 as a histogram which corresponds to the various clusters in Figure 8.4 and as a heat map which shows free energy as a function of the first two principal components.

The conformation most like the crystal structure is, as we stated earlier, a member of cluster FL03. FL03 is not the group of structures with the lowest free energy, in fact two clusters have lower minimum in the free energy landscape, shown in figure 8.9. Cluster FL01 is the global minima and cluster FL02 is the second lowest minimum, which indicates that the conformations within these two clusters are more predominant in aqueous solution than the crystal structure.

8.3.2 Active site structure and ribozyme mechanism

The hammerhead ribozyme catalyses an RNA self-cleavage phosphodiester isomerisation reaction reacts via nucleophilic attack of the C-17 2′O upon the adjacent scissile phosphate that produces two product RNA strands. the 5′ product possesses a 2′,3′-cyclic phosphate terminus, and the 3′ product has a 5′-OH terminus. The reaction is effectively reversible as the scissile phosphate remains a phosphodiester and may act as a substrate for ligation.

There are several known mechanisms for the self-cleavage reaction in ribozymes. Known mechanisms include the the base-catalysed abstraction of a proton from the $2'O$ nucleophile, the acid-catalysed donation of a proton to the $5'O$ leaving group, and also the direct electrostatic stabilisation of the pentacoordinate oxyphosphorane transition state [285]. The hammerhead ribozyme crystal structure resolved two invariant residues G-12 and G-8 that are positioned consistent with general acid base catalysis. G-12, in the crystal structure, is within hydrogen bonding distance of the nucleophile $2'O$ of C-17, and the ribose of G-8 hydrogen bonds to the leaving group $5'O$. It is still an ongoing concern as to the specific role, if any, the divalent Mg^{2+} ions play in the conformation and stability of the active site.

A second cluster analysis was performed, this time on the six conserved RNA residues which make up the hammerhead ribozyme active site. The clustering analysis was performed with a k-means radius of 1.5 Å as the active site conformations sampled differ less in RMSD than the full hammerhead ribozyme structure. The clustering of the active site was plotted as a function of the first two principal components of the full hammerhead ribozyme, in Figure 8.10, in order to observe the correlation between the active site conformation and the full length tertiary structure. Figure 8.10 indicates that the clustering of the active site has excellent correlation with the first two principal components of the hammerhead ribozyme which suggests that the active site structure is largely determined by the overall tertiary structure of the hammerhead ribozyme, a facet of the full-length hammerhead ribozyme which was proposed by the X-ray experiments of Scott *et al.* [266, 269]. They suggested that for optimal activity in the hammerhead ribozyme, the presence of the sequences in stem I and stem II are needed to form tertiary contacts which arrange the active site. This explains the lack of catalytic activity of the minimal crystal structure elucidated by Pley *et al* [280].

Because the cluster analysis of the active site correlates to the principal components it can be inferred that the cluster analysis of the active site also correlates with the clustering of the hammerhead ribozyme tertiary structure. The exception to this statement is that the independent cluster islands in Figure 8.6 of cluster FL03 and FL06 are assimilated in the clustering of the active site (Figure 8.10, cluster AS01). This means that the two islands have an active site conformation that has a RMSD value within 1.5 Åof each other. For this reason the active site conformation of cluster AS01, in Figure 8.10, is the most populated, having the lowest relative free energy of the active site conformations sampled.

The difference in RMSD of the active site within each cluster island is very small, suggesting that the six RNA residues which contribute to the active site are well ordered. For this reason we can display a meaningful average structure for each cluster instead

of the best member used for the full tertiary structure. The average structure of the active site from cluster FL03 and FL06 is shown in Figures 8.11 and 8.12 respectively. The average conformation of residues from cluster FL01, the lowest free energy cluster, show a significant deviation from the conformation within the crystal structure (2GOZ). Specifically, the guanine and adenine (G-10 and A-9) residues are orientated 90° from the z-x plane which the other residue are positioned in. This orientation has the effect of closing the active site pocket and arranging the RNA nucleotides in a closer, more compact conformation. The conformational change appears to retain all the important atom positions which participate in the self-cleavage reaction in place; the A-9 and C-1.1 phosphate groups and the G-10 N_7 all remain coordinated to the central Mg^{2+} cation. The base stacking interactions between G-10:A-9, G-8:C1.1 and G-12:C-17 are all still present. It is the pucker in the phosphate backbone around the G-10.1 and A-9 residues which brings about the structural rearrangement.

Figure 8.12 displays the average active site conformation for the structures belonging to cluster FL03 (see Figure 8.6). The arrangement of bases in this conformation have an RMSD value closest to that of the crystal structure which indicates that the crystal structure is a highly populated structure and a low free energy conformation, but it is not the lowest free energy structure. In other words, it is not the global minimum. This conformation differs from that in Figure 8.11 in that the G-10 and A-9 residues are rotated in the same plane as the rest of the active site residues. This arrangement has the effect of opening the active site pocket. This is seen in Figure 8.13 where the average distance from the centrally coordinated Mg^{2+} ion is larger than in the average structure from cluster FL06 (see Figure 8.13 (b) and (c)).

The arrangement of the active site is shown by displaying heavy atom distances as a probability distribution for significantly populated clusters from the initial cluster analysis in Figure 8.6. Clearly the distances between heavy atoms is different between clusters as observed in the probability distribution curve shift along the x-axis. The possible exception to this is the distances shown between the O2$'$ atom of A-9 and the O2$'$ atom of C-1.1. Each of the clusters displayed shows a distance of approximately 4.3 Å (values are given in Table 8.2). This distance is indicative of the coordination of these two negatively charged phosphate groups around a central Mg^{2+} cation in the active site pocket (which can be visualised in Figure 8.11 and 8.12). The coordination is also demonstrated in Figure 8.13 (c) where the O5$'$ group of C-1.1 coordinates around the Mg^{2+} ion. These features of the active site all point towards a single metal mechanism for the hammerhead ribozyme self-cleavage reaction. The proposed single metal mechanism suggests that a single Mg^{2+} ion bridges the A-9 and C-1.1 scissile phosphate in response to the negative charge that accumulates in the transition state from the proximity of the phosphate groups.

Table 8.2: Comparison of experimental and simulation heavy atom distances.

	X-ray		Simulation			
	2GOZ	2OEU	FL02	FL01	FL06	FL03
C1.1:OP2 – A9:OP2	4.33	4.28	4.302	4.318	4.302	4.134
G8:O2′ – C1.1:O5′	3.19	3.51	4.418	3.476	4.878	4.470
C17:O2′ – C1.1:P	3.18	3.30	4.181	3.753	4.090	4.200
G12:N1 – C17:O2′	3.54	3.26	7.335	4.799	5.013	5.297
A9:N6 – G12:N3	2.63	3.22	10.495	6.608	13.446	7.822
A9:N6 – G12:O2′	3.21	2.98	9.913	6.053	11.699	7.367
A9:N7 – G12:N2	2.90	2.90	13.050	8.973	14.022	9.256

Experimental evidence for the single metal cation mechanism has come from Wang *et al.* who observed the equilibrium and kinetic thio-effect at the A-9 phosphate and at the scissile phosphate which are consistent with this mechanism [282]. Experiments which have implicated the G-8 base as the general acid using pH dependence correlated with metal pK_a values [286].

One of the significant conformational changes between the active site in the crystal structure and the active site in the lowest free energy structure is the arrangement of the G-12 and C-17 residues. The crystal structure suggests the nitrogen in the 1 position of the guanine base donates electron density to the oxygen in the 2′ position of the ribose ring of the C-17 residue to initiate the self-cleavage reaction. The average structure from cluster FL06 shows a different arrangement of residues and indicates a different electron donor. The most probable electron donor in this case is the lone pair on the nitrogen in the 4 position of the guanine base as this is positioned in-line with the electron acceptor, the hydrogen atom bonded to the cytosine O2′.

The A-9 and G-12 residues exhibit a sheared base-pair hydrogen bonding network [287] in the crystal structure. In the average structure taken from the lowest free energy cluster it is found that the base-pair is disrupted, the residues having undergone a conformational rearrangement. The lowest free energy conformation actually shows the greatest distance between A-9 and G-12 residues out of all the highest populated conformations. This suggests that there is an energetic relaxation associated with the breaking of this base-pair and allowing the active site to adopt a different arrangement - the arrangement we discussed as having contacts which promote the self-cleavage mechanism.

8.4   Conclusions

The replica exchange simulations reported here are of an unprecedented scale. The large scale of the simulations in terms of size, simulation time and number of replicas allows

sampling of the complex conformational space of the large hammerhead ribozyme in aqueous solution. Previous simulation approaches for the study of the hammerhead ribozyme have employed classical molecular dynamics techniques which do not effectively sample the conformational space and are too susceptible to trapping in local energy minima. The replica exchange molecular dynamics techniques overcome this by exchanging temperatures in order to transcend local minima, at the cost of being much more expensive computationally. The tyranny of however, it is an ideal kind of application for deployment on petascale machines such as the ones we have used with hundreds of thousands of cores.

The hammerhead ribozyme has been studied in detail experimentally and theoretically but the precise conformation and mechanism of the self-cleavage reaction hitherto remains unknown. Simulations have shown that the full-length hammerhead ribozyme adopts very well defined and distinct structural conformations wherein it is observed that only a few populated misfolded structures and well populated low free energy conformations. The lowest free energy conformation of the hammerhead ribozyme is identified which is structurally distinct from the resolved crystal structure. The lowest free energy, and most highly populated conformation, is the native fold.

The simulations show that the contact made between secondary structural loops in the hammerhead ribozyme are important in arranging the active site for catalysis. It also shows that the active site conformational space correlates directly to the conformational landscape of the tertiary structure through the clustering analysis. This explains why the minimal hammerhead ribozyme crystal structure does not exhibit the catalytic potential the full-length model does.

Details of the active site conformation of the native fold which differs from that suggested by crystal structure studies are given. The simulation study reported here shows that the disputed mechanism corresponds to the previously proposed single metal reaction mechanism whereby a central divalent magnesium cation provides the electrostatic screening by means of which the active site residues coordinate around the central cation to facilitate the self-cleavage reaction. Holm discusses in detail the significance of Mg in prebiotic geochemistry and origins of life scenarios [288]. It is possible then to discount the proposed mechanistic scenarios which involve no mediation from metal cations and where the metal cation interacts only indirectly by coordinating with the G-10.1 residue.

It is shown in Chapters 7 and 6 that a clay mineral surface can have a significant impact on the structure and dynamics of RNA sequences. The montmorillonite surface investigated in Chapter 6 was shown to alter the motion and energetics of the RNA strand in a way which promoted the folding of the sequence into a well-defined secondary structure at the same time enhancing the kinetics of the system so that the native structure forms

more rapidly than in bulk water. It can be inferred that a more complex sequence like the hammerhead ribozyme would undergo similar physical-chemical changes when in a mineral environment. Experiments by Biondi *et al.* confirm these predictions inferred from previous simulations [289]. They showed that the montmorillonite presence protects the RNA against degradation and increases the rate of cleavage kinetics by approximately an order of magnitude. Like the case of the hammerhead ribozyme in bulk water, the self-cleavage mechanism of the ribozyme in a mineral environment is unknown as well as the mechanism for how the mineral environment alters the kinetics. This chapter demonstrates the ability of REMD to identify the major populated structures in the complex conformational landscape and to elucidate important catalytic mechanistic and structural details. This approach would lend itself to the study of the hammerhead ribozyme in a clay mineral environment and compliment experiments performed on this system.
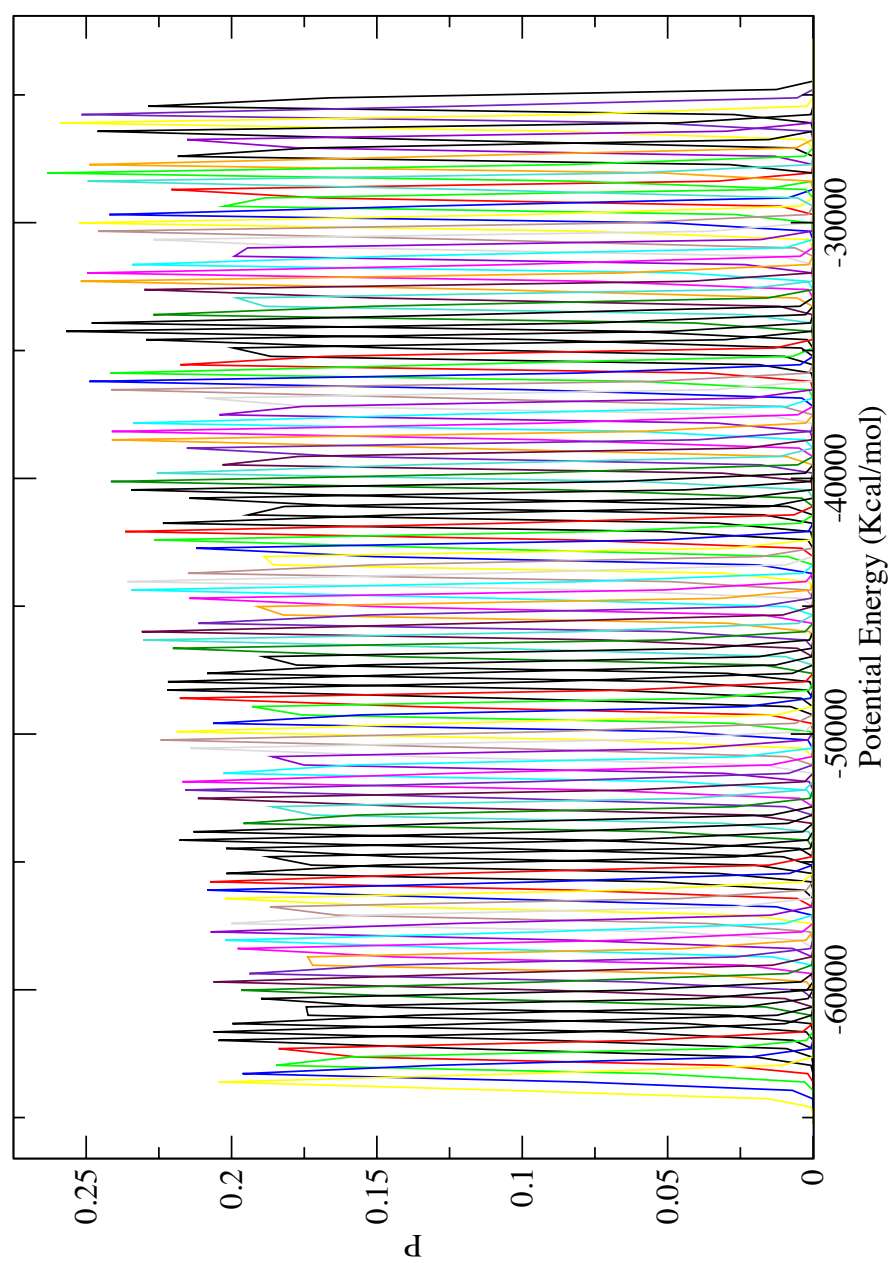
Figure 8.2: Probability distribution of total potential energy for all 100 replicas. The energy overlap between adjacent replicas provides an essentially uniform acceptance ratio of approximately 23%.
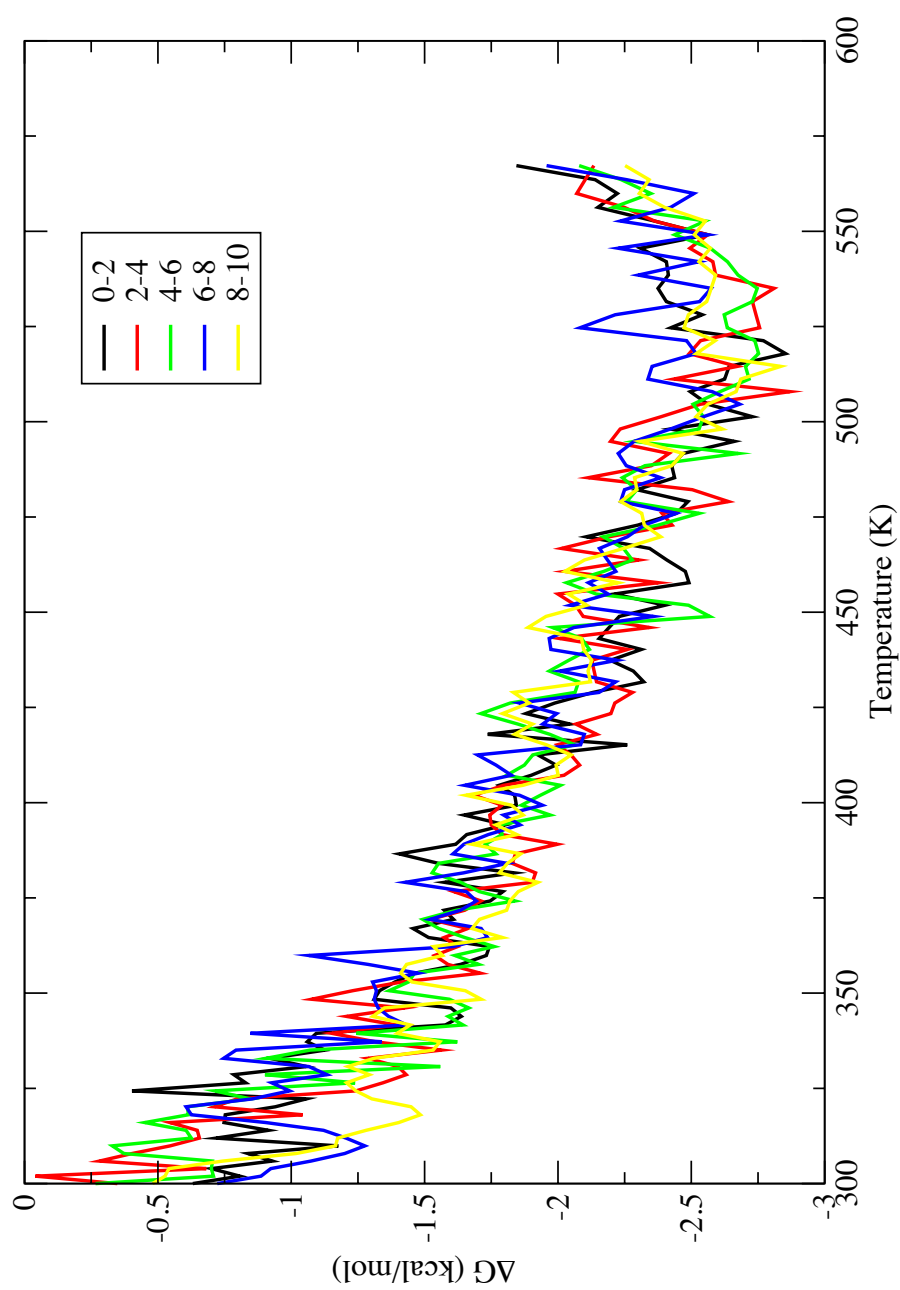
Figure 8.3: Difference of free energy between states and the lowest relative free energy state of the hammerhead ribozyme as a function of temperature based on block average data from the last 10 ns of replica exchange molecular dynamics. Each block average interval is shown in the legend, given in nanoseconds. The free energy differences between the blocks show excellent convergence of the replica exchange sampling.
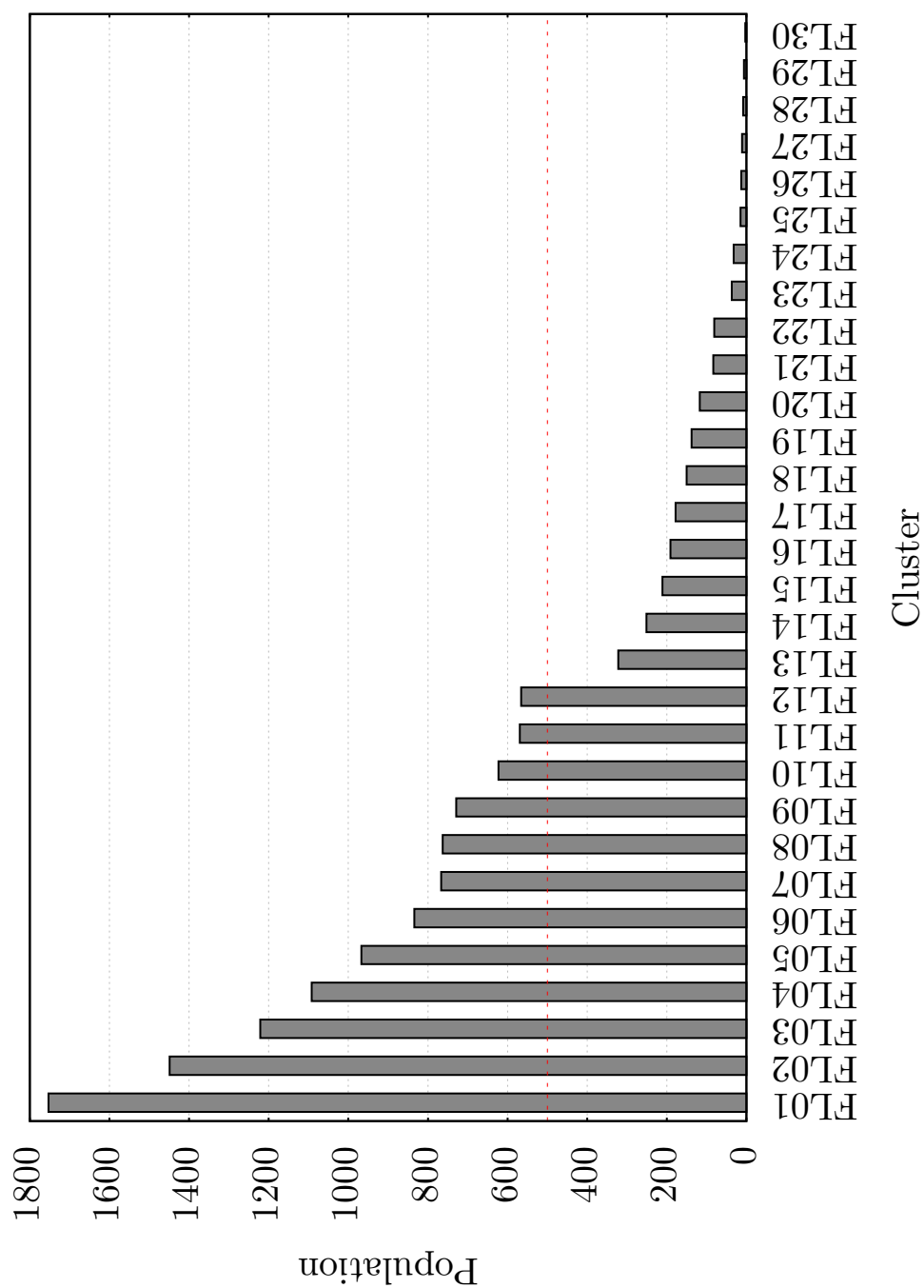
Figure 8.4: Histogram showing the number of ribozyme structures which belong to each structural cluster. Cluster FL01 is the most populated with 1753 structure that are within 4 Å of the centre of the cluster. Clusters with a population below the threshold of 500 were deemed to be insignificant and were not considered in the majority of the analysis.
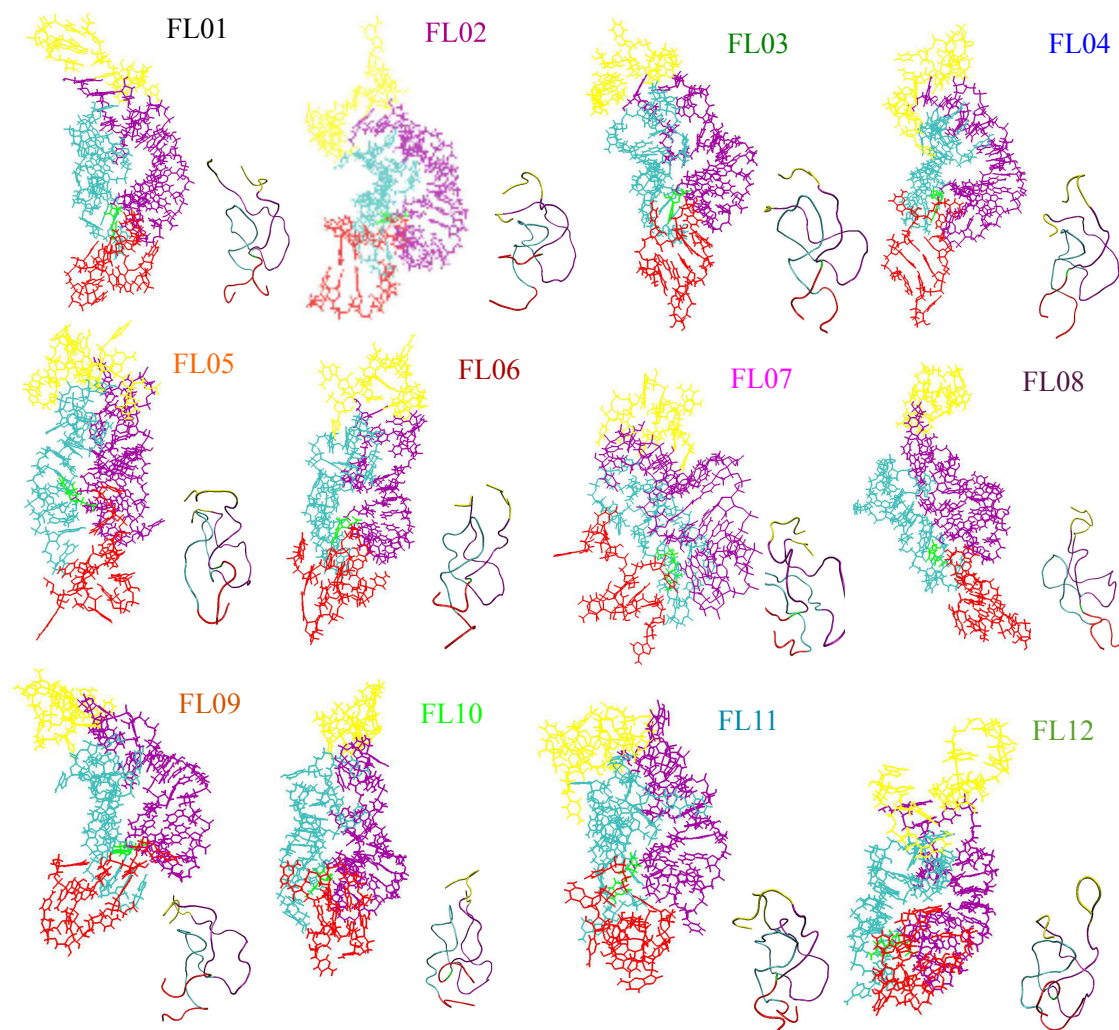
Figure 8.5: Structures with the lowest RMSD from the centre of their cluster (number shown next to molecule) which are referred to as the "best member" of the cluster island. Each structure is shown in a stick formation with a trace of its phosphate backbone beside. The trace more clearly displays the folded tertiary structure of each ribozyme.
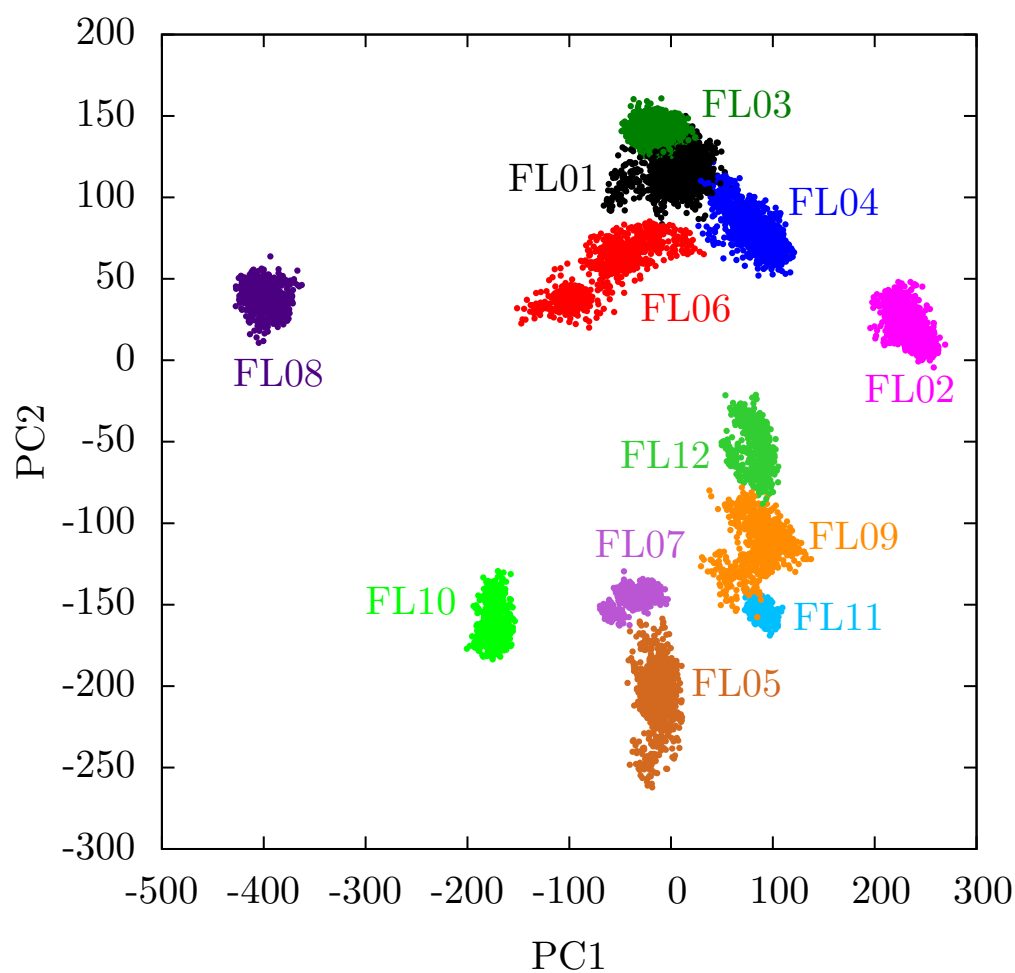
Figure 8.6: Principal component analysis where the first and second eigenvectors are projected onto the ribozyme molecule and plotted on the x and y axis respectively. Each structure is coloured according to which cluster it belongs to in Figure 8.4. The plot shows a well defined correlation between the principal components of the trajectories and the RMSD of the structure from the clustering process. Each structural "island" displays the cluster number.

(a)   (b)
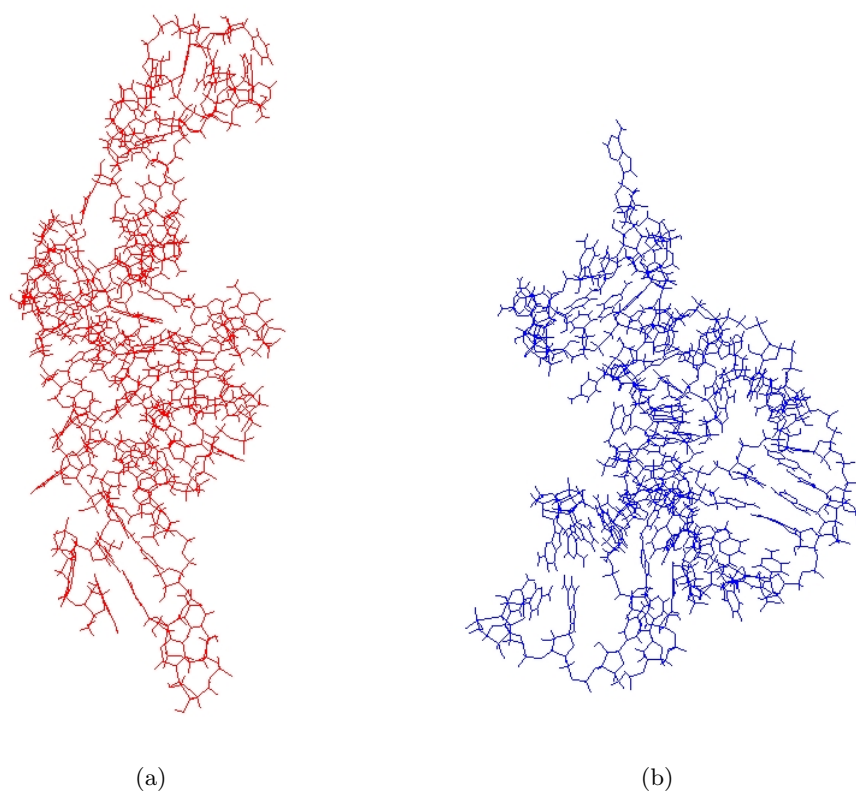
Figure 8.7: The dominant principal component (PC1) projected onto the ribozyme molecule. (a) the minimum eigenvalue, around -400 and (b) the maximum eigenvalue around 200. The motion is referred to as a junction-bend and is the dominant mode observed in standard RNA and DNA duplexes.
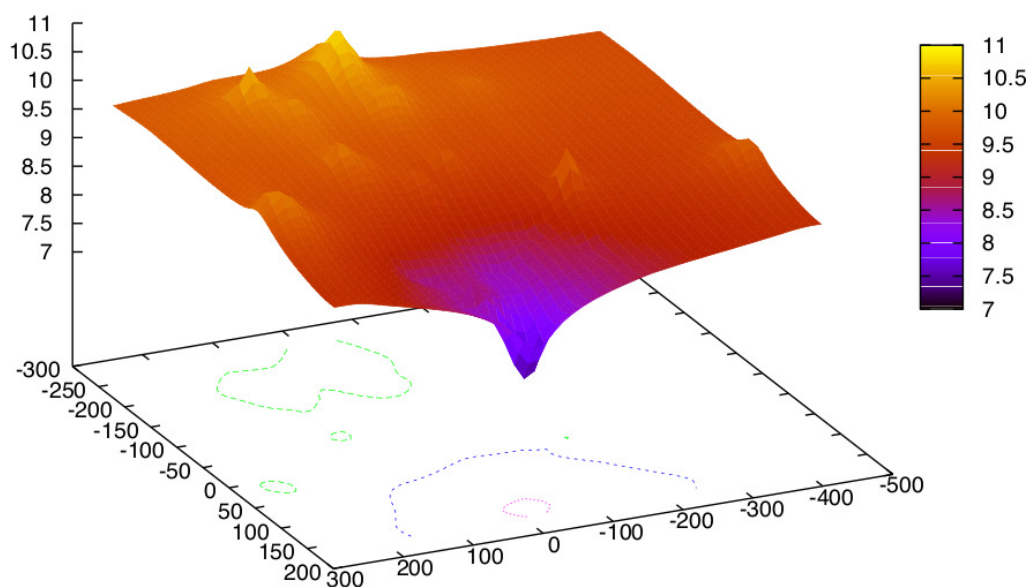
Figure 8.8: The first two principal components have been used to show the distinct cluster islands, as in Figure 8.6, plotted against an RMSD surface with reference to the crystal structure. The surface shows one distinct minimum of which the structure with the lowest RMSD (5.5 Å, shown on the right) from the crystal structure is positioned 1.33 Å from the centre of cluster FL03. It can be deduced from this that cluster FL03 is representative of the crystal structure. However, cluster FL03 is not the most populated cluster and therefore not the lowest free energy native folded hammerhead ribozyme.
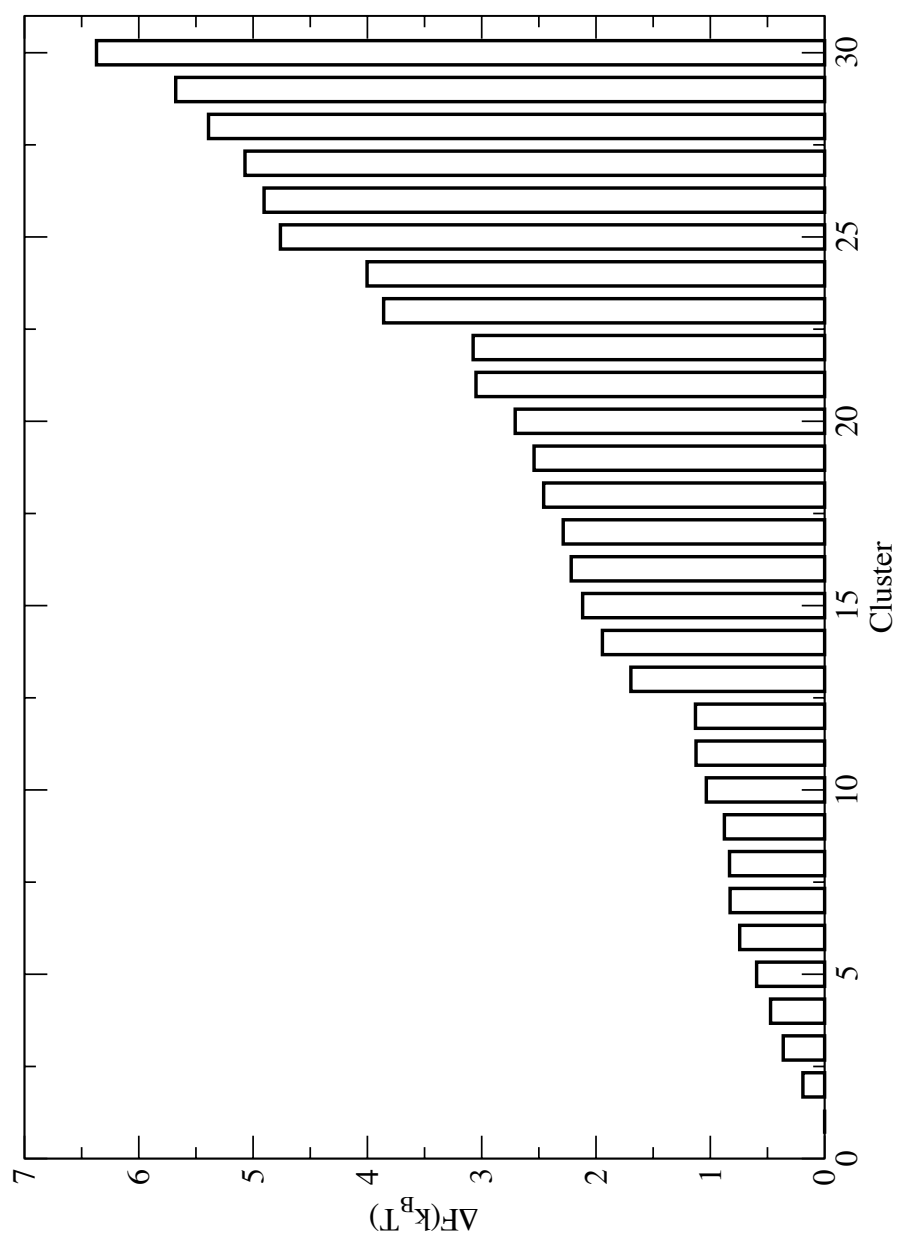
Figure 8.9: Histogram of the relative free energy for each cluster. The lowest free energy structures are those which are members of cluster FL01 with the best member having the lowest free energy.
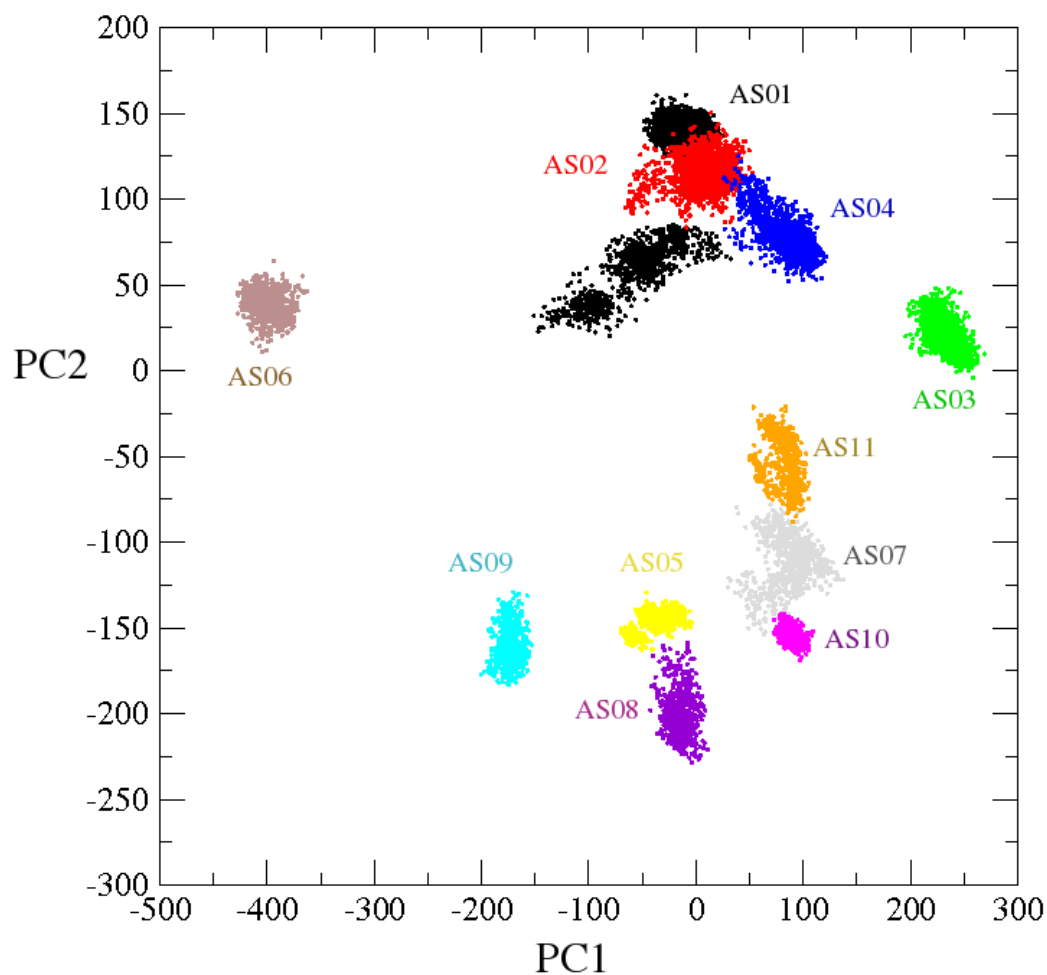
Figure 8.10: Principal component 1 and 2 of the full hammerhead ribozyme structure. Each structure is coloured according to the clustering of the active site (not the full tertiary structure as in Figure 8.6). The plot shows an excellent correlation between the principal motions of the full tertiary structure and the local conformation of the hammerhead ribozyme active site which is composed of six residues. This suggests that the active site structure, and therefore ribozyme mechanism, is dependent of the full tertiary structure.
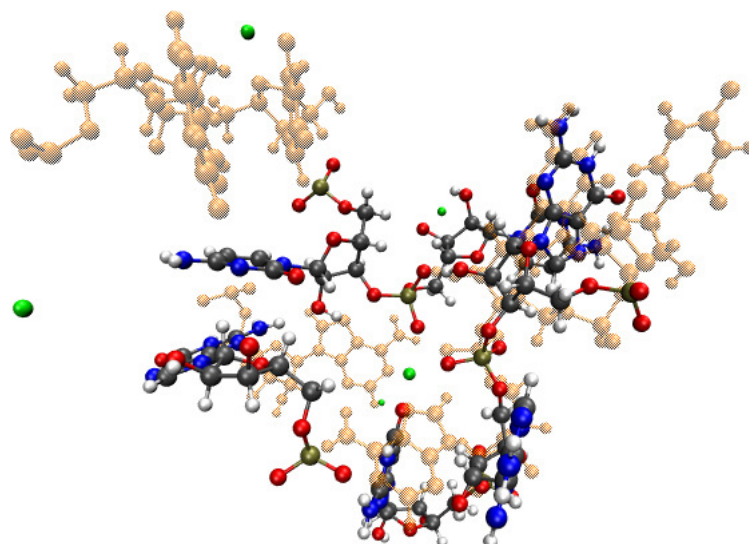
Figure 8.11: Average structure of the six conserved residues that participate in the self cleavage reaction, which are referred to as the active site, taken from cluster FL01 of the clustering analysis shown in Figure 8.10. The crystal structure is shown in orange. The arrangement is consistent with a single metal mechanism where the A-9 phosphate is positioned proximate to the C-17 scissile phosphate via electrostatic screening from the divalent metal cation positioned in the active site pocket.
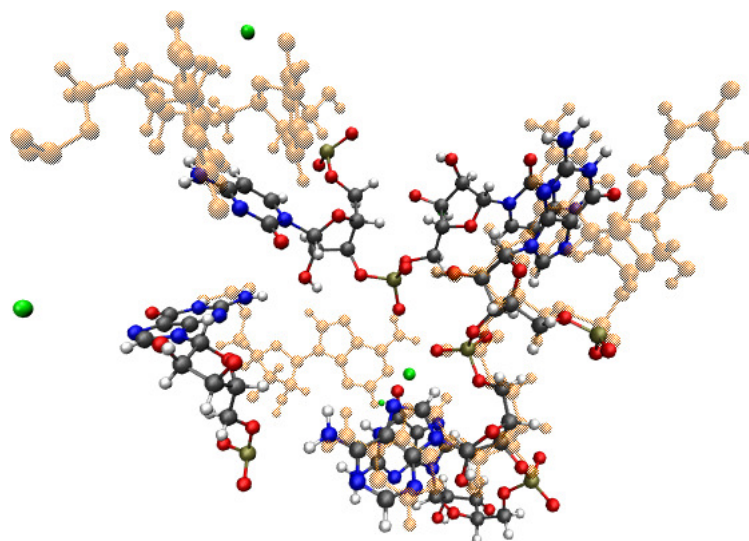
Figure 8.12: Average structure of the six conserved residues that participate in the self cleavage reaction, which are referred to as the active site, taken from cluster FL03 of the clustering analysis. The crystal structure is shown in orange.
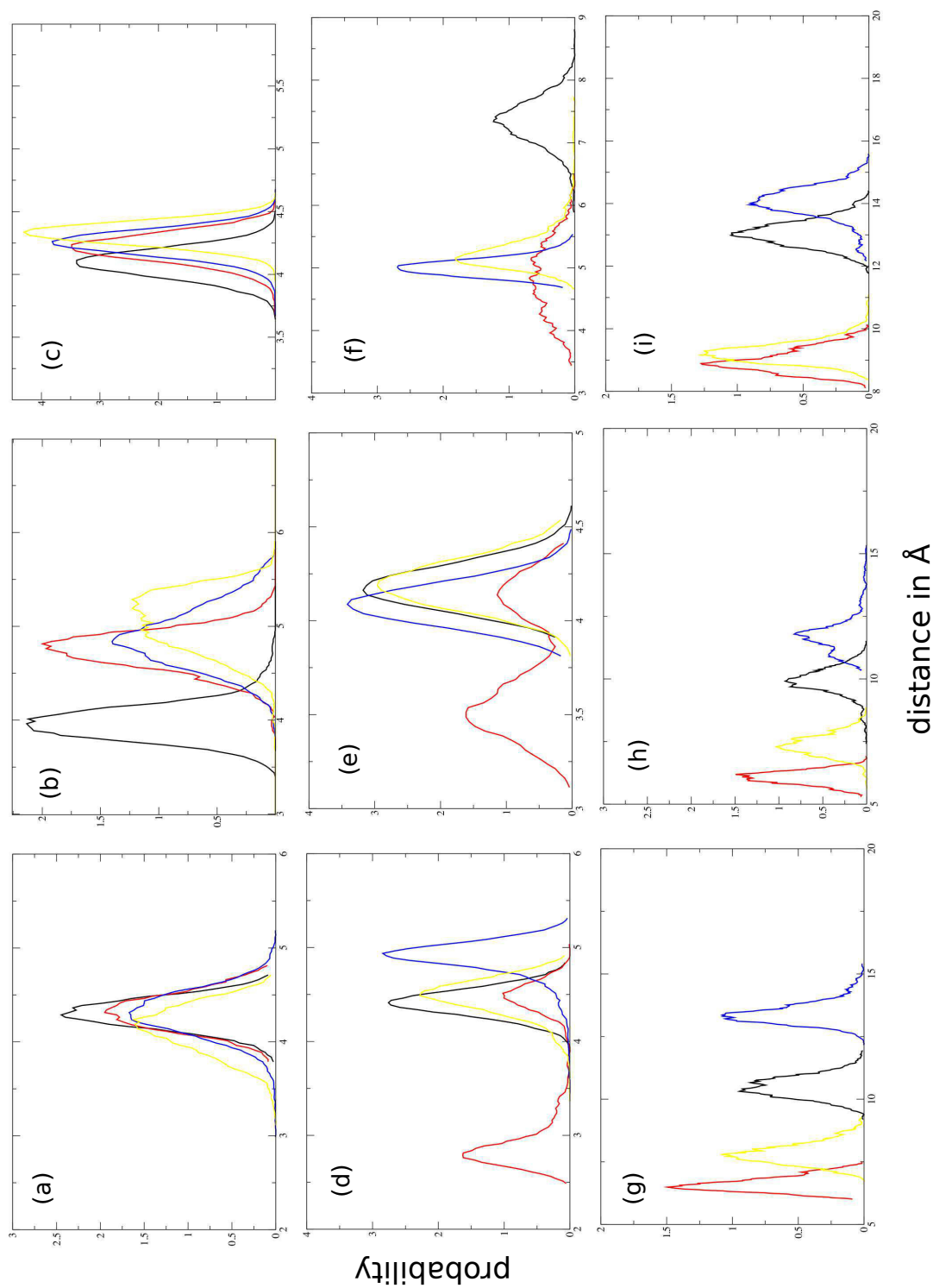
Figure 8.13: Probability distribution of selected heavy atom distances in the hammerhead ribozyme active site for the four most populated clusters. Cluster FL01 - Red; Cluster FL02 - Black; Cluster FL03 - Yellow and Cluster FL06 - Blue. The heavy atom distances are shown using the standard PDB notation for RNA: (a) C1.1:OP2-A9:OP2; (b) Mg-G8:O2′; (c) Mg-C1.1:O5′; (d) G8:O2′-C1.1:P; (e) C17:O2′-C1.1:O5′; (f) G12:N1-C17:O2′; (g) A9:N6-G12:N3; (h) A9:N6-G12:O2′; (i) A9:N7-G12:N2.
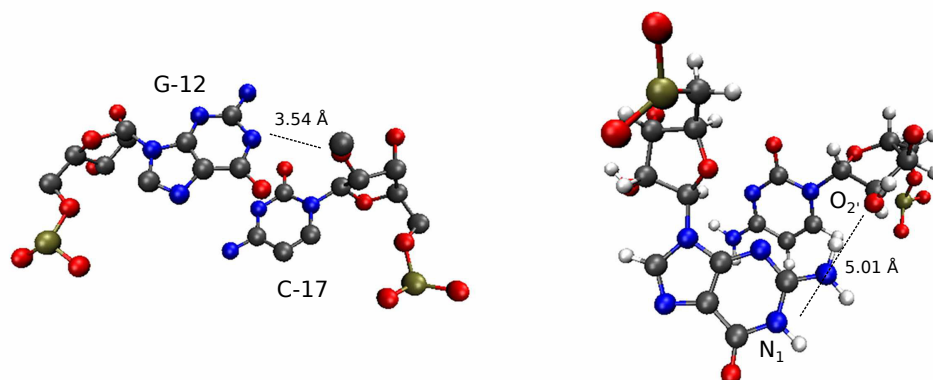
Figure 8.14: Distance between G12:N1 and C17:O2′ from the crystal structure and the average structure from cluster FL06. A previously proposed mechanism based on the crystal structure suggests the nitrogen in the 1 position of the guanine base donates electron density to the oxygen in the 2′ position of the ribose ring of the C-17 residue. The average structure from cluster FL06 shows a different arrangement of residues and indicates a different electron donor.
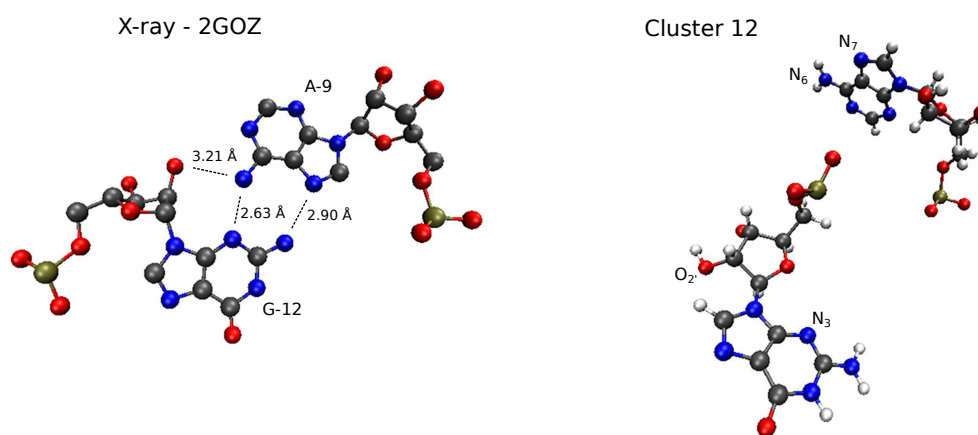
Figure 8.15: Distance between A9:N6 and G12:N3 from the crystal structure and the average structure from cluster FL06. The crystal structure shows a sheared type hydrogen bonding network between the adenine and guanine bases. The base pairing between A-9 and G-12 in the crystal structure is not observed in the average structure of the lowest free energy structure.

Chapter 9

Overall Conclusions

The work detailed in this thesis has used advanced simulation techniques to investigate how nucleic acid molecules interact with various charged clay mineral environments. The implications of our findings on the transition from chemistry to biology *de novo* on the early Earth, which we refer to as the origins of life, has been discussed in each Chapter.

Although there is a great deal of progress being made in the field of origins of life studies, there are still large gaps in our knowledge of the processes leading up to the emergence of life. Chapter 2 reviews theoretical and molecular modelling approaches that are now throwing considerable light on numerous challenging issues associated with the origin of life on Earth. As the field continues to gain respectability from within the scientific establishment, it is hoped that there will be stronger interactions between theoretical approaches and related experimental research, that can only serve to underpin further advances in our understanding of origins of life processes. I hope that the research presented here will go some way in promoting these interactions which may inspire new experiments.

Chapter 5 presents a comparative study of the structure, properties and stability of selected double-stranded sequences of RNA, DNA and PNA. The intention was to gain insight into which candidate genetic material arising at the time of the origin of life on Earth may have been preferentially favoured by the prevailing geochemistry, in particular when interacting with anionic clays. Of the three nucleic acids under investigation, DNA is observed to be the most stable when intercalated within an LDH, with respect to the percentage of Watson-Crick base-pairs. The opposite is found in bulk water where both PNA and RNA are more stable than DNA. Conceivably, a mineral mediated origin of life may have favoured DNA as the informational storage biomolecule over competing RNA and PNA, providing a route to modern biology from the RNA world.

One prevailing theory in origins of life studies is that RNA in prebiotic conditions may have been confined and restricted in motion by mineral surfaces, much as it is confined within the ribosome of living cells today. Due to the ubiquitous presence of clay minerals and their associated cations in nature, these aqueous environments may have provided prebiotic habitats for nascent nucleic acid polymers. Chapter 6 shows that single strands of RNA fold at a considerably enhanced rate when a montmorillonite surface is present compared with the same sequences in bulk water. The increased rate of folding at clay surfaces is of importance to the RNA world view, according to which early RNA polymers would have needed to produce well defined folded structures to support enzymatic activity in order to cleave self-replicating RNA strands, essential components in the first steps of molecular evolution. Accelerated folding kinetics considerably increases the parameter space that may be explored by clay-RNA complexes in constructing catalytically active structures under conditions which may have been far from benign.

Chapter 7 investigates the structure and arrangement of RNA at the LDH mineral surface. According to the simulations, while cationic montmorillonite clays promote folding of the RNA structure, LDH prevent folding. The folded tertiary structure induces the catalytic capacity of the RNA ribozymes; unfolding inhibits catalytic activity. From the simulations it is hypothesised that the LDH arranges the RNA in a way which might promote a templating mechanism, making LDHs a potential candidate for amplifying the population of RNA sequences. The interactions described here between RNA and LDH have relevance to origins of life studies, as the arrangement of single RNA strands in this manner might have made them amenable to Watson-Crick templating as part of an early hereditary process.

The research detailed in Chapter 8 is a natural progression from that in Chapter 6 where it takes a transition from looking at the enhanced rate of folding of relatively short RNA sequences to probing the structure and function of the large, more biologically relevant, RNA ribozyme molecule. The catalytic nature of RNA is a feature which has great importance in the origins of life and the hammerhead ribozyme is seen here as a prototype for understanding the catalytic nature of RNA. Catalytic, enzymatic-like, processes such as the self condensation of RNA for example would have been essential in prebiotic chemistry. The simulations show that the contacts made between secondary structural motifs in the hammerhead ribozyme are important in arranging the active site for catalysis. The results show that the active site conformational space correlates directly with the conformational landscape of the tertiary structure through clustering analysis. This explains why the minimal hammerhead ribozyme crystal structure does not exhibit the catalytic potential which the full-length model does and why in general the full tertiary structure is needed in ribonucleic enzyme activity.

The work contained within this thesis, and the research published as a result of the work displayed here, has made a significant contribution to the field of origins of life studies. The recent advances in computational power have resulted in the large-scale simulation of complex chemical systems which have been discussed here. The simulations have not only added insight to experimental work, but they also inform new directions of study. The work in this thesis illustrates the potential of charged mineral surfaces to radically alter the kinetics of chemical processes such as nucleic acid folding, which are likely to prove important in future years in underpinning research into one of the great challenges in prebiotic chemistry, namely to identify where and when genetic information first became an important discriminator. With increasingly powerful computational resources, and increasingly sophisticated large-scale e-infrastructure environments, simulations will grow in their capacity to address the fundamental scientific questions in the prebiotic chemistry domain.

Chapter 10

Future Work

Chapter 9 reviews the work contained in this thesis and presents this along with the important conclusions drawn from each research chapter, it is also discusses how these findings impact the overall field of origins of life studies. The origins of life is one of sciences most open-ended problems, it gives us an unending string of questions and plausible explanations. This can bring us a proliferation of riches in basic chemical understanding. So it is fitting that we close with a chapter on 'Future Work'. With this chapter briefly discussing the ways in which we can progress the ideas set out in this thesis and to hone methodologies so that we can move closer to ever more plausible origins of life scenario.

I have made every effort to compare the results of the simulations to corresponding observable experimental data in order to provide confidence in the selection of force field potentials we have chosen. A problem in validating force field potentials for mineral interfaces using experimental fitting is the lack of data which can be used. Relevant observable data are available for a limited number of systems, but they are most often insufficient for fitting purposes.

The force fields potentials should be validated against *ab initio* calculations for each mineral-organic system studied in this thesis. Clearly, it would be possible to obtain better agreement by refitting to the *ab initio* curve. However, such a procedure would lack the advantages of a more general procedure for transferability and ease of adding new interactions to the force field. There is inevitably a balance to be struck between accuracy and these other considerations.

It is important when setting up molecular dynamic simulations that the latest and most accurate force field is chosen. Much work is continuing to be done to make force field potentials as accurate as possible for many systems including inorganic-organic

hybrid systems like these. Current modifications include the 'Barcelona refinement' (PARMBSC0) from Orozco *et al.* that improves the description of $\alpha\beta$ concerted rotations in nucleic acids [125]. Jurečka *et al.* have refined the Cornell nucleic acid force field in order to reparameterize the glycosidic torsion, χ, in RNA [290]. When extending or reproducing any of the simulations reported here careful consideration should be taken in selecting the right force fields to simulation these systems.

Section 4.2.2 discussed briefly enhanced sampling methods. The enhanced configurational sampling that we get from these methods lead to better statistics from the time-averaged data we draw from molecular dynamics trajectories. Simply repeating simulations from alternative starting configurations (or initial atomic velocities) in an uncoupled ensemble will deliver better statistics and reduce standard deviations.

I hope that the ideas set forth in this chapter will help guide further theoretical research in origins of life studies and somewhat provide a map for progressing the work set out in this thesis. This chapter has discussed improvements and refinements that can be made to the methodologies used and how better sampling can lead to improved results.

Appendix A

Simulation Movies

The Movies listed below can be accessed at the following website:
http://ccs.chem.ucl.ac.uk/∼jacob/Movies/

Movie **5.1**: A-Mg$_2$Al, File Type: Mpeg.
Visualisation of the trajectories taken from fully atomistic molecular dynamics (MD) simulation of A-Mg$_2$Al (see Table 7.1 of Chapter 7) at 300 K and 1 atm. The simulation consists of a two layer Mg$_2$Al layered double hydroxide (LDH) clay surface and a single stranded RNA of 25 nucleotides.

Movie **5.2**: A-Mg$_3$Al, File Type: Mpeg.
Visualisation of the trajectories taken from fully atomistic molecular dynamics (MD) simulation of A-Mg$_3$Al (see Table 7.1 of Chapter 7) at 300 K and 1 atm. The simulation consists of a two layer Mg$_3$Al layered double hydroxide (LDH) clay surface and a single stranded RNA of 25 nucleotides.

Movie **5.3**: A-Mg$_6$Al, File Type: Mpeg.
Visualisation of the trajectories taken from fully atomistic molecular dynamics (MD) simulation of A-Mg$_6$Al (see Table 7.1 of Chapter 7) at 300 K and 1 atm. The simulation consists of a two layer Mg$_6$Al layered double hydroxide (LDH) clay surface and a single stranded RNA of 25 nucleotides.

Movie **5.4**: A-Bulk, File Type: Mpeg.

Visualisation of the trajectories taken from fully atomistic molecular dynamics (MD) simulation of A-Bulk (see Table 7.1 of Chapter 7) at 300 K and 1 atm. The simulation consists of a 24 base-pair DNA duplex of mixed base sequence in bulk water.

Movie **6.1**: Model **I**, File Type: Mpeg.

Visualisation of the trajectories taken from fully atomistic molecular dynamics (MD) simulation of Model **I** in Chapter 5 at 300 K and 1 atm. The simulation consists of a three sheet Mg_2Al layered double hydroxide (LDH) clay in which one of the interlayers is intercalated with a 24 base-pair RNA duplex of mixed base sequence.

Movie **6.2**: Model **II**, File Type: Mpeg.

Visualisation of the trajectories taken from fully atomistic molecular dynamics (MD) simulation of Model **II** (in Chapter 5 at 300 K and 1 atm. The simulation consists of a three sheet Mg_2Al layered double hydroxide (LDH) clay in which one of the interlayers is intercalated with a 24 base-pair DNA duplex of mixed base sequence.

Movie **6.3**: Model **III**, File Type: Mpeg.

Visualisation of the trajectories taken from fully atomistic molecular dynamics (MD) simulation of Model **III** in Chapter 5 at 300 K and 1 atm. The simulation consists of a three sheet Mg_2Al layered double hydroxide (LDH) clay in which one of the interlayers is intercalated with a 24 base-pair PNA duplex of mixed base sequence.

Movie **6.4**: Model **IV**: File Type: Mpeg.

Visualisation of the trajectories taken from MD simulation for Model **IV** in Chapter 5 at 300 K and 1 atm, which is comprised of a six sheet Mg_2Al LDH in which twenty four RNA duplexes are distributed evenly between the clay interlayers. The Movie shows that the motion of the nucleic acid has been restricted by the LDH when compared to the nucleic acid in bulk aqueous solution.

Movie **6.5**: Model **V**: File Type: Mpeg.

Visualisation of the trajectories taken from MD simulation for Model **V** in Chapter 5 at 300 K and 1 atm, which is comprised of a six sheet Mg_2Al LDH in which twenty four DNA duplexes are distributed evenly between the clay interlayers.

Movie **6.6**: Model **VI**: File Type: Mpeg.

Visualisation of the trajectories taken from MD simulation for Model **VI** in Chapter 5 at 300 K and 1 atm, which is comprised of a six sheet Mg_2Al LDH in which twenty four PNA duplexes are distributed evenly between the clay interlayers.

Movie **6.7**: Model **I**: File Type: Mpeg.

Visualisation of the projection of the motion of all RNA atoms onto the first eigenvector using principal component analysis for Model **I** in Chapter 5 which contains Mg_2Al LDH intercalated RNA taken from the last 1ns of simulation.

Movie **6.8**: Model **I**: File Type: Mpeg.

Visualisation of the projection of the motion of all DNA atoms onto the first eigenvector using principal component analysis for Model **II** in Chapter 5 which contains Mg_2Al LDH intercalated DNA taken from the last 1ns of simulation.

Movie **6.9**: Model **I**: File Type: Mpeg.

Visualisation of the projection of the motion of all PNA atoms onto the first eigenvector using principal component analysis for Model **III** in Chapter 5 which contains Mg_2Al LDH intercalated PNA taken from the last 1ns of simulation.

Movie **6.10**: Model **I**: File Type: Mpeg.

Visualisation of the projection of the motion of all RNA atoms onto the first eigenvector using principal component analysis for Model **VII** in Chapter 5 which contains RNA in bulk water taken from the last 1ns of simulation.

Movie **6.11**: Model **I**: File Type: Mpeg.

Visualisation of the projection of the motion of all DNA atoms onto the first eigenvector using principal component analysis for Model **VIII** in Chapter 5 which contains DNA in bulk water taken from the last 1ns of simulation.

Movie **6.12**: Model **I**: File Type: Mpeg.

Visualisation of the projection of the motion of all PNA atoms onto the first eigenvector using principal component analysis for Model **IX** in Chapter 5 which contains PNA in bulk water taken from the last 1ns of simulation.

Movie **7.1**: Model **V**, File Type: Mpeg.

Visualisation of the trajectories taken from 35ns of fully atomistic molecular dynamics (MD) simulation for Model **V** in Chapter 6. Model **V** consists of RNA of sequence **B** in the presence of Na^+ montmorillonite. Nucleosides and the phosphate backbone of RNA are coloured blue and yellow respectively. Surface silicon ions are magenta and Na^+ ions are grey. The visualisation shows the folding of the initial elongated RNA structure into a more compact secondary structure over the first 5ns of simulation which persists over the full 35ns of simulation. The animation also shows the ineffectiveness of the Na^+ ions in attracting the single-strand of RNA to the surface of the montmorillonite clay.

Movie **7.2**: Model **VI**: File Type: Mpeg.

Visualisation of the trajectories taken from 35ns of MD simulation for Model **VI** in Chapter 6. Model **VI** consists of RNA of sequence **B** in the presence of Ca^{2+} montmorillonite. Colour scheme as in animation **1**. with Ca^{2+} coloured grey. The visualisation shows the folding of the initial elongated RNA structure into a more compact secondary structure over the first 5ns of simulation and the subsequent tethering of the strand to the surface of the montmorillonite clay.

Movie **7.3**: Model **VII**: File Type: Mpeg.

Visualisation of the trajectories taken from 20ns of MD simulation for Model **VII** in Chapter 6. Model **VII** consists of RNA of sequence **B** in the presence of Na^+ ions in bulk water. Colour scheme as in animation **1**. The visualisation shows the absence of folding of the initial elongated RNA structure into a more compact secondary structure; the RNA molecule retains its helical structure over the entire duration of the 20ns simulation.

Movie **7.4**: Model **VIII**: File Type: Mpeg.

Visualisation of the trajectories taken from 20ns of MD simulation for Model **VIII** in Chapter 6. Model **VIII** consists of RNA of sequence **B** in the presence of Ca^{2+} ions in bulk water. Colour scheme as in animation **1**. The visualisation shows the

absence of folding of the initial elongated RNA structure into a more compact secondary structure;the RNA molecule retains its helical structure over the entire duration of the 20ns simulation. The Ca^{2+} ions are more tightly bound to negatively charged phosphate groups than Na^+ ions in model **VII**.

Bibliography

[1] Jacob B. Swadling, Peter V. Coveney, and H. Christopher Greenwell. Clay minerals mediate folding and regioselective interactions of RNA: A large-scale atomistic simulation study. *Journal of the American Chemical Society*, 132(39):13750–13764, 2010.

[2] Jacob B. Swadling, Peter V. Coveney, and H. Christopher Greenwell. Computer simulation of clay mineral-biomolecule interactions. *Mineralogical Magazine*, 75 (1):1972, 2011. Conference Goldschmidt 2011, Prague, JUN 14-19, 2011.

[3] Jacob B. Swadling, Peter V. Coveney, and H. Christopher Greenwell. Stability of free and mineral-protected nucleic acids: Implications for the RNA world. *Geochimica et Cosmochimica Acta*, 83:360–378, 2012.

[4] Peter V. Coveney, Jacob B. Swadling, Jonathan A. D. Wattis, and H. Christopher Greenwell. Theory, modelling and simulation in origins of life studies. *Chemical Society Reviews*, 41(16):5430–5446, 2012.

[5] Jacob B. Swadling, James L. Suter, H. Christopher Greenwell, and Peter V. Coveney. Influence of surface chemistry and charge on mineral-RNA interactions. *Langmuir*, 29(5):1573–1583, 2013.

[6] J.W. Szostak. Attempts to define life do not help to understand the origin of life. *Journal of Biomolecular Structure and Dynamics*, 29(4):599–600, 2012.

[7] W. Martin and M.J. Russell. On the origin of biochemistry at an alkaline hydrothermal vent. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1486):1887–1926, 2007.

[8] L.E. Orgel. The origin of life-a review of facts and speculations. *Trends in Biochemical Sciences*, 23(12):491–495, 1998.

[9] R.M. Hazen and D.A. Sverjensky. Mineral surfaces, geochemical complexities, and the origins of life. *Cold Spring Harbor Perspectives in Biology*, 2(5):1–21, 2010.

[10] B.M. Rode, A.H. Eder, and Y. Yongyai. Amino acid sequence preferences of the salt-induced peptide formation reaction in comparison to archaic cell protein composition. *Inorganica Chimica Acta*, 254(2):309–314, 1997.

[11] H. Trinks, W. Schröder, and C.K. Biebricher. Ice and the origin of life. *Origins of Life and Evolution of Biospheres*, 35(5):429–445, 2005.

[12] G.F. Joyce. The antiquity of RNA-based evolution. *Nature*, 418(6894):214–221, 2002.

[13] P.V. Coveney and P.W. Fowler. Modelling biological complexity: a physical scientist's perspective. *Journal of the Royal Society Interface*, 2(4):267–280, 2005.

[14] P.V. Coveney and R. Highfield. *The Arrow of Time.* Nature Publishing Group, 1991.

[15] P.V. Coveney and R. Highfield. *Frontiers of Complexity: The Search for Order in a Chaotic World.* Faber and Faber: London, 1995.

[16] S.L. Miller. A production of amino acids under possible primitive earth conditions. *Science*, 117(3046):528–529, 1953.

[17] J.F. Kasting. Earth's early atmosphere. *Science*, 259(5097):920–926, 1993.

[18] J.B.S. Haldane. The origin of life. *Rationalist Annual*, 148(1):3–10, 1929.

[19] J.A. Baross and S.E. Hoffman. Submarine hydrothermal vents and associated gradient environments as sites for the origin and evolution of life. *Origins of Life and Evolution of Biospheres*, 15(4):327–345, 1985.

[20] P. Walde, R. Wick, M. Fresta, A. Mangone, and P.L. Luisi. Autopoietic self-reproduction of fatty acid vesicles. *Journal of the American Chemical Society*, 116 (26):11649–11654, 1994.

[21] D.S. Kelley, J.A. Karson, D.K. Blackman, G.L. Früh-Green, D.A. Butterfield, M.D. Lilley, E.J. Olson, M.O. Schrenk, K.K. Roe, and G.T. Lebon. An off-axis hydrothermal vent field near the mid-atlantic ridge at 30N. *Nature*, 412(6843): 145–149, 2001.

[22] J.I. Drever. *The Chemistry of Weathering*, volume 149. Springer, 1985.

[23] M.J. Russell, R.M. Daniel, A.J. Hall, and J.A. Sherringham. A hydrothermally precipitated catalytic iron sulphide membrane as a first step toward life. *Journal of Molecular Evolution*, 39(3):231–243, 1994.

[24] R.F. Gesteland, T. Cech, and J.F. Atkins. *The RNA world: the nature of modern RNA suggests a prebiotic RNA world.* Cold Spring Harbor Laboratory Pr, 2006.

[25] T.R. Cech. A model for the RNA-catalyzed replication of RNA. *Proceedings of the National Academy of Sciences*, 83(12):4360, 1986.

[26] N. Ban, P. Nissen, J. Hansen, P.B. Moore, and T.A. Steitz. The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science*, 289(5481): 905–920, 2000.

[27] M.W. Powner, B. Gerland, and J.D. Sutherland. Synthesis of activated pyrimidine ribonucleotides in prebiotically plausible conditions. *Nature*, 459(7244):239–242, 2009.

[28] P.W. Atkins and J. De Paula. *Atkins Physical Chemistry, Volume 8. ed.* Oxford, 2006.

[29] L. Onsager. Reciprocal relations in irreversible processes. I. *Physical Review*, 37 (4):405, 1931.

[30] C. Huber and G. Wächtershäuser. Peptides by activation of amino acids with CO on (Ni, Fe) S surfaces: implications for the origin of life. *Science*, 281(5377): 670–672, 1998.

[31] G. Nicolis and I. Prigogine. *Self-Organization in Nonequilibrium Systems*, volume 316. John Wiley & Sons New York, 1977.

[32] A.M. Turing. The chemical basis of morphogenesis. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 237(641):37–72, 1952.

[33] J.W. Szostak, D.P. Bartel, and P.L. Luisi. Synthesizing life. *Nature*, 409(6818): 387–390, 2001.

[34] W. Ma and C. Yu. Intramolecular RNA replicase: Possibly the first self-replicating molecule in the RNA world. *Origins of Life and Evolution of Biospheres*, 36(4): 413–420, 2006.

[35] R. Lathe. Fast tidal cycling and the origin of life. *Icarus*, 168(1):18–22, 2004.

[36] U. Wiechert, A.N. Halliday, D.C. Lee, G.A. Snyder, L.A. Taylor, and D. Rumble. Oxygen isotopes and the moon-forming giant impact. *Science*, 294(5541):345–348, 2001.

[37] C. Fernando, G. Von Kiedrowski, and E. Szathmáry. A stochastic model of nonenzymatic nucleic acid replication: "Elongators" sequester replicators. *Journal of Molecular Evolution*, 64(5):572–585, 2007.

[38] C. Fernando and E. Di Paolo. The chemoton: a model for the origin of long RNA templates. *Proceedings of Artificial Life IX*, pages 1–8, 2004.

[39] T. Gánti, J. Griesemer, and E. Szathmáry. *The Principles of Life*. Oxford University Press Oxford, UK:, 2003.

[40] A.G. Cairns-Smith. *Genetic Takeover and the Mineral Origins of Life*. Cambridge University Press Cambridge, 1982.

[41] B. Obermayer and E. Frey. Error thresholds for self-and cross-specific enzymatic replication. *Journal of Theoretical Biology*, 267(4):653–662, 2010.

[42] J.A.D. Wattis and P.V. Coveney. The origin of the RNA world: a kinetic model. *The Journal of Physical Chemistry B*, 103(21):4231–4250, 1999.

[43] P.V. Coveney and J.A.D. Wattis. Coarse-graining and renormalization group methods for the elucidation of the kinetics of complex nucleation and growth processes. *Molecular Physics*, 104(1):177–185, 2006.

[44] R. Hanel, M. Pöchacker, and S. Thurner. Living on the edge of chaos: minimally nonlinear models of genetic regulatory dynamics. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 368(1933): 5583–5596, 2010.

[45] N. Lehman. A recombination-based model for the origin and early evolution of genetic information. *Chemistry & Biodiversity*, 5(9):1707–1717, 2008.

[46] S. Lonchin, P.L. Luisi, P. Walde, and B.H. Robinson. A matrix effect in mixed phospholipid/fatty acid vesicle formation. *The Journal of Physical Chemistry B*, 103(49):10910–10916, 1999.

[47] P.V. Coveney and J.A.D. Wattis. Becker–Döring model of self-reproducing vesicles. *Journal of the Chemical Society, Faraday Transactions*, 94(2):233–246, 1998.

[48] C.D. Bolton and J.A.D. Wattis. The size-templating matrix effect in vesicle formation i: a microscopic model and analysis. *The Journal of Physical Chemistry B*, 107(29):7126–7134, 2003.

[49] N. Berclaz, E. Blöchliger, M. Müller, and P.L. Luisi. Matrix effect of vesicle formation as investigated by cryotransmission electron microscopy. *The Journal of Physical Chemistry B*, 105(5):1065–1071, 2001.

[50] N. Berclaz, M. Müller, P. Walde, and P.L. Luisi. Growth and transformation of vesicles studied by ferritin labeling and cryotransmission electron microscopy. *The Journal of Physical Chemistry B*, 105(5):1056–1064, 2001.

[51] S.J. Mojzsis, T.M. Harrison, and R.T. Pidgeon. Oxygen-isotope evidence from ancient zircons for liquid water at the earth's surface 4,300 Myr ago. *Nature*, 409 (6817):178–181, 2001.

[52] P. Baaske, F.M. Weinert, S. Duhr, K.H. Lemke, M.J. Russell, and D. Braun. Extreme accumulation of nucleotides in simulated hydrothermal pore systems. *Proceedings of the National Academy of Sciences*, 104(22):9346, 2007.

[53] D. Braun and A. Libchaber. Thermal force approach to molecular evolution. *Physical Biology*, 1:1–8, 2004.

[54] J.F. Lambert. Adsorption and polymerization of amino acids on mineral surfaces: a review. *Origins of Life and Evolution of Biospheres*, 38(3):211–242, 2008.

[55] K. Plankensteiner, H. Reiner, and B.M. Rode. Prebiotic chemistry: The amino acid and peptide world. *Current Organic Chemistry*, 9(12):1107–1114, 2005.

[56] M.G. Schwendinger and B. M Rode. A monte carlo simulation of a supersaturated sodium chloride solution. *Chemical physics letters*, 155(6):527–532, 1989.

[57] N. Lahav, D. White, and S. Chang. Peptide formation in the prebiotic era: Thermal condensation of glycine in fluctuating clay environments. *Science*, 201(4350): 67–69, 1978.

[58] N. Lahav and D.H. White. A possible role of fluctuating clay-water systems in the production of ordered prebiotic oligomers. *Journal of Molecular Evolution*, 16(1): 11–21, 1980.

[59] J.R. Collins, G.H. Loew, B.T. Luke, and D.H. White. Theoretical investigation of the role of clay edges in prebiotic peptide bond formation. *Origins of Life and Evolution of Biospheres*, 18(1):107–119, 1988.

[60] A.J.A. Aquino, D. Tunega, M.H. Gerzabek, and H. Lischka. Modeling catalytic effects of clay mineral surfaces on peptide bond formation. *The Journal of Physical Chemistry B*, 108(28):10120–10130, 2004.

[61] B.M. Rode. Peptides and the origin of life. *Peptides*, 20(6):773–786, 1999.

[62] B.M. Rode and Y. Suwannachot. The possible role of Cu (II) for the origin of life. *Coordination Chemistry Reviews*, 190:1085–1099, 1999.

[63] K.R. Liedl and B.M. Rode. Ab initio calculations concerning the reaction mechanism of the copper (II) catalyzed glycine condensation in aqueous sodium chloride solution. *Chemical Physics Letters*, 197(1-2):181–186, 1992.

[64] M. Remko and B.M. Rode. Bivalent cation binding effect on formation of the peptide bond. *Chemical Physics Letters*, 316(5):489–494, 2000.

[65] A. Rimola, L. Rodriguez-Santiago, P. Ugliengo, and M. Sodupe. Is the peptide bond formation activated by Cu^{2+} interactions? insights from density functional calculations. *The Journal of Physical Chemistry B*, 111(20):5740–5747, 2007.

[66] N.N. Nair, E. Schreiner, and D. Marx. Glycine at the pyrite-water interface: The role of surface defects. *Journal of the American Chemical Society*, 128(42):13815–13826, 2006.

[67] L. Leman, L. Orgel, and M.R. Ghadiri. Carbonyl sulfide-mediated prebiotic formation of peptides. *Science*, 306(5694):283–286, 2004.

[68] C.H. Yu, S.Q. Newton, D.M. Miller, B.J. Teppen, and L. Schäfer. Ab initio study of the nonequivalence of adsorption of D-and L-peptides on clay mineral surfaces. *Structural Chemistry*, 12(5):393–398, 2001.

[69] M.A. Thyveetil, P.V. Coveney, H.C. Greenwell, and J.L. Suter. Computer simulation study of the structural stability and materials properties of DNA-intercalated layered double hydroxides. *Journal of the American Chemical Society*, 130(14): 4742–4756, 2008.

[70] D.C. Mathew and Z. Luthey-Schulten. Influence of montmorillonite on nucleotide oligomerization reactions: A molecular dynamics study. *Origins of Life and Evolution of Biospheres*, 40(3):303–317, 2010.

[71] W. Huang and J.P. Ferris. One-step, regioselective synthesis of up to 50-mers of RNA oligomers by montmorillonite catalysis. *Journal of the American Chemical Society*, 128(27):8914–8919, 2006.

[72] M.A. Thyveetil, P.V. Coveney, H.C. Greenwell, and J.L. Suter. Role of host layer flexibility in DNA guest intercalation revealed by computer simulation of layered nanomaterials. *Journal of the American Chemical Society*, 130(37):12485–12495, 2008.

[73] G.O. Arrhenius. Crystals and life. *Helvetica Chimica Acta*, 86(5):1569–1586, 2003.

[74] N.N. Greenwood and A. Earnshaw. *Chemistry of the Elements*. Pergamon Press, 1984.

[75] R.A. Schoonheydt, T. Pinnavaia, G. Lagaly, and N. Gangas. Pillared clays and pillared layered solids. *Pure Applied Chemistry*, 71(12):2367–2371, 1999.

[76] J. Swenson, GA Schwartz, R. Bergman, and WS Howells. Dynamics of propylene glycol and its oligomers confined in clay. *The European Physical Journal E: Soft Matter and Biological Physics*, 12(1):179–183, 2003.

[77] W.W. Kagunya. Properties of water adsorbed in anionic clays: A neutron scattering study. *The Journal of Physical Chemistry*, 100(1):327–330, 1996.

[78] D. Evans and R. Slade. Structural aspects of layered double hydroxides. *Layered Double Hydroxides*, pages 1–87, 2006.

[79] K. Yao, M. Taniguchi, M. Nakata, M. Takahashi, and A. Yamagishi. Nanoscale imaging of molecular adsorption of metal complexes on the surface of a hydrotalcite crystal. *Langmuir*, 14(9):2410–2414, 1998.

[80] C.J. Serna, J.L. Rendon, and J.E. Iglesias. Crystal-chemical study of layered $[ai_2li(oh)_6]^+X^- \cdot nH_2 0$. *Clay Minerals*, 30(3):180–184, 1982.

[81] M. Bellotto, B. Rebours, O. Clause, J. Lynch, D. Bazin, and E. Elkaim. A reexamination of hydrotalcite crystal chemistry. *The Journal of Physical Chemistry*, 100(20):8527–8534, 1996.

[82] A. Trave, A. Selloni, A. Goursot, D. Tichit, and J. Weber. First principles study of the structure and chemistry of Mg-based hydrotalcite-like anionic clays. *The Journal of Physical Chemistry B*, 106(47):12291–12296, 2002.

[83] A.S. Bookin and V.A. Drits. Polytype diversity of the hydrotalcite-like minerals. i. possible polytypes and their diffraction features. *Clays and Clay Minerals*, 41 (5):551–557, 1993.

[84] C. de la Calle, C.H. Pons, J. Roux, and V. Rives. A crystal-chemical study of natural and synthetic anionic clays. *Clays and clay minerals*, 51(2):121–132, 2003.

[85] D.R. Hines, S.A. Solin, U. Costantino, and M. Nocchetti. Physical properties of fixed-charge layer double hydroxides. *Physical Review B*, 61(17):11348, 2000.

[86] M.F. Thorpe. Layer rigidity and spacing in intercalation compounds. *Physical Review B*, 30(10370):l989, 1989.

[87] S.A. Solin. Clays and clay intercalation compounds: properties and physical phenomena. *Annual Review of Materials Science*, 27(1):89–115, 1997.

[88] V.R.L. Constantino and T.J. Pinnavaia. Basic properties of $mg_{1-x}^{2+}al_x^{3+}$ layered double hydroxides intercalated by carbonate, hydroxide, chloride, and sulfate anions. *Inorganic Chemistry*, 34(4):883–892, 1995.

[89] F. Kooli, I.C. Chisem, M. Vucelic, and W. Jones. Synthesis and properties of terephthalate and benzoate intercalates of Mg-Al layered double hydroxides possessing varying layer charge. *Chemistry of Materials*, 8(8):1969–1977, 1996.

[90] S.P. Newman, S.J. Williams, P.V. Coveney, and W. Jones. Interlayer arrangement of hydrated mgal layered double hydroxides containing guest terephthalate anions: Comparison of simulation and measurement. *The Journal of Physical Chemistry B*, 102(35):6710–6719, 1998.

[91] J.M. Cases, I. Bérend, G. Besson, M. Francois, J.P. Uriot, F. Thomas, and J.E. Poirier. Mechanism of adsorption and desorption of water vapor by homoionic montmorillonite. 1. the sodium-exchanged form. *Langmuir*, 8(11):2730–2739, 1992.

[92] E.J.M. Hensen and B. Smit. Why clays swell. *The Journal of Physical Chemistry B*, 106(49):12664–12667, 2002.

[93] N.T. Skipper, A.K. Soper, and M.V. Smalley. Neutron diffraction study of calcium vermiculite: hydration of calcium ions in a confined environment. *The Journal of Physical Chemistry*, 98(3):942–945, 1994.

[94] G. Sposito and R. Prost. Structure of water adsorbed on smectites. *Chemical Reviews*, 82(6):553–573, 1982.

[95] S.M. Shevchenko and G.W. Bailey. Modeling sorption of soil organic matter on mineral surfaces: wood-derived polymers on mica. *Supramolecular Science*, 5(1): 143–157, 1998.

[96] J.H. Harding, D.M. Duffy, M.L. Sushko, P.M. Rodger, D. Quigley, and J.A. Elliott. Computational techniques at the organic-inorganic interface in biomineralization. *Chemical Reviews*, 108(11):4823, 2008.

[97] A.D. McLaren and G.H. Peterson. Montmorillonite as a caliper for the size of protein molecules. *Nature*, 192(4806):960–961, 1961.

[98] L. Perezgasga, A. Serrato-Diaz, A. Negron-Mendoza, L.D.P. GalN, and FG Mosqueira. Sites of adsorption of adenine, uracil, and their corresponding derivatives on sodium montmorillonite. *Origins of Life and Evolution of Biospheres*, 35(2):91–110, 2005.

[99] J.P. Ferris, G. Ertem, and V.K. Agarwal. The adsorption of nucleotides and polynucleotides on montmorillonite clay. *Origins of Life and Evolution of Biospheres*, 19(2):153–164, 1989.

[100] D. Winter and G. Zubay. Binding of adenine and adenine-related compounds to the clay montmorillonite and the mineral hydroxylapatite. *Origins of Life and Evolution of the Biosphere*, 25(1-3):61–81, 1995.

[101] M. W. Powner, B. Gerland, and J. D. Sutherland. Synthesis of activated pyrimidine ribonucleotides in prebiotically plausible conditions. *Nature*, 459(7244):239–242, 2009.

[102] P.E. Nielsen, M. Egholm, R.H. Berg, and O. Buchardt. Sequence-selective recognition of DNA by strand displacement with a thymine-substituted polyamide. *Science*, 254(5037):1497, 1991.

[103] M. Egholm, O. Buchardt, L. Christensen, C. Behrens, S.M. Freier, D.A. Driver, R.H. Berg, S.K. Kim, B. Norden, and P.E. Nielsen. PNA hybridizes to complementary oligonucleotides obeying the Watson-Crick hydrogen-bonding rules. *Science (New York, NY)*, 365(1):566–568, 1993.

[104] William L Jorgensen. Quantum and statistical mechanical studies of liquids. 10. transferable intermolecular potential functions for water, alcohols, and ethers. application to liquid water. *Journal of the American Chemical Society*, 103(2):335–340, 1981.

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

[106] J.M. Wang, P. Cieplak, and P.A. Kollman. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *Journal of Computational Chemistry*, 21(12):1049–1074, 2000.

[107] P. Cieplak, J. Caldwell, and P. Kollman. Molecular mechanical models for organic and biological systems going beyond the atom centered two body additive approximation: aqueous solution free energies of methanol and N-methyl acetamide, nucleic acid base, and amide hydrogen bonding and chloroform/water partition coefficients of the nucleic acid bases. *Journal of Computational Chemistry*, 22(10):1048–1057, 2001.

[108] A.D. MacKerell Jr, D. Bashford, M. Bellott, R.L. Dunbrack Jr, J.D. Evanseck, M.J. Field, S. Fischer, J. Gao, H. Guo, and S. Ha. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *The Journal of Physical Chemistry B*, 102(18):3586–3616, 1998.

[109] S. Patel, A.D. Mackerell Jr, and C.L. Brooks III. CHARMM fluctuating charge force field for proteins: II protein/solvent properties from molecular dynamics

simulations using a nonadditive electrostatic model. *Journal of Computational Chemistry*, 25(12):1504–1514, 2004.

[110] C. Oostenbrink, A. Villa, A.E. Mark, and W.F. Van Gunsteren. A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets 53A5 and 53A6. *Journal of Computational Chemistry*, 25(13):1656–1676, 2004.

[111] D.J. Price and C.L. Brooks. Modern protein force fields behave comparably in molecular dynamics simulations. *Journal of Computational Chemistry*, 23(11): 1045–1057, 2002.

[112] B. Hess and N.F.A. van der Vegt. Hydration thermodynamic properties of amino acid analogues: a systematic comparison of biomolecular force fields and water models. *The Journal of Physical Chemistry B*, 110(35):17616–17626, 2006.

[113] M.G. Martin. Comparison of the AMBER, CHARMM, COMPASS, GROMOS, OPLS, traPPE and UFF force fields for prediction of vapor–liquid coexistence curves and liquid densities. *Fluid Phase Equilibria*, 248(1):50–55, 2006.

[114] R.T. Cygan, J.J. Liang, and A.G. Kalinichev. Molecular models of hydroxide, oxyhydroxide, and clay phases and the development of a general force field. *The Journal of Physical Chemistry B*, 108(4):1255–1266, 2004.

[115] HA Lorentz. Über die anwendung des satzes vom virial in der kinetischen theorie der gase. *Annalen der Physik*, 248(1):127–136, 1881.

[116] D. Berthelot. Sur le mélange des gaz. *Comptes Rendus*, 126:1703–1706, 1898.

[117] D. Boda and D. Henderson. The effects of deviations from lorentz–berthelot rules on the properties of a simple mixture. *Molecular Physics*, 106(20):2367–2370, 2008.

[118] Dorothy M Duffy and John H Harding. Simulation of organic monolayers as templates for the nucleation of calcite crystals. *Langmuir*, 20(18):7630–7636, 2004.

[119] Alastair N Cormack, Raymond Jess Lewis, and Alan H Goldstein. Computer simulation of protein adsorption to a material surface in aqueous solution: Biomaterials modeling of a ternary system. *The Journal of Physical Chemistry B*, 108 (52):20408–20418, 2004.

[120] J. Delhommelle and P. Millié. Inadequacy of the lorentz-berthelot combining rules for accurate predictions of equilibrium properties by molecular simulation. *Molecular Physics*, 99(8):619–625, 2001.

[121] J. Šponer and N. Špačková. Molecular dynamics simulations and their application to four-stranded dna. *Methods*, 43(4):278–290, 2007.

[122] J. Šponer, J.V. Burda, M. Sabat, J. Leszczynski, and P. Hobza. Interaction between the guanine-cytosine watson-crick dna base pair and hydrated group IIa (Mg2+, Ca2+, Sr2+, Ba2+) and group IIb (Zn2+, Cd2+, Hg2+) metal cations. *The Journal of Physical Chemistry A*, 102(29):5951–5957, 1998.

[123] J. Šponer, M. Sabat, L. Gorb, J. Leszczynski, B. Lippert, and P. Hobza. The effect of metal binding to the n7 site of purine nucleotides on their structure, energy, and involvement in base pairing. *The Journal of Physical Chemistry B*, 104(31): 7535–7544, 2000.

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

[125] A. Pérez, I. Marchán, D. Svozil, J. Sponer, T.E. Cheatham III, C.A. Laughton, and M. Orozco. Refinement of the amber force field for nucleic acids: improving the description of [alpha]/[gamma] conformers. *Biophysical Journal*, 92(11):3817–3829, 2007.

[126] T.E. Cheatham III and M.A. Young. Molecular dynamics simulation of nucleic acids: Successes, limitations, and promise. *Biopolymers*, 56(4):232–256, 2000.

[127] D.A. Case, T.A. Darden, T.E. Cheatham III, C.S. Simmerling, J. Wang, R.E. Duke, R. Luo, R.C. Walker, W. Zhang, K.M. Merz, B.P. Roberts, B. Wang, S. Hayik, A. Roitberg, G. Seabra, I. Kolossvai, K.F. Wong, F. Paesani, J. Vanicek, J. Liu, X. Wu, S.R. Brozell, T. Steinbrecher, H. Gohlke, Q. Cai, X Ye, J. Wang, M.-J. Hsieh, G. Cui, D.R. Roe, D.H. Mathews, M.G. Seetin, C. Sagui, V. Babin, T. Luchko, S. Gusarov, A. Kovalenko, and P.A. Kollman. Amber tools. *AMBER 11*, University of California, San Francisco, 2011.

[128] G.C. Shields, C.A. Laughton, and M. Orozco. Molecular dynamics simulation of a PNA DNA PNA triple helix in aqueous solution. *Journal of the American Chemical Society*, 120:5895–5904, 1998.

[129] J. Wang, A.G. Kalinichev, R.J. Kirkpatrick, and X. Hou. Molecular modeling of the structure and energetics of hydrotalcite hydration. *Chemistry of Materials*, 13 (1):145–150, 2001.

[130] J. Wang, A.G. Kalinichev, and R.J. Kirkpatrick. Effects of substrate structure and composition on the structure, dynamics, and energetics of water at mineral

surfaces: A molecular dynamics modeling study. *Geochimica et Cosmochimica Acta*, 70(3):562–582, 2006.

[131] K. Yu and J.R. Schmidt. Elucidating the crystal face-and hydration-dependent catalytic activity of hydrotalcites in biodiesel production. *The Journal of Physical Chemistry C*, 115:1887–1898, 2011.

[132] J.B. Swadling, P.V. Coveney, and H.C. Greenwell. Clay minerals mediate folding and regioselective interactions of RNA: A large-scale atomistic simulation study. *Journal of the American Chemical Society*, 132(39):13750–13764, 2010.

[133] J.B. Swadling, P.V. Coveney, and H.C. Greenwell. Stability of free and mineral-protected nucleic acids: Implications for the RNA world. *Geochimica et Cosmochimica Acta*, 1(83):360–378, 2012.

[134] W.L. Jorgensen and J. Tirado-Rives. Potential energy functions for atomic-level simulations of water and organic and biomolecular systems. *Proceedings of the National Academy of Sciences of the United States of America*, 102(19):6665–6670, 2005.

[135] C. Vega, J.L.F. Abascal, M.M. Conde, and J.L. Aragones. What ice can teach us about water interactions: a critical comparison of the performance of different water models. *Faraday Discussions*, 141:251–276, 2009.

[136] S. Plimpton. Fast parallel algorithms for short-range molecular dynamics. *Journal of Computational Physics*, 117(1):1–19, 1995.

[137] M.-A. Thyveetil, P.V. Coveney, H.C. Greenwell, and J.L. Suter. Computer simulation study of the structural stability and materials properties of DNA-intercalated layered double hydroxides. *Journal of the American Chemical Society*, 130:4742–4756, 2008.

[138] L. Verlet. Computer "experiments" on classical fluids. i. thermodynamical properties of Lennard-Jones molecules. *Physical Review*, 159(1):98, 1967.

[139] J.-P. Ryckaert, G. Ciccotti, and H.J.C. Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics*, 23(3):327–341, 1977.

[140] A.F. Voter, F. Montalenti, and T.C. Germann. Extending the time scale in atomistic simulation of materials. *Annual Review of Materials Research*, 32(1):321–346, 2002.

[141] A.F. Voter. Parallel replica method for dynamics of infrequent events. *Physical Review B*, 57(22):13985–13988, 1998.

[142] A.F. Voter. A method for accelerating the molecular dynamics simulation of infrequent events. *The Journal of Chemical Physics*, 106(11):4665–4677, 1997.

[143] M R. Sørensen and A F. Voter. Temperature-accelerated dynamics for simulation of infrequent events. *The Journal of Chemical Physics*, 112(1):9599, 2000.

[144] Y. Sugita and Y. Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters*, 314(1):141–151, 1999.

[145] J.L. Suter, R.L. Anderson, H.C. Greenwell, and P.V. Coveney. Recent advances in large-scale atomistic and coarse-grained molecular dynamics simulation of clay minerals. *Journal of Materials Chemistry*, 19(17):2482–2493, 2009.

[146] J.L. Suter, P.V. Coveney, H.C. Greenwell, and M.-A. Thyveetil. Large-scale molecular dynamics study of montmorillonite clay: Emergence of undulatory fluctuations and determination of material properties. *The Journal of Physical Chemistry C*, 111(23):8248–8259, 2007.

[147] H.C. Greenwell, W. Jones, P.V. Coveney, and S. Stackhouse. On the application of computer simulation techniques to anionic and cationic clays: A materials chemistry perspective. *Journal of Materials Chemistry*, 16(8):708–723, 2005.

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

[149] M.A. Thyveetil, P.V. Coveney, J.L. Suter, and H.C. Greenwell. Emergence of undulations and determination of materials properties in large-scale molecular dynamics simulation of layered double hydroxides. *Chemistry of Materials*, 19 (23):5510–5523, 2007.

[150] J. Li. Atomeye: an efficient atomistic configuration viewer. *Modelling and Simulation in Materials Science and Engineering*, 11(2):173–177, 2003.

[151] W. Humphrey, A. Dalke, and K. Schulten. VMD: visual molecular dynamics. *Journal of molecular graphics*, 14(1):33–38, 1996.

[152] P.V. Coveney. Scientific grid computing. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 363(1833):1707–1713, 2005.

[153] I. Budin and J.W. Szostak. Expanding roles for diverse physical phenomena during the origin of life. *Annual Review Biophysics*, 39:245–263, 2010.

[154] H. Follmann and C. Brownson. Darwins warm little pond revisited: from molecules to the origin of life. *Naturwissenschaften*, 96:1265–1292, 2009.

[155] C. Woese. *The Genetic Code.* Harper and Row, New York, 1967.

[156] G.F. Joyce. Booting up life. *Nature*, 418:214–221, 2002.

[157] L.E. Orgel. Prebiotic chemistry and the origin of the RNA world. *Critical Reviews in Biochemical Molecular Biology*, 39:99–123, 2004.

[158] L.K.L. Cheng and P.J. Unrau. Closing the circle: Replicating RNA with RNA. *Cold Spring Harbor Perspectives Biology*, 2:a002204, 2010.

[159] M. Wu and P.G. Higgs. Origin of self-replicating biopolymers: Autocatalytic feedback can jump-start the RNA world. *Journal of Molecular Evolution*, 69: 541–554, 2009.

[160] P.J. Unrau and D.P. Bartel. Combinatorial minimization and secondary structure determination of a nucleotide synthase ribozyme. *Nature*, 395:260–263, 1998.

[161] J.D. Sutherland. Ribonucleotides. *Cold Spring Harbor Perspectives Biology*, 2: a005439, 2010.

[162] P.C. Joshi, M.F. Aldersley, J.W. Delano, and J.P. Ferris. Mechanism of montmorillonite catalysis in the formation of RNA oligomers. *Journal of the American Chemical Society*, 131(37):13369–13374, 2009.

[163] A.M. Poole, D.C. Jeffares, and D. Penny. The path from the RNA world. *Journal of Molecular Evolution*, 46(1):1–17, 1998.

[164] D.C. Jeffares, A.M. Poole, and D. Penny. Relics from the RNA world. *Journal of Molecular Evolution*, 46(1):18–36, 1998.

[165] W.G. Scott, J.B. Murray, J.R.P. Arnold, B.L. Stoddard, and A. Klug. Capturing the structure of a catalytic RNA intermediate: the hammerhead ribozyme. *Science*, 274:2065–2069, 1996.

[166] P. E. Nielsen. Peptide nucleic acids (PNA) in chemical biology and drug discovery. *Chemistry and Biodiversity*, 4:1996–2002, 2007.

[167] K.E. Nelson, M. Levy, and S.L. Miller. Peptide nucleic acids rather than RNA may have been the first genetic molecule. *Proceedings from the National Academy of Science*, 97(8):3868, 2000.

[168] Y. Ura, J.M. Beierle, L.J. Leman, L.E. Orgel, and M.R. Ghadiri. Self-assembling sequence-adaptive peptide nucleic acids. *Science*, 325:73, 2009.

[169] W. Martin, J. Baross, D. Kelley, and M.J. Russell. Hydrothermal vents and the origin of life. *Nature Reviews: Microbiology*, 6(11):805–814, 2008.

[170] W. Martin and M.J. Russell. On the origins of cells: a hypothesis for the evolutionary transitions from abiotic geochemistry to chemoautotrophic prokaryotes, and from prokaryotes to nucleated cells. *Philosophical Transactions Biological Science*, 358:59–85, 2003.

[171] T.M. McCollom and E.L. Shock. Hydrothermal geochemistry and the deep biosphere. *Geochimica et Cosmochima Acta*, 61:4375–4391, 1997.

[172] Q. Fu, B. Sherwood Lollar, J. Horita, G. Lacrampe-Couloume, and W.E. Seyfried, Jr. Abiotic formation of hydrocarbons under hydrothermal conditions: Constraints from chemical and isotope data. *Geochimica et Cosmochima Acta*, 71:1982–1998, 2007.

[173] R.M. Hazen and D.A. Sverjensky. Mineral surfaces, geochemical complexities, and the origins of life. *Cold Spring Harbor Perspectives Biology*, 2:a002162, 2010.

[174] H.G. Hansma. Possible origin of life between mica sheets. *Journal of Theoretical Biology*, 266:175–188, 2010.

[175] S. Pitsch, A. Eschenmoser, B. Gedulin, S. Hui, and G. Arrhenius. Mineral induced formation of sugar phosphates. *Origins of Life and Evolution of the Biosphere*, 25 (4):297–334, 1995.

[176] G. Arrhenius. *Phosphate in Models for Chemical Evolution. In Chemical Evolution: Origin of life.* Deepak Publishers, Hampton, Virginia, U.S., 1993.

[177] G. Arrhenius, B. Sales, S. Mojzsis, and T. Lee. Entropy and charge in molecular evolution-the case of phosphate. *Journal of Theoretical Biology*, 187(4):503–522, 1997.

[178] R. Krishnamurthy, S. Pitsch, and G. Arrhenius. Mineral induced formation of pentose-2, 4-bisphosphates. *Origins of Life and Evolution of the Biosphere*, 26(3): 240–241, 1996.

[179] H.D. Bean, F.A.L. Anet, I.R. Gould, and N.V. Hud. Glyoxylate as a backbone linkage for a prebiotic ancestor of RNA. *Origins of Life and Evolution of the Biosphere*, 36(1):39–63, 2006.

[180] S.-Y. Kwak, W.M. Kriven, M.A. Wallig, and J.-H. Choy. Inorganic delivery vector for intravenous injection. *Biomaterials*, 25(28):5995–6001, 2004.

[181] J.-H. Choy, S.-Y. Kwak, J.-S. Park, Y.-J. Jeong, and J. Portier. Intercalative nanohybrids of nucleoside monophosphates and DNA in layered metal hydroxide. *Journal of the American Chemical Society*, 121(6):1399–1400, 1999.

[182] J.H. Choy, S.Y. Kwak, Y.J. Jeong, and J.S. Park. Inorganic layered double hydroxides as nonviral vectors. *Angewandte Chemie*, 39(22):4041–4045, 2000.

[183] J.-H. Choy, S.-Y. Kwak, J.-S. Park, and Y.-J. Jeong. Cellular uptake behavior of [γ-32p] labeled ATP–LDH nanohybrids. *Journal of Materials Chemistry*, 11:1671 – 1674, 2001.

[184] S.-Y. Kwak, J.-S. Park, Y.-J. Jeong, and J.-H. Choy. Bio-LDH nanohybrid for gene therapy. *Solid State Ionics*, 151(1-4):229–234, 2002.

[185] J.H. Choy, S.J. Choi, J.-M. Oh, and T. Park. Clay minerals and double layered hydroxides for novel biological applications. *Applied Clay Science*, 36:122–132, 2007.

[186] J.-H. Choy, J.-M. Oh, M. Park, K.-M. Sohn, and J.-W. Kim. Inorganic–biomolecular hybrid nanomaterials as a genetic molecular code system. *Advanced Materials*, 16(14):1181–1184, 2004.

[187] L. Desigaux, M.B. Belkacem, P. Richard, J. Cellier, P. Leone, L. Cario, F. Leroux, C. Taviot-Gueho, and B. Pitard. Self-assembly and characterization of layered double hydroxide/DNA hybrids. *Nano Letters*, 6(2):199–204, 2006.

[188] J.-M. Oh, S.-Y. Kwak, and J.-H. Choy. Intracrystalline structure of DNA molecules stabilized in the layered double hydroxide. *Journal of Physics and Chemistry of Solids*, 67(5-6):1028–1031, 2006.

[189] H.C. Greenwell and P.V. Coveney. Layered double hydroxide minerals as possible prebiotic information storage and transfer compounds. *Origins of Life and Evolution of Biospheres*, 36(1):13–37, 2006.

[190] H.C. Greenwell, A.A. Bowden, B. Chen, P. Boulet, J.R.G. Evans, P.V. Coveney, and A. Whiting. Intercalation and in situ polymerization of poly (alkylene oxide) derivatives within M^+-montmorillonite (M= Li, Na, K). *Journal of Materials Chemistry*, 16(11):1082–1094, 2006.

[191] M.A. Thyveetil, P.V. Coveney, H.C. Greenwell, and J.L. Suter. Role of host layer flexibility in DNA guest intercalation revealed by computer simulation of layered nanomaterials. *Journal of the American Chemical Society*, 130:12485–12495, 2008.

[192] S. Jha, P.V. Coveney, and C.A. Laughton. Force field validation for nucleic acid simulations: Comparing energies and dynamics of a DNA dodecamer. *Journal of computational chemistry*, 26(15):1617–1627, 2005.

[193] R. Soliva, E. Sherer, F.J. Luque, C.A. Laughton, and M. Orozco. Molecular dynamics simulations of PNA-DNA and PNA-RNA duplexes in aqueous solution. *Journal of the American Chemical Society*, 122:5997–6008, 2000.

[194] P.E. Nielsen and M. Egholm. An introduction to peptide nucleic acid. *Curr. Issues Mol. Biol.*, 1(1-2):89–104, 1999.

[195] S.R. Holbrook, C. Cheong, I. Tinoco, and S.H. Kim. Crystal structure of an RNA double helix incorporating a track of non-Watson–Crick base pairs. *Nature*, 353: 579–581, 1991.

[196] A. Pérez, A. Noy, F. Lankas, F.J. Luque, and M. Orozco. The relative flexibility of B-DNA and A-RNA duplexes: database analysis. *Nucleic Acids Research*, 32: 6144, 2004.

[197] E. Gavathiotis, G.J. Sharman, and M.S. Searle. Sequence-dependent variation in DNA minor groove width dictates orientational preference of hoechst 33258 in A-tract recognition: solution NMR structure of the 2:1 complex with d(CTTTTGCAAAAG)$_2$. *Nuc. Acids Res.*, 28:728, 2000.

[198] S.A. Harris, E. Gavathiotis, M.S. Searle, M. Orozco, and C.A. Laughton. Cooperativity in drug-DNA recognition: a molecular dynamics study. *JJournal of the American Chemical Society*, 123:12658–12663, 2001.

[199] R. Corradini, S. Sforza, T. Tedeschi, F. Totsingan, and R. Marchelli. Peptide nucleic acids with a structurally biased backbone: effects of conformational constraints and stereochemistry. *Current Topics in Med. Chem.*, 7:681–694, 2007.

[200] S. Sen and L. Nilsson. Molecular dynamics of duplex systems involving PNA: structural and dynamical consequences of the nucleic acid backbone. *JJournal of the American Chemical Society*, 120:619–631, 1998.

[201] J. Srinivasan, T.E. Cheatham III, P. Cieplak, P.A. Kollman, and D.A. Case. Continuum solvent studies of the stability of DNA, RNA, and phosphoramidate-DNA helices. *JJournal of the American Chemical Society*, 120:9401–9409, 1998.

[202] A. Noy, A. Pérez, C.A. Laughton, and M. Orozco. Theoretical study of large conformational transitions in DNA: the B A conformational change in water and ethanol/water. *Nucleic Acids Research*, 35:3330, 2007.

[203] E.S. Boek, P.V. Coveney, and N.T. Skipper. Monte carlo molecular modeling studies of hydrated Li-, Na-, and K-smectites: Understanding the role of potassium as a clay swelling inhibitor. *Journal of the American Chemical Society*, 117(50): 12608 – 12617, 1995.

[204] E.S. Boek, P.V. Coveney, and N.T. Skipper. Molecular modeling of clay hydration: a study of hysteresis loops in the swelling curves of sodium montmorillonites. *Langmuir*, 11(12):4629–4631, 1995.

[205] J.-H. Choy and Y.-H. Son. Intercalation of vitamer into LDH and their controlled release properties. *Bull. Korean Society*, 25(1):122–126, 2004.

[206] L. Zaikowski and J. Friedrich. *Chemical Evolution across Space and Time From the Big Bang to Prebiotic Chemistry*. OUP, Oxford, UK, 2008. ISBN 978-0-8412-7431-0.

[207] E.B. Watson and T.M. Harrison. Zircon thermometer reveals minimum melting conditions on earliest earth. *Science*, 308(5723):841–844, 2005.

[208] S.A. Wilde, J.W. Valley, W.H. Peck, and C.M. Graham. Evidence from detrital zircons for the existence of continental crust and oceans on the earth 4.4 gyr ago. *Nature*, 409(6817):175–178, 2001.

[209] M.D. Nikalje, P. Phukan, and A. Sudalai. Recent advances in clay-catalyzed organic transformations. *Organic Preparations and Procedures International*, 32(1): 1, 2000.

[210] J.P. Ferris and G. Ertem. Montmorillonite catalysis of RNA oligomer formation in aqueous-solution - a model for the prebiotic formation of RNA. *Journal of the American Chemical Society*, 115(26):12270–12275, 1993.

[211] K. Kawamura and J.P. Ferris. Clay catalysis of oligonucleotide formation: Kinetics of the reaction of the 5′-phosphorimidazolides of nucleotides with the non-basic heterocycles uracil and hypoxanthine. *Origins of Life and Evolution of the Biosphere*, 29(6):563–591, 1999.

[212] J.P. Ferris, A.R. Hill, R.H. Liu, and L.E. Orgel. Synthesis of long prebiotic oligomers on mineral surfaces. *Nature*, 381(6577):59–61, 1996.

[213] M. Franchi, J.P. Ferris, and E. Gallori. Cations as mediators of the adsorption of nucleic acids on clay surfaces in prebiotic environments. *Origins of Life and Evolution of the Biosphere*, 33(1):1–16, 2003.

[214] J.P. Dworkin, A. Lazcano, and S.L. Miller. The roads to and from the RNA world. *Journal of theoretical biology*, 222(1):127–134, 2003.

[215] T.R. Cech. The ribosome is a ribozyme. *Science*, 289(5481):878–879, 2000.

[216] P.C. Joshi, S. Pitsch, and J.P. Ferris. Selectivity of montmorillonite catalyzed prebiotic reactions of D, L-nucleotides. *Origins of Life and Evolution of Biospheres*, 37(1):3–26, 2007.

[217] D. Pastre, O. Pietrement, P. Fusil, F. Landousy, J. Jeusset, M.O. David, C. Hamon, E. Le Cam, and A. Zozime. Adsorption of DNA to mica mediated by divalent counterions: A theoretical and experimental study. *Biophysical Journal*, 85 (4):2507–2518, 2003.

[218] H.G. Hansma, I. Revenko, K. Kim, and D.E. Laney. Atomic force microscopy of long and short double-stranded, single-stranded and triple-stranded nucleic acids. *Nucleic acids research*, 24(4):713–720, 1996.

[219] V.M. Goldschmidt. Geochemical aspects of the origin of complex organic molecules on the earth, as precursors to organic life. *New Biology*, 12:97–105, 1952.

[220] J.D. Bernal. The physical basis of life. *Proceedings of the Physical Society of London Section A*, 62(357):537–558, 1949.

[221] P. Schuster. Evolution in silico and in vitro: the RNA model. *Biological chemistry*, 382(9):1301, 2001.

[222] J. Sponer and F. Lankas. *Computational Studies of RNA and DNA*, volume 2. Springer, 2006.

[223] J.P. Ferris. Montmorillonite catalysis of 30–50 mer oligonucleotides: Laboratory demonstration of potential steps in the origin of the RNA world. *Origins of Life and Evolution of Biospheres*, 32(4):311–332, 2002.

[224] R.A. Torres and T.C. Bruice. Molecular dynamics study displays near in-line attack conformations in the hammerhead ribozyme self-cleavage reaction. *Proceedings of the National Academy of Science of the United States of America*, 95 (19):11077–11082, 1998.

[225] A.R. Gruber, R. Lorenz, S.H. Bernhart, R. Neuböck, and I.L. Hofacker. The vienna RNA websuite. *Nucleic Acids Research*, 36(suppl 2):W70–W74, 2008.

[226] J.P. Ferris and G. Ertem. Oligomerization of ribonucleotides on montmorillonite - reaction of the 5′-phosphorimidazole of adenosine. *Science*, 257(5075):1387–1389, 1992.

[227] K. Reblova, N. Spackova, R. Stefl, K. Csaszar, J. Koca, N. B. Leontis, and J. Sponer. Non-Watson–Crick basepairing and hydration in RNA motifs: Molecular dynamics of 5S rRNA loop. *Biophysical Journal*, 84(06):3564–3582, 2003.

[228] R. Das, L.W. Kwok, I.S. Millett, Y. Bai, T.T. Mills, J. Jacob, G.S. Maskel, S. Seifert, S.G.J. Mochrie, P. Thiyagarajan, S. Doniach, L. Pollack, and D. Herschlag. The fastest global events in RNA folding: Electrostatic relaxation and tertiary collapse of the tetrahymena ribozyme. *Journal of Molecular Biology*, 332 (2):311–319, 2003.

[229] D. Thirumalai, N. Lee, S.A. Woodson, and D.K. Klimov. Early events in RNA folding. *Annual review of physical chemistry*, 52(1):751–762, 2001.

[230] S.L. Heilman-Miller, J. Pan, D. Thirumalai, and S.A. Woodson. Role of counterion condensation in folding of the tetrahymena ribozyme ii. counterion-dependence of folding kinetics 1. *Journal of molecular biology*, 309(1):57–68, 2001.

[231] E. Koculi, C. Hyeon, D. Thirumalai, and S.A. Woodson. Charge density of divalent metal cations determines RNA stability. *Journal of the American Chemical Society*, 129(9):2676–2682, 2007.

[232] E.J. Sorin, Y.M. Rhee, B.J. Nakatani, and V.S. Pande. Insights into nucleic acid conformational dynamics from massively parallel stochastic simulations. *Biophysical journal*, 85(2):790–803, 2003.

[233] M Zuker and P. Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research*, 9(1):133–148, 1981.

[234] Koculi E., N.K. Lee, D. Thirumalai, and S.A. Woodson. Folding of the tetrahymena ribozyme by polyamines: importance of counterion valence and size. *Journal of Molecular Biology*, 341(1):27–36, 2004.

[235] H.C. Greenwell, M.J. Harvey, P. Boulet, A.A. Bowden, P.V. Coveney, and A. Whiting. Interlayer structure and bonding in nonswelling primary amine intercalated clays. *Macromolecules*, 38(14):6189–6200, 2005.

[236] J.L. Suter and P.V. Coveney. Computer simulation study of the materials properties of intercalated and exfoliated poly (ethylene) glycol clay nanocomposites. *Soft Matter*, 5(11):2239–2251, 2009.

[237] Y. Li, D. Liu, H. Ai, Q. Chang, D. Liu, Y. Xia, S. Liu, N. Peng, Z. Xi, and X. Yang. Biological evaluation of layered double hydroxides as efficient drug vehicles. *Nanotechnology*, 21:105101, 2010.

[238] D.M. Dykxhoorn and J. Lieberman. Knocking down disease with sirnas. *Cell*, 126 (2):231–235, 2006.

[239] D. Reischl and A. Zimmer. Drug delivery of siRNA therapeutics: potentials and limits of nanosystems. *Nanomedicine: Nanotechnology, Biology and Medicine*, 5 (1):8–20, 2009.

[240] P.J. White. Barriers to successful delivery of short interfering RNA after systemic administration. *Clinical and Experimental Pharmacology and Physiology*, 35(11): 1371–1376, 2008.

[241] Y. Wong, K. Markham, Z.P. Xu, M. Chen, G.Q. Lu, P.F. Bartlett, and H.M. Cooper. Efficient delivery of siRNA to cortical neurons using layered double hydroxide nanoparticles. *Biomaterials*, 31(33):8770–8779, 2010.

[242] K. Ladewig, M. Niebert, Z.P. Xu, P.P. Gray, and G.Q.M. Lu. Efficient siRNA delivery to mammalian cells using layered double hydroxide nanoparticles. *Biomaterials*, 31(7):1821–1829, 2010.

[243] R.M. Hazen. Genesis: rocks, minerals, and the geochemical origin of life. *Elements*, 1(3):135–137, 2005.

[244] M. Sodupe J.-F. Lambert and P. Ugliengo. Themed issue: Prebiotic chemistry. *Chemistry Society Reviews*, 16(41):5365–5568, 2011.

[245] P.V. Coveney, J.B. Swadling, J.A.D. Wattis, and H.C. Greenwell. Theory, modelling & simulation in origins of life studies. *Chemical Society Reviews*, 1(41): 5430–5446, 2012. doi: doi:10.1039/C2CS35018A.

[246] S.V. Churakov. Ab initio study of sorption on pyrophyllite: structure and acidity of the edge sites. *The Journal of Physical Chemistry B*, 110(9):4135–4146, 2006.

[247] M.F. Aldersley, P.C. Joshi, J.D. Price, and J.P. Ferris. The role of montmorillonite in its catalysis of RNA synthesis. *Applied Clay Science*, 54(1):1–14, 2011.

[248] G.R. Williams, N.H. Rees, and D. O'Hare. Incorporation of phosphorus oxyacids into layered double hydroxides. *Solid State Sciences*, 11(7):1229–1238, 2009.

[249] R. Krishnamurthy, S. Pitsch, and G. Arrhenius. Mineral induced synthesis of ribose phosphates. *Origins of Life and Evolution of Biospheres*, 26(3):240–241, 1996.

[250] G. Ertem and J.P. Ferris. Synthesis of RNA oligomers on heterogeneous templates. *Nature*, 379(1):238–240, 1996.

[251] G. Ertem and J.P. Ferris. Sequence-and regio-selectivity in the montmorillonite-catalyzed synthesis of RNA. *Origins of Life and Evolution of Biospheres*, 30(5): 411–422, 2000.

[252] E. Biondi, S. Branciamore, M.C. Maurel, and E. Gallori. Montmorillonite pro-
tection of an UV-irradiated hairpin ribozyme: evolution of the RNA world in a
mineral environment. *BMC Evolutionary Biology*, 7(Suppl 2):S2, 2007.

[253] B. Chen, J.R.G. Evans, H.C. Greenwell, P. Boulet, P.V. Coveney, A.A. Bowden,
and A. Whiting. A critical appraisal of polymer–clay nanocomposites. *Chemical
Society Reviews*, 37(3):568–594, 2007.

[254] K.M. Tyner, S.R. Schiffman, and E.P. Giannelis. Nanobiohybrids as delivery ve-
hicles for camptothecin. *Journal of controlled release*, 95(3):501–514, 2004.

[255] H.D. Holland. *The Chemical Evolution of the Atmosphere and Oceans.* Princeton
University Press, 1984.

[256] R.A. Torres and T.C. Bruice. Molecular dynamics study displays near in-line
attack conformations in the hammerhead ribozyme self-cleavage reaction. *Pro-
ceedings of the National Academy of Sciences*, 95(19):11077, 1998.

[257] Z.P. Xu and G.Q. Lu. Layered double hydroxide nanomaterials as potential cellular
drug delivery agents. *Pure and Applied Chemistry*, 78(9):1771–1780, 2006.

[258] G.F. Joyce, T. Inoue, and L.E. Orgel. Non-enzymatic template-directed synthesis
on RNA random copolymers: Poly (C, U) templates. *Journal of Molecular Biology*,
176(2):279–306, 1984.

[259] S.P. Newman, T. Di Cristina, P.V. Coveney, and W. Jones. Molecular dynamics
simulation of cationic and anionic clays containing amino acids. *Langmuir*, 18(7):
2933–2939, 2002.

[260] G. Ertem. Montmorillonite, oligonucleotides, rna and origin of life. *Origins of Life
and Evolution of Biospheres*, 34(6):549–570, 2004.

[261] F.H.C. Crick. The origin of the genetic code. *Journal of Molecular Biology*, 38(3):
367–379, 1968.

[262] L.E. Orgel. Evolution of the genetic apparatus. *Journal of Molecular Biology*, 38
(3):381–393, 1968.

[263] S. Altman. *Nobel Lecture: Enzymatic Cleavage of RNA by RNA*. Nobel Prize
Committee, 1989.

[264] T.R. Cech. *Nobel Lecture: Self-Splicing and Enzymatic Activity of an Intervening
Sequence RNA from Tetrahymena*. Nobel Prize Committee, 1989.

[265] M. Martick, T.S. Lee, D.M. York, and W.G. Scott. Solvent structure and ham-
merhead ribozyme catalysis. *Chemistry & Biology*, 15(4):332–342, 2008.

[266] M. Martick and W.G. Scott. Tertiary contacts distant from the active site prime a ribozyme for catalysis. *Cell*, 126(2):309–320, 2006.

[267] A.C. Forster and R.H. Symons. Self-cleavage of virusoid RNA is performed by the proposed 55-nucleotide active site. *Cell*, 50(1):9–16, 1987.

[268] J. Perreault, Z. Weinberg, A. Roth, O. Popescu, P. Chartrand, G. Ferbeyre, and R.R. Breaker. Identification of hammerhead ribozymes in all domains of life reveals novel structural variations. *PLoS Computational Biology*, 7(5):e1002031, 2011.

[269] W.G. Scott, J.T. Finch, and A. Klug. The crystal structure of an all-RNA hammerhead ribozyme: A proposed mechanism for RNA catalytic cleavage. *Cell*, 81 (7):991–1002, 1995.

[270] L. Menten and MI Michaelis. Die kinetik der invertinwirkung. *Biochem Z*, 49: 333–369, 1913.

[271] G. Zuo, W. Li, J. Zhang, J. Wang, and W. Wang. Folding of a small RNA hairpin based on simulation with replica exchange molecular dynamics. *The Journal of Physical Chemistry B*, 114(17):5835–5839, 2010.

[272] A.E. Garcia and D. Paschek. Simulation of the pressure and temperature folding/unfolding equilibrium of a small RNA hairpin. *Journal of the American Chemical Society*, 130(3):815–817, 2008.

[273] T.-S. Lee, C.S. López, G.M. Giambasu, M. Martick, W.G. Scott, and M. Darrin. Role of Mg^{2+} in hammerhead ribozyme catalysis from molecular simulation. *Journal of the American Chemical Society*, 130(10):3053–3064, 2008.

[274] V.S. Pandé, I. Baker, J. Chapman, S.P. Elmer, S. Khaliq, S.M. Larson, Y.M. Rhee, M.R. Shirts, C.D. Snow, and E.J. Sorin. Atomistic protein folding simulations on the submillisecond time scale using worldwide distributed computing. *Biopolymers*, 68(1):91–109, 2002.

[275] H. Yang, F. Jossinet, N. Leontis, L. Chen, J. Westbrook, H. Berman, and E. Westhof. Tools for the automatic identification and classification of RNA base pairs. *Nucleic Acids Research*, 31(13):3450–3460, 2003.

[276] J.W. Pitera and W. Swope. Understanding folding and design: Replica-exchange simulations of "trp-cage" miniproteins. *Proceedings of the National Academy of Sciences*, 100(13):7587, 2003.

[277] G.R. Bowman, X. Huang, Y. Yao, J. Sun, G. Carlsson, L.J. Guibas, and V.S. Pandé. Structural insight into RNA hairpin folding intermediates. *Journal of the American Chemical Society*, 130(30):9676–9678, 2008.

[278] R.R. Johnson, A. Kohlmeyer, A.T.C. Johnson, and M.L. Klein. Free energy landscape of a DNA-carbon nanotube hybrid using replica exchange molecular dynamics. *Nano Letters*, 9(2):537–541, 2009.

[279] M. Feig, J. Karanicolas, and C.L. Brooks. MMTSB tool set: enhanced sampling and multiscale modeling methods for applications in structural biology. *Journal of Molecular Graphics and Modelling*, 22(5):377–395, 2004.

[280] H.W. Pley, K.M. Flaherty, and D.B. McKay. Three dimensional structure of a hammerhead ribozyme. *Nature*, 372(6501):68–74, 1994.

[281] K.F. Blount and O.C. Uhlenbeck. The structure-function dilemma of the hammerhead ribozyme. *Annual Review Biophysics Biomolecular Structure*, 34:415–440, 2005.

[282] S. Wang, K. Karbstein, A. Peracchi, L. Beigelman, and D. Herschlag. Identification of the hammerhead ribozyme metal ion binding site responsible for rescue of the deleterious effect of a cleavage site phosphorothioate. *Biochemistry*, 38(43):14363–14378, 1999.

[283] K. Suzumura, Y. Takagi, M. Orita, and K. Taira. NMR-based reappraisal of the coordination of a metal ion at the pro-r p oxygen of the A9/G10. 1 site in a hammerhead ribozyme. *Journal of the American Chemical Society*, 126(47): 15504–15511, 2004.

[284] J.B. Murray, C.M. Dunham, and W.G. Scott. A pH-dependent conformational change, rather than the chemical step, appears to be rate-limiting in the hammerhead ribozyme cleavage reaction. *Journal of Molecular Biology*, 315(2):121–130, 2002.

[285] J.E. Wedekind and D.B. McKay. Crystallographic structures of the hammerhead ribozyme: relationship to ribozyme folding and catalysis. *Annual Review of Biophysics and Biomolecular Structure*, 27(1):475–502, 1998.

[286] M. Roychowdhury-Saha and D.H. Burke. Extraordinary rates of transition metal ion-mediated ribozyme catalysis. *RNA*, 12(10):1846–1852, 2006.

[287] S. Rüdisser and I. Tinoco Jr. Solution structure of cobalt (iii) hexammine complexed to the GAAA tetraloop, and metal-ion binding to G·A mismatches. *Journal of Molecular Biology*, 295(5):1211–1223, 2000.

[288] N.G. Holm. The significance of Mg in prebiotic geochemistry. *Geobiology*, 10(4): 269–279, 2012.

[289] E. Biondi, S. Branciamore, L. Fusi, S. Gago, and E. Gallori. Catalytic activity of hammerhead ribozymes in a clay mineral environment: Implications for the RNA world. *Gene*, 389(1):10–18, 2007.

[290] M. Zgarbová, M. Otyepka, J. Šponer, A. Mládek, Pavel Baná, T.E. Cheatham III, and P. Jurecka. Refinement of the cornell et al. nucleic acids force field based on reference quantum chemical calculations of glycosidic torsion profiles. *Journal of chemical theory and computation*, 7(9):2886–2902, 2011.