COMPUTATIONAL APPROACHES TO TARGETING STRUCTURED GENOMIC REGIONS OF VIRAL GENOMES, THE CASE OF METAL-CONTAINING DRUGS

By

Lazaros Melidis

A thesis submitted to the University of Birmingham for the degree of DOCTOR OF PHILOSOPHY

School of Chemistry

College of Engineering and Physical Sciences

University of Birmingham

March 2022

Abstract

Infectious agents and especially viruses have had a tremendous impact on all stages of the evolution of the living matter. Being only composed of a few biomolecules they are often seen as a relic of processes from the early emergence of life on earth, but also they have contributed throughout evolution in all kingdoms of life. Throughout human history, infectious diseases have played significant role in the route of history, both regarding total death toll as well as policy changes. However, the highly dynamic nature of viral evolution in combination with overlapping mechanisms between the virus and the host poses a significant challenge for antiviral drug design based on traditional methods. This thesis examines the structural and dynamical properties of nucleic acid sequences from viruses as potentially broad-spectrum and robust antiviral target. The introductory chapter briefly introduces nucleic acid structures and previous work from the Hannon group on nucleic acid targeting and reviews viral processes throughout the different classifications. It also introduces theoretical and computational methods that could shine light on the generally elusive dynamical landscape of nucleic acid structures. After methodology is described in chapter a 2, fundamental theoretical understanding of dynamic molecular processes is discussed in chapter 3. Here, structure and dynamics of coordination compounds are discussed, in an effort to examine those away from crystal structure limitations. Employing Density Functional Theory (DFT), different spin states of complex molecules are discussed with the potential implications to metastable states within a chemical species. In chapter 4, work from the earlier chapters is applied in the understanding of structure and dynamics of HIV-1's TAR RNA, the most well-characterised RNA structure in literature. Having validated the Molecular Dynamics approach against known properties of TAR RNA, comparing metastable states identified through a Markov state model of the system with published NMR data, its potential interactions with supramolecular cylinders are examined resulting in a proposed dominant binding mode in agreement with previous experimental results. Chapter 5 introduces a pipeline for in silico optimisation of molecules targeting RNA given only the sequence of a novel virus (in this case SARS-CoV2). From sequence, secondary and tertiary structure are proposed and using molecular

dynamics the conformational landscape of some of its metastable states can be sampled, exposing potentially key regions that could be targeted and change the RNA's behaviour and thus potentially disturbing the viral replication cycle. The predictions of the pipeline have been experimentally verified in collaboration with the Grzechnik group and in cellulo effects are also being observed and reported. In Chapter 6, DNA structures are investigated this time, specifically G quadruplexes (G4s). After a very brief exploration of the interaction of previously characterized mono-nuclear complexes with human G4s (MYC and H-Telo) at the MD level of theory, the chapter focuses on the newly identified (2018) unique structure of HIV-1 LTR G4. This G4 structure features a stem on top of a guanine quartet potentially creating a high affinity target aiming to change the dynamics of the structure. The previously established methodology in mapping molecular interactions between ligands and nucleic acids is also applied here and a global minimum can be suggested although further simulations and coordination with experiments is needed for conclusive mapping of the interaction. Chapters 4 and 5 have been published and are presented here verbatim, as well as parts of Chapter 3 whereas Chapter 6 has not been published.

“Ludwig Boltzmann, who spent much of his life studying statistical mechanics, died in 1906, by his own hand. Paul Ehrenfest, carrying on the work, died similarly in 1933. Now it is our turn to study statistical mechanics.” — David L. Goodstein, States of Matter

Acknowledgements

I would like to thank the Physical Sciences for Health Centre for Doctoral Training and EPSRC for funding this hub where challenging ideas can be cultivated and tried. This work would have not been possible without the support of the Bluebear HPC systems administration team, especially Ed and Simon, whose help in navigating the ever changing computational environment can never be understated as well as the resources provided by Prof Jean-Baptis Cazier and the CaStLeS.

I would also like to thank my supervisors Prof Iain Styles and Dr Nik Hodges for keeping me balanced between different disciplines from mathematics and computational methods to cell biology and above all my primary supervisor Prof Mike Hannon who gave me an enormous amount of freedom to push the research field in new unexplored areas and his trust in the soundness of my results even when they came in opposition to existing dogma. The discussions with him were not only constructive and stimulating but also grew me as a scientist and opened new horizons in my communication of results across disciplines. I would be eternally grateful for the opportunity to work with him.

I also need to thank the mental health teams at the Barberry and Oleaster who did their best to keep me healthy both mentally and physically both during challenging times and in the day-to-day struggles.

This work would have not been possible if not for the support from Tasha and especially Jana who reminded me of human activities and life outside research.

Contents

Chapter 1

Introduction

This introduction is split into 2 sections, the first one, states the importance of the research conducted and lays the foundations of the biological target (nucleic acids) and biological processes within virology that can be targeted. The second section, lays the foundation of the theoretical framework of molecular dynamics, emphasizing the techniques used in this thesis.

Section 1 – Stating the biological challenge

1.1 Background and Importance

The aged population endures enhanced susceptibility to viral infections and subsequent superimposed bacterial infections. Such infections not only induce higher morbidity and mortality in older people but also appear to be increasing in number. Increasing antimicrobial resistance exacerbates this problem. New types of antiviral and antibiotic agents that act through new biological targets and mechanisms are needed to meet this challenge.

RNA and DNA are particularly attractive biological targets to fight both viruses and bacteria; their RNAs especially and their DNA have structural features such as bulges, junctions and folds that can be specifically recognised and used to block activity. The Hannon Group has developed a new class of nanosized supramolecular agents that recognise in a novel shape specific way some RNA and non-canonical DNA structures notably Y-shaped junctions and bulge structures [1],[2],[3]. For instance, the dinuclear triple helicate chemical agents, with cylindrical shape of diameter 1nm and length 2nm (cylinders) examined here recognise a specific bulge structure in the trans activation response region of HIV-1 RNA and inhibit HIV replication in mammalian cells without otherwise damaging the cells[4].

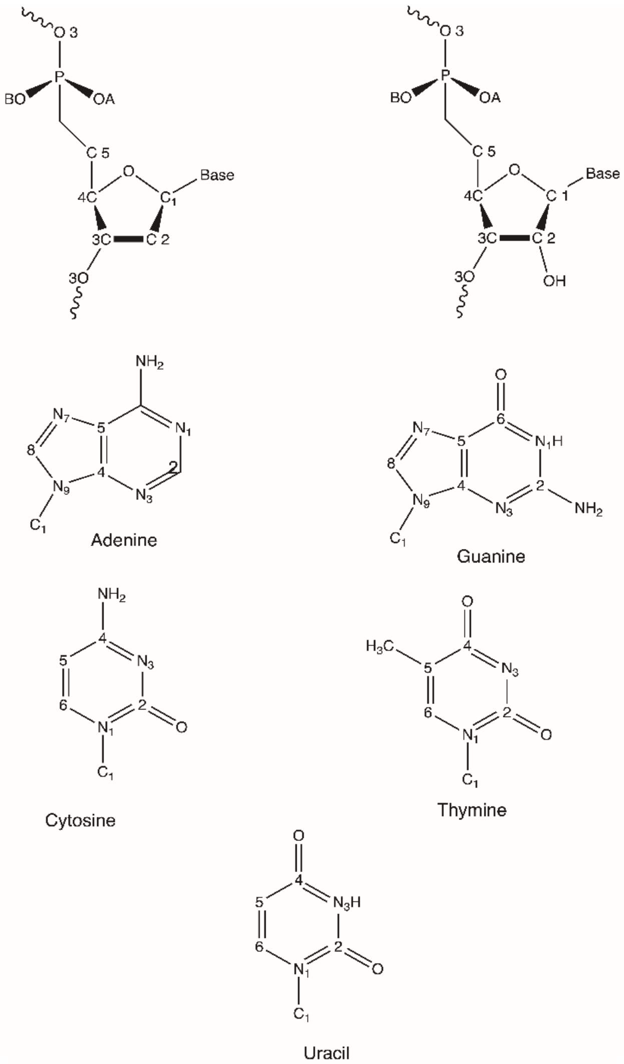
Over the past years, it has been increasingly crucial to develop new ways to create antiviral and antibiotic action but underpinning that is a need within the scientific community to enhance our ability to study drugs inside cells to understand better their mode of action. Pushing the frontiers of techniques for this application will be a feature of this project.

All pharmaceutical companies invest in computational studies either for drug design or for finding new targets[5]. A lot of the DNA-binding molecules contain metals, which raises a challenge for computational studies, since the classical molecular dynamics (MD) and docking methods[6],[7] do not consider the electronic structure of the molecule. Given the presence of transition metals in the cylinder, Density Function Theory or Molecular dynamics/DFT hybrid needs to be employed for geometry optimisation and more importantly noncovalent interactions in extended molecular systems. Once the optimisation and in silico characterisation of the series of cylinder variants is done, one can start simulating the interaction with RNA and DNA structures. As explained extensively in a RNA-ligand interaction review in 2017[8] an accurate charge distribution on the ligand (cylinder) is vital for the success of the simulation. The problem with quantum molecular calculations (like density functional theory) is the computational cost, especially in larger structures and interactions. Therefore, a layered simulation can be employed by using DFT on one scale and MD on larger scales[9]. Another approach is to consider the ligand (cylinder) to have a fixed surface electron density and treat it as one object. Following the current trend in the scientific community, there have been attempts to incorporate machine learning to improve the performance on MD[10] and less so on DFT[11]. On the other hand, machine learning has been valuable in docking simulations and database searches[12].

1.1.2 Introduction to Nucleic acid structures

Nucleic acids were discovered as a biological molecule in 1868 by Friedrich Miescher from the nucleus of white blood cells[13], but it wasn’t until the 1940s that it was first suggested as the genetic material. In 1952, it was proven to be so by Hershey and Chase through a series of historic experiments using radiolabelled DNA[14] and proteins on a virus-bacteria model. A year later, with preliminary results from Rosalind Franklin suggesting helical conformation, Watson and Francis Crick created the model of DNA that is most commonly used[15].

Nucleic acids are polynucleotides and the monomer unit (nucleotide) consists of a planar aromatic base; Adenine (A), Cytosine (C), Guanine (G), Thymine (T), Uracil (U) attached to a deoxyribose (DNA) or ribose(RNA) unit, a furanose-ring sugar moiety with a 5’-phosphate group. Each monomer is connected to the next via bonding of each 3’-carbon of its sugar to the 5’-carbon of the next. Hence the chain has one free 5’ position on one end and one free 3’ position on the other. Information is stored in the sequence of bases.



Adenine

Guanine

Cytosine

Thymine

Uracil

Figure 1.1 *Deoxyribose and ribose backbone of Nucleic acids (top) Nitrogenous bases in nucleic acids (bottom).*

1.1.3 Double helix

The first crystal structure of nucleic acids was of the iconic double stranded helix showing that two strands can form a double helix with two antiparallel strands stabilised by hydrogen bonds between complementary bases (A:T (2), G:C (3), A:U(2) for RNA). The double helix is further stabilised due to stacking between the aromatic rings of the bases. DNA is largely found in right-handed double helical complexes, with 2 anti-parallel strands, forming almost exclusively Watson-Crick (WC) base pairs. The double helix, in physiological conditions can be seen in B-DNA conformation, as described by Watson and Crick, and Rosalind Franklin's XRD data. Franklin also described A-DNA, which is more compact and can be stable under dehydration conditions. A-DNA has been shown to also form in vivo[16] and is especially interesting in the context of DNA-protein and DNA-ligand interaction.

The 2'OH-hydroxyl group of the ribose gives rise to the profound differences between DNA and RNA in terms of structure and dynamics[17], as it can contribute to further hydrogen bonding interactions. The hydroxyl also plays a part in non-canonical (non-Watson-Crick) pairing[18]. This opens up the conformational landscape of a single RNA strand to multiple possible local minima within a given temperature[19],[20],[21] and thus RNA secondary and tertiary structure should be considered in terms of an ensemble of metastable states rather than a single state. Additionally, it allows for the creation of rigid pockets which can recruit metals and enable catalytic activity[22] which has been theorised to have played crucial role in the RNA world hypothesis[23]. Moreover, the ability to hold different metastable states has been utilised by viruses, by using the same genomic region for multiple functions in different stages of the replication.

Another difference between RNA and DNA, is the substitution of uracil (U) with thymine (T), which is U methylated at the C5' position, in the base position. That methyl group improves the helical stability for DNA[24],[25] and potentially contributes to DNA repair in the case of spontaneous conversion of C to U[26] (deamination of cytosine to uracil). However, U can still be found in bacteriophages (ssDNA) and DNA viruses (poxviruses, herpes-viruses)[27] which code their own UDG enzyme indicating importance of U in their process. Deamination and chemical modification of C to U has also been extensively research in efforts to sequence modified C bases in the genome and modified bases in general[28].

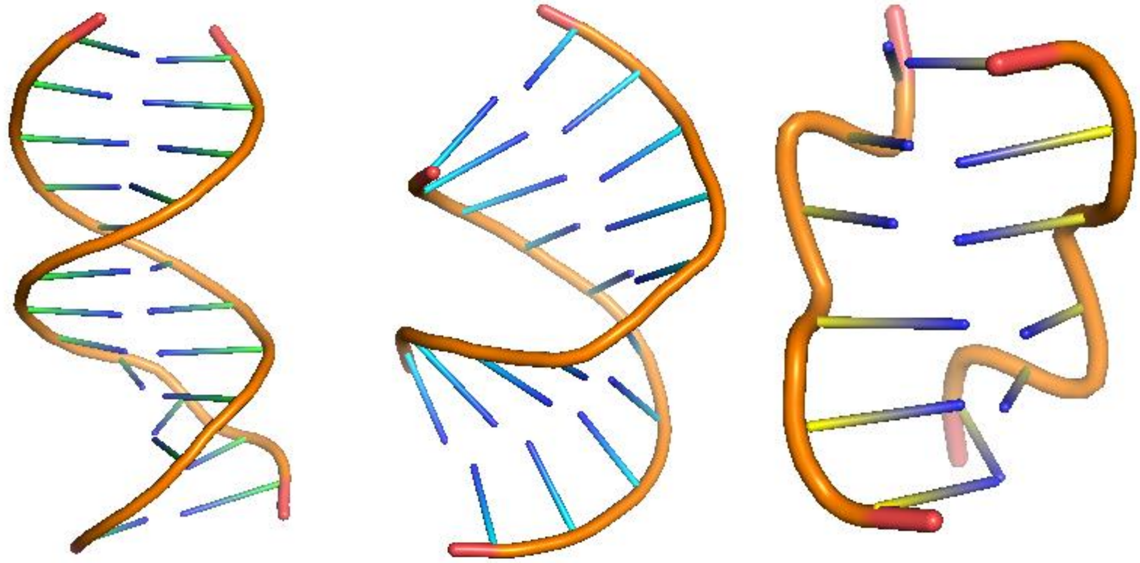


Figure 1.2 *B, A and Z DNA, PDB 1BNA, 4IZQ, 4HIF.*

1.1.4 Stems – Loops – Bulges

A single strand can fold on itself forming base pairing between nucleosides of the same strand. This has been seen to naturally occur for RNA, with a lot of folded structures playing crucial roles in biological processes. Untranslated regions of viral genomes are great examples to study the structure-function relationship of RNA stem loops. In many of the stem loops studied, there are regions where complementarity breaks but transient base interactions along nucleotides either side of the opposite position create interesting and unique structural and dynamic signatures for the macromolecule. Additionally, when base paring is missing all together, -additional nucleotides added in one or both strands, bulges can form which mechanically allow for large-scale conformational changes of the molecule and can be attractive targets for drugs.

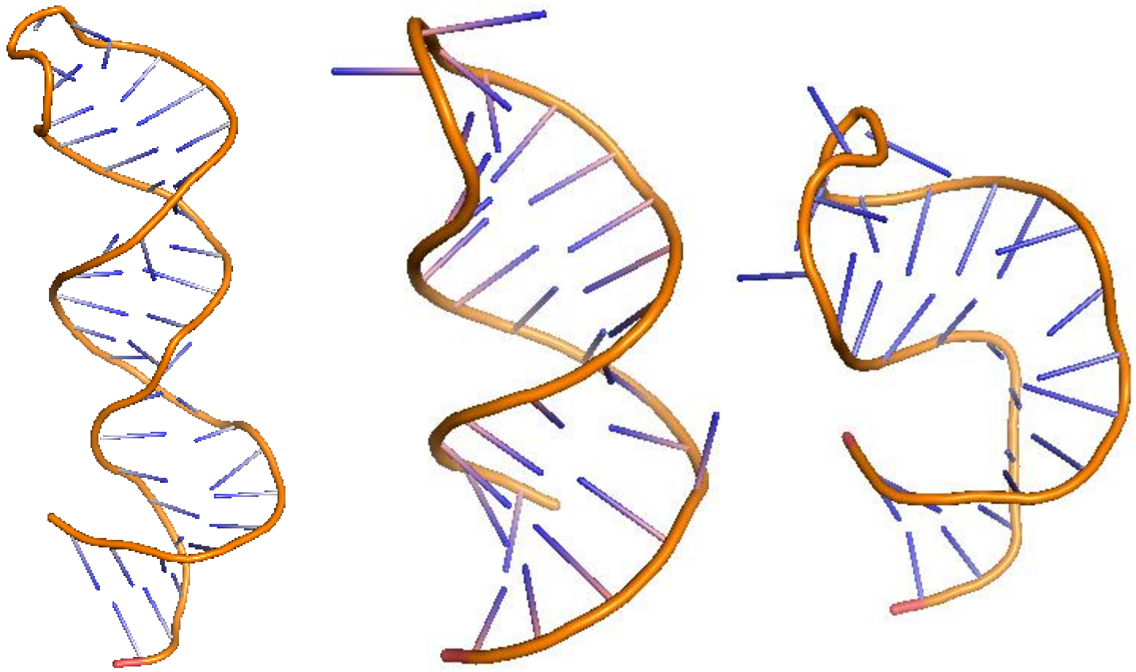


Figure 1.3 *examples of stem loops PDB; 2FEY, 2K5Z and bulged stem loop 1ANR.*

1.1.5 G-quadruplex

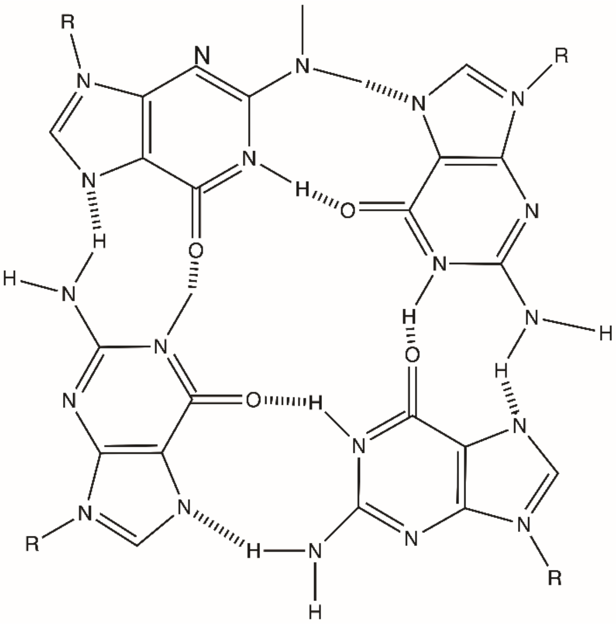


Figure 1.4 *G4 quartet.*

Guanine derivatives have been shown to self-aggregate, stabilised by Hoogsteen hydrogen bonds forming G quartets[29] (Figure 1.4). These Gs can be in the same strand or come together from different strands, adding to the topological landscape of the nucleic acids. Although single strand folding has been shown to have biological functions[30],[31],[32], multi strand formations of G quartets have been theorised in the past but only recently reports suggest they form in vivo[33].

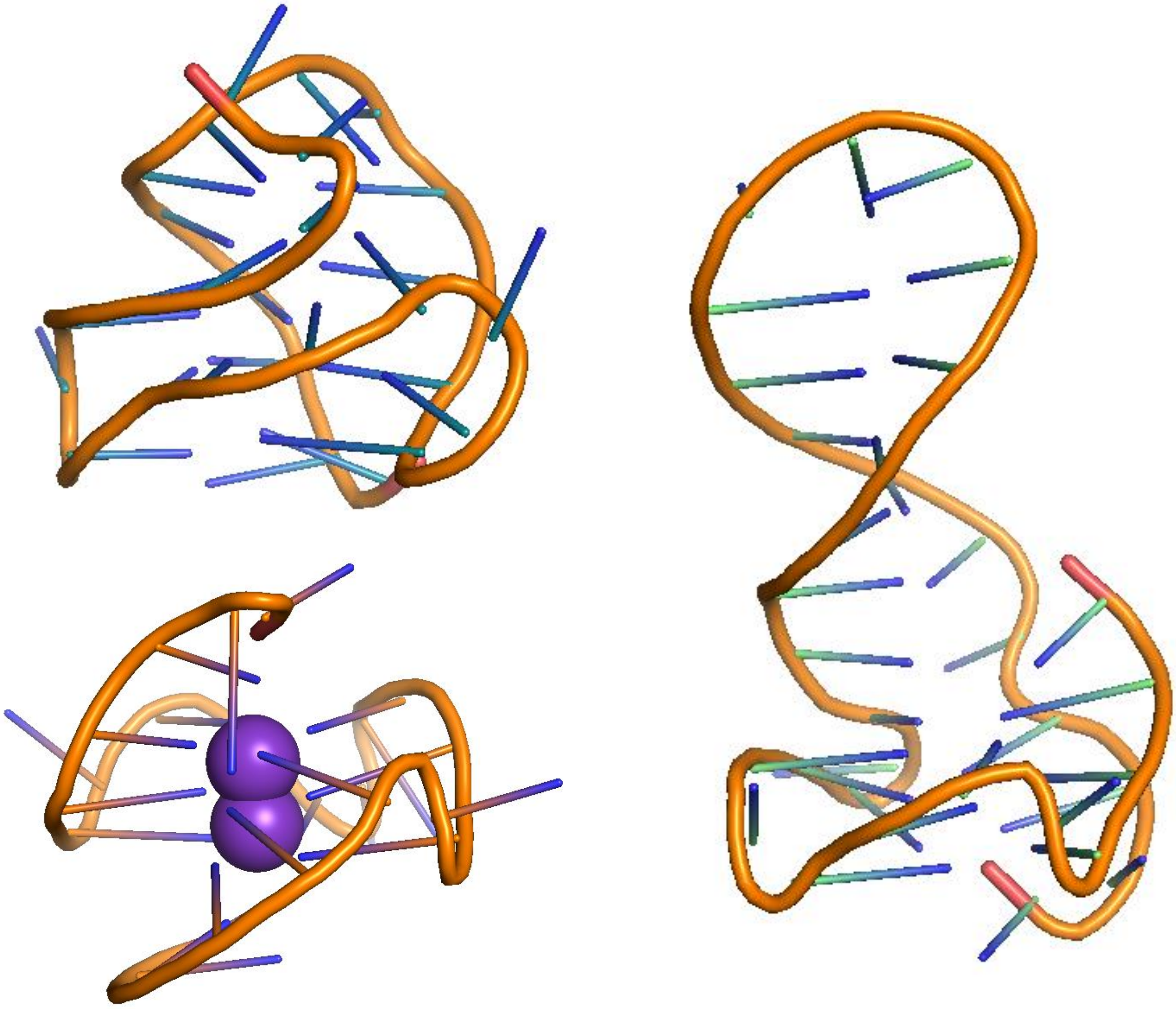


Figure 1.5 *Parallel G4 in the promoter of MYC(1XAV), antiparallel G4 (2MBJ), stem-loop-G4 (7CLS).*

1.1.6 Pseudoknots

Finally, nucleic acids and especially RNA can form more complex topologies on their own, as in ribozymes, where there is a rigid pocket acting as the catalytic site, or with help of proteins can be as complicated as ribosomes. The easiest way to visualise the folding of a pseudoknot is to consider the free bases of a large loop folding back and base pairing with free bases down or upstream of the stem.

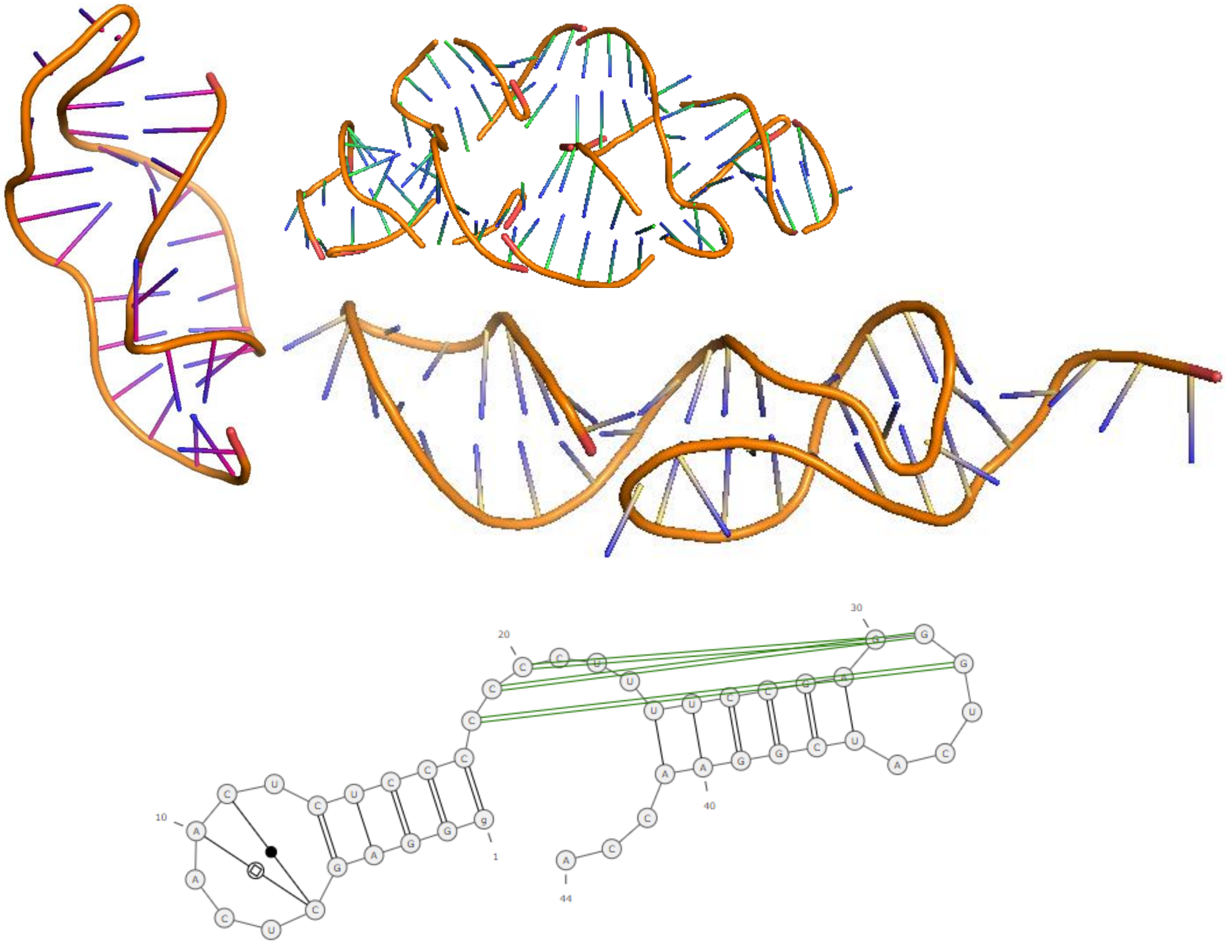


Figure 1.6 *Pseudoknot 1RNK, ribozyme 1MME, pseudoknot 1A60.*

1.1.7 Introduction to Ligands

Helicates were introduced by Jean-Marie Lehn[34] as a novel, chiral class of supramolecular structures, formed spontaneously after coordination of 2 or 3 transition metals to pyridine species of different organic strands; the pyridines are separated by bridges/spacer and the chiral axis is then defined by the line connecting the metal centres. The HannonGroupp has introduced a class of helicates (cylinders) with three ligands with pyridylimine binding motifs and diphenylmethylene as a spacer. This results a 3D shape similar to a cylinder, with a length on 2nm and diameter of 1nm. This is the parent molecule for this study and the template to produce further modifications to increase and tailor biological activity. Previously there have been crystallographic studies that show the interaction of the cylinder with DNA[35] and RNA[36]. This has been crucial work as it 1. demonstrates the ability of a molecule to drive the conformational equilibrium of nucleic acids to a less abundant or a new state and 2. Demonstrates the binding interaction with cavities formed by nucleosides, along with binding to terminal base pair. Additionally, to these crystal structures, there have been biophysical studies showing interaction with stem loops (TAR- HIV1) causing biophysically observable conformational changes further away from the binding site[37].

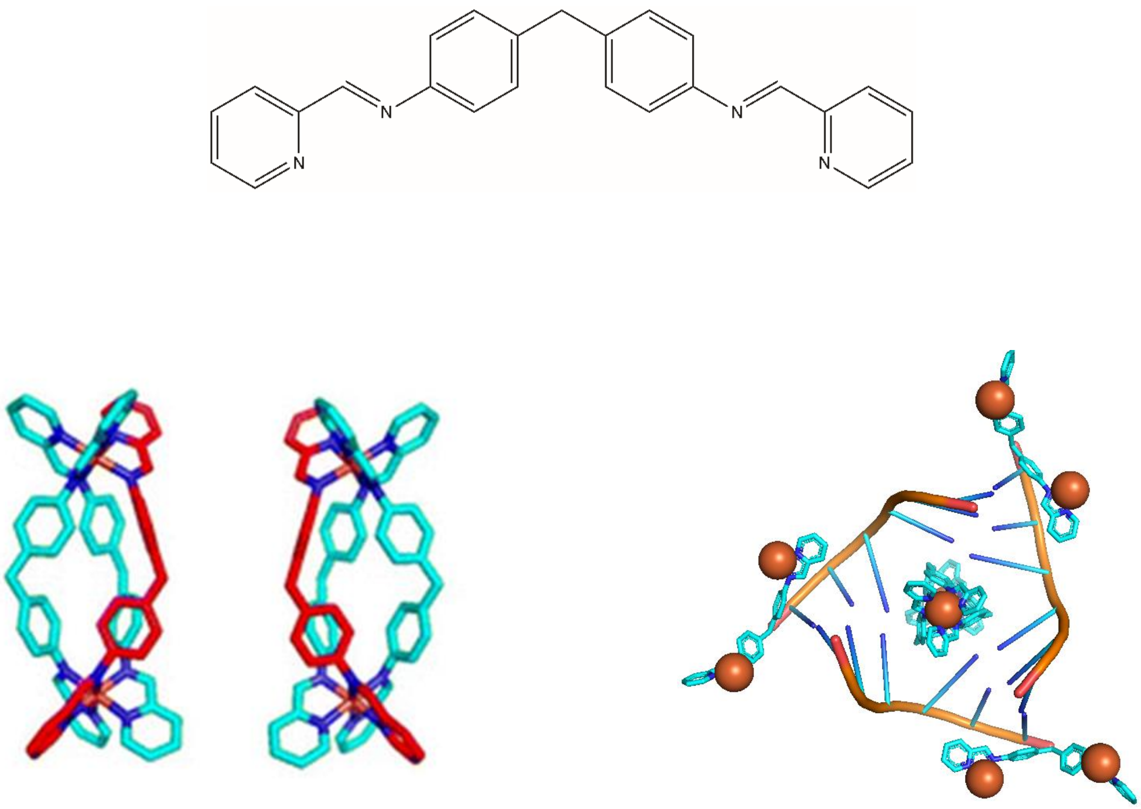


Figure 1.7 *Parent Hannon ligand (top).Two enantiomers of parent Hannon cylinder. Co-crystal structure of RNA 3-way-junction (PDB 4JIY) demonstrating binding modes.*

1.1.8 Introduction to Virology

This thesis aims to develop strategies for antiviral drug design and therefore one should start by defining what a virus is and explore the biological space and potentially common bottlenecks in the replication cycle of most viruses. For this thesis, a virus is defined as "An infectious, obligate intracellular parasite comprising genetic material (DNA and/or RNA) surrounded by a protein coat and/or a membrane." Under this definition one can describe viruses as two-state objects, i) their existence outside a cell is that of a quite stable particle (virion) and ii) their parasitic existence when they enter a cell which leads to production of more particles. For the virion to infect a cell the cell must be susceptible, *i.e.* provide a functional receptor or other entry mechanism for a given virus so that the virus can enter the cell membrane AND permissive, *i.e.* the cell has the capacity to replicate virus. Viruses can be classified in many ways regarding their genetic material, DNA or RNA; single or double stranded; linear or circular; positive or negative sense or both, their shell structure as virions (lipid enveloped or not enveloped) but the universal behaviour that defines their replication is that: Viral genomes must make mRNA that can be read by the host's ribosomes. Based on that fact, David Baltimore created the Baltimore scheme of classification of viruses[38] as follows.

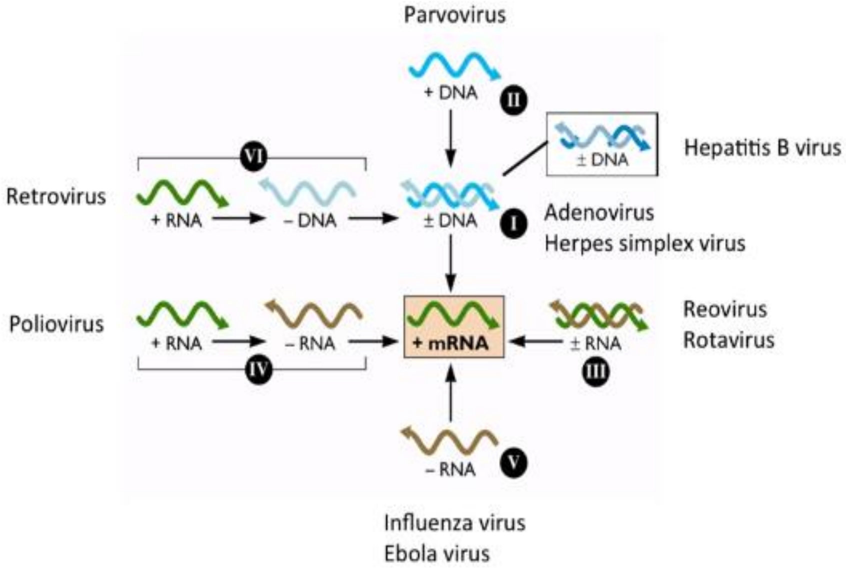


Figure 1.8 *Baltimore scheme of virus classification. (permission from Prof V Raccaniello).*

In this nomenclature, positive RNAs or DNAs are in the orientation of translation, but notably not all +RNAs are translatable by the ribosome (in the case of retroviruses, like HIV-1 the +RNA is reverse transcribed into -DNA which then forms dsDNA which is incorporated to the hosts' genome). In this classification the terms single, double stranded or gapped are related to the flow of genetic information, structurally, viral DNA and RNA genomes exhibit tremendous diversity and dynamic nature. This diversity of genome structure is largely related to unstable/metastable nature of RNA. It has been shown that RNA genomes appeared first in evolution (RNA world[39]) but the only RNA genomes today are viral. In the context of this thesis, a key characteristic that viral genomes have is that they are not chromatinised, which frees them to adopt different shapes that are used to regulate stages of infection. No matter how the virus stores the genetic information, in order to multiply all viruses need to encode for the following; 1. Replication of the viral genome 2. assembly and packaging of the genome 3. regulation and timing of the replication cycle, 4. modulation of host defences 5. spread to other cells and hosts.

1.1.9 Attachment and cell entry

The first stage of the infection cycle for all viruses is the attachment to a susceptible cell. Unlike fungi viruses (which have no extracellular phases) and plant viruses (which enter the cells after mechanical damage) animal viruses recognise receptors on the cell membrane of the susceptible cell. Those receptors are membrane proteins often with additional polyhydrate chains with vital function for the cell and often specific for the cell type, which leads to the specificity for the viral infection. The same receptor can be used by different viruses[40]. For 20-hedral capsids the attachment occurs between the receptor and one of the capsid proteins and for enveloped viruses the attachment takes place between glycoproteins of the envelope and is the first step towards the fusion of the two membranes (hairpin) creating an entry point for the inner capsule. Also, larger particles can enter via endocytosis in an endosome which ultimately fuses with lysosomes. There, as the endosome moves into the cell, proton pumps on the surface of the endosome drop the pH (as low as 5.5). The low pH causes proteolytic cleavage which activates the fusion protein for cleavage of the endosome membrane (class I) or it activates the cleavage of a second protein which then activates the fusion protein (class II), in either case pH induced conformational changes are important. Almost always DNA viruses are too big to pass through the membranes so, in the case of DNA viruses, that need to release the genome in the nucleus a second attachment of the late-stage particle to a membrane is necessary, this time on the nucleus pore. There the pH difference releases the DNA to the nucleus.

1.1.10 RNA viruses.

RNA-dependent RNA polymerase (RdRp) may initiate synthesis de novo, can start from the 3'end of the template continues to the end or can require a primer (cap-RNA -for mRNA or a terminal protein and the first triphosphate) but also complementary initiation protein complex will be needed from the cell machinery[41]. It also uses the 2-metal mechanism of polymerase catalysis (usually Mg 2+).

In class V, (-)RNA, genomes are coated with protein and they carry their own polymerase RdRp since the cells do not have a mechanism of RNA replication. The structure of the viral particle is usually a helical capsid, although Vesicular Stomatitis Virus (VSV) is rod-like and influenza has envelope capsid. In further detail, the influenza virus attaches to the cell and follows the class I pH regulated disassembly by the end of which the 8 (-)RNA fragments and the enclosed proteins enter the nucleus. One of the proteins cleaves the capped end of cellular mRNA and those are used as primers for mRNA synthesis[42]. The mRNAs created in this way are shorter compared to the native but once the viral protein expression reaches a sufficient amount the nuclear protein expressed prevents early termination of the RNA synthesis and the whole genome is replicated (first to +RNA and then to -RNA). The various sizes and pH dependent structure of influenza viruses' nucleic acid process can make it a good candidate for inhibition of replication using RNA binding compounds.

In the case of VSV the polymerase creates segmented mRNAs which code for proteins that then allow the polymerase to continue creating bigger pieces of mRNA[43]. The expression of the first part of the genome then regulates the expression of the genes further down. Only when the last part of the genome is expressed the -RNA can be copied as a whole and the viral particles can be assembled. Well studied regulatory processes like this can provide an opportunity to understand broad spectrum antiviral activity of a compound as well as introduce chemically controlled multimodal therapeutics.

(+)RNA genomes are naked (no protein coating) since they can be translated immediately (not for retrovirus and coronavirus). In the Flavivirus genus (Zika, Dengue etc), viruses have a capped 5' RNA and poly-adenylated 3' which makes the genome fully formed mRNAs. The whole mRNA is translated in one piece and the multiple proteins are created by cleaving of the amino acid chain by two or three proteins depending on the virus. The one-step translation can be an interesting case for RNA binding drugs. For most of these viruses the cell destroys the ER and cell membranes creating new vesicles that are the replication sites of the new viral RNA. In this class, it is worth mentioning another genus; Alphaviruses[44], since they use a -RNA intermediate. Specifically, in the case of Togaviridae, only the first part of the mRNA is translated since there is a stop codon right after the first protein, the polymerase. That starts replication of the RNA and the same protein starts again the (+)RNA of the rest of the genome after passing through (-)RNA first.

A notable, well-studied example of (+)ssRNA virus is Hepatitis C (HepC). The virion is an enveloped spherical particle about 50nm in diameter, with two virus-encoded membrane proteins (E1 and E2) surrounding the capsid protein which encloses the genomic ss(+)RNA (~9.5Kb)[45],[46]. Attachment is mediated between the envelope proteins and host receptors which initiates clathrin-mediated endocytosis, and after fusion of the virus membrane with the host endosomal membrane, the genome is released into the cytoplasm. The whole of the genome is translated into a polyprotein which is then cleaved into structural and non-structural proteins (yielding replication proteins). Replication takes place on the surface of the ER in cytoplasmic viral factories. Host miRNA (mi-122) which is specific to hepatic cells binds to the 5' non-coding region[47],[48],[49] of the genome preventing exonuclease Xrn1 from degrading the genome, hence allowing replication. miRNA is often the target of drugs as will be discussed later[50]. Then the dsRNA genome is produced and transcribed/replicated to produce mRNA and (+)ssRNA respectively. When a critical mass of structural proteins is reached the assembly of the virus takes place at the ER with the help of viral ionic channel p7. The particles move to the Golgi apparatus before they are released from the cell by exocytosis.

dsRNA (for example Reoviridae) genomes are naked (not coated in protein) but carry RdRp so that mRNA can be produced from the double-stranded RNA. Entry is achieved with endocytosis, the membrane is degraded in endosome-lysosome which opens up the core particle. In every 5-fold axis of symmetry there is an RNA polymerase which would only allow mRNAs out of the particle as it produces them, keeping the original double-stranded genome in the core particle.

1.1.11 DNA and retroviruses

Looking back at the Baltimore scheme, categories I, II, VI and VII go through or start from double strand DNA and then use the cell's transcription machinery, except for poxvirus and giant viruses that replicate in the cytoplasm and carry or encode their own RNA polymerase. If the virus needs to enter the nucleus, it needs to be able to be transcribed and replicated in the environment of the host, which includes sequence-specific DNA binding proteins and co-activators (which can either be native to the host or included in the virus) that are used to promote or silence transcription. Retroviruses with simple genomes can achieve that with just the host's machinery but more complicated retroviruses (HIV-1), papillomaviruses, and parvoviruses need one additional protein that is either packed or expressed first. Even bigger DNA viruses (adenoviruses and herpesviruses) have more than one viral protein to stimulate transcription whereas poxviruses carry everything needed to replicate and therefore can stay in the cytoplasm (still need at least one cellular protein to start the cycle).

In the nucleus of somatic cells, DNA is mostly packed in chromatin and every new piece of DNA that enters will be chromatinised. The degree of packing of the DNA to the chromatin is a regulatory mechanism (the tighter it is the more difficult for the polymerase complex to access it). When new dsDNA fragment is introduced in the nucleus it is chromatinised quickly unless the virus already has nucleosomes (SV40-polyomavirus)[51].

Viral regulatory mechanisms include i) positive or negative autoregulatory loop during which the cellular machinery recognises and expresses one gene which encodes a protein that enhances the expression of that gene (or silences it) and cascade regulation when the protein promotes the expression of a previously silent gene[52]. For our work, it is important to remember that all viruses in this category need to express their first protein to move rapidly to the next stage, which creates a bottleneck in the replication process and an opportunity for drug interventions.

In this class of viruses, it is worth mentioning splicing[53],[54],[55], the process of maturation of pre-mRNA by removing introns, cleaving the pre-mRNA. The spliceosome is composed of small RNA fragments and proteins that together produce the mRNA. Most importantly, the proteins are not essential, and splicing can only happen with the small RNA fragments. Although after splicing a lot of the proteins remain attached to the mRNA and allow others to bind too, those proteins are necessary for the export of the mRNA to the cytoplasm. Retroviruses need to export their whole, unspliced, RNA genome to be put to the new viral particles that are assembling in the cytoplasm and for that reason they express at least one protein earlier in the cycle that would bind to the genome molecule and enable the export pathway – in the case of HIV the protein is Rev. The RNA-protein interaction during splicing can potentially produce off-target effects, when targeting viral RNA-protein or viral RNA-RNA interactions, therefore understanding the cellular localisation of the potential drug can be crucial for its rapid development.

1.1.12 Impact on the host cell

All viruses, pathogenic or not, interact with the signalling pathways of the host during attachment (disrupting the actin filament layer and allowing endocytosis) all the way to release. Notably, it is common for the Pi3k-mTor[56] relay to be activated (adenovirus, hepatitis C, HBV) which is regulating apoptotic pathways. Important for the purposes of understanding the role of untranslated regions and their targeting is how flaviviruses block Akt activation to induce apoptosis (which is important for the release of the new particles). The 5’ end (+)RNA genome is de-capped by host proteins and exonuclease (Xrn1) starts to degrade the genome until it reaches the 3’UTR. The UTR is a highly structured short RNA called sfRNA (subgenomic flavi) which blocks the Akt activation and induces apoptosis. Viruses with mutated sfRNA fail to induce apoptosis and form plaques making the sfRNA a good drug target for RNA binding drugs. Viruses change the overall signalling pathways of the host cell not only utilising the coded proteins but also structural features of their genome. At the same time viruses that rely heavily on the host's replication machinery have evolved to inhibit expression of the host's genes in favour of translation and then replication of their own genome. That happens either in disturbing splicing or disturbing 5'-end-dependent initiation complex for translation, either by removing the cap (if their own mRNA is not capped) or by using proteases to cleave relevant proteins. Also important, the structure of the 5'-noncoding region (defined as RNA sequence before first translated AUG) can be used to bind initiation complex (with eIF proteins). Disturbing that structure can inhibit the binding of the translation initiation complex[57].

Most common cellular defence mechanism against viral infection is the expression of PKR[58],[59], which can be present but inactive in healthy cells but is activated and over-expressed in viral infection (by interferon). PKR is an interferon-induced enzyme that is activated by dsRNA, leading to phosphorylation of eIF2alpha and inhibition of translation and apoptosis. All viruses have evolved mechanisms to inactivate the PKR pathway often with short subgenomic RNAs either in the form of the original genome or as by-products of the translation mechanism. Deactivating the binding ability of those fragments can potentially be a drug target with broad antiviral activity.

Lastly, the impact on cellular metabolism is enormous. Needing 4 ATP to make each single peptide bond for protein synthesis, and a higher number of nucleic acids and amino acids the cell starts to consume more glucose. Some viruses promote glucose uptake (e.g. HCMV) or inhibit it (e.g. HSV). Although these are both herpesviruses they have different behaviour with respect to their impact on glycolysis, shifting the equilibrium of the pathway towards the most efficient production of the needed metabolites.

1.1.13 Human Immunodeficiency Virus

Given that there has been documented antiviral activity of the cylinder with HIV-1 virus[4] it is worth examining this virus in more detail. HIV was first isolated in 1983 from the lymph node of a patient with lymphadenopathy in Paris by Montagnier and Barre-Sinoussi (Nobel 2008) and a year later the first blood test was developed. It is classified in the orthoretrovirinae subfamily of retroviridae as Lentivirus (which includes HIV-1 and HIV-2). It is a typical retrovirus with envelope with glycoproteins and a capsid with 2 copies of a (+)ssRNA genome. After reverse transcription, the dsDNA (called provirus) merges with the host DNA and codes for the following groups of proteins: gag, pol, tat, rev, vpu, vif and envelope. The large number of different proteins that can be expressed from a 10Kb genome is a result of splicing and the smaller proteins are regulatory agents of the replication process. Sequence analysis shows that HIV-1 is originated from SIVcpz (simian immunodeficiency virus). Further combinatorial sequence analysis traced the virus back to SIV, slightly different for each monkey species (SIVmon for Mona monkey, SIVrcm for red-capped mangabey, SIVsmm for Sooty mangabey and more). SIVcpz (chimpanzee) is a recombination of SIVmon and SIVrcm and it is pathogenic to chimpanzees[60]. Transmission from chimps to humans created HIV-1 M and N, and it also passed to Western gorillas, creating SIVgor which when then passed to humans created HIV-1 P and O. M and O crossovers can be traced to early 20th century whereas N and P are more recent. HIV-2 on the other hand, is a crossover from SIVsmm to humans. It has 30-40% identity with HIV-1 and is less virulent - infections do not progress to AIDS and it is less transmissible. HIV-1 M is the dominant type (99% worldwide) and it has 9 subtypes (A-K). Biological differences between subtypes are minimal (though D kills host humans faster, and C has higher shedding in the female genital tract) but high-risk individuals are often infected with multiple subtypes which can make treatment more difficult and cause new recombinations.

HIV-1 attaches to 2 receptors on the target cell (CD4+ T-lymphocytes), CD4 and α or β-chemokine receptor, which initiates fusion. Leaving the envelope behind, the capsid containing 2 copies (+)ssRNA, tRNA primers, viral protease, retro-transcriptase and integrase is now in the cytosol. As the capsid disintegrates reverse transcription is initiated resulting in a mostly double stranded DNA with terminal direct repeats and blunt ends (dscDNA). Before it reaches the nucleus the pre-integration complex (PCI) is formed, composed of double-stranded complementary DNA (dscDNA), integrase, matrix protein, retrotranscriptase, viral protein R (Vpr) and various host proteins, such as high-mobility group protein B1 (HMG1) or lens epithelium derived growth factor (LEDGF). As discussed previously, viral trafficking is mediated by microtubules. The formation of the PCI enables the entrance to the otherwise intact nuclear envelope of the resting cell (non-replicating). The pro-viral DNA can be integrated on the cell chromosomal DNA (facilitated by emerin -improving localisation to chromatin and LEDGF -transcriptional activator that binds integrate and helps promote viral integration) or be circularized as one or two long terminal repeats (LTR). After the integration, the LTR-flagged provirus acts like any other eukaryotic gene with polyadenylation and termination side from LTR regions. Interestingly, the LTR contains sequence able to form G-quadruplexes[61], which is another potential structured nucleic-acid target and will be closely examined later. Activation of the T-cell facilitates the binding of the transcriptional pre-initiation complex to enhancer elements in the 5' LTR proximal promoter which gathers both host and viral elements. Important host elements are; nuclear factor-κB, nuclear factor of activated T-cell (NFAT) ad SP1, these enhancer proteins belong to the general transcription machinery and promote the binding of RNA-polymerase II (RNAPII) to the TATA box (repeating T A cis regulatory element) to initiate mRNA production. On the viral side, a 59-nucleotide stem-loop structure termed the trans-activation response element (TAR) is formed on the 5' end of the viral transcript, providing a binding site for the viral trans-activator Tat. Tar-Tat complex enables the RNAPII to elongate the transcription process enabling transcription to mRNA of the whole viral genome. The RNA fragment is then further processed by Rev (HIV protein) factor which also regulates the nucleo-cytosolic transport and splicing of viral mRNA species. In the cytoplasm, the mRNA is translated as a whole and the resulting poly-protein is post-processed to separate the proteins. After critical mass is attained, mature viral proteins assemble into the capsid structure and after the 2 copies of ss(+)RNA with the proteins mentioned earlier are enclosed, the capsid exits the cell through budding, taking with it the envelope from the membrane of the host cell.

The infection of CD4 T-cells cascades through the immune cell equilibrium causing problems to all cells taking part in the immune response[62] (adaptive and innate immunity). The lack of immune response leads to a series of secondary effects that can manifest as neurological symptoms and cancers (40% of infected individuals) either by the viral proteins or by metabolic products of viral replication or by the generalised inflammation. Other dormant viruses that are already present in the individual eg. EBV, HHV8, HPV can be activated by the inappropriate cell proliferation caused by higher levels of cytokines triggering oncogenesis. The difficulty to produce an efficient vaccine[63] for HIV is due to rapid antigenic drift which causes changes on the envelope proteins making the traditional vaccines unlikely to be successful, nevertheless chemically targeting multiple different processes in the replication pathway has been shown to be extremely beneficial therapeutically[64].

1.1.14 Coronaviruses

Coronaviruses are a member of the Coronavirinae subfamily of the Coronaviridae family in the order of Nidovirales according to the International Committee on Taxonomy of Viruses[65]. The subfamily has four genera, defined on the basis of phylogenetic relationship; Alphacoronavirus, Betacoronavirus, Gammacoronavirus and Deltacoronavirus. The first two can only be found in mammalian species whereas gamma- and delta- can have a wider range which can include avian species. In all cases, pathogenesis results in respiratory and enteric diseases. The family was first characterised in the 1960s[66], as a cause of a substantial proportion of upper respiratory tract infections in children, and it owes its name to the characteristic corona around the spherical virus particles as seen by electron microscopy. During the 20th century the study of coronaviruses was limited to the HCoV-229E and HCov-OC43 strains in humans, with differences defined by serological studies and symptomatology. These two, with the more recently identified HCoV-NL63 and HCoV-HKU1 are the endemic human coronaviruses that cause seasonal and usually mild respiratory infections[67],[68]. During the 21st century interest in coronavirus virology has increased dramatically with the arrival of Middle East Respiratory Syndrome Coronavirus (MERS-CoV) and Severe Acute Respiratory Syndrome (SARS-CoV) in humans. Both newly identified viruses can cause life-threatening respiratory pathologies and lung damage. In both cases bats are considered to be the main host reservoir with camels and possibly civets an intermediate host before the slipover to humans[69],[70]. Although both viruses can cause disease with fatality ratios, over 50% in certain age groups, they lack high transmissibility among humans which kept the number of infections globally low. After the SARS-CoV epidemic in 2003 several groups attempted to study further the molecular virology and genetics of coronaviruses as well as the immune response by humans and other animals [69],[71],[72],[73],[74],[75],[76],[77],[78],[79],[80].

Coronaviruses are enveloped, positive sense, single strand RNA viruses with characteristically long genomes (~30kb)[81]. Their long (+)ssRNA contains multiple open reading frames(ORFs), with a protease being encoded in the ORF1. The specific architecture of those reading frames differs between different genera but subgenomic RNAs during the replication process allow for recombination events[82] which in turn can lead to phylogenetic jumps and zoonosis[83]. Viral particles are spherical in diameter varying from 80 to 160nm and within the membrane envelope there are 3 proteins; spike (S), membrane(M) and envelope(E). Inside the particle, the genomic RNA is capped and wrapped with phosphorylated nucleocapsid(N) proteins[70].

1.1.15 SARS-CoV2

In December 2019 a cluster of pneumonia cases of unknown origin was reported to China National Health Commission. The pathogenic agent was isolated on the 7th of January 2020 and WHO received the whole genome sequence of the novel virus on the 12th of January 2020, identified as a coronavirus and named 2019-nCoV[84],[85],[86],[87]. WHO declared public health emergency of International Concern on the 30th of January and declared a pandemic on the 11th of March 2020. After the sequence became publicly available[88] a global effort was launched to understand the fundamental molecular processes[89] of viral replication and identify potential antiviral targets in parallel to the vaccine design effort. Here, the results of this global effort are briefly reviewed, focusing on the RNA structure-function relationship as a domain of potential novel antiviral design.

1.1.16 Viral entry

SARS-CoV2 reference sequence has 79% homology to SARS-CoV, with only 75% for S protein[90], but it is important to state that these percentages fluctuate as evolution occurs. Despite the relatively low nucleic acid sequence homology the two viruses (higher homology with some coronaviruses found in bats[91]) have very similar amino acid composition and protein structure with similarities in the replication and expression profiles, which include the frameshifting in Orf1ab which encodes 16 non-structural proteins, a potential target for RNA binding molecules. Both viruses have a receptor binding domain (RBD) in S recognising the angiotensin converting enzyme 2 (ACE2) receptor on cells which activates a TMPRSS2 mediated viral invasion[92],[93]. Crucially, there is more than one mechanism of entry[94] depending on the expression of TMPRSS2. In absence of TMPRSS2, a late endosomal pathway is activated as described before which slows down the entry process. Protease efficiency and endosomal pathways can be important when the replication of SARS-CoV2 is studied in different cell lines (Vero, Caco-2, Calu-3). The combination of receptor binding and proteolytic cleavage is crucial for efficiency of membrane fusion[95],[96].

1.1.17 Protein expression and RNA interactions

As part of the nidovirus family, SARS-CoV2 employs a multiple ORF translation strategy[97]. Specifically for SARS-CoV2[98], the genome can be expressed in 10 major ORFs with accessory genes producing proteins or RNA that is potentially functional in the replication process and/or host interaction/pathogenicity[99]. It is worth mentioning that the expression, regulation and order is mediated by RNA-RNA interactions[100] and RNA-protein interactions raising an opportunity for RNA binding drugs. For the purposes of this work, it is particularly interesting to study the structure and dynamics of the 5' UTR[101], and the frameshifting element[102],[103],[104]. The 5' UTR is known to have multiple roles in mediating the replication and mRNA production[57]. It includes the transcription regulatory sequence (TRS) which is conserved among all coronaviruses[57] and a series of stem loops, including a branched system-loop which will be examined further in chapter 5.

1.1.18 Antivirals.

There are under 100 different licensed antivirals with about 50 that can be used on humans (varies between countries). With the exception of influenza viruses[105],[106] which cause an acute infection, research has focused on viruses causing persistent infection (HIV, HCV, herpes simplex)[107],[108] and lately there is an interest for viruses that can often become pathogenic in immunodeficient patients (secondary infection in HIV/HepC patients)[109]. Antiviral strategies can be developed for each one of the stages of replication described above, and antiviral agents can be either small organic or metallo-organic molecules (none licensed), small peptides[110], small proteins (interferon or other cytokines whose mechanism of action was briefly mentioned earlier) or longer polymeric chains with or without metal ions. We can categorise antivirals broadly into 3 categories: the ones with higher affinity for viral proteins and structures, those that target host structures and those that can influence both. The major challenge in designing antiviral drugs is to stop the replication cycle while not inducing cytotoxic effects on the host cells. Moreover, the drug needs to be highly potent, since the large number of particles that can be produced even during a day in combination with the high error rate of viral polymerases, especially in RNA viruses, would make the production of a resistant mutant very likely. Finally, most pharmaceutical regulatory bodies (including FDA) need two animal models to approve drug trials in humans which is difficult to produce due to the specificity of most viruses. This creates a difficult regulatory landscape for approval of new compounds, although during outbreaks licensing is often fast-tracked under emergency use, as was shown during the SARS-CoV2 pandemic with the UK becoming the first country to authorize antivirals Molnupiravir and Paxlovid[111]. Here, I will summarise the current state of antiviral research with respect to stages in the replication cycle often drawing examples from HIV-1, since it is the most studied.

Starting with attachment, one can inhibit the attachment of the virus to its selected receptor by blocking the binding side of either of the first membrane protein inducing conformational changes and thus destroy the initial binding. In the case of HIV-1, there are two drugs for this stage, enfuvirtide; a fusion inhibitor that binds a region of gp41 and disturbs the conformational changes in membrane fusion, and maraviroc, a CCR5 antagonist.

Going further, after fusion in most cases the endosome starts to acidify and the uncoating is mediated by lower pH. In the case of influenza virus, the lower pH activates the capsid - ion channel protein M2 which drops the pH in the viral interior and in that way weakens the electrostatic interactions that hold the capsid intact and starts to release the ribonucleoproteins and the RNA into the cytosol. Amantadine, one of the first antiviral drugs blocks the M2 channel and inhibits the release of the viral genome and proteins into the cytosol. Amantadine was approved even before the mechanism of action was known and after it was studied it showed the importance of pH change mediated conformational change. However, overuse of amantadine in combination of the genomic versatility of the influenza virus made most of the current viruses immune.

Next stop in the replication pathway is the nucleic acid synthesis, where polymerases (RNA, DNA pol, or reverse transcriptase) are very attractive targets for antivirals. The first class of inhibitors that showed clinical efficacy here is the nucleoside analogues (NAs). Currently they are used against most viruses, HBV, HSV-1,2, HIV-1. The analogue is used instead of the canonical base and inhibits the attachment of the adjacent one, stopping the process. NAs have effective antiviral activity even in concentrations a lot smaller than needed to create problems with cell machinery leading to toxicity. They are often given as pro-drugs in the form of triphosphates so that they can be more selective to viral mechanisms or to enhance delivery. Specifically, for HIV-1 there has been a series of drugs that have been approved, targeting the viral polymerase which can be classified into 3 groups. Nucleoside reverse transcriptase inhibitors (NRTIs) such as zidovudine, nucleotide reverse transcriptase inhibitors (NtRTIs) of which the only approved is tenofovir, and non-nucleoside reverse transcriptase inhibitors (NNRTIs)[112],[113] (Efavirenz and Etravirine) which have a completely different mode of action since it binds to the allosteric pocket of the p66 subunit of the RT and not the active site.

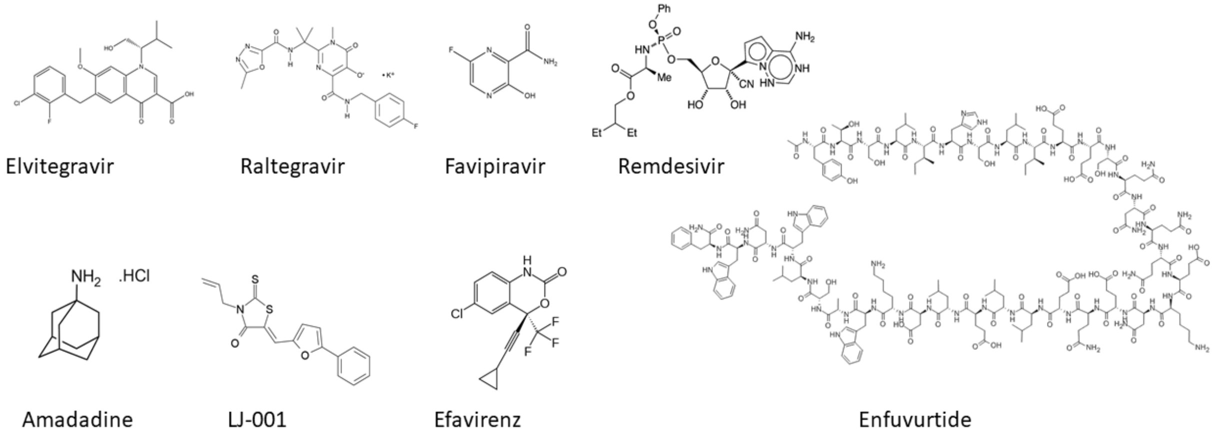


Figure 1.9 *structures of commercial antivirals.*

All the above compounds interfere via binding pockets on proteins. An alternative target is the structural elements of viral nucleic acids. Specifically for viruses, targeting 5’ and 3’ non coding units’ structure can be advantageous as an antiviral target due to their highly conserved nature, in contrast to the rest of the genome[114]. Misfolding of these regions would lead to inability of the genome to regulate expression and replication. In HIV-1, similar to the small 5’-LTR end (TAR), a more complex HIV-1 RNA target is the Rev-responsive element (RRE). RRE is a cis-acting RNA element in all intron-retaining viral mRNAs. It is the main binding site of Rev which is produced from the fully spliced mRNA. Initial binding of Rev to RRE and the level of oligomerization with additional Rev molecules changes the overall affinity of the complex up to 500fold[115] thus regulating the splicing and extraction of unspliced mRNA to the cytoplasm. More interestingly, the RRE can adopt two different configurations, resulting in different conformations of the complex which in turn regulates the replication kinetics of HIV as a response to various immunological cues[115]. Further review of nucleic acid structures targeting will follow below.

On the subject of broad spectrum antivirals, there are a few notable examples studied in the last few years that need to be mentioned in order to discuss the necessary requirements and limitations for candidates as discussed in detail before[116]. First LJ001[117], showed remarkable initial success inhibiting a series of viruses by attacking/deforming the envelope of the virion, hence inhibiting the ability of enveloped viruses to attach to cells. It is important to remember that some damage is also done to eukaryotic cell membranes but cells can regenerate their membrane whereas virions cannot. The success of LJ-001 was short lived since further study showed that LJ-001 disturbs membranes by causing oxidation of the membrane lipids and free oxygen radical when in an excited state, practically means under light irradiation (where all of the experiments took place)[118] . The second example of broad spectrum antiviral activity lately is favipiravir[105],[119] (T-705), the most studied in the class of RNA polymerase inhibitors, along with remdesivir[108] showing promising results in mice and hamster models which also went into trial for SARS-CoV-2 pandemic. For the first part of the pandemic remdesivir was the only antiviral drug that was used, although later double blind controlled studies showed little to no effect[120]. This also highlights the importance of timely treatment with antiviral drugs.

Overall, the research for antivirals is focusing on protein inhibitors or membrane targets. What is missing is targeting nucleic acids themselves, which is not surprising since structural analysis of non-coding RNA and DNA and its link to biological function is a relatively new and fast growing research field. During the SARS-CoV2 pandemic there have been numerous attempts to repurpose or design new small molecules targeting all phases of the viral replication with various degrees of success as it will be discussed further in chapter 4.

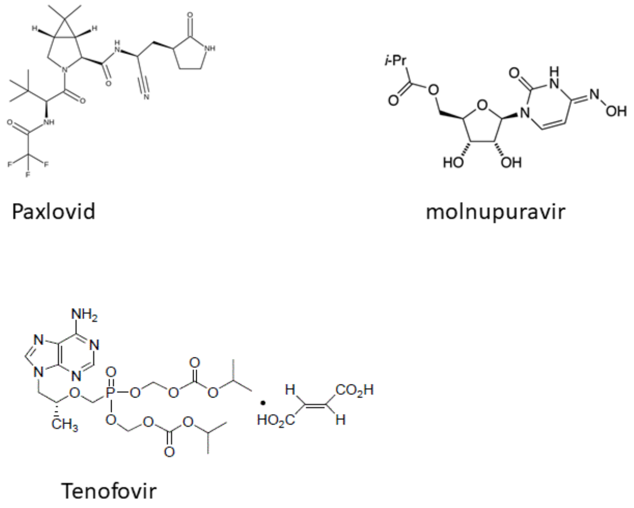


Figure 1.10 *structures of commonly used antivirals.*

1.1.19 Metal-containing antivirals

Metal-containing compounds were shown to exhibit antiviral properties as early as 1985[121]. In that case, the agent involved coordination of a metal ion to the ring nitrogen atoms and exocyclic oxygen of known antiviral agent ribavirin. Research interest in such a strategy is still active[122] and focuses mainly on enzymatic inhibition and use of nucleotide analogues. Another application of metal-containing compounds is as inhibitors of either attachment or fusion. In the case of HIV-1, many poly-ionic substances have been tested including polysulfates such as polyacetylal polysulfate (PAPS) and polyvinylalcohol sulfate (PVAS)[123]. The polyanionic substances' antiviral activity is broader than HIV, with activity against other enveloped viruses, like herpesviruses, cytomegalovirus, influenza A and RSV, in all cases in effective concentrations orders of magnitude below the cytotoxic concentration (although both are measured at 50%).

The major driving force of research in metal-containing drugs is their success as anti-cancer chemotherapy agents[124],[125] The common characteristics between cancers and viral infections (increased nucleic acid production, changes in apoptosis regulation) led to a few metal-containing compounds being studied for antiviral activity, almost always against HIV. In 2016, an Italian group tried to use known anti-HIV drugs specifically targeting the HIV integrase (Elvitegravir and Raltegravir) as well as an antibacterial targeting DNAgyrase (Quinolones) to form ruthenium complexes.[126] Their methodology and experimental analysis follows a 2006 paper regarding the chemical alteration of the drugs in order to have increased stability as ruthenium complex, but they claimed the first use of ruthenium compound as an anti-HIV-integrase drug. More relevant to this work is the 2008 paper; ”Binding of dinuclear ruthenium(II) complex to TAR region of HIV-AIDS viral RNA”[127]. It is the first to include modelling aspects (without dynamics) and explore the binding between the Ru-complex and TAR using NMR but lacks any biological significance. On the other hand, a very small number of review papers summarised anti-HIV activity while reviewing metallodrugs[128]. There, one of the ruthenium-based complexes (Ru(Bu_2bpy)$_2$(2-amino-4-phenylamino- 6-(2-pyridyl)-1,3,5-triazine))(2+) is reported to have antiviral activity against HIV-1 but not HepBV. Finally, the most thorough investigation with regards to antiviral activity of metalo-drugs was undertaken with Cobalt compounds[129]. Cobalt has very few known biochemical roles - most important among them is Vitamin B12. Epstein and co-workers reported that cobalt complex CTC-96 was very potent treatment for herpes simplex virus type 1 and early in the 2000s other compounds of the same CTC family should succeed against adenovirus both in culture and rabbit model and lead to be developed and sold as Doxovir (TM). There are 2 modes of action; CTCs are known to bind strongly histidine residues, which are often found in viral maturation protease and a serine protease, with particularly high stability in the axial position[130].

The next Cobalt containing example of antiviral is Cohex (Hexamminecobalt(III) chloride), a classical Werner complex. In contrast to CTC ligands, the ammonia ligands are inert towards ligand exchange but it does have the ability to form hydrogen bonds with nitrogenous bases of nucleotides and the phosphate backbone of DNA[131]. Although the mechanism of action is not fully understood, it has been shown to reduce protein synthesis, which would mean that its effect on the cycle is further upstream than translation. Also, very important in the context of this research, Cohex disrupts the interaction of Sindbis virus glycoproteins with highly negatively charged polysulfonated heparan sulphate receptors, disrupting the viral entry. Based on the success with those characteristics a larger cobalt compound which was produced in the 60s[132] seemed a worthy candidate for antiviral. Cingler and Rezacova have used cobalt bis(1,2-carbollides) as antiviral specifically targeting HIV protease[133] with a series of compounds and XRD of 2 molecules of the simplest compound in the binding site of HIV protease.

Since Cohex is easy to synthesise, there has been an attempt to study different metal centres of Cohex, specifically Ni and Ru[134],[131] but without any success in improving antiviral activity against Sindis virus (+)ssRNA -nonretroviral. In that study, they also reported that Cohex with Ru centre at 0.325mM showed an increase in viral protein expression (measured as fluorescence intensity of reporter GFP) and an increase in survival rate at the same concentration.

During the SARS-CoV2 pandemic, there has been a substantial effort to identify possible metallodrugs interfering with the replication cycle of SARS-CoV2 or treating Covid-19[135],[136],[137],[138],[139],[140], this will be examined further in a dedicated chapter 6. Fundamentally, as it can be seen in this small review here, the challenge for traditionally designed and synthesized coordination compounds is the lack of selectivity. This work aims to understand the potential interactions, identify the potentially therapeutic targets, and propose modifications to the cylinder that would enhance those against the off-target binding.

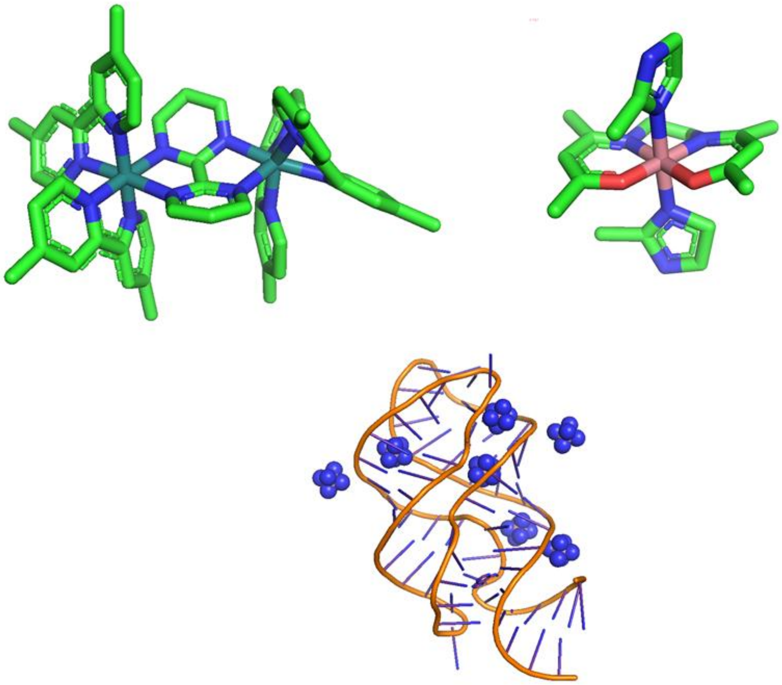


Figure 1.11 *metal complexes with some antiviral activity, from left to right; dinuclear Ruthenium by Keene, CTC96-Cobalt complex, cohex co-crystal with B. Subtilis riboswitch.*

1.1.20 RNA targeting Compounds

The RNA targeting field has expanded rapidly over the past 8-10 years, especially in the field of oncology and cancer related long-noncoding RNAs[141],[50],[142]. The Disney group has been in the forefront of the field[143],[144] with a series of dimer molecules targeting the bulge loop regions of RNA stem loops[145],[146]. The optimization process for these molecules employs massively parallel assays[147] and computational methods that move the bioinformatics of drug design forward. The existence in vivo and functional role of RNA G quadruplexes has been studied extensively both in mammalian cells[148],[149],[30],[150] as well as in relation to viral genomes[151],[152],[153]. Although RNA was first considered to be a valuable antiviral target 20 years ago[114], it has only been in the last 5-10 years that research on RNA structure has boomed[154]. Overall, out of the estimated ~85% of the human genome that is transcribed only ~3% is translated [155], leaving the rest of produced RNAs “functionally” orphan. A lot of those orphan RNAs are now found to have highly structured regions that play crucial roles in regulating gene expression including splicing[156],[157], tRNA, microRNAs which have been found to be related to certain cancers[158]. In a series of publications in “MicroRNA Biogenesis and Cancer”[159] miRNAs are characterised as Highly Conserved molecules, which is not at all surprising if we revisit the introduction to virology earlier, even in short viral genomes the diversity can be found mostly in the coding regions since disturbing the structure of binding sites has a higher penalty. Given the vast number of host miRNAs and other structurally unique nucleic acids in the cell, one might wonder if targeting highly structured nucleic acids can be a viable option for antivirals. As it is discussed in the previous section, inhibiting the initial stage of the virus would limit the abundance of the targeted viral genome to the multiplicity of infection and statistically that can be done at far lower concentration of that causing cell toxicity.

1.1.21 Closing remarks I

This section summarised the minimal understanding of nucleic acid structures, viral replication and tries to identify potential bottlenecks where chemical intervention can disturb the replication cycle. The relatively small landscape of antiviral drug strategies has been discussed including metal-containing compounds overall keeping the focus on the early stages of cell infection and stages where coordination compounds based on the Hannon cylinder can be optimised to disturb the viral replication cycle.

Section 2 – Introduction to theory and computation

1.2.1 Theory and computation

This section discusses the theoretical and computational tools one can employ to study the impact of compounds in the replication cycle of viruses, focusing on the genomic and other helper nucleic acid oligomers as targets for antivirals - specifically, how conformational changes induced by charged metal-containing compounds and ions can be studied computationally.

The increased interest in RNA has also boosted the theory and computational methods used for nucleic acids. Starting decades ago, physicists and mathematicians studied RNA and DNA topology proposing alternative structures[160] and dynamics. After a large gap, in 2008 Bon et al [161], tried to redefine the relevance of topology introducing a new topological classification of RNA structures similar to those used in protein folding, and used PDB entries to classify structures and form a statistical model. Classifications based on topology can be a first step in separating potential targets for any drug targeting non-canonical RNA and DNA structures by minimising the dynamical conformational structures to topological ranks.

It is worth exploring the theoretical frameworks involved in this thesis from the shorter time and length scale (QM-DFT) to the longest MD-thermodynamics.

1.2.2 Density Functional Theory

If one defines a molecule as a closed and stable group of nuclei and electrons interacting mostly with each other, then a structure for this group can be proposed by minimising the energy of their interaction. Even for given positions of nuclei and considering them static (with Born-Oppenheimer approximation), this is a many-body quantum mechanical minimisation problem of a N-electron wavefunction, impossible to solve for almost all applications.

Density Functional Theory attempts to reformulate and reduce the N-electron wavefunction to a function of 3N variables, allowing the energy and properties calculation of ground state bypassing solving the molecular wavefunction. In 1964, Hahenburg and Kohn proved that "the electron density determines the external potential" and in 1965 Kohn and Sham published the Kohn-Sham (KS) equation[162]

$$E[\rho] = T_s[\rho] + V_{ext}[\rho] + V_H[\rho] + E_{xc}[\rho] \quad \textit{Eq. 1.1}$$

$$\left[-\frac{1}{2}\nabla^2 + V_{ext}(r) + \int \frac{\rho(\boldsymbol{r}')}{|\boldsymbol{r}-\boldsymbol{r}\prime|} d\boldsymbol{r}' + v_{xc}(\boldsymbol{r})\right] \varphi(\boldsymbol{r}') = \varepsilon_i \varphi_i(\boldsymbol{r}) \quad \textit{Eq.1. 2}$$

$$\text{With } v_{xc}(r) = \frac{\delta E_{\chi\psi}[\rho]}{\delta\rho} \qquad \textit{Eq. 1.3}$$

Which recasts the Schrödinger problem of interacting electrons to a single electron problem moving in an effective potential (the energy cloud caused by nuclei and other electrons). The total energy is the sum of the kinetic energy of the electrons, the total potential due to Coulombic interactions with each other and the nuclei, and the quantum mechanical contribution for exchange and correlation of the electrons and is a function of electron density ρ(r), therefore a functional. This electron density, the space where the minimisation of energy is solved, can be described by different mathematical functions, but for the purposes of this thesis only Gaussian basis functions have been considered as they are more appropriate for single molecule systems[163].

There is a long list of energy functionals published over the last years with a great variation. Many introduce empirical parameters aiming to increase accuracy in the specific field of interest. An inherent limitation factor of DFT is the lack of one universal functional that could be used in different property calculations and different systems producing consistent results. 10 years ago, even in the same theory level and same family of functionals, their creators would suggest slight deviations depending on the property needed to be calculated (thermochemistry, kinetics, excitations)[164]. During the last few years, extensive work has been done particularly at UC Berkley on increasing the parametrization of a family of functionals, resulting to a new improvement on their functional family[165] introducing 14 parameterisations after screening through trillions. (This paper continues the divide described[166] between theory and further numerical parametrization). But even then, this research is limited to main group chemistry. Generally, functionals are classified and ranged on a Jacob's ladder of DFT in six major groups; local (spin) density approximation (LDA), generalised gradient approximation (GGA), meta-GGA, hybrid Density Functionals, double-hybrid Dfs, and range-separated Dfs.

Over the past decade, following the dramatic increase of frameworks and applications of machine learning algorithms, several groups started taking advantage of their efficiency to solve ground state (DFT-optimisation[167]) as well as molecular dynamics[168],[169].

1. Local Density Approximation (LDA) and Local Spin Density Approximation (LSDA) are based on homogeneous electro-gas model which works fine for geometry and vibrational analysis, when density varies slowly with position but fails slightly in molecular atomisation energies.

2. Generalised Gradient Approximation (GGA) Df, corrects the density variation by introducing the gradient of the electron density for two spins (contributing to electron density inhomogeneity).

3. Meta-GGA Dfs, also include the Laplacian of the density and/or the orbital kinetic energy, which is a logical next step, going down the terms of the Taylor expansion for KS- orbital kinetic energy densities.

4. Hybrid functionals; A major step forward took place when Becke[170] proposed a mix of GGA functionals with exact exchange (HF), separating the exchange and correlation component as GGAs do but adding a weighted separation. Weighting the separation coefficients gave rise to a huge amount of functionals which are tailored for the needs of specific systems and don't transfer well across.

5. Double hybrid Dfs, here the exchange term is still calculated as a percentage of the HF exchange and the correlation energy term is substituted in part with ab initio correlation energy e.g. MP2-second order perturbation treatment of KS orbitals.

6. Range-separated Dfs; is a combination of two functionals from above to describe exchange and correlation in different ranges (long-short).

After choosing a density functional, one needs to choose a basis set, a collection of vectors which spans a space in which a problem is solved. In quantum chemistry, the basis set refers to the set of non-orthogonal one-particle functions used to build molecular orbitals. The list of options is not as long as it is for functionals but nonetheless the combination of basis-functional is not straightforward (for a given system, the optimum basis set differs between functionals). Bypassing plane-wave basis sets which are more appropriate for periodic systems, here, only Gaussian-Type orbitals (GTOs) have been examined, where the width of the orbitals can be easily computed by using the Gaussian product theorem. On the other hand, for Slater-Type orbitals (STOs) the exponent is linear with r making it more difficult to compute but providing better results. Merging the two, one can use multiple gaussians to mimic STO behaviour which gave rise to STO-nG , n the number of gaussians used. Here, STO-3G has been used for software testing but for any other geometry optimisation and properties calculation double and triple zeta basis, usually polarised with or without split valence have been used (split valence- effective core potential for the case of Ruthenium). For studying anions, it is important to include the effects of diffusion in the functions (smaller ζ), meaning that the electron can be farther away from the nucleus. In the case of the cylinder, we can also use these functions since we are aiming to study the effect on van der Waal's non-covalent long-range interactions when we move from DFT to Molecular Dynamics and QM/MD.

1.2.2.1 DFT Functionals

B3LYP

B3LYP has been the most widely used standard of density functionals over the last 20 years[171]. Originally developed to study vibrational absorption and circular dichroism it offered a good compromise between computational cost, coverage, and accuracy of results. The reason it is used here is that being so popular one can compare one's results with that of a long list of publications.

CAM-B3LYP

A major improvement of B3LYP came with incorporation of Coulomb attenuation methods[172]. Although the CAM-B3LYP corrects the enormous underestimation of long-range effects like charge transfer, by introducing range-dependent parameters for exchange - correlation, with only the standard error function for smoothing the boundary it has not caught on as much as it should.

Minnesota functionals

In recent years the most successful family of functionals (due to fast implementation of computer software) has been the Minnesota functionals[173], after that the group updates the series almost every 4 years while including occasional revisions of the most popular ones[174]. As reviewed by Head-Gordon[175] and others, Minnesota functionals are heavily parametrised and optimised for speed while offering specific series for metal-organic compounds and transition metal chemistry[176]; Mx-L are the local functionals series which is optimised to give very accurate thermochemistry results but not the state of the art for systems with high self-interaction induced errors. On the other hand, the hybrid meta-GGA M06-2x and M11 can provide a good balance, the former often[176] replacing B3LYP or PBE0 for parametrization of non-standard residues for molecular dynamics (AmberTools18, CHARMM). The -L shortcomings regarding self-interaction can be overcome by increasing the size of basis set while introducing non- xc functional - dispersion corrections (in nwchem[177]).

SSB-D

Given the size of the cylinders, it is important to use both short and long-range energy contributions and one simple way to do so is employing range separation functionals. In the case of SSB-D[178] , this is achieved by a smooth transition between PBE (a GGA functional) and OPBE[179] (which adds Handy's optimised exchange to PBE) at a predefined and optimised point P of the reduced density gradient s. This results in a more accurate description of reaction barriers and spin-state energies (from OPBE), as well as respectively good results from PBE, ie hydrogen bonding and π-π stacking/interactions.

1.2.2.2 Basis sets

The choice of basis set has taken a secondary role since it is often the case that overall, the errors in the functional will decrease with increasing complexity of the basis set. For that reason, 6-31G* has been chosen for everything in the first round of simulations. Once some functionals have been identified as highly efficient, lan2 and def2 families of basis sets have been used.

1.2.2.3 Dispersion correction

Dispersion correction introduced in the NWCHEM code is based on the preliminary work of [180],[164],[181],[182] Grimme et al. which includes (geometry) dependent optimisation as well as fractional coordination numbers (smearing). There is a further range separation classification to medium-range correlation (2-4A) and long-range (London dispersion[180]) above that.

1.2.2.4 Solvent effects.

On top of corrections regarding long-range interactions within the molecule, corrections to address the interaction of the molecule with its solvent surrounding can be approximated, with a Universal Solvation Model [183], SMD. In this model, the observable solvation free energy is split into two main components, 1) bulk electrostatics from a self-consistent reaction field, corresponding to the solution of the non-homogeneous Poisson equation in terms of integral-equation-formalism polarizable continuum model (IEF-PCM) where bulk cavity is defined as volume left after superposition of nuclei centred spheres and 2) the cavity – dispersion solvent structure term arising from the short-range (local) interaction between solute and solvent in the first solvation shell. Minnesota hybrids and B3LYP were the main functionals for the training set but in the NWCHEM implementation the cavity can be defined by any geometry, corresponding electron density and corresponding polarisation to create a domain for the non-homogeneous Poisson equation.

During the last 10-15 years, research bodies have been increasingly investing in computational resources to decrease the overall cost of research while increasing efficiency. Specifically, for transition metals, comprehensive reviews were published in 2009 and 2012[183],[184], pointing out the increasing demand for the development of application-specific functionals and demonstrating the catalyst community as the driving force in the field. As the field is driven by applications in solid state and electrocatalysis, there is substantial room for research in electronic dynamics for single molecule in solution. Early steps towards addressing the challenge can be found in two reviews in 2016 and 2018 on excited states[185],[186] and overall discussion of the current state of the art of DFT[187]. The field expanded from blind computational screening of tens of thousands of compounds and targets in high-throughput screening generation[188],[189] to incorporating big data algorithms[190] and using big data approaches to correct errors in results at the quantum mechanical level[191], recognising the importance of thorough analysis in this level in the success of any simulation or docking study in larger scales.

1.2.3 Molecular dynamics

Moving up in the time and length scales, the definition of a molecule is reduced to a closed group of atoms, connected through bonds. Molecular dynamics investigations of biomolecules have been a crucial and increasingly important part of structural biology and drug discovery over the past half a century. In Molecular dynamics (MD), atoms are now treated as spheres of different mass and diameter and their motion is described by Newtonian physics. The Forces applied to each sphere are given by the potential energy function:

$$V = \frac{1}{2}\sum_{bonds} k_b(b-b_0)^2 + \frac{1}{2}\sum_{bond\ angles} k_\theta(\theta-\theta_0)^2 + \frac{1}{2}\sum_{dihedral} k_\varphi[1+\cos(n\varphi-\delta)] + \frac{1}{2}\sum_{non\ bonded}\left[\frac{A}{r^{12}} - \frac{C}{r^6} + \frac{q_1 q_2}{\varepsilon r}\right], \qquad \textit{Eq. 1.4}$$

The first three terms describe the bonded interactions, specifically energy with regards to length of the bond (distance of the spheres that are not ancestrally elemental atoms), the energy associated with the angle between 3 atoms about an equilibrium and the torsional rotation of 4 atoms about a central respectively, whereas the fourth describes the non-bonded interactions with respect to van der Waals (using Lennard-Jones 6-12 in this case) and electrostatic/Coulombic interactions respectively. Parameters that would describe specific bonds can either be out of experimental data[192] (IR , NMR, low temperature crystallography, or quantum mechanical optimisation of the structures[193]. For the Coulombic term, the charges are located in the atomic (sphere's) centre and have fixed value, remain constant regardless local or global conformational changes and do not respond to external electric fields, including those originated by the solvent.

This theory can describe very successfully protein interactions as well as membrane behaviour. On the other hand, additive models are particularly inadequate for modelling base stacking, hydrogen bonding, and ion interactions which are all crucial stabilising factors in nucleic acids[194]. Additionally, especially for the purposes of this thesis, as shown by studies of DNA on graphene[195], aromatic rings have quantifiable molecular polarizability due to the delocalization of π electrons, which increases the thermal stability of the duplex specifically for DNA[194].

Complementary to forcefields, there is another important aspect of macromolecular simulations to be considered; jiggling and wiggling[196]. Macromolecules themselves, and each system that is studied in this scale, are thermodynamic systems[197]. Copper in 1984 [198] coined the thermodynamic uncertainty principle to refer to the inherent uncertainty about the particular state of a macromolecule. From there, Wolynes et al[199] expanded and analysed the statistical treatment of macromolecules and most importantly the landscape view for proteins[199],[200], preliminary studies for the nucleic acids (RNA) came along later[201]. In this "new view"[202], one needs to describe some terms.

Phases are generally associated with local free energy minima and can be described by sufficient parameter(s) that would explicitly describe the system in those minima. Stability; each phase responds differently to perturbations in external parameters (temperature, pH, solvent). If the interconversion time of conformations within a minimum is fast compared with the transition rate to other minima we may view relatively high free energy minimum as “metastable” phase. Transitions are referring to the free energy barrier between two minima signals distinct phases that are related by a first order (cooperative) phase transition, ie. when the two minima exchange relative stabilities, the equilibrium value(s) of the order parameter(s) change discontinuously. In contrast, continuous transitions are described by smooth shifts in the locations of a single minimum with changing external conditions, or the splitting of one minimum into two. That had a major influence on the way molecular dynamics has been done over the past twenty years, with increasing efforts to sample greater space of phases and hence including greater area in the conformational space.

This is particularly important for biomolecules since they often have local minima separated by high energy barriers[199], making it easy to fall and get trapped to a local minimum with unrelated function for the study or worse no function, no matter how long the simulated time is.

These are four approaches to overcome high energy barriers between local minima;

1. Replica-exchange, developed by Sugita and Okamoto[203], employs several parallel simulations of the system at different temperatures and introduces state exchanges between replicas based on Monte Carlo weights defined by the temperature gap. It is also worth mentioning, Hamiltonian and multidimensional Replica exchange methods, where sampling can be on other dimensions as well as temperature[204].

2. Metadynamics; Parrinello’s group introduced a strategy to accelerate sampling of the energy landscape by discouraging previously visited states[205]. Doing so one gets higher on any well thus allowing transition to another minimum[206].

3. Generalised Simulated Annealing; based on the annealing principles in metallurgy, GSA methods depend on lowering an artificial temperature that decreases during the simulation with the expectation that the trajectory would find the minimum[207] . Similarly genetic algorithms would transform the energy minimum to an optimisation problem that can be solved with machine learning and "genetic" combinatorial functions[208].

4. Gaussian acceleration[209],[210]; Developed relatively recently by the Miao group, Gaussian accelerated molecular dynamics can enhance the sampling space while calculating the corresponding free energy landscape by using harmonic functions to apply a boost potential adaptive to the biomolecule. The boost can be applied to the dihedral angles or to the whole biomolecule and it is following a near-Gaussian distribution which allows for re-weighting and mapping between the conformation and the corresponding free energy of a state. The main advantage of these methods is that boosting is not applied along a specific coordinate reducing the applied bias compared to other methods (metadynamics, umbrella sampling) and does not require multi node computational infrastructure (REMD). In a newer variant of GaMD, Ligand GaMD, the boost is only applied to the interaction between the atoms of a ligand and the biomolecule which directly measures the energy required for dissociation between the two again without induced bias in a single coordinate.

Finally the fundamental theory behind macromolecular processes is still studied with increasing interest lately on the fundamental bounds of predictions[211],[212]. The vast majority of computational molecular dynamics work has been done with proteins and in a 2012 review[213] only two paragraphs were devoted to nucleic acids explaining the difficulty in producing results due to lack of solved structures and inadequate force fields in the past as well as the inability to translate the success of sampling methodologists developed for proteins to nucleic acids[214]. It was the latest improvements of the AMBER forcefield from the Mathews' group in 2017[215] and Shaw group in NYC[216] in 2018, that produced force fields comparable with proteins for RNA, based on long-range correction of DFT and MP2 levels of QM theory. The Sponer group has published extensively reviews of the state of RNA MD simulations along with evaluation of current forcefields and comparison with their efforts to create parameters[217],[218],[219]

1.2.4 Principles of DNA and RNA dynamics.

The two forms of double-stranded helix (A and B) and transitions from one to the other is an opportunity to test MD parameters and simulations. The stability of DNA has been largely overestimated by older force fields[220]. It also allowed for false confidence in multi-scale simulations. In coarse-graining models the system is simplified from atomic to pseudo-atom per base, to plane per base to cylinder per each base[221] or ellipse describing the base pair[222],[223]. Even the landscape of transitions from A to B DNA and back has been challenging. Lai and Schatz[224], calculated the different conformations in water and 85% ethanol using the Amber16 tool kit with umbrella sampling concluding a uniformly downhill transition to B-DNA but also the presence of another minimum (local) corresponding to A-DNA that they associated with aggregation. Addressing those problems, a lot of groups turned to quantum mechanical calculations, to understand the different energy contribution of the stacking[24] and backbone[225] in the different conformations. In a comparison of the last 3 AMBER forcefields, including the bsc1 and OL15[226] modifications, none of them provided satisfactory results of the B to A transition (compared to NMR data)[227]. Finally, a single strand of DNA, rich in guanines, can form tertiary structures with 4 guanines forming tetrads which when stacked form a G-quadruplex[228],[229].

Besides that, it is worth mentioning the studies of DNA condensation at mesoscale level induced by multivalent ions[230],[231],[232] where they use Cohex(3+) (an antiviral described earlier, but also commonly used as a DNA crystallisation aid and also known to promote Z-DNA). In this case, parameters for Cohex were obtained by Carr-Parrinello MD, and observed B-DNA to A-DNA transition on 36bp DNAs and in higher concentration formation of bundle like structures. Later, they reduced the macromolecules to a coarse grain description, with effective sites of groups of atoms and effective potentials fitted from the underline atomistic simulation. Finally, the longest study describing DNA and ligand interaction describes the interaction of B-DNA with ethidium bromide[233], where they conducted multiple very long molecular dynamics simulations analysing the interaction between ethidium bromide and double stranded B-DNA. This creates a benchmark for methodology and demonstrates how the simulations can be validated against experimental data.

Although the landscape in mesoscopic research has not changed much in the last years, the interest in DNA non-helical structures has sky-rocketed, further revealing the inadequacies of older forcefields. To improve on that researchers returned to the principles of the forcefield and examined the interactions both with MM and QM DFT-D3 (with dispersion - discussed later), or introduce empirical polarizable forcefield (Drude oscillator model)[234]. G-quadruplexes have been a prime testing ground for DNA forcefields, with Gkionis[235] and Song, Ji and Zhang[236] both highlighting the importance of polarisation in the force field and illustrating that by using the state of the art (ca 2014) non-polarised force fields (amber10 and amber12) and polarised field (PNC) one can create new QM calculated parameters for the charges of the atoms of the G-DNA quadruplex.

The corrections and additions on DNA force fields are small compared to the almost new field of RNA dynamics [217]. RNA structure can be classified as primary (1D), secondary (2D) and tertiary (3D) structure. Although formally single-stranded, RNA can fold on itself creating regions of anti-parallel WC base pairing which are particularly stable (1-3kcal/mol/base). RNA secondary structure prediction has drawn a lot of attention [237]. Most algorithms use free energy minimisation protocols[238] or scanning /statistical methods with partition functions[239] and there is a long list of software available for each case. Tertiary structure refers to the non- WC and long-range interactions of the 3D structure and poses a very interesting mathematical problem [240].

It has been profoundly difficult to understand and simulate ion effects on the structure and functionality of RNA to the point that for years the topic was described as the elephant in the room[241]. In principle, every folding and conformational change of nucleic acids involve the relocation of two or more strands of negatively charged phosphate groups which contribute towards an unfavourable energy contribution of the structure. Metastable regions on the energy landscape can be created by the presence of salts. Monovalent atoms are considered to play a more dominant role in stabilising the electrostatics surrounding the structure (local ionic atmosphere) whereas divalent (Mg2+) have been shown to be chelated to sites involving the phosphate groups thus mediating the electrostatic interactions between those groups[242]. Publications in this area mainly focus on understanding the 3D structure of RNAs[243] but the few that attempted to review RNA-ligand interactions[231], urge the importance of electrostatic distribution of the ligand, solvation effects on both sides and the largely kinetic nature of the interaction, which usually leads to higher induced affinity, and almost always to a more complicated energy landscape, Additionally on should also consider the effects of molecular crowding to the in vivo RNA conformational landscape as described by Chen[244].

Overall, the RNA field is more vibrant than ever and for the first time RNA forcefields have reached similar accuracy with that of proteins[245]. Shaw's group managed that by using MP2 level of theory for the parametarisation of the forcefield. The importance of sampling the energy landscape is even higher than that of the proteins given that structural features are a lot more diverse (even for a given sequence) and the kinetics from one state to others are tightly connected to the regulatory role of DNA and RNA structures. Regarding interactions with ligands, the effect of a ligand on the nucleic acid is as crucial as to the ionic atmosphere around the complex (crowding-local ion concentration).

Fundamentally, this thesis examines transitions between metastable states and for the molecular dynamics part of this thesis this is achieved by reducing the dimensionality of the trajectories to a space where Markov state modelling can be used to quantify the transition efficiency[246],[247],[248],[249],[250].

As opposed to the acceleration techniques described earlier, this approach attempts to examine the landscape as well as the transitions between local minima within it. This is more appropriate for the study of flexible systems although requires very long simulations and has a high computational cost.

There are two ways dimensionality reduction and Markov state modelling has been used during this work.

1. To present the results of a single very long simulation in a digestible format, which does not quantify necessarily the characteristics of the system.

2. To combine the results of multiple simulations which can lead to a closer representation of the dynamics of the system. Markov state modelling starts by reducing the dimensionality of the system from 4 times the number of atoms to a set of coordinates that would ideally capture the relative states of the undergone transformations through the sampled space. This smaller space can undergo traditional clustering (k-means, k-centres etc) using a structural metric. Principal component analysis and time-lagged component analysis are ideal to further reduce the dimensionality of the problem while retaining the kinetic characteristics and nature of the data set. A transition matrix can be assembled with the help of Bayesian probabilities to link non-directly connected microstates.

1.2.5 Docking studies

There are various software packages that attempt to quantify the binding affinity of chemical compounds to biomolecules [251],[252],[253]. Most platforms work with organic molecules targeting well-defined structurally rigid pockets in biomolecules and evaluating the binding affinity based on a scoring function. These platforms can usually compute the conformational space of the small organic molecule based on its flexible and rotatable dihedral angles and some can allow for this type of flexibility on the target. Machine learning algorithms have been used to increase the efficiency[12] and quantum mechanical approximations to increase the accuracy of the scoring function[254],[255],[256]. In some cases, effects of solvation have been taken into account explicitly[257]. Docking has been hugely successful in computational drug screening[258],[7],[259] and played an important role in early response to the Covid19 pandemic[260],[261],[261]. Fundamentally, docking studies work best (i.e. their results are closer to the experimental values) when the target site is well defined, not very flexible and with good solvent accessibility. Ideally, there would be very little induced conformational changes on the target upon ligand binding. However, over the past few years there have been attempts to apply docking to nucleic acid structures as well[262],[263],[259]. This is usually helped by structural characterisation of nucleic acids based on NMR or Cryo-EM which results in an ensemble of proposed structures rather than one crystal structure from crystallographic techniques. The multiple conformations of the solutions allow for greater conformational space to be sampled within the docking experiments and can potentially create more reliable results in cases where the substrate molecule remains close to the local minima identified.

Of the many platforms and software packages available for docking studies, we chose Autodock Vina. It is the most popular software that can take advantage of multi-core processing. Traditionally used with ligands and proteins, it allows for more than 8 rotational bonds (for small molecules) [258]. Although it has been widely and successfully used for organic ligands and protein targets, Vina has also been used to generate the starting point of MD simulations of ligand-Nucleic acid interaction, including charged ligands. The binding affinity is calculated based on a scoring function, which is defined as the weighted average of parameters calculated by machine learning, including hydrogen bonding and local electrostatics. Here, it is only considered as an indication of relative affinity to different structures, since it does not take any consideration of induced affinity and does not allow user-defined partial charges, which play a major role. Overall, owing to their size and exposed aromatic rings, the cylinders rank high on docking affinity in comparison to smaller molecules. Specifically for DNA and RNA structures, it is often the case that the bending radius is very close to the geometry of the cylinder providing a pocket. At the same time, the cylinder often finds a bed between two (parallel or antiparallel) backbone strands.

1.2.6 Concluding remarks, aim and objectives

This chapter lays the foundations of broad-spectrum antiviral drug design. It introduces the biological space of the challenge, with the classification of viruses based on their path to produce mRNA and replicate (Baltimore scheme) and gives a fundamental overview of the theoretical/computational approach to the dynamics of nucleic acids and their interaction with potential drug candidates with a special interest in the design of a coordination compound as a broad-spectrum antiviral agent. Finally, the common theme throughout this work is the transition between metastable states across scales, from the long scale of viral replication to the mesoscale of biomolecular dynamics and even approaching the sub-molecular space.

The aim of this thesis is the use of computational methods to understand the conformational landscape of small nucleic acid fragments often seen in viruses with the goal to elucidate the underpinning mechanisms of targeting them with metal-containing compounds. To that end, in chapter 3, I examine the electronic structure of coordination compounds and prepare them for molecular dynamics simulation in the following chapters. In chapter 4, the use of multi-microsecond long molecular dynamics simulations of RNA fragments is shown to retain experimentally observed characteristics of the structures, when the state-of-the-art RNA forcefields are employed and specifically in the case of TAR-RNA, the binding mechanism with supramolecular cylinders is captured. In chapter 5, the theoretical understanding of the previous chapters is applied to inhibit the replication pathway of a novel coronavirus, providing a computational pipeline for rapid response to novel pathogens. In chapter 6, the same pipeline is applied to DNA structures, namely G-quadruplexes. In this case, the structures have been shown to play a role both in eukaryotic chromatin as well as in virus-host interactions.

1.3 References

[1] J. Malina, M. J. Hannon, V. Brabec, *Chem. - A Eur. J.* **2015**, *21*, 11189–11195.

[2] A. Oleksi, A. G. Blanco, R. Boer, I. Usón, J. Aymamí, A. Rodger, M. J. Hannon, M. Coll, *Angew. Chemie - Int. Ed.* **2006**, *45*, 1227–1231.

[3] A. C. G. Hotze, B. M. Kariuki, M. J. Hannon, *Angew. Chemie - Int. Ed.* **2006**, *45*, 4839–4842.

[4] L. Cardo, I. Nawroth, P. J. Cail, J. A. McKeating, M. J. Hannon, *Sci. Rep.* **2018**, *8*, 13342.

[5] I. Khanna, *Drug Discov. Today* **2012**, *17*, 1088–1102.

[6] H. Zhao, A. Caflisch, *Eur. J. Med. Chem.* **2015**, *91*, 4–14.

[7] Y. C. Chen, *Trends Pharmacol. Sci.* **2015**, *36*, 78–95.

[8] L. Sun, D. Zhang, S. Chen, *Annu. Rev. Biophys* **2017**, *46*, 227–46.

[9] K. Spiegel, A. Magistrato, *Org. Biomol. Chem.* **2006**, *4*, 2507–2517.

[10] L. Zhang, J. Tan, D. Han, H. Zhu, *Drug Discov. Today* **2017**, *00*, 1–6.

[11] V. Botu, R. Ramprasad, *Int. J. Quantum Chem.* **2015**, *115*, 1074–1083.

[12] S. Ekins, A. A. Godbole, G. Kéri, L. Orfi, J. Pato, R. S. Bhat, R. Verma, E. K. Bradley, V. Nagaraja, *Tuberculosis* **2017**, *103*, 52–60.

[13] R. Dahm, *Hum. Genet.* **2008**, *122*, 565–581.

[14] A. D. HERSHEY, M. CHASE, *J. Gen. Physiol.* **1952**, *36*, 39–56.

[15] F. Crick, J. Watson, *Nat*. **1953,***171***(***4356***)** 737-738.

[16] D. R. Whelan, T. J. Hiscox, J. I. Rood, K. R. Bambery, D. McNaughton, B. R. Wood, *J. R. Soc. Interface* **2014**, *11*, 3–8.

[17] J. B. Swadling, K. Ishii, T. Tahara, A. Kitao, *Phys. Chem. Chem. Phys.* **2018**, *20*, 2990–3001.

[18] N. B. Leontis, *Nucleic Acids Res.* **2002**, *30*, 3497–3531.

[19] J. A. Cruz, E. Westhof, *Cell* **2009**, *136*, 604–609.

[20] R. Russell, X. Zhuang, H. P. Babcock, I. S. Millett, S. Doniach, S. Chu, D. Herschlag, *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 155–160.

[21] A. Umuhire Juru, N. N. Patwardhan, A. E. Hargrove, *ACS Chem. Biol.* **2019**, *14*, 824–838.

[22] C. Wang, S. Y. Le, N. Ali, A. Siddiqui, *Rna* **1995**, *1*, 526–537.

[23] D. C. Jeffares, A. M. Poole, D. Penny, *J. Mol. Evol.* **1998**, *46*, 18–36.

[24] D. Svozil, P. Hobza, J. Šponer, *J. Phys. Chem. B* **2010**, *114*, 1191–1203.

[25] F. A. Momany, U. Schnupf, W. M. C. Sameera, F. Maseras, S. K. Kim, J. Schermann, T. Ha, E. B. Cagmat, J. Szczepanski, W. L. Pearson, H. Powell, J. R. Eyler, N. C. Polfer, P. Chem, C. A. Morgado, P. Jure, D. Svozil, P. Hobza, M. Pincu, B. Brauer, R. B. Gerber, S. Bari, R. Hoekstra, T. Schlathölter, B. J. Mccullough, J. M. Kalapothakis, W. Chin, K. Taylor, D. J. Clarke, H. Eastwood, D. Macmillan, J. Dorin, E. Perdita, F. A. Faber, L. Hutchison, B. Huang, J. Gilmer, S. S. Schoenholz, G. E. Dahl, O. Vinyals, S. Kearnes, P. F. Riley, O. A. von Lilienfeld, P. Li, K. M. Merz, Y. Li, H. Li, F. C. Pickard, B. Narayanan, F. G. Sen, M. K. Y. Chan, S. K. R. S. Sankaranarayanan, B. R. Brooks, B. Roux, L. S. Stelzl, G. Hummer, W. Yang, B. T. Riley, X. Lei, B. T. Porebski, I. Kass, A. M. Buckle, S. McGowan, H. Zhang, Y. Jiang, H. Yan, C. Yin, T. Tan, D. Van Der Spoel, F. A. Momany, U. Schnupf, J. L. Willett, U. Schnupf, H. Abdizadeh, A. R. Atilgan, C. Atilgan, B. Dedeoglu, G. M. Hocky, T. Dannenhoffer-Lafage, G. A. Voth, O. Andreussi, I. G. Prandi, M. Campetella, G. Prampolini, B. Mennucci, J. Kreutzer, P. Blaha, U. Schubert, D. W. Watkins, J. M. X. Jenkins, K. J. Grayson, N. Wood, J. W. Steventon, K. K. Le Vay, M. I. Goodwin, A. S. Mullen, H. J. Bailey, M. P. Crump, F. MacMillan, A. J. Mulholland, G. Cameron, R. B. Sessions, S. Mann, J. L. R. Anderson, K. E. Ranaghan, W. G. Morris, L.

Masgrau, K. Senthilkumar, L. O. Johannissen, N. S. Scrutton, J. N. Harvey, F. R. Manby, A. J. Mulholland, C. Barolo, M. K. Nazeeruddin, S. Fantacci, D. Di Censo, P. Comte, P. Liska, G. Viscardi, P. Quagliotto, F. De Angelis, S. Ito, M. Grätzel, *Carbohydr. Res.* **2011**, *346*, 619–630.

[26] B. G. Vértessy, J. Tóth, *Acc Chem Res* **2009**, *42*, 97–106.

[27] H. E. Krokan, F. Drabløs, G. Slupphaug, *Oncogene* **2002**, *21*, 8935–8948.

[28] M. J. Booth, M. R. Branco, G. Ficz, D. Oxley, F. Krueger, W. Reik, S. Balasubramanian, *Science (80).* **2012**, *336*, 934–937.

[29] J. Spiegel, S. Adhikari, S. Balasubramanian, *Trends Chem.* **2020**, *2*, 123–136.

[30] D. Varshney, J. Spiegel, K. Zyner, D. Tannahill, S. Balasubramanian, *Nat. Rev. Mol. Cell Biol.* **2020**, *21*, 459–474.

[31] K. G. Zyner, D. S. Mulhearn, S. Adhikari, S. M. Cuesta, M. Di Antonio, N. Erard, G. J. Hannon, D. Tannahill, S. Balasubramanian, *Elife* **2019**, *8*, 1–40.

[32] K. G. Zyner, A. Simeone, S. M. Flynn, C. Doyle, G. Marsico, S. Adhikari, G. Portella, D. Tannahill, S. Balasubramanian, *Nat. Commun.* **2022**, *13*, 142 DOI 10.1038/s41467-021-27719-1.

[33] D. Liano, S. Chowdhury, M. Di Antonio, *J. Am. Chem. Soc.* **2021**, *143*, 20988–21002.

[34] J. M. Lehn, A. Rigault, J. Siegel, J. Harrowfield, B. Chevrier, D. Moras, *Proc. Natl. Acad. Sci. U. S. A.* **1987**, *84*, 2565–2569.

[35] J. Malina, M. J. Hannon, V. Brabec, *Chem. - A Eur. J.* **2008**, *14*, 10408–10414.

[36] S. Phongtongpasuk, S. Paulus, J. Schnabl, R. K. O. Sigel, B. Spingler, M. J. Hannon, E. Freisinger, *Angew. Chemie - Int. Ed.* **2013**, *52*, 11513–11516.

[37] J. Malina, M. J. Hannon, V. Brabec, *Sci. Rep.* **2016**, *6*, 29674.

[38] E. V. Koonin, M. Krupovic, V. I. Agol, *Microbiol. Mol. Biol. Rev.* **2021**, *85*, DOI 10.1128/mmbr.00053-21.

[39] P. G. Higgs, N. Lehman, *Nat. Rev. Genet.* **2015**, *16*, 7–17.

[40] M. A. Sommerfelt, M. Marsh, *Adv. Drug Deliv. Rev*. **1988,** *4,*1-26

[41] C. T. Ranjith-Kumar, Y.-C. Kim, L. Gutshall, C. Silverman, S. Khandekar, R. T. Sarisky, C. C. Kao, *J. Virol.* **2002**, *76*, 12513–12525.

[42] R. and C. Lamb Purnell, *Ann. Rev. Biochem.* **1983**, 467–506.

[43] S. Barik, A. K. Banerjea, **1992**, *428*, 417–428.

[44] J. H. Strauss, E. G. Strauss, *Microbiol. Rev.* **1994**, *58*, 491–562.

[45] Y. Song, P. Friebe, E. Tzima, C. Jünemann, R. Bartenschlager, M. Niepmann, *J. Virol.* **2006**, *80*, 11579–11588.

[46] M. J. Ross, I. Fidai, J. A. Cowan, *ChemBioChem* **2017**, *18*, 1743–1754.

[47] Y. P. Li, J. M. Gottwein, T. K. Scheel, T. B. Jensen, J. Bukh, *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 4991–4996.

[48] C. L. Jopling, *Biochem. Soc. Trans.* **2008**, *36*, 1220–1223.

[49] J. A. Wilson, S. M. Sagan, *Curr. Opin. Virol.* **2014**, *7*, 11–18.

[50] A. Donlic, A. E. Hargrove, *Wiley Interdiscip. Rev. RNA* **2018**, *9*, 1–21.

[51] J. F. Yang, J. You, *Viruses* **2020**, *12*, 1–18.

[52] M. Charman, C. Herrmann, M. D. Weitzman, *FEBS Lett.* **2019**, *593*, 3531–3550.

[53] N. Mueller, N. van Bel, B. Berkhout, A. T. Das, *Virology* **2014**, *468*, 609–620.

[54] D. F. Purcell, M. A. Martin, *J. Virol.* **1993**, *67*, 6365–6378.

[55] T. S. Su, C. J. Lai, J. L. Huang, L. H. Lin, Y. K. Yauk, C. M. Chang, S. J. Lo, S. H. Han, *J. Virol.* **1989**, *63*, 4011–4018.

[56] G. R. Campbell, R. S. Bruckman, S. D. Herns, S. Joshi, D. L. Durden, S. A. Spector, *J. Biol. Chem.* **2018**, *293*, 5808–5820.

[57] M. Mohammadi-Dehcheshmeh, S. M. Moghbeli, S. Rahimirad, I. O. Alanazi, Z. S. Al Shehri, E. Ebrahimie, *Cells* **2021**, *10 (2)*, DOI 10.3390/cells10020319.

[58] S. Bannwarth, A. Gatignol, *Curr. HIV Res.* **2005**, *3*, 61–71.

[59] T. S. Fung, M. Huang, D. X. Liu, *Virus Res.* **2014**, *194*, 110–123.

[60] J. Hemelaar, *Trends Mol. Med.* **2012**, *18*, 182–192.

[61] E. Butovskaya, B. Heddi, B. Bakalar, S. N. Richter, A. T. Phan, *J. Am. Chem. Soc.* **2018**, *140*, 13654–13662.

[62] M. Stevenson, **2003**, *9*, 853–860.

[63] D. H. Barouch, *Nature* **2008**, *455*, 613–619.

[64] E. J. Arts, D. J. Hazuda, E. F. D. Bushman, G. J. Nabel, R. Swanstrom, *Cold Spring Harb. Perspect. Med.* **2012**, *2*, 1–23.

[65] P. J. Walker, S. G. Siddell, E. J. Lefkowitz, A. R. Mushegian, E. M. Adriaenssens, D. M. Dempsey, B. E. Dutilh, B. Harrach, R. L. Harrison, R. C. Hendrickson, S. Junglen, N. J. Knowles, A. M. Kropinski, M. Krupovic, J. H. Kuhn, M. Nibert, R. J. Orton, L. Rubino, S. Sabanadzovic, P. Simmonds, D. B. Smith, A. Varsani, F. M. Zerbini, A. J. Davison, *Arch. Virol.* **2020**, *165*, 2737–2748.

[66] J. S. Kahn, K. McIntosh, *Pediatr. Infect. Dis. J.* **2005**, *24*, 223–227.

[67] P. V’kovski, A. Kratzel, S. Steiner, H. Stalder, V. Thiel, *Nat. Rev. Microbiol.* **2021**, 19, 155-170

[68] R. A. M. Fouchier, N. G. Hartwig, T. M. Bestebroer, B. Niemeyer, J. C. De Jong, J. H. Simon, A. D. M. E. Osterhaus, *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101*, 6212–6216.

[69] N. Zhang, J. Tang, L. Lu, S. Jiang, L. Du, *Virus Res.* **2015**, *202*, 151–159.

[70] K. Stadler, V. Masignani, M. Eickmann, S. Becker, S. Abrignani, H. D. Klenk, R. Rappuoli, *Nat. Rev. Microbiol.* **2003**, *1*, 209–218.

[71] N. S. Ogando, F. Ferron, E. Decroly, B. Canard, C. C. Posthuma, E. J. Snijder, *Front. Microbiol.* **2019**, *10*, 1–17.

[72] W. A. Cantara, E. D. Olson, K. Musier-Forsyth, *Virus Res.* **2014**, *193*, 24–38.

[73] P. V'kovski, H. Al-Mulla, V. Thiel, B. W. Neuman, *Virus Res.* **2015**, *202*, 33–40.

[74] D. X. Liu, T. S. Fung, K. K. L. Chong, A. Shukla, R. Hilgenfeld, *Antiviral Res.* **2014**, *109*, 97–109.

[75] R. Madhugiri, M. Fricke, M. Marz, J. Ziebuhr, *Virus Res.* **2014**, *194*, 76–89.

[76] C. R. Reid, A. M. Airo, T. C. Hobman, *Viruses* **2015**, *7*, 4385–4413.

[77] B. W. Neuman, P. Chamberlain, F. Bowden, J. Joseph, *Virus Res.* **2014**, *194*, 49–66.

[78] M. Frieman, M. Heise, R. Baric, *Virus Res.* **2008**, *133*, 101–112.

[79] B. Hsue, P. S. Masters, *J. Virol.* **1997**, *71*, 7567–7578.

[80] R. W. Fulton, J. F. Ridpath, L. J. Burge, *Vaccine* **2013**, *31*, 886–892.

[81] S. N. Neerukonda, *Pathogens* **2020**, *9*, 1–22.

[82] V. Ntafis, V. Mari, N. Decaro, M. Papanastassopoulou, D. Pardali, T. S. Rallis, T. Kanellos, C. Buonavoglia, E. Xylouri, *Infect. Genet. Evol.* **2013**, *16*, 129–136.

[83] J. Cui, F. Li, Z. L. Shi, *Nat. Rev. Microbiol.* **2019**, *17*, 181–192.

[84] F. Wu, S. Zhao, B. Yu, Y. M. Chen, W. Wang, Z. G. Song, Y. Hu, Z. W. Tao, J. H. Tian, Y. Y. Pei, M. L. Yuan, Y. L. Zhang, F. H. Dai, Y. Liu, Q. M. Wang, J. J. Zheng, L. Xu, E. C. Holmes, Y. Z. Zhang, *Nature* **2020**, *579*, 265–269.

[85] C. Huang, Y. Wang, X. Li, L. Ren, J. Zhao, Y. Hu, L. Zhang, G. Fan, J. Xu, X. Gu, Z. Cheng, T. Yu, J. Xia, Y. Wei, W. Wu, X. Xie, W. Yin, H. Li, M. Liu, Y. Xiao, H. Gao, L. Guo, J. Xie, G. Wang, R. Jiang, Z. Gao, Q. Jin, J. Wang, B. Cao, *Lancet* **2020**, *395*, 497–506.

[86] N. Zhu, D. Zhang, W. Wang, X. Li, B. Yang, J. Song, X. Zhao, B. Huang, W. Shi, R. Lu, P. Niu, F. Zhan, X. Ma, D. Wang, W. Xu, G. Wu, G. F. Gao, W. Tan, *N. Engl. J. Med.* **2020**, *382*, 727–733.

[87] J. T. Wu, K. Leung, G. M. Leung, *Lancet* **2020**, *395*, 689–697.

[88] R. Lu, X. Zhao, J. Li, P. Niu, B. Yang, H. Wu, W. Wang, H. Song, B. Huang, N. Zhu, Y. Bi, X. Ma, F. Zhan, L. Wang, T. Hu, H. Zhou, Z. Hu, W. Zhou, L. Zhao, J. Chen, Y. Meng, J. Wang, Y. Lin, J. Yuan, Z. Xie, J. Ma, W. J. Liu, D. Wang, W. Xu, E. C. Holmes, G. F. Gao, G. Wu, W. Chen, W. Shi, W. Tan, *Lancet* **2020**, *395*, 565–574.

[89] D. Benvenuto, M. Giovanetti, M. Salemi, M. Prosperi, C. De Flora, L. C. Junior Alcantara, S. Angeletti, M. Ciccozzi, *Pathog. Glob. Health* **2020**, *114*, 64–67.

[90] J. A. Jaimes, N. M. André, J. S. Chappie, J. K. Millet, **2020,** 432, 3309-3325.

[91] A. Grifoni, J. Sidney, Y. Zhang, R. H. Scheuermann, B. Peters, A. Sette, *Cell Host Microbe* **2020**, *27*, 671-680.e2.

[92] W. Tai, L. He, X. Zhang, J. Pu, D. Voronin, S. Jiang, Y. Zhou, L. Du, *Cell. Mol. Immunol.* **2020**, *17*, 613–620.

[93] M. Hoffmann, H. Kleine-Weber, S. Schroeder, N. Krüger, T. Herrler, S. Erichsen, T. S. Schiergens, G. Herrler, N. H. Wu, A. Nitsche, M. A. Müller, C. Drosten, S. Pöhlmann, *Cell* **2020**, *181*, 271-280.e8.

[94] J. Koch, Z. M. Uckeley, P. Doldan, M. Stanifer, S. Boulant, P. Lozach, *EMBO J.* **2021**, *40*, 1–20.

[95] G. R. Whittaker, S. Daniel, J. K. Millet, *Curr. Opin. Virol.* **2021**, *47*, 113–120.

[96] N. Murgolo, A. G. Therien, B. Howell, D. Klein, K. Koeplinger, L. A. Lieberman, G. C. Adam, J. Flynn, P. McKenna, G. Swaminathan, D. J. Hazuda, D. B. Olsen, *PLoS Pathog.* **2021**, *17*, 1–18.

[97] I. Sola, F. Almazán, S. Zúñiga, L. Enjuanes, *Annu. Rev. Virol.* **2015**, *2*, 265–288.

[98] Y. Finkel, O. Mizrahi, A. Nachshon, S. Weingarten-Gabbay, D. Morgenstern, Y. Yahalom-Ronen, H. Tamir, H. Achdout, D. Stein, O. Israeli, A. Beth-Din, S. Melamed, S. Weiss, T. Israely, N. Paran, M. Schwartz, N. Stern-Ginossar, *Nature* **2021**, *589*, 125–130.

[99] C. J. Michel, C. Mayer, O. Poch, J. D. Thompson, *bioRxiv* **2020**, 1–13.

[100] O. Ziv, J. Price, L. Shalamova, T. Kamenova, I. Goodfellow, F. Weber, E. A. Miska, *Mol. Cell* **2020**, *80*, 1067-1077.e5.

[101] I. Manfredonia, C. Nithin, A. Ponce-Salvatierra, P. Ghosh, T. K. Wirecki, T. Marinus, N. S. Ogando, E. J. Snijder, M. J. van Hemert, J. M. Bujnicki, D. Incarnato, *Nucleic Acids Res.* **2020**, *48*, 12436–12452.

[102] P. R. Bhatt, A. Scaiola, G. Loughran, M. Leibundgut, A. Kratzel, R. Meurs, R. Dreos, K. M. O'Connor, A. McMillan, J. W. Bode, V. Thiel, D. Gatfield, J. F. Atkins, N. Ban, *Science (80-.).* **2021**, *372*, 1306–1313.

[103] C. Roman, A. Lewicka, D. Koirala, N.-S. Li, J. A. Piccirilli, *ACS Chem. Biol.* **2021**, *16*, 1469–1481.

[104] Y. Sun, L. Abriola, R. O. Niederer, S. F. Pedersen, M. M. Alfajaro, V. S. Monteiro, C. B. Wilen, Y. C. Ho, W. V. Gilbert, Y. V. Surovtseva, B. D. Lindenbach, J. U. Guo, *Proc. Natl. Acad. Sci. U. S. A.* **2021**, *118*, (26) e2023051118.

[105] H. L. Yen, *Curr. Opin. Virol.* **2016**, *18*, 126–134.

[106] J.-G. Park, G. Ávila-Pérez, A. Nogales, P. Blanco-Lobo, J. C. de la Torre, L. Martínez-Sobrido, *J. Virol.* **2020**, *94*, DOI 10.1128/jvi.02149-19.

[107] M. Akram, I. M. Tahir, S. M. A. Shah, Z. Mahmood, A. Altaf, K. Ahmad, N. Munir, M. Daniyal, S. Nasir, H. Mehboob, *Phyther. Res.* **2018**, *32*, 811–822.

[108] A. Shannon, N. T. T. Le, B. Selisko, C. Eydoux, K. Alvarez, J. C. Guillemot, E. Decroly, O. Peersen, F. Ferron, B. Canard, *Antiviral Res.* **2020**, *178*, 104793.

[109] A. M. Peters Van Ton, T. J. G. Gevers, J. P. H. Drenth, *J. Viral Hepat.* **2015**, *22*, 965–973.

[110] T. G. Villa, L. Feijoo-Siota, J. L. R. Rama, J. M. Ageitos, *Biochem. Pharmacol.* **2017**, *133*, 97–116.

[111] E. Mahase, *Bmj* **2021**, 375n2697.

[112] I. Usach, V. Melis, J.-E. Peris, *J. Int. AIDS Soc.* **2013**, *16*, 1–14.

[113] M. Vanangamudi, S. Kurup, V. Namasivayam, *Curr. Opin. Pharmacol.* **2020**, *54*, 179–187.

[114] K. L. McKnight, B. A. Heinz, *Antivir. Chem. Chemother.* **2003**, *14*, 61–73.

[115] C. Sherpa, J. W. Rausch, S. F. J. Le Grice, M. L. Hammarskjold, D. Rekosh, *Nucleic Acids Res.* **2015**, *43*, 4676–4686.

[116] J. Da Zhu, W. Meng, X. J. Wang, H. C. R. Wang, *Front. Microbiol.* **2015**, *6*, 1–15.

[117] M. C. Wolf, A. N. Freiberg, T. Zhang, Z. Akyol-Ataman, A. Grock, P. W. Hong, J. Li, N. F. Watson, A. Q. Fang, H. C. Aguilar, M. Porotto, A. N. Honko, R. Damoiseaux, J. P. Miller, S. E. Woodson, S. Chantasirivisal, V. Fontanes, O. A. Negrete, P. Krogstad, A. Dasgupta, A. Moscona, L. E. Hensley, S. P. Whelan, K. F. Faull, M. R. Holbrook, M. E. Jung, B. Lee, *Proc. Natl. Acad. Sci.* **2010**, *107*, 3157–3162.

[118] F. Vigant, J. Lee, A. Hollmann, L. B. Tanner, Z. Akyol Ataman, T. Yun, G. Shui, H. C. Aguilar, D. Zhang, D. Meriwether, G. Roman-Sosa, L. R. Robinson, T. L. Juelich, H. Buczkowski, S. Chou, M. A. R. B. Castanho, M. C. Wolf, J. K. Smith, A. Banyard, M. Kielian, S. Reddy, M. R. Wenk, M. Selke, N. C. Santos, A. N. Freiberg, M. E. Jung, B. Lee, *PLoS Pathog.* **2013**, *9*, DOI 10.1371/journal.ppat.1003297.

[119] L. Sun, A. Meijer, M. Froeyen, L. Zhang, H. J. Thibaut, J. Baggen, S. George, J. Vernachio, F. J. M. Van Kuppeveld, P. Leyssen, R. Hilgenfeld, J. Neyts, L. Delang, *Antimicrob. Agents Chemother.* **2015**, *59*, 7782–7785.

[120] Y. Wang, D. Zhang, G. Du, R. Du, J. Zhao, Y. Jin, S. Fu, L. Gao, Z. Cheng, Q. Lu, Y. Hu, G. Luo, K. Wang, Y. Lu, H. Li, S. Wang, S. Ruan, C. Yang, C. Mei, Y. Wang, D. Ding, F. Wu, X. Tang, X. Ye, Y. Ye, B. Liu, J. Yang, W. Yin, A. Wang, G. Fan, F. Zhou, Z. Liu, X. Gu, J. Xu, L. Shang, Y. Zhang, L. Cao, T. Guo, Y. Wan, H. Qin, Y. Jiang, T. Jaki, F. G. Hayden, P. W. Horby, B. Cao, C. Wang, *Lancet* **2020**, *395*, 1569–1578.

[121] D. W. Hutchinson, *Antiviral Res.* **1985**, *5*, 193–205.

[122] M. Carcelli, E. Fisicaro, C. Compari, L. Contardi, D. Rogolino, C. Solinas, A. Stevaert, L. Naesens, *Polyhedron* **2017**, *129*, 97–104.

[123] E. De Clercq, *Met. Based. Drugs* **1997**, *4*, 173–192.

[124] N. Graf, S. J. Lippard, *Adv. Drug Deliv. Rev.* **2012**, *64*, 993–1004.

[125] J. L. García-Giménez, J. Hernández-Gil, A. Martínez-Ruíz, A. Castiñeiras, M. Liu-González, F. V. Pallardó, J. Borrás, G. Alzuet Piña, *J. Inorg. Biochem.* **2013**, *121*, 167–178.

[126] M. Carcelli, A. Bacchi, P. Pelagatti, G. Rispoli, D. Rogolino, T. W. Sanchez, M. Sechi, N. Neamati, *J. Inorg. Biochem.* **2013**, *118*, 74–82.

[127] D. P. Buck, C. B. Spillane, J. G. Collins, F. R. Keene, *Mol. Biosyst.* **2008**, *4*, 851–854.

[128] R. W. Y. Sun, D. L. Ma, E. L. M. Wong, C. M. Che, *Dalt. Trans.* **2007**, 4884–4892.

[129] E. L. Chang, C. Simmers, D. A. Knight, *Pharmaceuticals* **2010**, *3*, 1711–1728.

[130] T. Takeuchi, A. Böttcher, C. M. Quezada, T. J. Meade, H. B. Gray, *Bioorganic Med. Chem.* **1999**, *7*, 815–819.

[131] J. B. Delehanty, J. E. Bongard, D. C. Thach, D. A. Knight, T. E. Hickey, E. L. Chang, *Bioorganic Med. Chem.* **2008**, *16*, 830–837.

[132] M. F. Hawthorne, T. D. Andrews, *J. Am. Chem. Soc.* **1965**, *87*, 2496.

[133] P. Cigler, M. Ko i ek, P. eza ova, J. Brynda, Z. Otwinowski, J. Pokorna, J. Ple ek, B. Gruner, L. Dole kova-Mare ova, M. Ma a, J. Sedla ek, J. Bodem, H.-G. Krausslich, V. Kral, J. Konvalinka, *Proc. Natl. Acad. Sci.* **2005**, *102*, 15394–15399.

[134] D. Andrew Knight, T. E. Hickey, J. E. Bongard, D. C. Thach, R. Yngard, E. L. Chang, *J. Inorg. Biochem.* **2010**, *104*, 592–598.

[135] S. Yuan, R. Wang, J. F. W. Chan, A. J. Zhang, T. Cheng, K. K. H. Chik, Z. W. Ye, S. Wang, A. C. Y. Lee, L. Jin, H. Li, D. Y. Jin, K. Y. Yuen, H. Sun, *Nat. Microbiol.* **2020**, *5*, 1439–1448.

[136] R. E. F. De Paiva, A. Marçal Neto, I. A. Santos, A. C. G. Jardim, P. P. Corbi, F. R. G. Bergamini, *Dalt. Trans.* **2020**, *49*, 16004–16033.

[137] D. Cirri, A. Pratesi, T. Marzo, L. Messori, *Expert Opin. Drug Discov.* **2021**, *16*, 39–46.

[138] C. Chuong, C. M. Duchane, E. M. Webb, P. Rai, J. M. Marano, C. M. Bernier, J. S. Merola, J. Weger-Lucarelli, *Viruses* **2021**, *13(6)*,980

[139] B. Balasubramaniam, Prateek, S. Ranjan, M. Saraf, P. Kar, S. P. Singh, V. K. Thakur, A. Singh, R. K. Gupta, *ACS Pharmacol. Transl. Sci.* **2021**, *4*, 8–54.

[140] S. W. Jaros, J. Król, B. Bazanów, D. Poradowski, A. Chrószcz, D. S. Nesterov, A. M. Kirillov, P. Smoleński, *Molecules* **2020**, *25*, 1–13.

[141] S. P. Velagapudi, M. G. Costales, B. R. Vummidi, Y. Nakai, A. J. Angelbello, T. Tran, H. S.

Haniff, Y. Matsumoto, Z. F. Wang, A. K. Chatterjee, J. L. Childs-Disney, M. D. Disney, *Cell Chem. Biol.* **2018**, *25*, 1086-1094.e7.

[142] J. T. Hua, S. Chen, H. H. He, *Trends Genet.* **2019**, *35*, 840–851.

[143] S. P. Velagapudi, M. D. Cameron, C. L. Haga, L. H. Rosenberg, M. Lafitte, D. R. Duckett, D. G. Phinney, M. D. Disney, *Proc. Natl. Acad. Sci.* **2016**, *113*, 5898–5903.

[144] M. G. Costales, J. L. Childs-Disney, H. S. Haniff, M. D. Disney, *J. Med. Chem.* **2020**, *63*, 8880–8900.

[145] B. M. Suresh, W. Li, P. Zhang, K. W. Wang, I. Yildirim, C. G. Parker, M. D. Disney, *Proc. Natl. Acad. Sci. U. S. A.* **2020**, *117*, 33197–33203.

[146] S. M. Meyer, C. C. Williams, Y. Akahori, T. Tanaka, H. Aikawa, Y. Tong, J. L. Childs-Disney, M. D. Disney, *Chem. Soc. Rev.* **2020**, *49*, 7167–7199.

[147] S. Vezina-Dawod, A. J. Angelbello, S. Choudhary, K. W. Wang, I. Yildirim, M. D. Disney, *ACS Med. Chem. Lett.* **2021**, *12*, 907–914.

[148] C. K. Kwok, S. Balasubramanian, *Angew. Chemie - Int. Ed.* **2015**, *54*, 6751–6754.

[149] G. Biffi, M. Di Antonio, D. Tannahill, S. Balasubramanian, *Nat. Chem.* **2014**, *6*, 75–80.

[150] R. C. Monsen, J. O. Trent, *Biochimie* **2018**, *152*, 134–148.

[151] A. M. Fleming, Y. Ding, A. Alenko, C. J. Burrows, *ACS Infect. Dis.* **2016**, *2*, 674–681.

[152] D. Ji, M. Juhas, C. M. Tsang, C. K. Kwok, Y. Li, Y. Zhang, *Brief. Bioinform.* **2021**, *22*, 1150–1160.

[153] C. Zhao, G. Qin, J. Niu, Z. Wang, C. Wang, J. Ren, X. Qu, *Angew. Chemie - Int. Ed.* **2021**, *60*, 432–438.

[154] C. M. Connelly, M. H. Moon, J. S. Schneekloth, *Cell Chem. Biol.* **2016**, *23*, 1077–1090.

[155] M. J. Hangauer, I. W. Vaughn, M. T. McManus, *PLoS Genet.* **2013**, *9*, (6), e1003569.

[156] R. Welty, K. B. Hall, *J. Mol. Biol.* **2016**, *428*, 4490–4502.

[157] M. E. Dinger, K. C. Pang, T. R. Mercer, M. L. Crowe, S. M. Grimmond, J. S. Mattick, *Nucleic Acids Res.* **2009**, *37*, 122–126.

[158] E. H. Blackburn, C. W. Greider, J. W. Szostak, *Nat. Med.* **2006**, *12*, 1133–1138.

[159] J. Winter, S. Diederichs, *Methods Mol. Biol.* **2011**, *676*, 3–22.

[160] S. Wadsworth, R. J. Jacob, B. Roizman, *J. Virol.* **1975**, *15*, 1487–1497.

[161] M. Bon, G. Vernizzi, H. Orland, A. Zee, *J. Mol. Biol.* **2008**, *379*, 900–911.

[162] W. Kohn, L. J. Sham, *Phys. Rev.* **1965**, *140*, A1133-A1138.

[163] and M. J. Sergio Filipe Sousa, Pedro Alexandrino Fernandes, Ramos, *J.Phys.chem A* **2007**, *111*, 10439–10452.

[164] L. Goerigk, S. Grimme, *J. Chem. Theory Comput.* **2011**, *7*, 291–309.

[165] N. Mardirossian, M. Head-Gordon, *J. Chem. Phys.* **2018**, *148*, 241736-1-241736-14.

[166] M. Korth, *Angew. Chemie - Int. Ed.* **2017**, *56*, 5396–5398.

[167] Y. Yang, O. A. Jiménez-Negrón, J. R. Kitchin, *J. Chem. Phys.* **2021**, *154*, 234704-1-9

[168] F. Noé, A. Tkatchenko, K. R. Müller, C. Clementi, *Annu. Rev. Phys. Chem.* **2020**, *71*, 361–390.

[169] B. T. Afflerbach, L. Schultz, J. H. Perepezko, P. M. Voyles, I. Szlufarska, D. Morgan, *Comput. Mater. Sci.* **2021**, *199*, 110728.

[170] A. D. Becke, *J. Chem. Phys.* **1993**, *98*, 5648–5652.

[171] J. Tirado-Rives, W. L. Jorgensen, *J. Chem. Theory Comput.* **2008**, *4*, 297–306.

[172] T. Yanai, D. P. Tew, N. C. Handy, *Chem. Phys. Lett.* **2004**, *393*, 51–57.

[173] Y. Zhao, D. G. Truhlar, *Theor. Chem. Acc.* **2008**, *120*, 215–241.

[174] Y. Wang, X. Jin, H. S. Yu, D. G. Truhlar, X. He, *Proc. Natl. Acad. Sci.* **2017**, *114*, 8487–8492.

[175] N. Mardirossian, M. Head-Gordon, *J. Chem. Theory Comput.* **2016**, *12*, 4303–4325.

[176] C. J. Cramer, D. G. Truhlar, *Phys. Chem. Chem. Phys.* **2009**, *11*, 10757.

[177] J. R. Hammond, *Computing* **2009**, 1–24.

[178] M. Swart, M. Solà, F. M. Bickelhaupt, *J. Chem. Phys.* **2009**, *131*, 094103-1-9

[179] M. Swart, A. W. Ehlers, K. Lammertsma, *Mol. Phys.* **2004**, *102*, 2467–2474.

[180] S. Grimme, *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2011**, *1*, 211–228.

[181] S. Grimme, M. Steinmetz, *Phys. Chem. Chem. Phys.* **2013**, *15*, 16031–16042.

[182] S. Grimme, J. Antony, S. Ehrlich, H. Krieg, *J. Chem. Phys.* **2010**, *132*, 154104-1-154104-19

[183] A. V Marenich, C. J. Cramer, D. G. Truhlar, *J. Phys. Chem. B.* **2009**, *113*, 6378–6396.

[184] A. J. Cohen, P. Mori-Sánchez, W. Yang, *Chem. Rev.* **2012**, *112*, 289–320.

[185] M. Swart, M. Gruden, *Acc. Chem. Res.* **2016**, *49*, 2690–2697.

[186] S. Mai, F. Plasser, J. Dorn, M. Fumanal, C. Daniel, L. González, *Coord. Chem. Rev.* **2018**, *361*, 74–97.

[187] P. Verma, D. G. Truhlar, *Trends Chem.* **2020**, *2*, 302–318.

[188] R. P. Hertzberg, A. J. Pope, *Curr. Opin. Chem. Biol.* **2000**, *4*, 445–451.

[189] K. H. Bleicher, H. J. Böhm, K. Müller, A. I. Alanine, *Nat. Rev. Drug Discov.* **2003**, *2*, 369–378.

[190] N. Schneider, D. M. Lowe, R. A. Sayle, M. A. Tarselli, G. A. Landrum, *J. Med. Chem.* **2016**,

59, 4385–4402.

[191] R. Ramakrishnan, P. O. Dral, M. Rupp, O. A. Von Lilienfeld, *J. Chem. Theory Comput.* **2015**, *11*, 2087–2096.

[192] M. Karplus, G. a Petsko, *Nature* **1990**, *347*, 631–639.

[193] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, P. A. Kollman, *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.

[194] H. Rosemeyer, F. Seela, S. M. Freier, B. J. Burger, D. Alkema, T. Neilson, D. H. Turner, D. H. Turner, N. Sugimoto, R. Kierzek, S. D. Dreiker, P. J. Romaniuk, D. W. Hughes, R. J. Gregoire, T. Neilson, R. A. Bell, N. Sugimoto, R. Kierzek, D. H. Turner, S. M. Freier, D. Alkema, A. Sinclair, T. Neilson, T. H. Turner, M. Senior, R. A. Jones, K. J. Breslauer, E. T. Kool, S. Bommarito, N. Peyret, J. J. SantaLucia, T. Sugiyama, E. Schweinberger, Z. Kazimierczuk, N. Ramzaeva, H. Rosemeyer, F. Seela, M. J. Doktycz, T. M. Paner, M. Amaratunga, A. S. Benight, H. Grosjean, D. G. Söll, D. M. Crothers, S. Ball, M. A. Reeve, P. S. Robinson, F. Hill, D. M. Brown, D. Loakes, J. C. Williams, S. C. Case-Green, K. U. Mir, E. M. Southern, Y. Benenson, T. Paz-Elizur, R. Adar, E. Keinan, Z. Livneh, E. Shapiro, A. J. Gutierrez, T. J. Terhorst, M. D. Matteucci, B. C. Froehler, K. M. Guckian, B. A. Schweitzer, R. X.-F. Ren, C. J. Sheils, P. L. Paris, D. C. Tahmassebi, E. T. Kool, L. E. Xodo, G. Manzini, F. Quadrifoglio, G. A. van der Marel, J. H. van Boom, T. J. Povsic, P. B. Dervan, Y. S. Sanghvi, G. D. Hoke, S. M. Freier, M. C. Zounes, C. Gonzales, L. Cummins, H. Sasmor, P. D. Cook, S. Wang, E. T. Kool, B. C. Froehler, S. Wadwani, T. J. Terhorst, S. R. Gerrard, J. Sagi, A. Szemzo, K. Ebinger, A. Szabolcs, G. Sagi, E. Ruff, L. Ötvös, L. C. Sowers, B. R. Shaw, W. D. Sedwick, K. M. Guckian, B. A. Schweitzer, R. X.-F. Ren, C. J. Sheils, D. C. Tahmassebi, E. T. Kool, F. Seela, M. Zulauf, N. Ramzaeva, F. Seela, F. Seela, G. Becher, F. Seela, H. Berg, H. Rosemeyer, N. Ramzaeva, C. Mittelbach, F. Seela, N. Ramzaeva, F. Seela, J. A. McDowell, D. H. Turner, F. Seela, R. Kröschel, Y. He, F. Seela, S. Lampe, *J. Chem. Soc. Perkin Trans. 2* **2002**, *22*, 746–750.

[195] S. Gowtham, R. H. Scheicher, R. Ahuja, R. Pandey, S. P. Karna, *Phys. Rev. B - Condens.*

Matter Mater. Phys. **2007**, *76*, 2–5.

[196] R. Feynman, R. Leighton, M. Sands, **1964**., Feynman Lectures on Physics

[197] J. R. Perilla, B. C. Goh, C. K. Cassidy, B. Liu, R. C. Bernardi, T. Rudack, H. Yu, Z. Wu, K. Schulten, *Curr. Opin. Struct. Biol.* **2015**, *31*, 64–74.

[198] A. Cooper, *Prog. Biophys. Mol. Biol.* **1984**, *44*, 181–214.

[199] J. N. Onuchic, Z. LutheySchulten, P. G. Wolynes, Z. Luthey-Schulten, P. G. Wolynes, *Annu. Rev. Phys. Chem.* **1997**, *48*, 545–600.

[200] K. A. Dill, H. S. Chan, *Nat. Struct. Biol.* **1997**, *4*, 10–19.

[201] S. J. Chen, K. a Dill, *Proc. Natl. Acad. Sci. U. S. A.* **2000**, *97*, 646–51.

[202] V. S. Pande, A. Y. Grosberg, T. Tanaka, D. S. Rokhsar, *Curr. Opin. Struct. Biol.* **1998**, *8*, 68–79.

[203] Y. Sugita, Y. Okamoto, *Chem. Phys. Lett.* **1999**, *314*, 141–151.

[204] D. R. Roe, C. Bergonzo, T. E. Cheatham, *J. Phys. Chem. B* **2014**, *118*, 3543–3552.

[205] A. Laio, M. Parrinello, *Proc. Natl. Acad. Sci. U. S. A.,* **2002**, *99 (20)*, 12562-12566.

[206] E. Darve, A. Pohorille, *J. Chem. Phys.* **2001**, *115*, 9169–9183.

[207] C. Tsallis, D. A. Stariolo, *Phys. A Stat. Mech. its Appl.* **1996**, *233*, 395–406.

[208] H. Zang, S. Zhang, K. Hapeshi, *J. Bionic Eng.* **2010**, *7*, S232–S237.

[209] Y. Miao, V. A. Feher, J. A. McCammon, *J. Chem. Theory Comput.* **2015**, *11*, 3584–3595.

[210] Y. Miao, A. Bhattarai, J. Wang, *J. Chem. Theory Comput.* **2020**, *16*, 5526–5547.

[211] R. Marsland, J. England, *Reports Prog. Phys.* **2018**, *81*, DOI 10.1088/1361-6633/aa9101.

[212] P. Pietzonka, A. C. Barato, U. Seifert, **2016**, DOI 10.1088/1742-5468/2016/12/124004.

[213] R. O. Dror, R. M. Dirks, J. P. Grossman, H. Xu, D. E. Shaw, *Annu. Rev. Biophys.* **2012**, *41*, 429–452.

[214] P. S. Nerenberg, T. Head-Gordon, *Curr. Opin. Struct. Biol.* **2018**, *49*, 129–138.

[215] A. H. Aytenfisu, A. Spasic, A. Grossfield, H. A. Stern, D. H. Mathews, *J. Chem. Theory Comput.* **2017**, *13*, 900–915.

[216] D. Tan, S. Piana, R. M. Dirks, D. E. Shaw, *Proc. Natl. Acad. Sci. U. S. A.* **2018**, *115*, E1346–E1355.

[217] J. Šponer, G. Bussi, M. Krepl, P. Banáš, S. Bottaro, R. A. Cunha, A. Gil-Ley, G. Pinamonti, S. Poblete, P. Jurečka, N. G. Walter, M. Otyepka, *Chem. Rev.* **2018**, *118,8*,4177-4338.

[218] K. Mráziková, V. Mlýnský, P. Kührová, P. Pokorná, H. Kruse, M. Krepl, M. Otyepka, P. Banáš, J. Šponer, *J. Chem. Theory Comput.* **2020**, *16*, 7601–7617.

[219] J. Šponer, M. Krepl, P. Banáš, P. Kührová, M. Zgarbová, P. Jurečka, M. Havrila, M. Otyepka, *Wiley Interdiscip. Rev. RNA* **2017**, *8*, 1–17.

[220] T. E. Cheatham, *Curr. Opin. Struct. Biol.* **2004**, *14*, 360–367.

[221] A. Y. L. Sim, P. Minary, M. Levitt, *Curr. Opin. Struct. Biol.* **2012**, *22*, 273–278.

[222] J. J. Uusitalo, H. I. Ingólfsson, S. J. Marrink, I. Faustino, *Biophys. J.* **2017**, *113*, 246–256.

[223] J. A. Joseph, A. Reinhardt, A. Aguirre, P. Y. Chew, K. O. Russell, J. R. Espinosa, A. Garaizar, R. Collepardo-Guevara, *Nat. Comput. Sci.* **2021**, *1*, 732–743.

[224] C.-T. Lai, G. C. Schatz, *J. Phys. Chem. B* **2018**, 122,33,7990-7996

[225] J. Šponer, A. Mládek, J. E. Šponer, D. Svozil, M. Zgarbová, P. Banáš, P. Jurečka, M. Otyepka, *Phys. Chem. Chem. Phys.* **2012**, *14*, 15257–15277.

[226] M. Zgarbová, J. Šponer, M. Otyepka, T. E. Cheatham, R. Galindo-Murillo, P. Jurečka, *J.*

Chem. Theory Comput. **2015**, *11*, 5723–5736.

[227] M. Zgarbová, P. Jurečka, J. Šponer, M. Otyepka, *J. Chem. Theory Comput.* **2018**, *14*, 319–328.

[228] S. Burge, G. N. Parkinson, P. Hazel, A. K. Todd, S. Neidle, *Nucleic Acids Res.* **2006**, *34*, 5402–5415.

[229] M. Kogut, C. Kleist, J. Czub, *Nucleic Acids Res.* **2016**, *44*, 3020–3030.

[230] T. Sun, A. Mirzoev, N. Korolev, P. Alexander, L. Nordenskiöld, *Nucleic Acids Res* **2019,** 47,11 5550-5562

[231] L.-Z. Sun, D. Zhang, S.-J. Chen, *Annu. Rev. Biophys.* **2017**, *46*, 227–246.

[232] T. Sun, A. Mirzoev, N. Korolev, A. P. Lyubartsev, L. Nordenskiöld, *J. Phys. Chem. B* **2017**, *121*, 7761–7770.

[233] R. Galindo-Murillo, T. E. Cheatham, *Nucleic Acids Res.* **2021**, *49*, 3735–3747.

[234] A. Savelyev, A. D. MacKerell, *J. Comput. Chem.* **2014**, *35*, 1219–1239.

[235] K. Gkionis, H. Kruse, J. A. Platts, A. Mládek, J. Koča, J. Šponer, *J. Chem. Theory Comput.* **2014**, *10*, 1326–1340.

[236] J. Song, C. Ji, J. Z. H. Zhang, *Phys. Chem. Chem. Phys.* **2013**, *15*, 3846–3854.

[237] J. Fallmann, S. Will, J. Engelhardt, B. Grüning, R. Backofen, P. F. Stadler, *J. Biotechnol.* **2017**, *261*, 97–104.

[238] D. H. Mathews, D. H. Turner, *Curr. Opin. Struct. Biol.* **2006**, *16*, 270–278.

[239] S. H. Bernhart, H. Tafer, U. Mückstein, C. Flamm, P. F. Stadler, I. L. Hofacker, *Algorithms Mol. Biol.* **2006**, *1*, 1–10.

[240] M. F. Sloma, D. H. Mathews, *PLoS Comput. Biol.* **2017**, *13*, 1–23.

[241] S. Vangaveti, S. V. Ranganathan, A. A. Chen, *Wiley Interdiscip. Rev. RNA* **2017**, *8*,2, e1396.

[242] M. T. Panteva, G. M. Giambaşu, D. M. York, *J. Comput. Chem.* **2015**, *36*, 970–82.

[243] Z. Miao, R. W. Adamiak, M. Antczak, R. T. Batey, A. J. Becka, M. Biesiada, M. J. Boniecki, J. M. Bujnicki, S. J. Chen, C. Y. Cheng, F. C. Chou, A. R. Ferré-D'Amaré, R. Das, W. K. Dawson, F. Ding, N. V. Dokholyan, S. Dunin-Horkawicz, C. Geniesse, K. Kappel, W. Kladwang, A. Krokhotin, G. E. Łach, F. Major, T. H. Mann, M. Magnus, K. Pachulska-Wieczorek, D. J. Patel, J. A. Piccirilli, M. Popenda, K. J. Purzycka, A. Ren, G. M. Rice, J. Santalucia, J. Sarzynska, M. Szachniuk, A. Tandon, J. J. Trausch, S. Tian, J. Wang, K. M. Weeks, B. Williams, Y. Xiao, X. Xu, D. Zhang, T. Zok, E. Westhof, *Rna* **2017**, *23*, 655–672.

[244] T. Yu, Y. Zhu, Z. He, S.-J. Chen, *J. Phys. Chem. B* **2016**, *120*, 8837–8844.

[245] D. Tan, S. Piana, R. M. Dirks, D. E. Shaw, *Proc. Natl. Acad. Sci.* **2018**, *115(7)*,E1346-E1355.

[246] F. Noé, C. Clementi, *J. Chem. Theory Comput.* **2015**, *11*, 5002–5011.

[247] B. Reuter, M. Weber, K. Fackeldey, S. Röblitz, M. E. Garcia, *J. Chem. Theory Comput.* **2018**, *14*, 3579–3594.

[248] M. K. Scherer, B. Trendelkamp-Schroer, F. Paul, G. Pérez-Hernández, M. Hoffmann, N. Plattner, C. Wehmeyer, J. H. Prinz, F. Noé, *J. Chem. Theory Comput.* **2015**, *11*, 5525–5542.

[249] Y. Meng, C. Gao, D. K. Clawson, S. Atwell, M. Russell, M. Vieth, B. Roux, *J. Chem. Theory Comput.* **2018**, *14*, 2721–2732.

[250] B. E. Husic, V. S. Pande, *J. Am. Chem. Soc.* **2018**, *140*, 2386–2396.

[251] N. S. Pagadala, K. Syed, J. Tuszynski, *Biophys. Rev.* **2017**, *9*, 91–102.

[252] X. Tao, Y. Huang, C. Wang, F. Chen, L. Yang, L. Ling, Z. Che, X. Chen, *Int. J. Food Sci. Technol.* **2020**, *55*, 33–45.

[253] K. K. Chaudhary, N. Mishra, *JSM Chem* **2016**, *4*, 1029.

[254] A. A. Adeniyi, M. E. S. Soliman, *Drug Discov. Today* **2017**, *22*, 1216–1223.

[255] P. Asadi, G. Khodarahmi, H. Farrokhpour, F. Hassanzadeh, L. Saghaei, *Res. Pharm. Sci.* **2017**, *12*, 233–240.

[256] C. N. Cavasotto, M. G. Aucar, *Front. Chem.* **2020**, *8*, 1–10.

[257] P. Śledź, A. Caflisch, *Curr. Opin. Struct. Biol.* **2018**, *48*, 93–102.

[258] M. M. Jaghoori, B. Bleijlevens, S. D. Olabarriaga, *J. Comput. Aided. Mol. Des.* **2016**, *30*, 237–249.

[259] D. Morgnanesi, E. J. Heinrichs, A. R. Mele, S. Wilkinson, S. Zhou, J. L. Kulp, *Antiviral Res.* **2015**, *123*, 204–215.

[260] M. Nand, P. Maiti, T. Joshi, S. Chandra, V. Pande, J. C. Kuniyal, M. A. Ramakrishnan, *Sci. Rep.* **2020**, *10*, 1–12.

[261] M. Tsuji, *FEBS Open Bio* **2020**, *10*, 995–1004.

[262] L. R. Ganser, J. Lee, A. Rangadurai, D. K. Merriman, M. L. Kelly, A. D. Kansal, B. Sathyamoorthy, H. M. Al-Hashimi, *Nat. Struct. Mol. Biol.* **2018**, *25*, 425–434.

[263] D. M. Krüger, J. Bergs, S. Kazemi, H. Gohlke, *ACS Med. Chem. Lett.* **2011**, *2*, 489–493.

Chapter 2

Methods

1.1. Computational resources

All computational resources for this projects were provided by the University of Birmingham, through Birmingham Environment for Academic Research (BEAR) with additional resources obtained through Compute and Storage for Life Sciences (CaStLeS)[1] which include access to POWER9 nodes (144 logical threads with 4 GPUs each), CaStLeS GPU node, and a virtual machine (18 CPUs), along with priority access to blueBear Nodes. Overall, there have been 4 project accounts; edmondac-power9-testing (Power9 nodes), hannonmj-supramolecular-cylinders (castles resources), stylesib-cylinders (blueBEAR), stylesib-zinc-cylinders (blueBEAR) that used 393628, 115994, 9854178 and 1087 CPU hours respectively. The approximate environmental impact of 10363800 CPU hours, based in the United Kingdom, has a carbon footprint of 124.54 T CO2e, which is equivalent to 135860.96 tree-months calculated usinggreen-algorithms.org v2.1[2] (711,600km in a passenger car).

2.1.1 Basics of computer science parallelisation – GPU acceleration.

Different parallelisation protocols have been used depending on the way memory is shared along the cpus.

a. OpenMP; is a shared memory protocol multithreading protocol, used on a single node.
b. tMPI; is also used on a single node and can increase the efficiency of gromacs (measured in ns/day) relative to OpenMP.
c. MPI; is designed for distributed memory, allowing computations to take advantage of cores in multiple nodes (NWCHEM 6.8 and multinode gromacs)
d. GPU; Gromacs and AMBER18, AMBER20 can use GPUs either on one or multiple nodes. Balancing between numbers of GPU and CPU threads in Gromacs 2019 can

be challenging and largely system dependent therefore before every long simulation the balance has been optimised with trial runs.

2.1.2 Semi-empirical calculations

All molecules lacking a crystal structure were designed with Avogadro[3] and later the structure underwent semi empirical optimisation performed on MOPAC[4] version 2019 using the Castles VM with OpenMP across 18 cores. Starting with initial geometry, the inverse Hessian was calculated along with the second order Taylor expansion of the energy around the point. Geometry changes were proposed across the steepest gradient and the Self-Consistent Field Calculation starts by constructing the density matrix, the Fock matrix and calculates the eigenvectors of the diagonalised Fock accepting only geometry updates that would lower the overall energy as calculated by the sum of total electronic energy and the core-core repulsion energy. MOPAC calls the energy of the converged structure "Heat of Formation" and defines it as "the calculated gas-phase heat of formation at 298K of one mole of a compound from its elements in their standard state". Given the large contribution of intermolecular non-bonded interaction in the studied systems, the latest PM7[5] model for the energy Hamiltonian calculation was chosen. Helping the geometry optimisation and depending on the metal the following keywords were used. 1. UHF, allowing for unrestricted Hartree-Fock Hamiltonian to be used which results in faster calculation but produces a not spin quantized wavefunction. 2. LET, overrides certain safety checks and specifically in geometry optimisation adopts the GNORM value but most importantly allows local rise of a step's energy level, which potentially can lead to a lower overall minimum of geometry (Caution needs to be taken when LET is used as it can lead to unnatural bond formation). 3. LARGE, allowing for the 20 M.O.s around the HOMO-LUMO to be printed. In conclusion, outcomes from semi-empirical calculations have only been used as an input for DFT calculations and quick geometry checks concerning steric clashing of potentially novel molecular interactions (e.g. in the case of CB10).

2.1.3 DFT

2.1.3.1 NWCHEM Achitecture

All DFT optimisations were carried out using NWCHEM[6] v6.6.2[7] an v6.7 using blueBEAR nodes, with the exception of the DFT related to the parametarisation of cylinders via the MCPB.py platform, that used GAUSSIAN09[8] for software consistence.

2.1.3.2 Geometry

Initial geometries were either taken from crystal structures or semi-empirically optimised as described earlier in 2.1.2.

2.1.3.3 Charge and spin

The charge of each system is given usually given by the charge of the metal centres unless imidazoles are present when multiple protonation states, and therefore multiple charges are considered to potentially evaluate the effect of the overall molecular geometry. Spin multiplicity for diamagnetic molecules is set to 1 (closed shell calculation) but for most paramagnetic systems, geometry optimisation is calculated over multiple spin multiplicities to examine the geometry end energy landscape around the theoretically minimum. Additionally, as discussed earlier, ligand field theory often miscalculates[9,10] the suggested spin state in complex systems and therefore a more detailed analysis of the molecular orbitals configuration with different spin multiplicities resulting in a landscape of geometries within the accessible free energy.

2.1.3.4 Functionals

Although initial DFT optimisations employed commonly used and well established functionals (B3LYP, PBE(0) and Minnesota functionals) the bulk of DFT geometry optimisations employed meta-GGA functionals with additional range separated weighting to accommodate the relatively large system and non-bonded contributions within the molecule[11]. For this reason the most commonly used functional within this work has been SSB-D[12] and TPSS family[13,14], which are fully incorporated into NWCHEM code. SSB-D combines the good spin-state splitting of OPBE and good estimation of weak interactions from PBE and also includes Grimme's dispersion correction. Other, newly introduced meta-parameterisations of functionals that are not formally included in the code, can be manually added. For example

CAM-B3LYP:

```
 xc xcamb88 1.00 lyp 0.81 vwn_5 0.19 hfexch 1.00

cam 0.33 cam_alpha 0.19 cam_beta 0.46
```

2.1.3.5 Basis Sets

Due to the size of the system and the computational limitations only relatively small basis sets have been used. Lanl2-dz and def2-svp are the most commonly used but triple zeta versions have been explored more thorough examinations.

2.1.3.6 Dispersion

Given the size of the system and intermolecular non-bonded interactions that can occur there is a potential of van der Waals interactions which can contribute to the stability and behaviour of the system. NWCHEM allows for additional long-range contributions to be calculated, by triggering DFT-D3BJ[15,16] dispersion model in selected supported functionals.

2.1.3.7 Smearing

As defining the spin multiplicity of the complex multi-centered transition metal coordination compounds can be challenging and environment dependent, allowing for partial occupancy of the spin states can accelerate the optimisation and produce results that are closer to that of crystal structures. NWCHEM allows for partial occupation of orbitals as described by Warren and Dunlap[17] by triggering smearing (key word smear). By allowing the gaussian broadening of a spin state one increases the total energy of the system.

2.1.4 Molecular dynamics parameters and software

Molecular dynamics simulations were performed in AMBER18 AMBER20 and GROMACS packages in different versions as described below with the corresponding computational set up.

2.1.4.1 GROMACS[18]

The classical molecular dynamics simulations in this work were carried out using different versions of the GROMACS package (2018.4, 2019.2, 2020.3) run on all computational resources available (VM, multi-node CPU blueBEAR, power9 IBM node(s)). Minimisation and equilibration used GROMACS running on either OpenMP or tMPI (depending on module installation) on the castles VM (single node). All minimisations employed steepest descent to maximum force < 1000 kJ/mol/m. Equilibration of the solvent is in two steps, 1. Canonical (Constant Number of particles, Volume and Temperature), (NVT) for 100ps with ligand and nucleic acid coupled together against solvent and ions for temperature coupling, at 310K with Berendsen thermostat (V-rescale)[19]. The equilibration of pressure (isothermal-isobaric) ensemble was a simulation of additional 100ps with added Parrinello Rahman[20] pressure coupling. Production runs used a 2fs time step, although 1fs and 4fs were also tried. 4fs deemed not refined enough for RNA systems causing infinities and crashing in all the systems, 2fs has been found to be the best compromise between stability and computational expense.

tMPI was used with Cascadelake nodes (40cores) and the IBM power9 nodes (144 logical cores) on single run. It is worth noting that depending on the size of the system, tMPI MD runs that use only the CPUs of power9 nodes can outperform the use of tMPI with GPU acceleration using the optimum combination of CPUs and GPUs on the same node.

Optimum use of the Broadwell – 2x NVIDIA Tesla P100 GPU is achieved using 4 thread-MPI ranks with 2 ranks per GPU and 5 OpenMP threads per rank.

Optimum use of the IBM power9 GPU nodes is achieved by the use of 4 thread-MPI ranks, one per GPU, with 8 OpenMP threads per rank, but it should always be compared to the performance when only CPUs are used.

2.1.4.2 AMBER[21]

Amber18 and Amber20 has been used on VM for minimisation and equilibration and on the Broadwell – 2x NVIDIA Tesla P100 GPU for longer classical – initiation MD and Gaussian Acceleration MD with default-automated settings regarding parallelisation within the single node.

2.1.4.3 Forcefields -Nucleic acids

Throughout this work we used two forcefield set of parameters for DNA, both based on Amber14sb with additional corrections; BSC1[22] and OL15[23] has been taken from the GROMACS forcefield forum (including the correction of the Na Joung-Cheatham parameters).

The majority of simulation of RNA have used the Mathews' forcefield[24]. Initial validation of this forcefield compared it to the RNA parameters within the amber14sb package (chiOL3 corrections for RNA). Comparison between forcefields parameters and their ability to capture experimental results are discussed in the corresponding chapters.

2.1.4.4. Small molecules

All small molecules and ligands for coordination compounds were parameterised within the AmberTools suits with GAFF or GAFF2 as specified in each section.

2.1.4.5 MCPB – parametarization of novel residues - links to DFT

Parametarization of molecules that include metals was carried out using the well-established MCPB.py[25] pipeline. Licence limitations on GAUSSIAN09 in blueBEAR limits the use of the program on a single node, so the initial geometries are the results of DFT optimisation using NWCHEM as previously described. Since all the DFT calculations in GAUSSIAN09 are carried on a single node, memory limitations can be overcome by using the relatively small basis set (6-31G* or LANL2-DZ) and the current directory for scratch memory. Nevertheless, an element of consistency can be retained by using wB97D3[26] functional (available in G09) as it is the functional used for the parametarisation by Mathews et al. . If needed, MCPB.py output geometry and topology were translated to GROMACS format using parmed (https://parmed.github.io/ParmEd/html/index.html). It is worth noting that when multiple different novel residues are introduced to the same system, the name of each atom needs to be unique.

2.1.5 Classical MD

The core of this research is to understand the dynamics between metastable states that lead to an ensemble understanding of nucleic acid structures and their interaction with small molecules and ligands. In order to increase the sampled space, within the computational limits classical MD simulations were carried at 310K, with Na^+ as a positive counter ion and in relatively low salt concentration. There are several runs with K^+ and higher concentrations that show the overall effect in the size of sampled space, but the behaviour of the ionic environment around the nucleic acid has not been studied in detail.

2.1.6 GaMD[27]- LiGaMD[28]

A relatively new methodology that has been developed for accelerated molecular dynamics by the Miao group called Gaussian acceleration, where additional energy “boosting” is applied to the system in specific energy terms allowing the system to overcome energy barriers in transitions. This has the advantage of not applying any constraints in the system or examining the energy perturbations across pre-defined coordinates or collective coordinates. If the boost is applied by a harmonic potential, the resulted gaussian distribution of the observed potential can be used in reweighting the resulted landscape to map it back to that of the original temperature of the system. Similarly, when the harmonic boost is applied in energy terms of the interaction between a ligand and a target, the reweighted result can produce the energy landscape of the interaction in a non-biased manner and include the resulting conformation-energy changes of the biomolecule.

2.2 MD Data analysis

2.2.1 PYEMMA[29] -MSM[30]

The whole compressed trajectory is loaded into PyEMMA with 10ps time steps, without concatenating the independent simulations of the same system.

1. The dimensionality of the simulation is reduced by feature selection and using Principal Component Analysis (PCA). This step aims to capture as much of the kinetic variance of the simulation as possible in fewer dimensions. PCA gives a broad picture of the overall kinetics of the simulations, highlighting minima in a throughout-the-simulation manner. The distribution of the principal component eigenvalues reveals that significant variance in the data is described by higher order features beyond the traditional first two components and we therefore choose to plot the projections onto the pairwise combinations of the first four components.

2. To extend the representation into the time domain we use Time-lagged Independent Component Analysis to examine the similarity between time points with a specified offset (lag time). This was used to select a lag time that produced the fewest number of independent components that incorporated 95% of the overall kinetic variance. For visualisation purposes, we plot the projections onto the pairwise combinations of the first four components.

3. The data are projected onto the first 2 ICs for the selected lag time (number of steps usually between 200 and 500) and are then clustered using k-means. The number of clusters is chosen as the starting point of the plateau in the VAMP2 graph. This allows the problem to be mapped onto a set of discrete states (clusters) which then allows Markov State Models (MSMs) to be constructed that describe transitions between pairs of clusters. MSM calculates every pairwise transition at a given lag-time for the MSM (which independent of the TICA lag time). It is essential that the selected lag time is long enough to ensure Markovian dynamics (ie. to have a stable transition matrix), but short enough to resolve the transition dynamics. To achieve this the smallest possible lag-time (in the range 1to 1200 steps) that shows convergence of the underlying processes is chosen. We validate the model with the Chapman-Kolmogorov test, ensuring that the

model also describes longer time-scales. If the resulted model includes all states, the original space can then be split into the longest living meta-stable states with care taken to ensure sure that those states are well defined on the discretised surface. For each state a sample of 10 pdb structures is extracted and represented in Leontis-Westholf notation using Barnaba[31]. Finally spectral clustering using PCCA++ algorithm coarse-grains the space of the metastable distributions and approximates the stationary probabilities and relative free energies for that particular set of data and not for the overall system.

2.3 RNA Structure prediction

2.3.1 Secondary Structure prediction

Over the past 20 years there has been extensive work in the field of secondary structure prediction for RNA fragments, with largely successful results for fragments under 100 bases, with efforts mostly focusing on creating scoring functions based on hydrogen bonding strength, thermodynamic principles[32] even neighbouring base pairing[33] and input of SHAPE experiments[34]. More recently, deep learning techniques have been employed[35] which can potentially increase the accuracy in longer fragments.

2.3.2 FARFAR2[36]

Tertiary RNA structure prediction is even more elusive, and a single solution is largely irrelevant. As it has been discussed the dynamic nature of RNA structure can be described as a series of metastable states which in turn can be predicted using clustering methodology and energy functions similar to that in secondary structure prediction. The SARS-COV2 sequence fragments were introduced to FARFAR2 to create structures for molecular dynamics simulations, of the RNA alone, or the RNA with the cylinder.

2.4 Links to experiments

Comparison of experimental techniques for input structure (CRYO-EM, NMR, XRD)

Although in the past XRD has been the gold standard for structural studies of biomolecules the field is moving towards ensemble-based techniques like NMR and Cryo-EM. This allows 1. To create experimental structures of flexible and therefore difficult to crystalize molecules, and 2. at the same time deconvolute the metastable states within the landscape of the structure[37,38,39].

2.5. References

1 S. J. Thompson, S. E. M. Thompson and J. Cazier, 2019, 1–6.

2 L. Lannelongue, J. Grealey and M. Inouye, *Adv. Sci.*, 2021, **8**, 1–10.

3 W. A. De Jong, A. M. Walker and M. D. Hanwell, *J. Cheminform.*, 2013, DOI:10.1186/1758-2946-5-25.

4 James J. P. Stewart, *J. Comput. Aided. Mol. Des.*, 1990, **4**, 1–105.

5 R. L. M. Gieseking, M. A. Ratner and G. C. Schatz, *J. Phys. Chem. A*, 2018, **122**, 6809–6818.

6 E. Aprà, E. J. Bylaska, W. A. De Jong, N. Govind, K. Kowalski, T. P. Straatsma, M. Valiev, H. J. J. Van Dam, Y. Alexeev, J. Anchell, V. Anisimov, F. W. Aquino, R. Atta-Fynn, J. Autschbach, N. P. Bauman, J. C. Becca, D. E. Bernholdt, K. Bhaskaran-Nair, S. Bogatko, P. Borowski, J. Boschen, J. Brabec, A. Bruner, E. Cauët, Y. Chen, G. N. Chuev, C. J. Cramer, J. Daily, M. J. O. Deegan, T. H. Dunning, M. Dupuis, K. G. Dyall, G. I. Fann, S. A. Fischer, A. Fonari, H. Früchtl, L. Gagliardi, J. Garza, N. Gawande, S. Ghosh, K. Glaesemann, A. W. Götz, J. Hammond, V. Helms, E. D. Hermes, K. Hirao, S. Hirata, M. Jacquelin, L. Jensen, B. G. Johnson, H. Jónsson, R. A. Kendall, M. Klemm, R. Kobayashi, V. Konkov, S. Krishnamoorthy, M. Krishnan, Z. Lin, R. D. Lins, R. J. Littlefield, A. J. Logsdail, K. Lopata, W. Ma, A. V. Marenich, J. Martin Del Campo, D. Mejia-Rodriguez, J. E. Moore, J. M. Mullin, T. Nakajima, D. R. Nascimento, J. A. Nichols, P. J. Nichols, J. Nieplocha, A. Otero-De-La-Roza, B. Palmer, A. Panyala, T. Pirojsirikul, B. Peng, R. Peverati, J. Pittner, L. Pollack, R. M. Richard, P. Sadayappan, G. C. Schatz, W. A. Shelton, D. W. Silverstein, D. M. A. Smith, T. A. Soares, D. Song, M. Swart, H. L. Taylor, G. S. Thomas, V. Tipparaju, D. G. Truhlar, K. Tsemekhman, T. Van Voorhis, A. Vázquez-Mayagoitia, P. Verma, O. Villa, A. Vishnu, K. D. Vogiatzis, D. Wang, J. H. Weare, M. J. Williamson, T. L. Windus, K. Woliński, A. T. Wong, Q. Wu, C. Yang, Q. Yu, M. Zacharias, Z. Zhang, Y. Zhao and R. J. Harrison, *J. Chem. Phys.*, **2020**, *152*,184102-1-26

7 M. Valiev, E. J. Bylaska, N. Govind, K. Kowalski, T. P. Straatsma, H. J. J. Van Dam, D. Wang, J. Nieplocha, E. Apra, T. L. Windus and W. A. De Jong, *Comput. Phys. Commun.*, **2010**, *181*, 1477–1489.

8 Gaussian 09, Revision A.02, M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, T. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, and D. J. Fox, Gaussian, Inc., Wallingford CT, 2016.

9 S. K. Singh, J. Eng, M. Atanasov and F. Neese, *Coord. Chem. Rev.*, 2017, **344**, 2–25.

10 L. Lang, M. Atanasov and F. Neese, *J. Phys. Chem. A*, 2020, **124**, 1025–1037.

11 W. Zhang, D. G. Truhlar and M. Tang, *J. Chem. Theory Comput.*, 2013, **9**, 3965–3977.

12 M. Swart, M. Solà and F. M. Bickelhaupt, *J. Chem. Phys.*, 2009, DOI:10.1063/1.3213193

13 K. P. Kepp, *Inorg. Chem.*, 2016, **55**, 2717–2727.

14 S. Kossmann, B. Kirchner and F. Neese, *Mol. Phys.*, 2007, **105**, 2049–2071.

15 S. Grimme, *J. Comput. Chem.*, 2004, **25**, 1463–1473.

16 S. Grimme, *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 2011, **1**, 211–228.

17 R. W. Warren and B. I. Dunlap, *Chem. Phys. Lett.*, 1996, **262**, 384–392.

18 H. J. C. Berendsen, D. van der Spoel and R. van Drunen, *Comput. Phys. Commun.*, 1995, **91**, 43–56.

19 S. Pronk, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov, M. R. Shirts, J. C. Smith, P. M. Kasson, D. Van Der Spoel, B. Hess and E. Lindahl, *Bioinformatics*, 2013, **29**, 845–854.

20 S. Pronk, S. Páll, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov, M. R. Shirts, J. C. Smith, P. M. Kasson, D. Van Der Spoel, B. Hess and E. Lindahl, *Bioinformatics*, 2013, **29**, 845–854.

21 R. Salomon-Ferrer, D. A. Case and R. C. Walker, *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 2013, **3**, 198–210.

22 R. Galindo-Murillo, J. C. Robertson, M. Zgarbová, J. Šponer, M. Otyepka, P. Jurečka and T. E. Cheatham, *J. Chem. Theory Comput.*, 2016, **12**, 4114–4127.

23 M. Zgarbová, J. Šponer, M. Otyepka, T. E. Cheatham, R. Galindo-Murillo and P. Jurečka, *J. Chem. Theory Comput.*, 2015, **11**, 5723–5736.

24 A. H. Aytenfisu, A. Spasic, A. Grossfield, H. A. Stern and D. H. Mathews, *J. Chem. Theory Comput.*, 2017, **13**, 900–915.

25 P. Li and K. M. Merz, *J. Chem. Inf. Model.*, 2016, **56**, 599–604.

26 Y. Minenkov, Å. Singstad, G. Occhipinti and V. R. Jensen, *Dalt. Trans.*, 2012, **41**, 5526–5541.

27 Y. Miao, V. A. Feher and J. A. McCammon, *J. Chem. Theory Comput.*, 2015, **11**, 3584–3595.

28 Y. Miao, A. Bhattarai and J. Wang, *J. Chem. Theory Comput.*, 2020, **16**, 5526–5547.

29 M. K. Scherer, B. Trendelkamp-Schroer, F. Paul, G. Pérez-Hernández, M. Hoffmann, N. Plattner, C. Wehmeyer, J. H. Prinz and F. Noé, *J. Chem. Theory Comput.*, 2015, **11**, 5525–5542.

30 B. Reuter, M. Weber, K. Fackeldey, S. Röblitz and M. E. Garcia, *J. Chem. Theory Comput.*, 2018, **14**, 3579–3594.

31 S. Bottaro, G. Bussi, G. Pinamonti, S. Reiber, W. Boomsma and K. Lindorff-Larsen, *Rna*, 2019, **25**, 219–231.

32 F. Meng, V. N. Uversky and L. Kurgan, *Cell. Mol. Life Sci.*, 2017, **74**, 3069–3090.

33 M. F. Sloma and D. H. Mathews, *PLoS Comput. Biol.*, 2017, **13**, 1–23.

34 K. Zarringhalam, M. M. Meyer, I. Dotu, J. H. Chuang and P. Clote, *PLoS One*, 2012, 7, 10, e45160

35 J. Singh, J. Hanson, K. Paliwal and Y. Zhou, *Nat. Commun.*, **2019**, **10,** 5407, 13395-9

36 A. M. Watkins, R. Rangan and R. Das, *Structure*, 2020, **28**, 963-976.e6.

37 N. J. Baird, S. J. Ludtke, H. Khant, W. Chiu, T. Pan and T. R. Sosnick, *J. Am. Chem. Soc.*, 2010, **132**, 16352–16353.

38 J. Giraldo-Barreto, S. Ortiz, E. H. Thiede, K. Palacio-Rodriguez, B. Carpenter, A. H. Barnett and P. Cossio, *Sci. Rep.*, 2021, **11**, 1–15.

39 P. Ge and Z. H. Zhou, *Proc. Natl. Acad. Sci. U. S. A.*, 2011, **108**, 9637–9642.

Chapter 3

Electronic Structure through Density Functional Theory and Molecular Dynamics of coordination compound complexes

Comments on the chapter

The final goal of this thesis is to study the interaction of metal complexes and RNA at the atomic level. However, elements of electronic structure of these complexes have been studied 1. As part of the classical parametarization pipeline 2. To draw theoretical inside and aid synthesis and characterisation of the compounds.

The underlining hypothesis of this thesis is also applied here, although only in a preliminary level. In this chapter, along with the ground state, metastable or transient electronic states around them are also examined, to build an energy landscape that can describe the dynamics that would lead to different behaviour in larger scales.

Major motivation for this chapter is the different effect of Ni and Ru parent cylinders on cells. It has been shown previously, that although identical in structure the Ni centred cylinder has very little to no cytotoxic effects in cell lines where Ru cylinder does. Following the DFT analysis of the parent compound with different metals using different levels of theory, different ligands are also examined, changing the coordinating pyridine with imidazoles.

Finally, the interaction of cylinders with CB10 is studied, proposing a ground state of the interaction using semi-empirical and DFT optimisation. The resulting structures are parameterised and moved to classical molecular dynamics simulations where the complex's dynamics can be examined in longer timescales in solution.

Results from DFT for the cylinders in this chapter have been published in Chem. Sci., 2021,12, 7174-7184 (parent Fe and Ru), J. Am. Chem. Soc. 2020, 142, 49, 20651–20660 (imidazole and CB10 rotaxane and pseudorotaxane) and Angew. Chem. Int. Ed. **2019**, 60, 33, p 18144-18151. The results on the stability of partially capped rotaxanes will be included in a manuscript that is in the writing process.

3.1 Introduction

Coordination complexes are those that include “coordinate covalent bond” between a ligand (Lewis base) and a central metal atom (Lewis acid). The Lewis base can supply (donate) electrons to the metal, therefore in the case of coordinate covalent bonding electrons are supplied only by the donor atom, as opposed to covalent bonds where electrons originate from both involved atoms.

Theoretical understanding of these bonds started developing shortly after the introduction of quantum mechanics with the development of crystal field theory, which only describes electrostatic interactions between metal ions and ligands. Substantial extension and improvement on crystal field theory continued in the 1950s with the development of Ligand Field Theory, which describes both the electrostatic and the covalent nature of the coordination bonds[1]. In the same period, fundamental understanding of molecular electronic structure[2] was also developing. Critical work in combining the two approaches was developed by the Ballhausen group for many years starting with molecular orbital description of square planar complexes[3].

As early as the 1930s lower symmetry structure solutions were acknowledged to have lower energy. The Jahn-Teller theorem[4] examines elegantly the effect of orbital degenerate states and the symmetry of the stable polyatomic structure. Although initially the importance of the effect and the implication to the dynamics of the system was only described in the solid state and vibrational modes[5], ab initio molecular dynamics made possible to start examining the solution structures[6][7][8] and although the cylinders and especially the less used Cu centered cylinder would be a great and challenging molecule to investigate under this theoretical framework, it is beyond the scope of this chapter and thesis but should be considered for future work.

The field progressed massively after publication of the Kohn-Sham equation and development of numerical and computational tools to use DFT (as it was briefly discussed in the introduction). All of this work always focuses on the ground states and is validated using crystallographic data for bond lengths, Infrared and Raman Spectroscopy for vibrational modes and Electron Paramagnetic Resonance (EPR)[9] for spin states.

Interest on single molecule spin states spiked after the 1990s when single molecule magnets started developing[10][11][12]. This also lead to the overall increased interest in thermodynamic effects on spin states[13][14][15] with special interest in Fe centers[16]. Literature is more sparse and mostly focused on surface or bulk effects[17] but some published work over the past 5 years tries to combine DFT with thermodynamics and spectroscopy on a single molecule level[18][19][20].

Other groups are also adopting a combined experimental and theoretical tool kit to study the dynamics of metal complexes[21][22][23][24], with Jean-Marie Lehn leading the field on dynamic bond characterization[25][26].

3.2 Parent cylinder

3.2.1 Iron – ruthenium

The first attempt to create a model for the parent cylinder used a purely de novo approach, creating three-dimensional structure of the ligand and coordinating three of them to the two metal centres. The resulting structure was minimized using the Universal Force Field (UFF) and the resulting structure was further optimized at the DFT ssb-d/Def2-svp level of theory. Interestingly the resulting structure was not a helicate. This structure has been theorized before in the group and can potentially exist in the racemic mixture after synthesis.

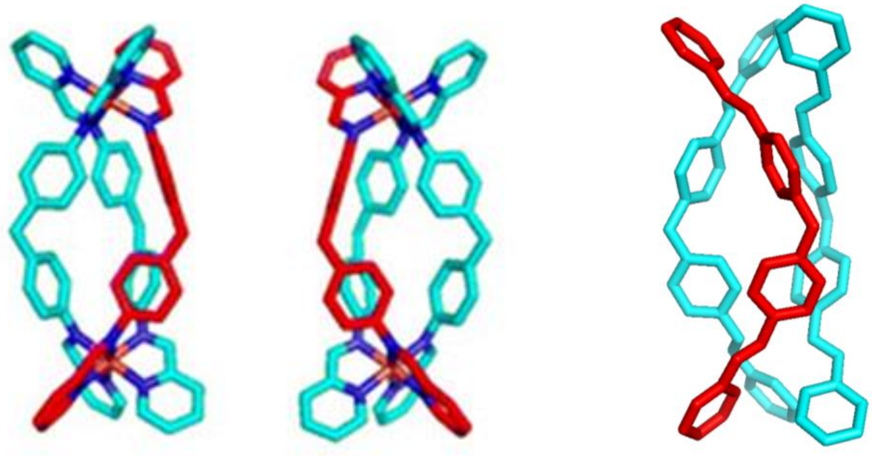


Figure 3.1 *From right to left, P M and non-helical enantiomers of Hannon parent supramolecular cylinders.*

Examining the above structures at the DFT level of theory one can examine the potential differences and potentially their co-existence in solution. Since ruthenium cylinders are the most stable and almost certainly in the lower spin coordination, here only the Ruthenium version of the non-helical cylinder is examined and compared with the helical.

In terms of total DFT energy of the cylinder multiple combinations of functionals and basis sets have been used to optimize the structures.

Ruthenium Cylinder converged energy in au (1au = 27.2114eV)

TPPSH D4 DEF2TZVP	-3741.97398650
CAM-B3LYP LANL2DZ	-3735.37293530
TPPSH D4 DEF2-SVP	-3738.25134756
SSBD LANL2TZ-ECP	-3772.113856862878
SSBD LANL2DZ	-3771.66839692
B97D4 LANL2TZ	-3737.03937479
SSBD DEF2-SVP	-3771.56328560

non-helical Ruthenium Cylinder converged energy

SSBD LANL2DZ	-3771.66039413
SSBD LANL2DZ	-3771.55105066
B97D4 LANL2TZ-6311g	-3737.03937479

The total DFT energy difference between the two structure is less than 3 eV and it would be reasonable to hypothesize that this conformation would be present in a molecular mixture, it is worth examining closer the difference between them in the shape of the resulting molecular orbitals at the optimized structure.

Helical Ruthenium cylinder

In the case of the helical structure the highest occupied molecular orbital is on the metal centre whereas for the a-helical is is on the ligand. This is probably due to the higher deformation of the ligand in order to accommodate the complex structure.

It is clear than in all combinations of basis set and functional, the molecular orbital on Ru is a well formed d_z^2, which is interesting when compared to on row higher element iron(Fe).

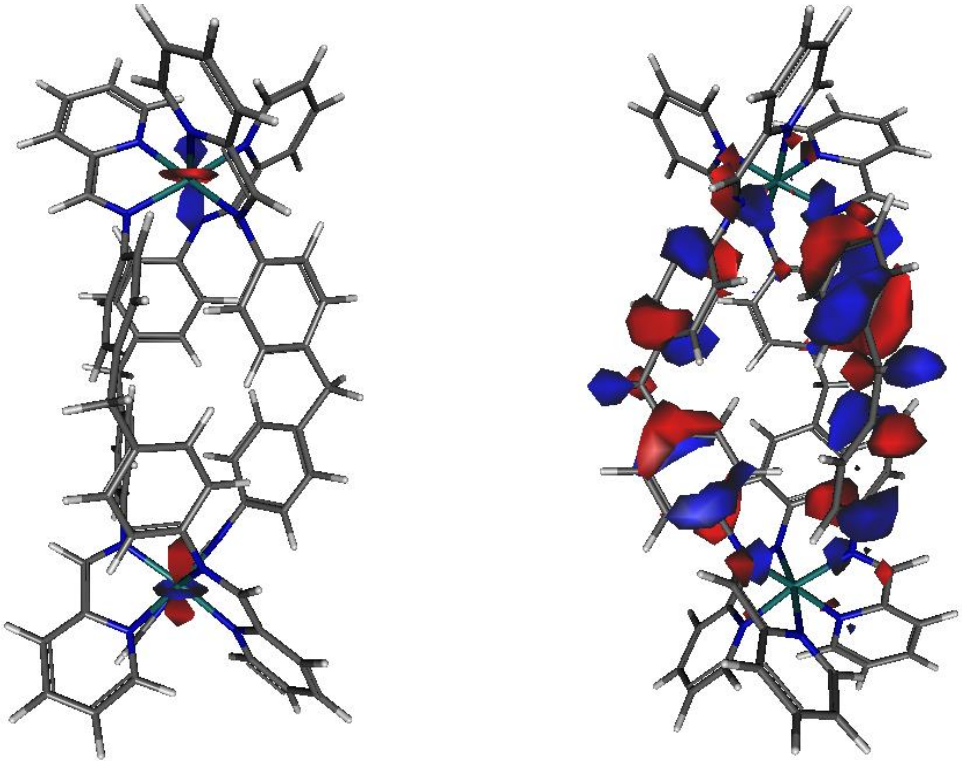


Figure 3.2 *HOMO of helical (left) and a-helical optimized structures (in TPPSH/DEF2-SVP).*

Examining the supramolecular cylinders with Fe metal centers one observes that the resulting orbital is less well formed under the same level of theory.

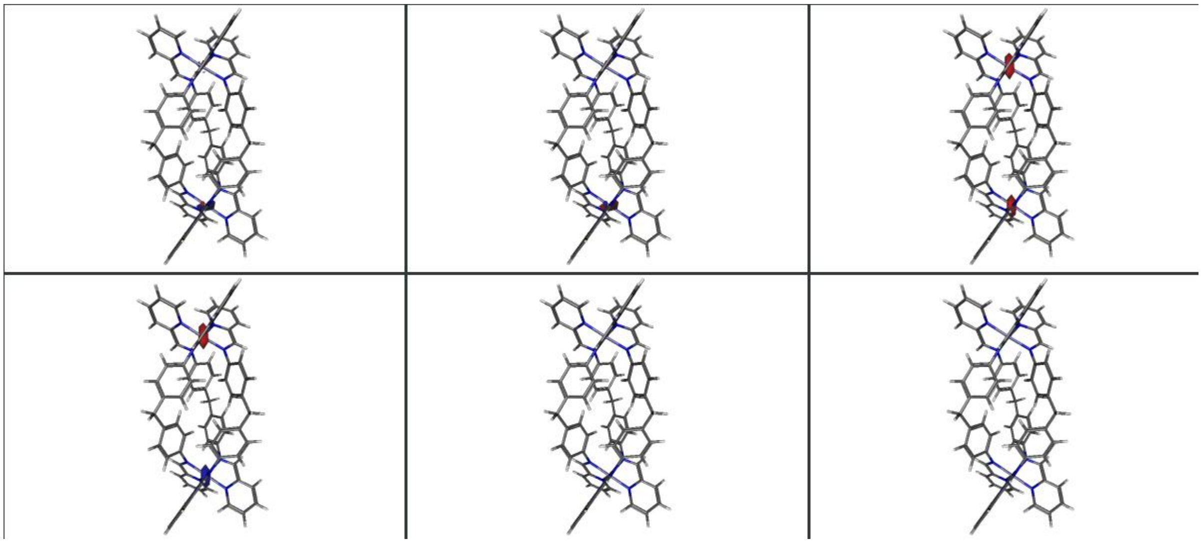


Figure 3.3 *HOMO-4 to LUMO+2 molecular orbitals of Fe P enantiomer. the (TPPSH / DEF2-SVP)*

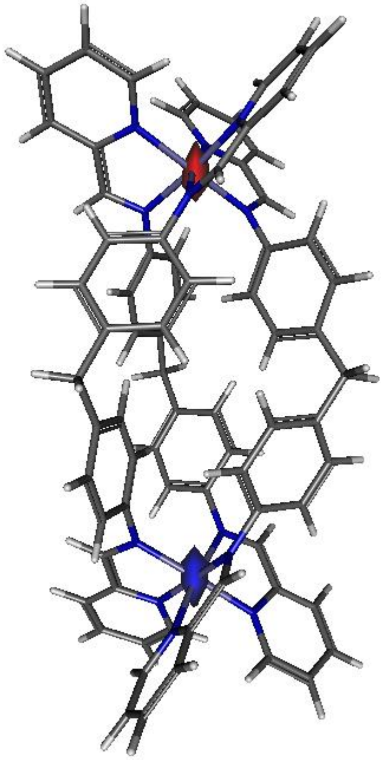


Figure 3.4 *HOMO of Fe P enantiomer supramolecular cylinder at the (TPPSH / DEF2-SVP).*

3.3 Nickel

The huge difference in the in vivo effect between nickel and ruthenium versions has been the inspiration for this chapter and indeed examining the electronic structure that results from DFT optimization one can observe differences that shine light on that. Nickel cylinders are paramagnetic and therefor characterization with NMR can been challenging. Different spin states have been examined (m =1 , 5 , 9) as a starting point for the DFT optimization of Nickel cylinder in an effort to explore the wider landscape of spin and conformation states in solution and in the presence of other molecules (in cellulo). Convergence was achieved in all cases under strict conditions (in open shell calculations that allow smearing) with difference in energy levels smaller than 3eV, but molecular spin of 5 is consistently the lower energy solution, therefore the most abundant species within the solution.

SSBD D def2-SVP m1 Total DFT energy = -6603.413268025022

SSBD D def2-SVP m5 Total DFT energy = -6603.486954217697

SSBD D def2-SVP m9 Total DFT energy = -6603.337324123771

TPPSH D4 def2-SVP m1 Total DFT energy = -6564.630426949247

TPPSH D4 def2-SVP m5 Total DFT energy = -6564.725478727734

tpsshD4 -def2svp m9 Total DFT energy = -6564.577083270906

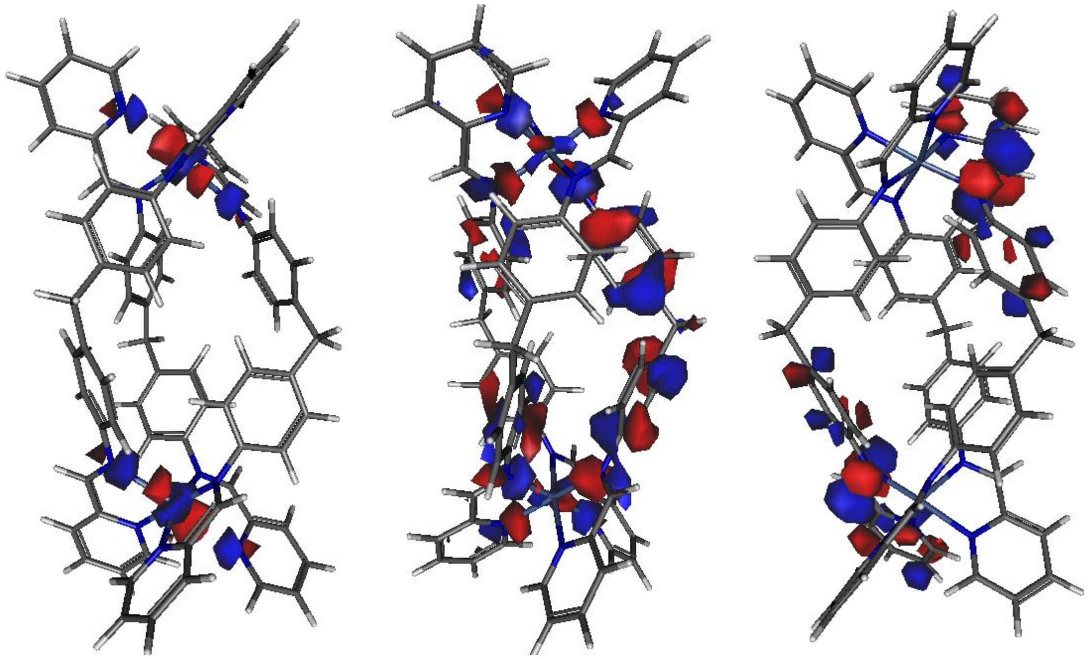


Figure 3.5 *Structure and HOMO of molecular spin 1, 5 and 9 (left to right) at the (TPPSH / DEF2-SVP).*

It is worth mentioning the effect that spin has on the molecular conformation of the optimised structure and the metal centre geometry. Specifically, elongation of the distance between the metal centre and one of the pyridines is observed, similar to the Jahn-Teller effect discussed earlier.

Additionally, the distribution of partial charges within the molecule is drastically changing with Ni being able to withdraw more charge from the ligands resulting in a more positively charged surface.

Only Ru and Fe cylinder were parameterized for MD using the MCPB.py[27] pipeline with Gaussian09 using wB97XD9/6-31G* closed cell level of theory, and were used in all the publications resulted from this thesis[28][29][30].

3.4 Imidazole ligands and interaction with CB10

In an effort to create functionalisable cylinders the pyridine of the parent Hannon ligand (L) is changed with imidazole, which resulted in drastic re-configuration of the electronic structure. With the three pyridine nitrogens being replaced with imidazole nitrogens, coordination to Fe is now less stable (multiple spin states produced converged solutions) but coordination to Ni becomes more stable (the lowest energy solution of spin 5). These computational results also correlated with synthetic challenges of Fe centered L'

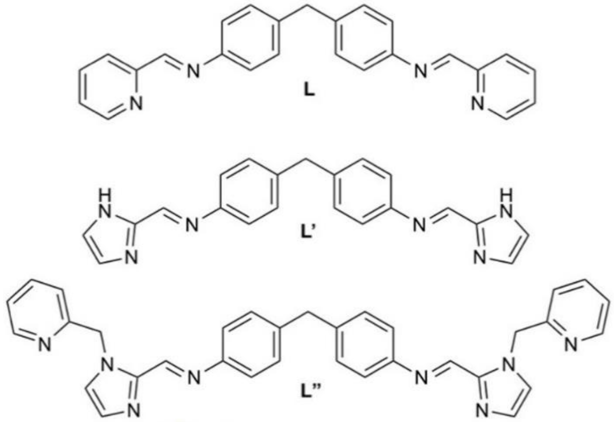


Figure 3.6 *Ligands used for cylinder optimization.*

As they were the most stable experimentally, only Ni centered cylinders with L’ and L’’ were parametarized for MD simulations using the pipeline with open shell DFT this time and multiplicity of 5.

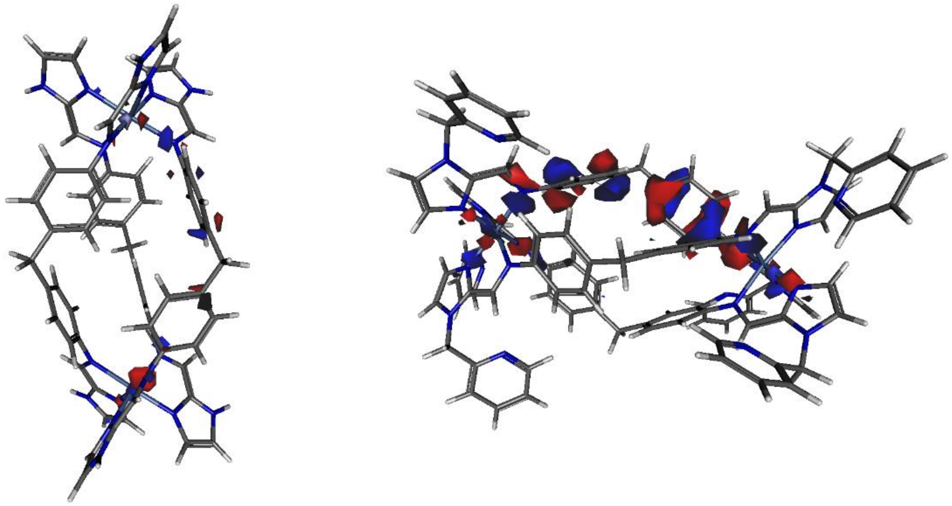


Figure 3.7 *HOMO of Ni L’ (left) and L’’ (right) cylinders, at SSBD-def2-SVP level of theory.*

Interesting cases of driving equilibria at the single molecule level can rise by using molecules with different levels of solubility. The cucurbitinil CB(n)[31] family of molecules are macrocycles with 5 7 8 or 10 glycoluril units. CB(n)s have very low solubility in water[31][32] but can become soluble upon hosting another molecule in their cavity[33]. There is a long list of molecules and reviews on the interactions with CB(n)[34] but the vast majority of research has focused on the smaller macrocycles (<8 units). In the Hannon group however we used CB10 to host supramolecular cylinders[28], creating novel rotaxane and pseudo-rotaxane species by using the imidazole Ni cylinder with 6 methyl pyridine caps.

Crystallization and structural characterization of these complexes has been challenging and therefore DFT and Molecular dynamics were used to gain any insight into the complex. First step was the characterization and parameterization of CB10 using DFT at the same level of theory as the parameterization of the cylinders (SSB-d/ Def2-SVP).

Preliminary molecular dynamics simulations of the CB10 in water solution suggested possible distorted states. To evaluate to what effect that state could be real and not an arartefact of the parameterization, the resulted structure was optimized again with DFT at the same level of theory and a new local minimum was uncovered, with energy difference between the two 0.01 Hartree or 0.27eV.

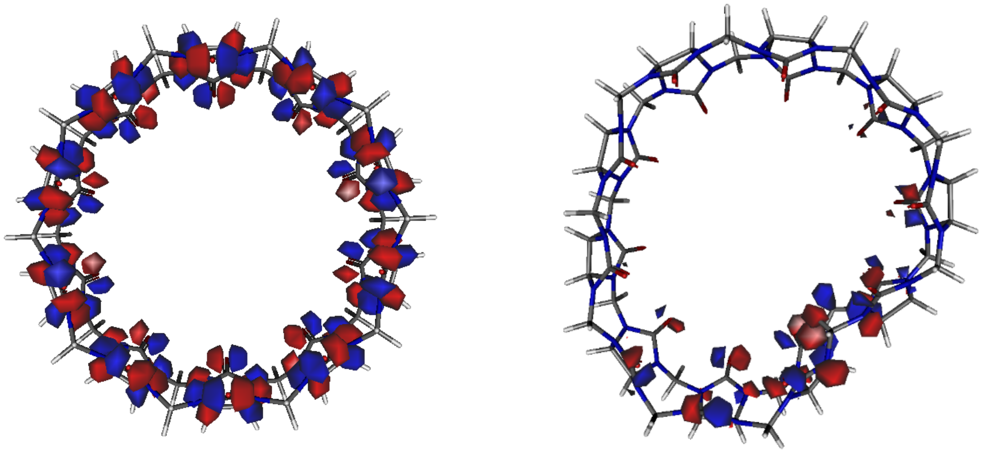


Figure 3.8 *HOMO of CB10 optimized on crystal structure (left) and local minimum (right) identified after MD the (TPPSH / DEF2-SVP)*

Interestingly, similar deformation of CB10 has also been observed in DFT optimization with CB10 hosting cylinders. This breaking of point symmetry could contribute to the difficulties in crystallization of the complex. Additionally, starting the optimization of the cylinder-CB10 complex with the cylinder placed in the middle of CB10 with the dinuclear axis perpendicular to the CB10 plane, always results in optimized structure that tilts the axis. Molecular dynamics of the system revealed that the cylinder is rotating within CB10 in the pseudo-rotaxane case, but after rotaxanation the methyl-pyridine caps interact with CB10 stabilizing, or slowing down this movement.

This rotaxane and the pseudorotaxane are small and dynamic systems that present a great opportunity to employ PCA and TiCA representations to describe long dynamic trajectories and reveal differences between the two systems, while understanding the meaning of the parameters used in the process.

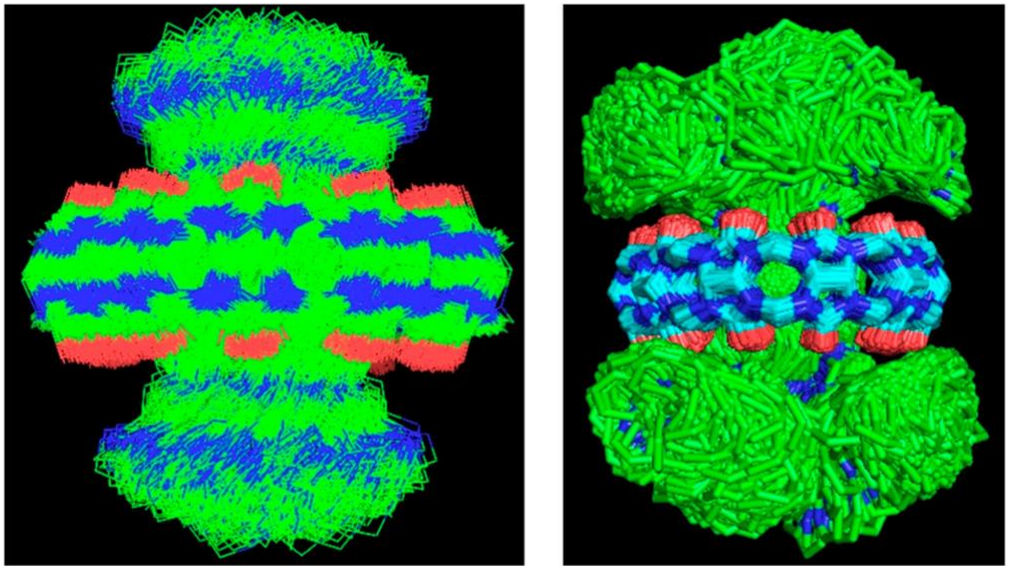


Figure 3.9 *Combined and superimposed views of Molecular Dynamics simulations of the pseudo-rotaxanated* $[Ni_2L'_3{\cdot}CB10]^{4+}$ *(left) and proper rotaxanated* $[Ni_2L''_3{\cdot}CB10]^{4+}$ *(right) complexes showing the free rotation of the* $[Ni_2L'_3]^{4+}$ *cation in the CB[10] and the more restricted motion caused by the picolyl groups for* $[Ni_2L''_3]^{4+}$*; in the 1 μs time scale the rotaxanated cylinder does not rotate. Hydrogens are omitted for clarity. Supplementary videos of the simulations are also available. There is a tendency for the CB[10] ring to undergo fluxional distortions during the simulation, including a buckling of the ring at the* CH_2 *groups that create a heart-shaped ring and allowing it to close up around the cylinder. This is also seen at very low frequency (<0.1%) in simulations of the free CB[10] but is more prevalent in the rotaxane. The paramagnetic broadening in the* ^{1}*H NMR does not allow this feature to be confirmed experimentally but reflects the greater flexibility of CB[10] compared to lower order cucubit[n]urils. (as published* [28]*).*

During the 1 microsecond long simulation of the pseudorotaxane the cylinder is rotating within CB10, however as the timescale of the simulation is orders of magnitude longer compared to the kinetics of the system PCA averages all the trajectories to a single point in the landscape. However, applying time lagged steps on the autocorrelation of the trajectory data transient metastable states can be revealed. This presents a great example of the effect of the choice of lag time in TiCA representation. In this case, the first two eigenvectors capture the circular motion of the cylinder within CB10.

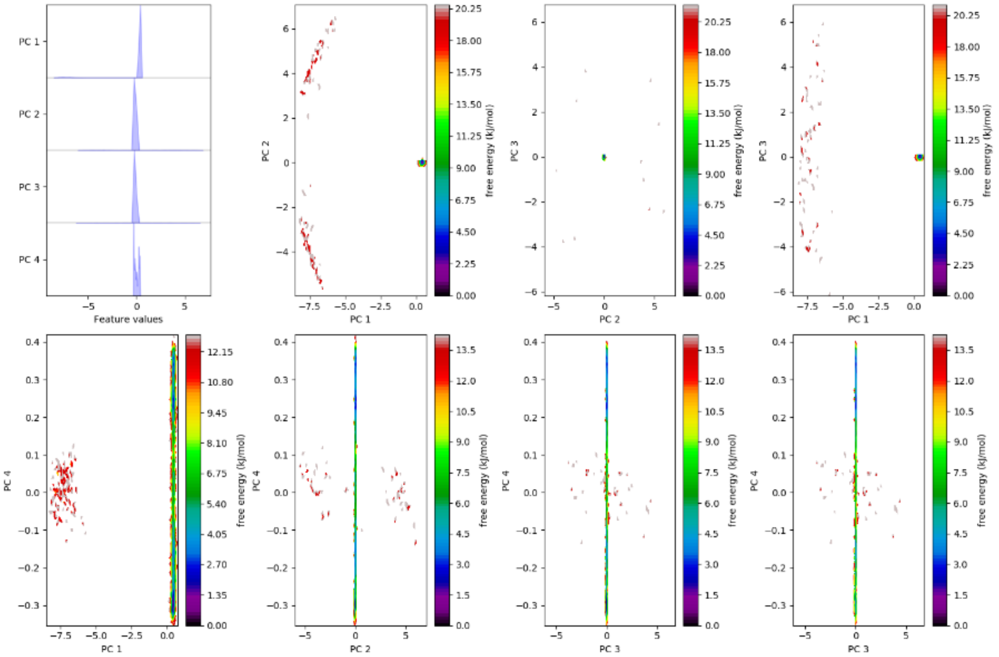


Figure 3.10 *PCA of pseudorotaxane 1μs MD simulation.*

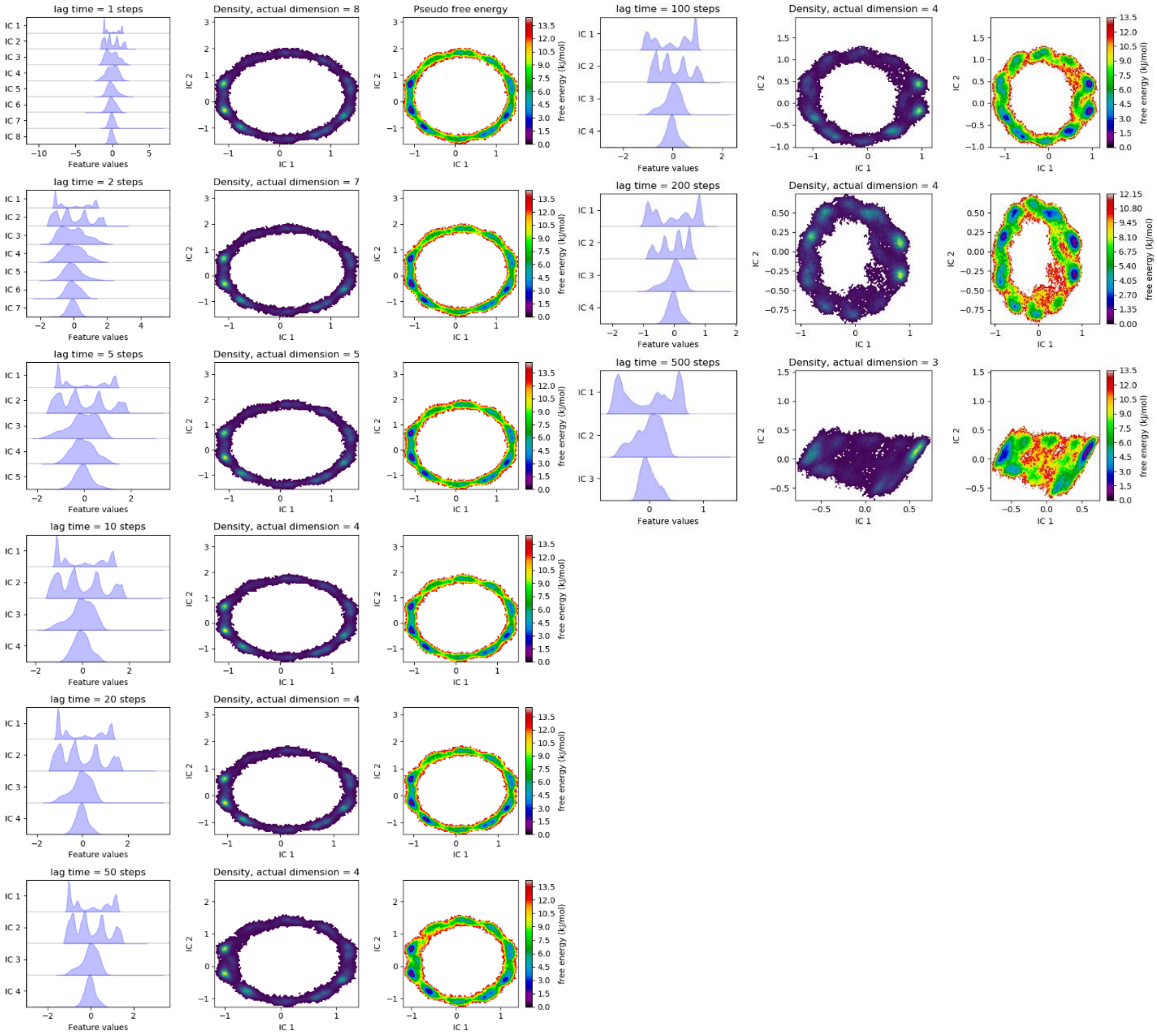


Figure 3.11 *TiCA representation and landscape with different lagging times (1-500 10ps steps).*

On the other hand, in the case of the rotaxane the cylinder does not rotate in CB10 rather, it is the movement of the caps that dominate the kinetic variability of the trajectory. This time PCA identifies two broadly defined regions which remain close together, whereas TiCA can capture metastable states of the caps transiently making contact with CB10.

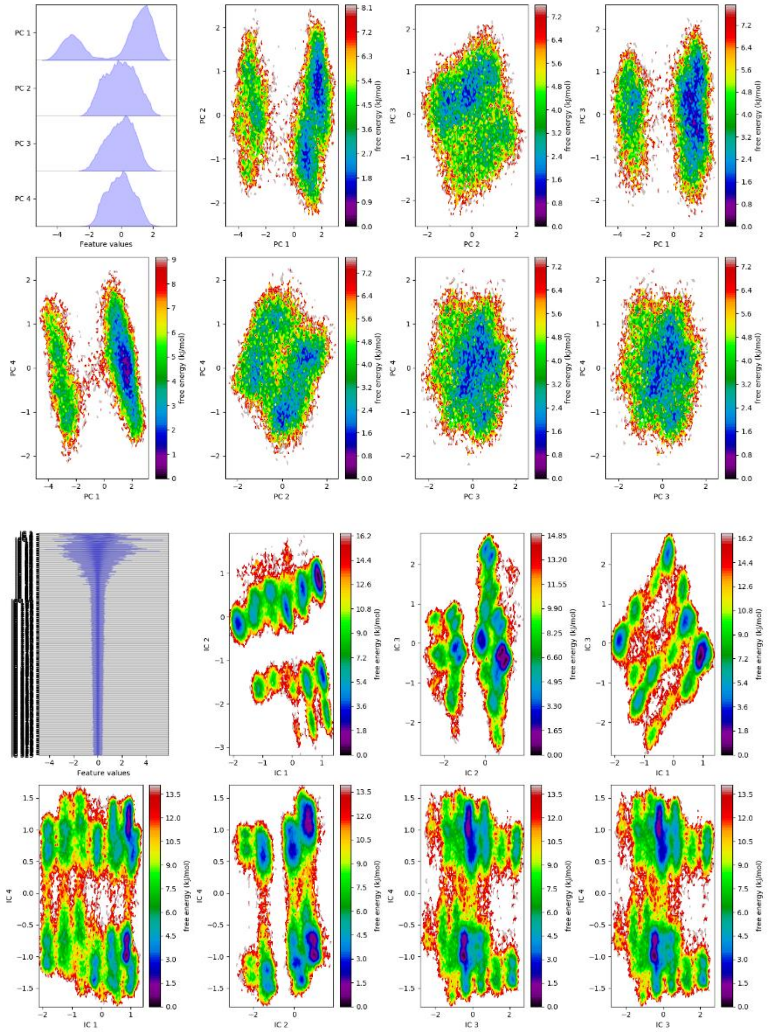


Figure 3.12 *PCA (top) and TiCA (bottom) of 1µs simulation of rotaxane.*

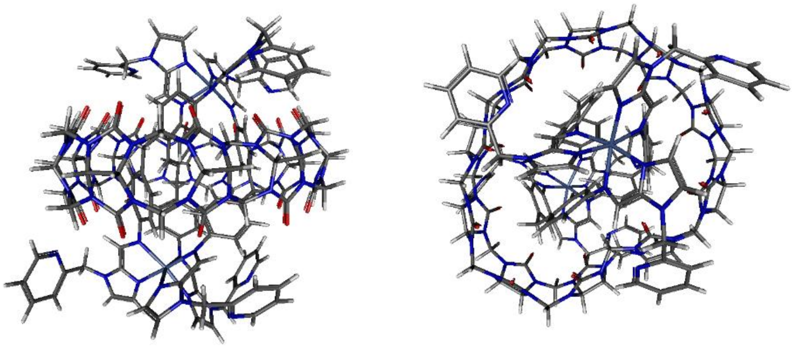


Figure 3.13 *DFT optimization of capped cylinder – CB10 complex the (TPPSH / DEF2-SVP).*

Following the first publication[28], methyl-pyridine caps were replaced with DNB caps, in order to gain control over the capping. Mass spectroscopy revealed that partial capped cylinder – CB10 complexes also exist in solution and MD simulations were performed to evaluate the force needed to dissociate the capped and partially capped DNB cylinders from CB10. The fully capped DNB cylinder was parameterized as previously described along with combinations of partially capped cylinders (3 caps on one metal center and 2 or 1 on the other).

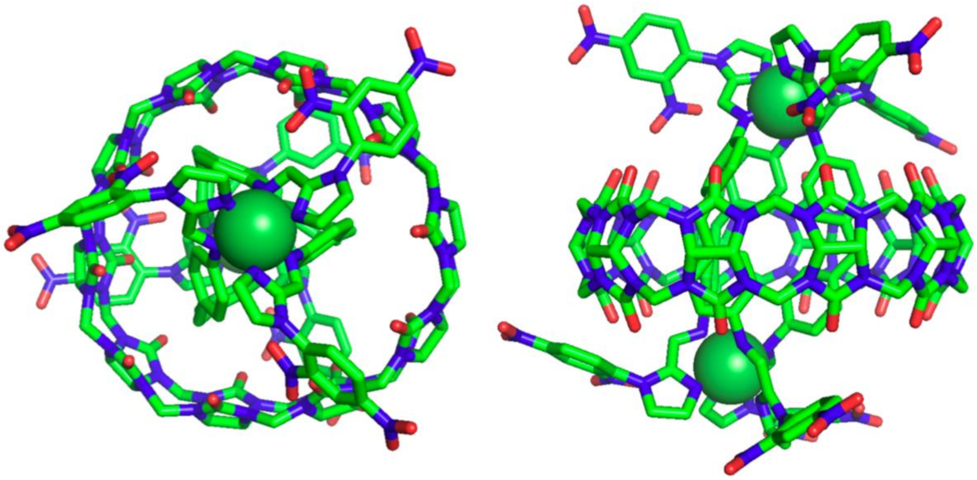


Figure 3.14 *DFT optimized structure of DNB cylinder.*

The complex was placed on an orthogonal box in water, with added NaCl and Na+ for charge neutralization and final salt concentration of 50mM. Minimization and equilibration of the system followed the protocol described in chapter 2 but with 1fs time step instead of 2fs. The CB10 plane defines the xy plane and after 100ps of traditional MD pulling force is applied along the z axis with a spring constant of 1000 kJ /mol /nm^2 to separate the two centers of mass (cylinder and CB10) during a 300ps trajectory. No additional constraints are applied on rotations on either molecule to avoid inducing bias and the process is replicated 100 times in each case to accommodate stochastic variability. The same process was applied to the pseudo-rotaxane (Ni L') and the methyl-pyridine rotaxane Ni L'' for broader comparison.

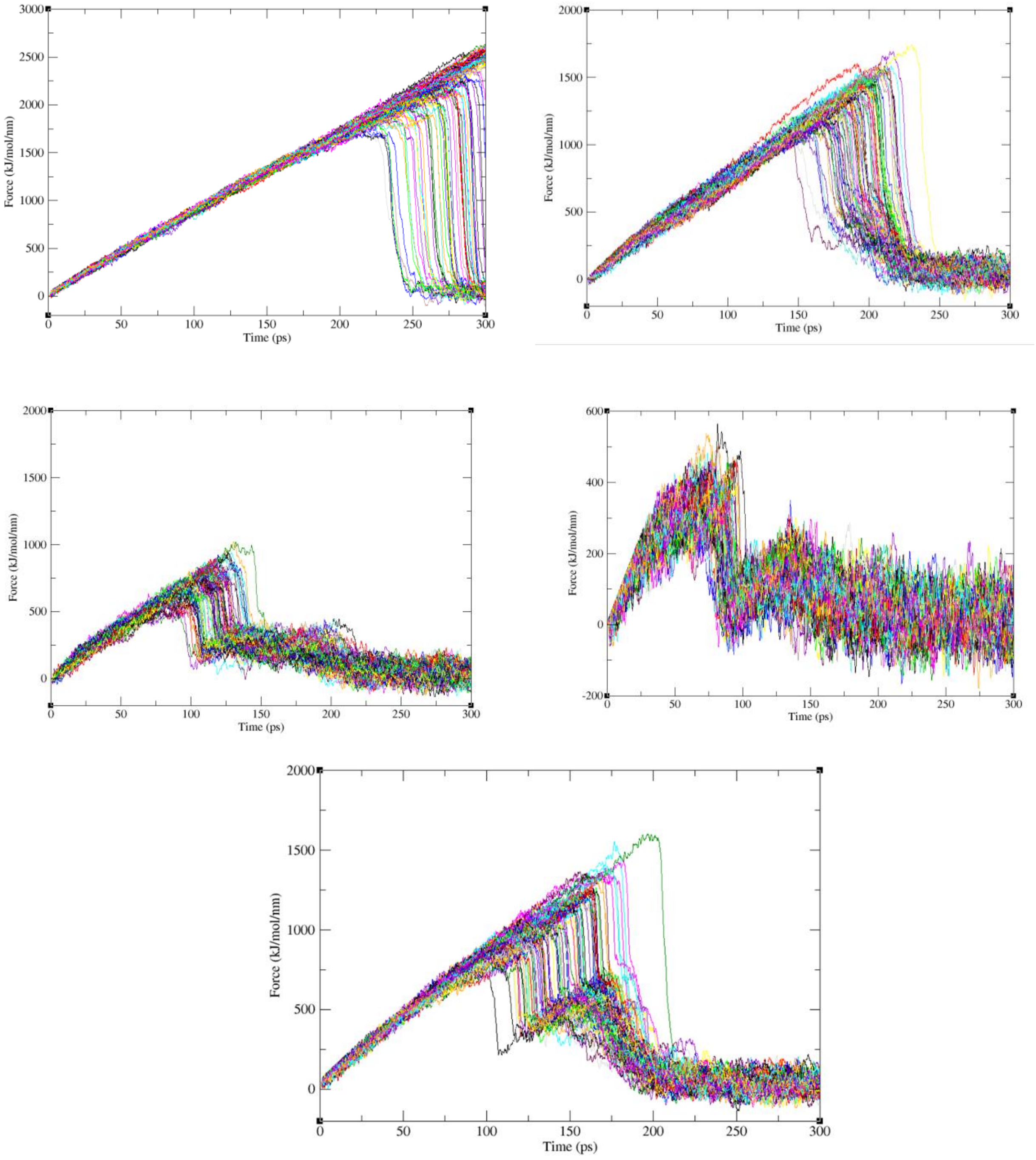


Figure 3.15 *Force vs time results of the pulling experiments in sequential order; fully capped DNB rotaxane, 3-2 DNB cylinder-CB10, 3-1 DNB cylinder CB10, pseudorotaxane (Ni L'), original rotaxane (Ni L'').*

3.5 Hannon G4 Ligands (bis-biisoquinoline Pt and Pd)

Back in 2014 the Hannon group developed bis-biisoquinoline metal complexes with Pd and Pt centers that showed to be excellent G quadruplex binding[35] agents in vitro. Although Pt and Pd centers have almost identical crystal structures, different metal centers showed different binding behaviour, and this provided a great opportunity to explore the binding under a computational lens.

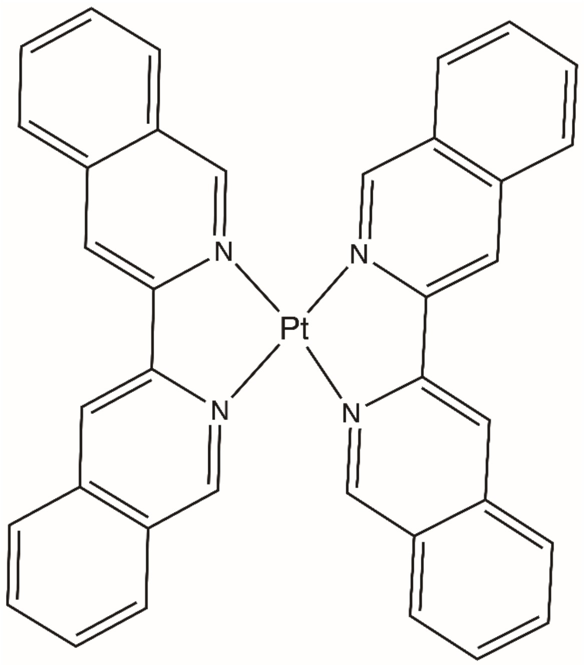


Figure 3.16 *bis-biisoquinoline Pt complex.*

Starting from the crystal structures the two compounds (Pt and Pd) have been optimized under SSB-d/Def2-SVP level of theory before they were parameterized as previously described. The interaction of these compounds with G quadruplexes and the limitations of the theoretical framework used will be further discussed in chapter 6.

For the purposes of this chapter it would be worth presenting the top 6 HOMO for each metal center and the attached molecular dynamics simulation of the structures.

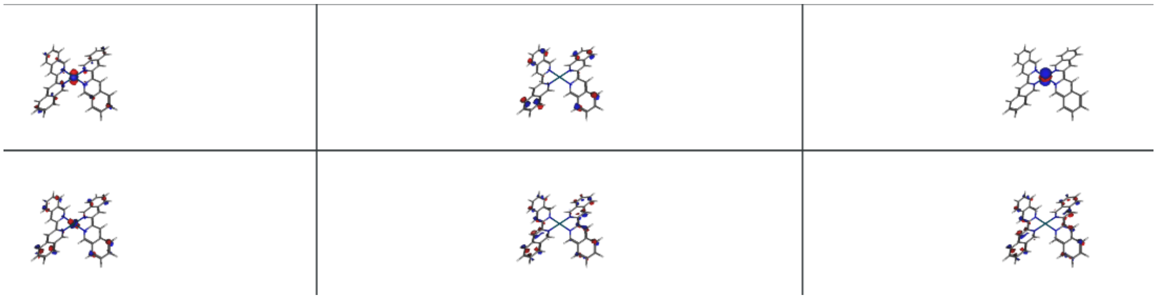


Figure 3.17 *top 6 HOMO of Pd biisoquinoline.*

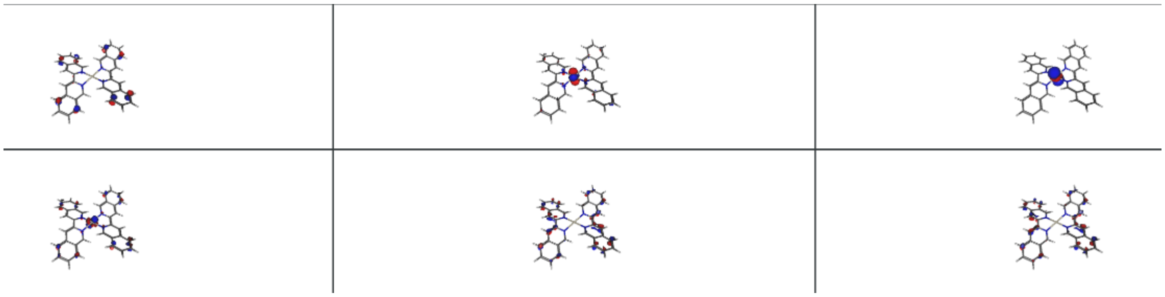


Figure 3.18 *top 6 HOMO of Pt biisoquinoline.*

3.6 Conclusions

This chapter presents the results of DFT parameterization of the compounds for MD which are used throughout the thesis and opens the discussion on how the dynamic landscape of electronic orbital states correlates to the in cellulo behaviour of a compound (where the probability of chemical interactions with other species is increased). CB10 and its interaction with the cylinders has also been discussed with an effort to quantify the ease of disassociation in partially capped cases. With the exception of the published data, deeper understanding and further simulations in combination with experiments is needed to fully understand the electronic and dynamic structure of these compounds on their own merit.

3.7 References

[1] B. J. S. Griffith, L. E. Orgel, *Q. Rev. Chem Soc.* **1957**.

[2] C. C. J. Roothaan, *Rev. Mod. Phys.* **1951**, *23*, 69–89.

[3] H. B. Gray, C. J. Ballhausen, *Inorg. Chem.* **1962**, *85*, 260–264.

[4] H.A. Jahn and E. Teller, *Proc. R. Soc. London* **1937**, *161*, 220–235.

[5] J. C. Slonczewski, *Phys. Rev.* **1963**, *131*, 1596–1610.

[6] C. F. Schwenk, B. M. Rode, *ChemPhysChem* **2003**, *4*, 931–943.

[7] S. V. Streltsov, D. I. Khomskii, *Phys. Rev. X* **2020**, *10*, 31043.

[8] K. R. Nandipati, O. Vendrell, *Phys. Rev. Res.* **2021**, *3*, 0–4.

[9] M. Munzarova, M. Kaupp, *J. Phys. Chem. A* **1999**, *103*, 9966–9983.

[10] B. Lapo, W. Wolfgang, *Nat. Mater.* **2008**, *7*, 179.

[11] M. Atanasov, D. Aravena, E. Suturina, E. Bill, D. Maganas, F. Neese, *Coord. Chem. Rev.* **2015**, *289–290*, 177–214.

[12] G. Aromí, E. K. Brechin, *Struct. Bond.* **2006**, *122*, 1–67.

[13] P. Dyszel, J. T. Haraldsen, *Magnetochemistry* **2021**, *7*, 1–20.

[14] A. Michalowicz, J. Moscovici, J. Charton, F. Sandid, F. Benamrane, Y. Garcia, *J. Synchrotron Radiat.* **2001**, *8*, 701–703.

[15] R. Jakobi, H. Spiering, P. Gütlich, *J. Phys. Chem. Solids* **1992**, *53*, 267–275.

[16] G. Battistuzzi, M. Bellei, M. Zederbauer, P. G. Furtmüller, M. Sola, C. Obinger, *Biochemistry* **2006**, *45*, 12750–12755.

[17] G. Félix, W. Nicolazzi, M. Mikolasek, G. Molnár, A. Bousseksou, *Phys. Chem. Chem. Phys.* **2014**, *16*, 7358–7367.

[18] S. G. Tabrizi, A. V. Arbuznikov, M. Kaupp, *J. Phys. Chem. A* **2016**, *120*, 6864–6879.

[19] C. J. Schattenberg, T. M. Maier, M. Kaupp, *J. Chem. Theory Comput.* **2018**, *14*, 5653–5672.

[20] E. J. Reijerse, V. Pelmenschikov, J. A. Birrell, C. P. Richers, M. Kaupp, T. B. Rauchfuss, S. P. Cramer, W. Lubitz, *J. Phys. Chem. Lett.* **2019**, *10*, 6794–6799.

[21] F. Plasser, S. Mai, M. Fumanal, E. Gindensperger, C. Daniel, L. González, *J. Chem. Theory Comput.* **2019**, *15*, 5031–5045.

[22] T. Uelisson da Silva, E. Tomaz da Silva, K. de Carvalho Pougy, C. Henrique da Silva Lima, S. de Paula Machado, *Inorg. Chem. Commun.* **2022**, *135*, DOI 10.1016/j.inoche.2021.109120.

[23] M. Abedi, G. Levi, D. B. Zederkof, N. E. Henriksen, M. Pápai, K. B. Møller, *Phys. Chem. Chem. Phys.* **2019**, *21*, 4082–4095.

[24] J. P. Zobel, L. González, *JACS Au* **2021**, *1*, 1116–1140.

[25] M. He, J. M. Lehn, *Chem. A Eur. J.* **2021**, *27*, 7516–7524.

[26] R. Gu, J. M. Lehn, *J. Am. Chem. Soc.* **2021**, *143*, 14136–14146.

[27] P. Li, K. M. Merz, *J. Chem. Inf. Model.* **2016**, *56*, 599–604.

[28] C. A. J. Hooper, L. Cardo, J. S. Craig, L. Melidis, A. Garai, R. T. Egan, V. Sadovnikova, F. Burkert, L. Male, N. J. Hodges, D. F. Browning, R. Rosas, F. Liu, F. V. Rocha, M. A. Lima, S. Liu, D. Bardelang, M. J. Hannon, *J. Am. Chem. Soc.* **2020**, *142*, 20651–20660.

[29] L. Melidis, H. Hill, N. Coltman, S. Davies, K. Winczura, T. Chauhan, J. Craig, A. Garai, C. Hooper, R. Egan, J. McKeating, N. Hodges, Z. Stamataki, P. Grzechnik, M. Hannon, *Angew. Chemie Int. Ed.* **2021**, 133, 18292-18299.

[30] L. Melidis, I. B. Styles, M. J. Hannon, *Chem. Sci.* **2021,** 12, 7174-7184.

[31] J. W. Lee, S. Samal, N. Selvapalam, H. J. Kim, K. Kim, *Acc. Chem. Res.* **2003**, *36*, 621–630.

[32] X. Ling, S. Saretz, L. Xiao, J. Francescon, E. Masson, *Chem. Sci.* **2016**, *7*, 3569–3573.

[33] R. Joseph, A. Nkrumah, R. J. Clark, E. Masson, *J. Am. Chem. Soc.* **2014**, *136*, 6602–6607.

[34] S. J. Barrow, S. Kasera, M. J. Rowland, J. Del Barrio, O. A. Scherman, *Chem. Rev.* **2015**, *115*, 12320–12406.

[35] H. L. Pritchard, **2015**. Recognition agents for DNA and RNA quadruplex structures

Chapter 4

Informative statement regarding the chapter

This chapter has been published in Chemical Science; 2021 Apr 5;12(20):7174-7184. doi: 10.1039/d1sc00933h.

In this chapter, multiple multi-microsecond simulations starting from different NMR solutions of an RNA structure (HIV-1 TAR) are performed for the first time allowing a non-biased enrichment of the sampled conformational space. Having validated that the forcefield used can retain experimental observations in conformations away from the first minimum the cylinder is introduced in the system. For consistency RNA and ligand parameters are created at the same level of theory. Through 100s of microseconds all the possible interactions between RNA and ligand are mapped and quantified using Markov state modelling.

In the appendix, extensive docking studies are presented using all the NMR solutions as well as closer examination of the results and the methodology developed.

Equally important RNA structures from other viruses have been analysed to a lesser extent (in terms of simulated time) but under the same consideration regarding the dynamics of base paring in ambiguous regions.

Supplementary information referred to in the chapter are attached as appendix I in this thesis.

Targeting structural features of viral genomes with a nano-sized supramolecular drug

Abstract

RNA targeting is an exciting frontier for drug design. Intriguing targets include functional RNA structures in structurally-conserved untranslated regions (UTRs) of many lethal viruses. However, computational docking screens, valuable in protein structure targeting, fail for inherently flexible RNA. Herein I harness MD simulations with Markov state modeling to enable nanosize metallo-supramolecular cylinders to explore the dynamic RNA conformational landscape of HIV-1 TAR untranslated region RNA (representative for many viruses) replicating experimental observations. These cylinders are exciting as they have unprecedented nucleic acid binding and are the first supramolecular helicates shown to have anti-viral activity in cellulo: the approach developed in this study provides additional new insight about how such viral UTR structures might be targeted with the cylinder binding into the heart of an RNA-bulge cavity, how that reduces the conformational flexibility of the RNA and molecular details of the insertion mechanism. The approach and understanding developed represents a new roadmap for design of supramolecular drugs to target RNA structural motifs across biology and nucleic acid nanoscience.

4.1 Introduction

Infectious disease represents one of the greatest current threats to humans as demonstrated by the frequency of recent lethal viral outbreaks: 4 out of the 10 greatest threats identified by the World Health Organization are viral related. While vaccines offer long-term eradication or suppression, they are bespoke to the disease and their development and implementation across a global population is slow. There is therefore a pressing need for a new generation of drugs that could hold an emerging disease at bay while bespoke solutions are created; broad-acting anti-viral agents having different molecular designs and molecular targets, offering a diverse platform that maximizes the potential preventative effect against new diseases.

Modern drug research tends to focus primarily on the protein targets as the effectors of disease. However, to target broad classes of disease, drugs that target the nucleic acids[1][2][3][4][5] (DNA, RNA) of the infectious agents are of particular interest with RNA increasingly recognized as a druggable target.[6][7] The rapid emergence of infections, and subsequent rapid evolution of viral genetic sequences, means that drugs that target a specific sequence are unsuitable. However, agents that target a specific nucleic-acid structure could be much more interesting. In particular, the untranslated regions (UTR) at both 3′ and 5′ ends of many viral genomes are not only highly structured but often share common structural elements[6][7][8][9][10] that are functionally essential and so conserved as the virus evolves (drifts) genetically[10][11] Indeed, structure-affecting mutations in the UTR have been used to create live attenuated or inactivated vaccine strains[12][13] .UTRs have been mostly studied in RNA viruses, such as HIVs,[7][10][14][15] coronaviruses,[16][17][18] dengue[11] [19][20] zika[21] and other flaviviruses[22] https://pubs.rsc.org/en/content/articlelanding/2021/sc/d1sc00933h - cit22 and, in every studied case, functional involvement of the UTR has been shown in either initiation of replication[16][20] (by recruiting proteins or by direct interaction with the ribosome) or regulation of the replication cycle. The most studied example is the retrovirus HIV-1 which contains a bulge in the first stem loop of the 5′ UTR of its RNA genome[23][24][25][26][27][28][29][30], the structure and dynamics of which are crucial for initiation of viral replication. Similar bulges are found in UTRs of other RNA viruses including coronaviruses and SARS-COV-2. These UTR structures represent exciting potential antiviral targets.

Structure-based recognition of RNA (and DNA) by drugs is still very much in its infancy[4][31][32][33][34] .The molecular structural information needed for such recognition is not yet available for most viruses, and crystal structures of drugs bound to RNA structures are rare (and not necessarily representative); new molecular-level understanding of such binding is a critical need. Structural studies on RNA are further complicated by the inherent flexibility of RNA molecules, which requires an understanding of their dynamics not just their ground state conformation. Consequently, simple molecular docking will not suffice; by contrast molecular dynamics potentially allows the energy landscape and structural flexibility to be probed. Herein we employ molecular dynamics to explore in detail, for the first time, a nano-scale drug inserting into a bulge in a UTR viral RNA, replicating experimental observations and gaining fundamental new insight into the dynamics of the RNA and of the drug entry process; crucial intelligence to inform design of new UTR-structure-targeting drugs. The nano-scale drugs studied are supramolecular cylinders, which not only have unprecedented RNA bulge-binding ability but are the first in class of metallo-supramolecular architectures to show potent anti-viral activity in cellular assays[35]. There is a growing interest in the application of metallo-supramolecular architectures in biology[36][37][38][39][40].

4.2 Results and discussion

As a suitable UTR structure for our studies we chose HIV-1 TAR RNA which is both experimentally well described and representative of wider viral UTR structural motifs. As a drug we chose a nanoscale metallo-supramolecular cylinder because it is unique as a nano-drug that has previously been crystallographically characterised when bound within an RNA cavity (a perfect three-way junction (3WJ)) (Figure 1)[41][42]. It is also unique in threading through an RNA cavity, interacting with all of the internal structure. These cylinders also bind bulge structures in RNA, prevent TAT protein from recognizing the binding site in the TAR sequence of HIV[35][43] and arrest HIV replication in mammalian cells. The strong evidence of binding and in-cell efficacy, makes this an ideal test-bed to investigate whether molecular dynamics simulations can identify the processes that underpin the kinetics of targeting highly flexible RNA strands. At the same time, it provides a suitable challenging size of drug, and one with large, nanoscale, 3-dimensional molecular surfaces whose match and strong binding to the 3D shape of RNA structural motifs should collapse the RNA's conformational landscape to a non-functional (impotent) state. The cylinder exists in two enantiomeric forms, both of which bind RNA bulges. Experimental X-ray crystal structures are also available for unbound cylinders;[44][45] the calculated DFT structures herein are almost identical.

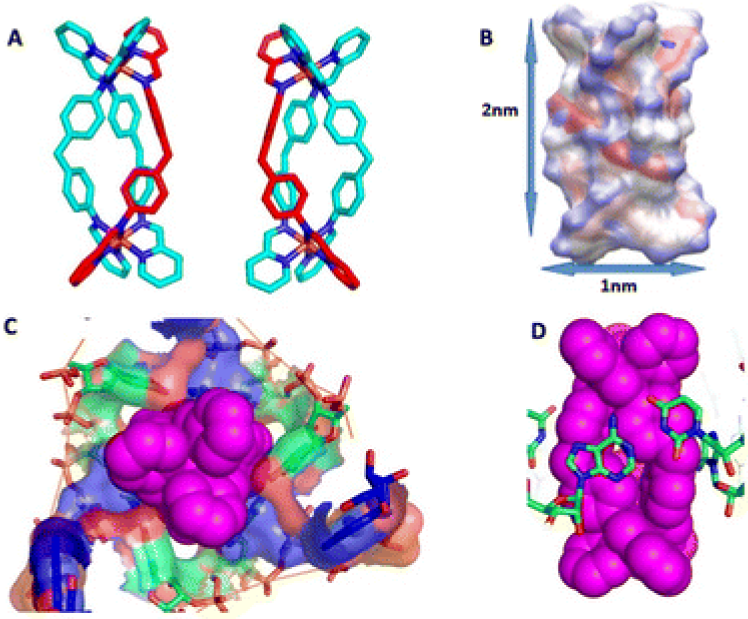


Figure 4.1 *(A) P and M enantiomers of the iron cylinder [Fe_2L_3]$^{4+}$ optimized by DFT. (B) Distribution of partial charges for P enantiomer as calculated by ssb-d-D3/Def2-SVP level of DFT theory, visualized by VMD, and also showing approximate cylinder size. (C) Surface of the RNA 3-way junction cavity stabilized by the M enantiomer of the cylinder from the crystal structure pdb 4JIY.[43] (D) Stacking of RNA bases to the cylinder in the centre of the 3-way junction in pdb 4JIY.[43] Analogous stacking is also seen with cylinders located at the terminal base pairs of the strands (see ESI part B‡). Hydrogen are omitted for clarity in A, C and D.*

4.3 Simulations of RNAs (uncomplexed)

For multi-microsecond simulations, classical MD forcefields describing the dynamics of both RNA and DNA have until very recently[46][47][48][49][50][51][52] been found to be unsatisfactory – over such timescales they induced structures not seen experimentally. With longer simulations being available, the conformational space sampled can deviate further from the absolute minimum energy point and explore the importance of non-covalent interaction dynamics as pi-stacking and hydrogen bonding [52][53][54][55][56][57] . However new forcefields[50][48]https://pubs.rsc.org/en/content/articlelanding/2021/sc/d1sc00933h - cit50 have become available and I show now that the Rochester-Mathews forcefield[50][52] can be used to simulate RNA over long timescales, reproducibly, not only for free RNA but for drug–bound complexes. The Rochester-Mathews forcefield is publicly available giving it the potential to be accessed and implemented by all. It uses the same underpinning level of DFT theory as that applied to metal-containing cylinder coordination compounds creating an overall consistency. Moreover, there are ways to accurately model NMR ensembles of RNA structures without the need of extensive MD simulations[58]. Collectively we accumulated over 200 µs of simulated time; such long and data-rich simulations on a flexible RNA system, brought new challenges in analysis. we address these by applying Markov state modeling[59][60] to the problem and show that this enables us to identify stable and metastable conformations among the millions of frames.

Overall, we have performed 123 simulations of at least 1 µs and up to 10 µs, overall ∼200 µs including several shorter runs with varying initial conditions. To analyse the vast volume of data, over 200 000 000 coordinate frames, we employ the PyEmma workflow[60] and Markov State Modelling (MSM). This involves reducing the dimensionality by choosing appropriate features of the simulation and identifying macrostates of each simulations using MSM and extracting those metastable structures with Perron-cluster cluster analysis (PCCA). Those extracted structures and the whole simulation are also presented in the Leontis–Westholf[61] nomenclature using Barnaba[62]. A detailed explanation of this workflow is included in ESI.‡

To confirm the ability of the forcefield[49][50][51] to conserve structural features of viral stem-loop RNAs (as observed, dynamically, in NMR), and to establish the effectiveness of our approach to analysis, we first explored the dynamics of poliovirus stem loop (pdb: 2GRW)[63] coxsackievirus stem loop (pdb: 1RFR)[64] and HIV2-TAR (pdb: 1AJU)[65] RNA with no bound drugs. The simulations reliably reproduced NMR observations for the stem loops (including regions of non-Watson–Crick pairing) and the predicted effect of a small bound ligand on the HIV2 TAR. Indeed for poliovirus stem-loop, the MD simulations reveal and explain features that are observed in the NMR structural data, but have not previously been satisfactorily captured in the deposited conformations, and for HIV2-TAR shows how the ligand-free RNA structure deviates from the conformation of the bound state, demonstrating the effect a binding molecule can have on an RNA structure: a detailed analysis of these free RNA simulations is included in ESI.‡

4.4 HIV1-TAR

We now turned to a more in depth study of the dynamics of our test UTR stem-loop, the HIV-1 TAR RNA. While in the coxsackievirus, poliovirus and HIV-2 simulations we had focused on the proposed ground state of the RNA as the starting point for the simulations, now we expanded our attention beyond the ground state to look also at other conformations within the experimentally suggested (NMR; pdb 1ANR) structures. In an effort to avoid introducing biases and acceleration methods to the simulation we chose to explore the conformation landscape by starting simulations from different local minima as described in the original HIV-1 TAR NMR solution structure[66]. There are 20 NMR solutions proposed and we started from five such minima (first, third, fourth, seventh and twelfth). For each of these higher energy solutions a 2 µs simulation retained the characteristics consistent with the NMR structure and did not deviate into unnatural (loosely bound) conformations. From each starting point similar features can be observed as the simulation proceeds which indicates that the forcefield can reproduce transitions within the landscape of a few µs per solution. These unbiased MD simulations capture the conformational changes of the RNA across the energy landscape for the first time, and clearly reveal the variation possible in the RNA structure and the range of conformations sampled (and which a drug could encounter and sample). Importantly, time-lagged independent component analysis (TICA) of the trajectories (Figure 2) revealed a broad energy minimum in the ground state which shows that small perturbations in the conformation have minimal effect on the energy. Moreover, a single 10 µs long simulation (as well as an independent 6 µs long simulation) of the ground state reveals the conformational richness near to the minimum. These observations highlight the limitation of a simple docking approach for flexible RNAs.

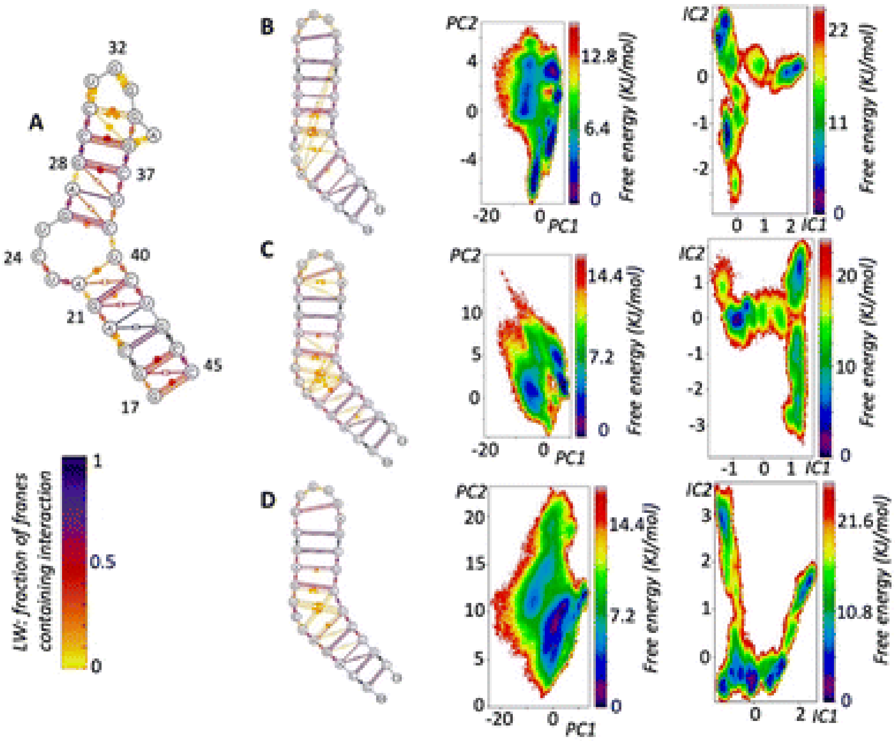


__Figure 4.2__ (A) Summary of the 1ANR 20 NMR solutions presented in Leontis Westhof (LW) nomenclature. (B) LW nomenclature of 10 µs simulation and PCA and TICA free energy surfaces, demonstrating: how the simulation reproduces 1ANR NMR structure but also reveals transient pairings (LW yellow) not well defined by (but nevertheless noted in) NMR; the greater richness of information in TICA analysis over PCA; the many conformations (TICA minima) that are accessible in the simulation at this temperature (310 K). (C) LW nomenclature of 6 µs simulation and PCA and TICA free energy surfaces. (D) Combined results of seven 2 µs simulations (see ESI Methods‡) starting with different NMR solutions.

Across the simulations, the helical regions remain relatively stable with strong WC base pairing. The only stem base pair not retaining the WC pairing is A22:U40, which often drifts apart as the

U40 retains strong stacking with C39. It is often the case that U40 seems to be in the 2nd rather than the first stem.

While a variety of transient base-pairings of all types were observed in the bulge region, as expected from the experimental NMR observations, no new stable base pairings were observed apart from that between C30 and G33/4 which is not observed by NMR but is observed in gel electrophoresis. On the un-bulged strand stacking is strong and continuous, but this is a lot less evident on the bulge strand. Of the three bulged nucleotides, U23 and C24 are more likely to stack whereas U25 is the most likely to be fully outside the helix and can even create long range interactions with the loop nucleotides (G33 and A35) creating a transient folding up of the second stem. Such a folding was not observed in the HIV-2 TAR simulation.

The loop region is characterised by limited stacking between bases and common WC pairing between C30 and G34. Transient non-WC pairing can include C30 cis or trans WC/Hoogsteen to A35.

Examining the runs starting from the different local energy minima; the first simulation starting from 1anr1 identified 3 distinct states, that can be recognised even by the PCA analysis. All are energetically and conformationally close together as seen by the RMSD and ERMSD. PCCA analysis shows one to be in much higher occupancy, clearly the ground state. A second simulation also starting from 1anr1 sampled a wider conformational space. Base pairing of stems was retained although stacking between C19:G43 and A20:U21 was not, although it is observed in the NMR. After that, stacking does continue all the way to the loop. At the loop a few different conformations were sampled that mostly gave rise to the different MSM states identified. C30 base pairs with either G33 or G34. In the bulge region U40 is stacked strongly with C39 but not always to C41 and transient, short lived pairing takes place between all bulge residues and either of C39 and U40, with pairing types including both sugar and Hoogsteen edges as well as in the trans position. MSM analysis gave 6 different states.

The simulation starting from the third NMR solution, 1anr3, yielded 5 local minima in the TiCA projections and CK test allowed for 5 states in MSM analysis. Overall, stacking and pairing throughout the stems is conserved and transient pairing within the bulge region is similar to that of the previous run. Most importantly the second state is very reminiscent of the ground state.

As we planned to apply significant external forces to the RNA structure, by introducing the cylinder into the system, we also tested the behaviour of the forcefield with higher NMR energy solutions. An experimental analysis of higher energy RNA conformations (when in the presence of a bound ligand) has been discussed by Orlovsky et al.[67] In that work, 3 nucleotide bulges are observed to adopt multiple conformations; we replicate these multiple conformations in our simulations (Figure 2B–D, ESI‡) providing further experimental validation of our model. Going up the energy ladder from the starting conformation one might expect to encounter more structures that deviate significantly from the ground state. Nevertheless starting from the fourth solution, 1anr4, most of the important structural features were retained. Pairing and stacking remains consistent with the exception of the U23 to C24 stacking. PCA revealed 3 stationary points which become 5 with TiCA. Also notable is that from this state up, examining the first 4 TiCA vectors instead of just two showed much higher diversity. In the loop, pairing C30:G34 is seen again, as well as the usual transient non-traditional pairing, but now interactions between U23 and U38 and trans Hoogsteen to sugar between U23 and C39 are observed. Stacking of U40 to C39 remains strong but stacking of U23 to C24 was less prevalent.

The seventh, 1anr7, and twelfth, 1anr12, structures are quite different from the ground state and this brings challenges for the simulation: specifically, the loss of A helix structures which is characterised by the overall elongation of G17 to G33 distance can be testing to any forcefield. Nevertheless, starting from 1anr7, the stacking and pairing remains consistent. PCA identified 2 states whereas TiCA suggested 6 states and the CK test is also passed with 6 states. The first 4 states are reminiscent of the ground state with different loop configurations, namely sugar to Hoogsteen between C30 and A35, or less often trans WC to Hoogsteen. In the other two states, U25, which generally points outside the bulge can create temporary long-range interactions with loop residue G33.

Starting from 1anr12, which is also very elongated with a sharp backbone kink in the bulge area, also retrieved most of the properties of the ground state. Pairing and stacking remain consistent for the stems. In the loop the common C30 to G34 pairing is stable along with a transient Hoogsteen to sugar between A35 and C30. In the bulge region stacking between C39 and U40 is strong and most of the transient non traditional base pairings are also seen. PCA revealed 2 states whereas TiCA revealed 5.

The results demonstrate that the forcefield can satisfactorily retain characteristics of the structure as described by the NMR experimental constraints. In addition to the unbound 1anr structure, there are some TAR RNA structures with various different bound drugs, and so for comparison we also explored as a starting point one such structure (the only solution of pdb; 1UUI)[68] from which we had removed the drug. The structure, after removing the ligand, has some differences with the 1anr structure: pairing on the stems is the same, but stacking is disturbed before the bulge, probably since U23 is WC paired with A27.

When using this as the starting point for a 2 μs simulation, the loop folded back onto the bulge (from which the ligand had been removed) forming interactions from U23 and C24 to A35, and the stem remained folded for much of the simulation. The bulge stacking did not return to the transient pairings seen in the earlier simulations. PCA analysis of the simulation revealed 3 states and TiCA 6, which was also passed the CK test on with MSM with the sixth state being ground state of this run. The simulation demonstrates how ligand binding can modify the structure and dynamics of the TAR RNA and again highlights that docking, while a useful guide, may miss key features and opportunities. The Rochester forcefield[50] behaved well for every case of RNA molecular dynamics, even in cases outside the ground state of the structure in question.

4.5 Cylinders binding to HIV1-TAR

Docking studies

Disney has recently used docking to screen libraries of small molecules binding to RNAs including TAR[51][3]. I initially undertook simple docking calculations as described in methods using all 20 structures from pdb; 1anr TAR RNA NMR study. The results are dominated by different forms of bulge region binding. While the two enantiomers do show slightly different binding energies, the Autodock Vina[69] as other docking software (used as it is one of few that allow incorporation of first row d-block metal centres) as other docking software tends to underestimate the electrostatic contribution when a charged molecule is involved. Nevertheless the docking scores are high compared to other small molecule drugs assessed by this method reflecting the larger available surface of the cylinder.

It is interesting to compare the results of docking with overall results of subsequent MD simulations. In particular in the MD simulations, capping of the open terminal bases is a transient, but relatively stable (more than 2 µs) location seen with both enantiomers. Although only a local minimum in the interaction of cylinders with TAR it highlights the limitations of docking in targeting nucleic acids because, across all 20 NMR solutions of TAR RNA, the terminal bases are coplanar only in one (the ninth). Consequently only in this structure solution does the docking reveal the end capping as a potential binding site. So docking outcomes are constrained by the rigid RNA structure(s) used in the docking, whereas in reality – as we shall see – RNAs are highly fluxional and dynamic molecules that access much structural space. Thus, while such simple docking studies are valuable for high throughput screening they might be more suited to small molecules where the molecule is less likely to have a major effect on RNA conformation. For the larger cylinders the size of the binding surface means that induced conformational change is more likely and so more sophisticated MD can offer greater insight into the interaction. Crucially, while the docking showed bulge region binding, bulge insertion by the cylinder was not observed.

4.6 Molecular dynamics simulations

To examine the interaction between TAR and the cylinders, simulations (112) started with the cylinder (DFT optimized – Figure 1A and B) in random places 1 nm away from the RNA as well as from sites identified by docking studies with initial TAR conformations derived from multiple experimental 1ANR solutions examined earlier.
The size of the cylinder restricts how rapidly it will move between sites (local minima) in the simulations' timescale. Consequently a single simulation would fail to explore all binding sites and conformations. Instead I take the quite different approach of using multiple simulations (1–10 µs) from different starting points which allows the cylinder to explore a much greater range of RNA conformations and to encounter multiple potential binding sites. By combining this with Markov state modelling analysis we are now able to explore effectively the dynamic conformational landscape of the TAR RNA – cylinder complex.

The simulations show the cylinder moving up, down and around the DNA exploring different sites and positions, and moving between them, until it ultimately inserts into the 3-base bulge. Such a dynamic exploration of different positions is what is anticipated for such a polycation with a sophisticated RNA polyanion in these timescales. There are a number of different, kinetically-accessible, positions that the cylinder explores and occupies transiently en route, of which some represent local minima with longer residence times (though still transient) and are identified from the MSM analysis (Figure S8 and S9‡). We will describe these briefly before turning to the 3-base bulge that is the ultimate binding site.

4.7 Transient end-stacking interactions

Often the cylinder (both enantiomers) found a local minimum, which it occupied for at least 1 µs at a time, and in which it capped the terminal G17:C45 bases (Figure 4.3). Some RNA forcefields have been suggested to over-emphasise base-stacking[70][71]. However, in this RNA system this binding position is among the most accessible kinetically and, since such cylinder binding has also been observed in X-ray crystal structures[41][42], it demonstrates that the simulation is replicating an experimentally validated binding location. To assess how well the forcefield and the parameterisation (now including the cylinder) reproduces this binding as captured by the crystal structure we extracted the G17:C45 bases and the cylinder from a frame of the longest lived position and I then optimised that structure at the ssb-d-D3/LANL2DZ (DFT and semi-empirical (PM7)) and superimposed it on the binding mode extracted from a crystal structure. The overlap (Figure 4.3) is extremely good, implying that the forcefield is working as desired, and that the crystallographically observed binding is reproduced. This end capping is to some extent a feature of using a shortened oligonucleotide both in these simulations and in X-ray crystal structures: it certainly does demonstrate the affinity of the cylinder for extended planar pi-surfaces, but such end capping sites are not so common in biologically active RNAs.

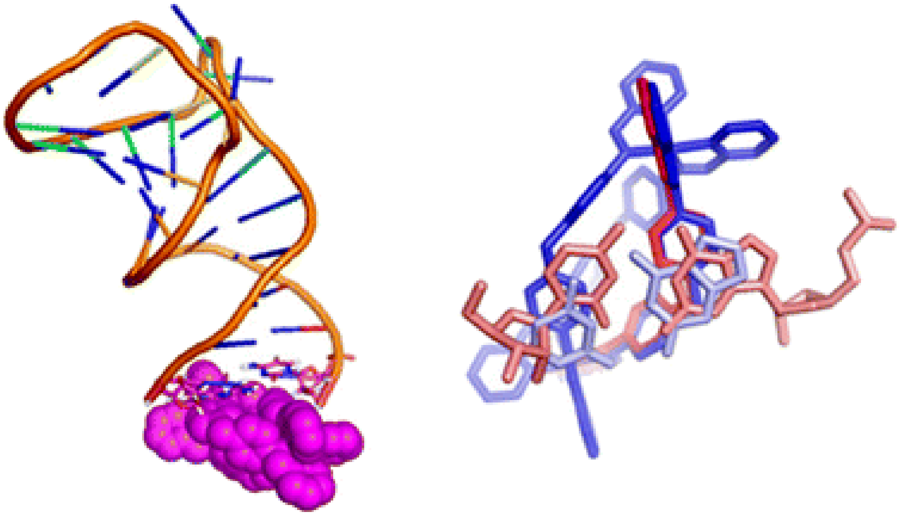


Figure 4.3 *Left: end-capping of the cylinder observed in an MD simulation. Right: the end-stacking experimentally observed in crystal structure 4JIY[42] (red), overlain with that observed in an MD simulation followed by DFT optimisation (blue).*

4.8 Transient groove interactions

The cylinder is commonly observed exploring the RNA grooves, primarily the groove of the first stem. The residence time for the M enantiomer on average is longer than for the P implying that the M enantiomer may have a higher affinity for the grooves although the kinetics were not adequately sampled to quantify difference.

4.9 Transient loop interactions

The cylinder can take advantage of unpaired open bases of the loop and interact transiently there (also seen in simulations with the coxsackievirus stem), but this is less commonly observed in the simulation compared to other locations. Loops are a common feature in RNA structures (and indeed in non-canonical DNA structures such as G-quadruplexes and i-motifs) but seem not to be a particular target for the cylinder, consistent with our experimental observations.

4.10 Transient interactions in the bulge area

The cylinder is most frequently found around or on the bulge (Figure 4) in the simulation (and as confirmed by experimental data[35][43]), with M and P being very similar in their preference for this location. RNA conformations that involve the loop bridging to the bulge (U25) can be stabilised for longer (compared to free TAR), with the cylinder sitting on top of the bridge or mediating stacking. In the absence of the bridge, the cylinder can also sit between the bulge and the opposite RNA strand, in a position in which it opens up the base pairing protecting the TAT binding site. In the case that the cylinder sits on the bulge nucleotides, it stabilises the transient base pairing and dislocates the counter ions that would normally reside there which leads to an overall elongated structure of the RNA with minimal helicity.

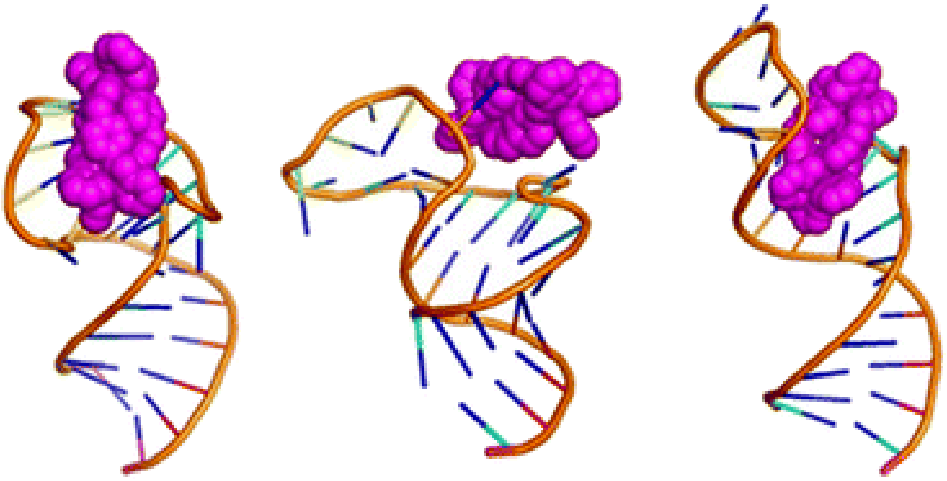


Figure 4.4 *Exemplar bulge-binding interactions observed in the simulations, en route to bulge insertion, including bridging from bulge to loop. The right hand figure has the cylinder in the position where a cyclic peptide has been observed to bind to TAR.*

In this context it is noteworthy that Keene and Collins have explored the binding of a dinuclear ruthenium polypyridyl agent (but of quite different shape to the cylinders) to a TAR-like RNA and proposed that it might bind around the groove near the bulge[72][73]. Given that the bulge-area is the most frequent location for the cylinder prior to bulge-insertion, it seems likely that this region could also be a preferred area of binding for other dinuclear complexes that cannot insert inside the bulge; for example differently shaped metallo-helices have been reported to not remain bound to TAR in electrophoresis[74], in contrast to the bulge-inserting cylinders herein,[35][43] and might be more loosely associated outside the bulge.

4.11 Bulge insertion

For both M and P enantiomers, insertion into the bulge is observed; once in the bulge the cylinder is strongly bound and remains there. In this unique binding mode, the cylinder sits in a V-shaped cleft (Figure 4.5) that resembles the 3WJ structure (Figure 4.1C). The effect of the binding is to restrict/collapse the conformational flexibility of the RNA, prevent the transient loop–bulge interactions and lessen the helicity of the stems. It is striking that, although this is the most stable binding mode in simulations, it fails to be identified in docking studies from any of the 20 1ANR solutions, because docking does not account for RNA flexibility. The bulge insertion and its effects are consistent with and explain both experimental RNase A footprinting results[75] and the ability of this cylinder to remain bound in electrophoresis when other metallo-helices do not[74].

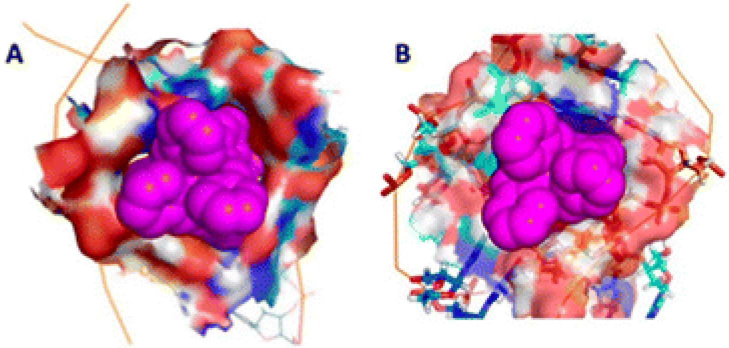


Figure 4.5 *The bulge insertion mode: the surface of the RNA cavity shows the extremely high contact surface for (A) the M enantiomer and (B) the P enantiomer, and the similarity to each other and to the 3WJ-binding (compare Figure 1C).*

The MD simulations also provide intriguing molecular-level insight into how an insertion is possible:

4.12 Entry mechanism for M enantiomer (Movies S1, S2; ‡Figure 6A–E)

The cylinder first associates with the RNA outside the bulge (Figure 4.6A and B). It interacts with the two base pairs at the bulge; A22–U40 and G26–C39. The G26–C39 base pair stacks onto a pair of phenyls (drawn from different strands of the cylinder; Figure 4.6C). The A22–U40 pair is transient and we see it both paired and unpaired and interacting (stacking) with the cylinder with the U40 having a particular tendency to stack on a phenyl even when not paired (Figure 4.6C and D). From here the mechanism of entry proceeds by two very similar processes, differing primarily in whether the A22–U40 is paired during entry or not. The entry process seems to be quicker when A22–U40 is paired, but entry can take place without this pairing (Figure 4.6E). The stacking of the paired bases A22–U40, along with the stacking of paired G26–C39 to the cylinder is effectively a V-shaped cleft about the cylinder and is reminiscent of the stacking observed in the 3WJ structure. The bulge itself is initially folded (rather than open) (Figure 4.6D) and neutralised by sodium cations, implicating the kinetic contribution of the ionic environment.

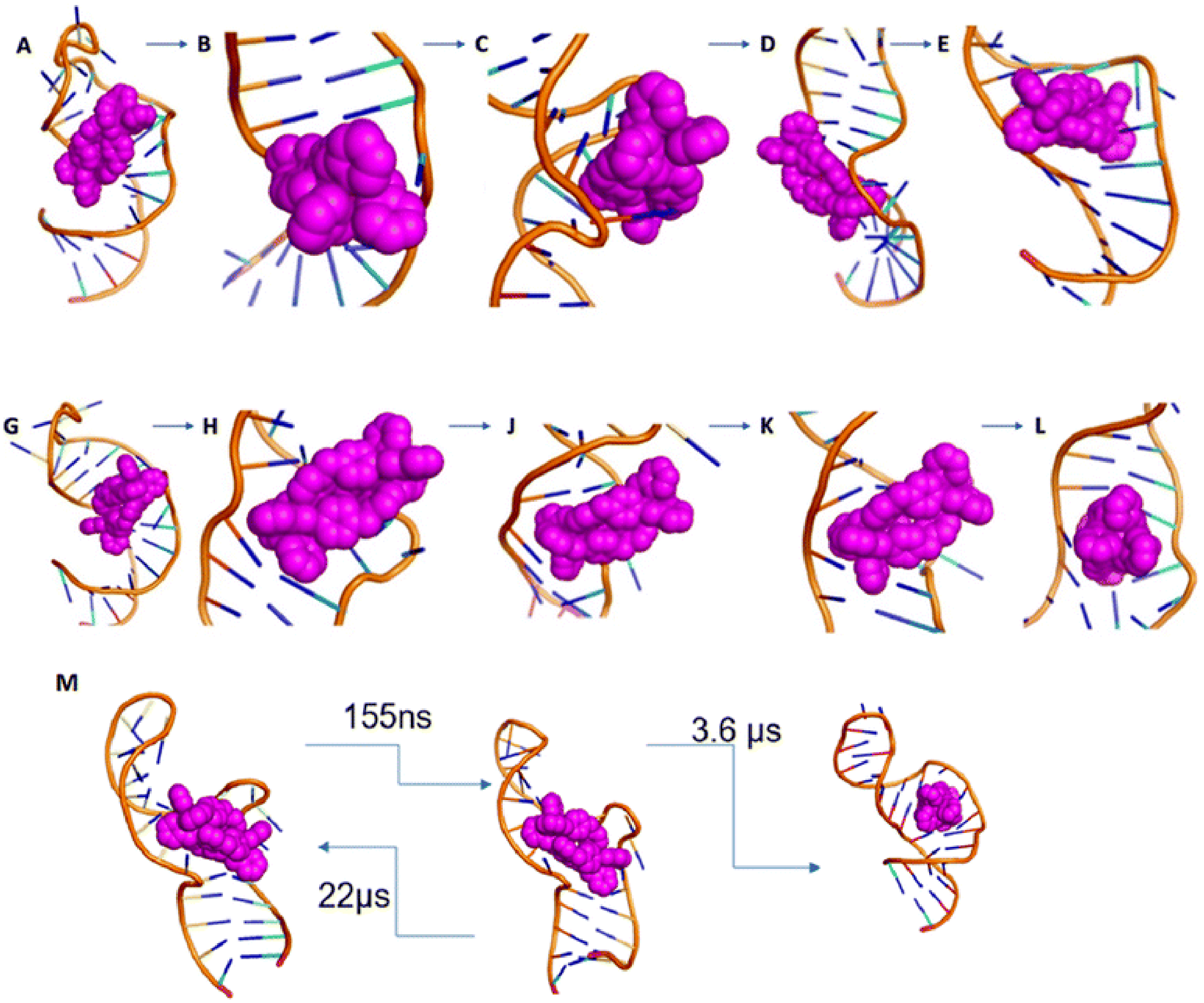


Figure 4.6 *(A–E) Entry of M enantiomer: (A) starting position of M cylinder on 1ANR1. (B) Cylinder rotates to split the U25 G26 and (C) aligns in parallel to the G26:C39 base pair (order of microseconds). (D) After relaxation of the backbone (order of microseconds), (E) the cylinder is inserted into the cavity (order of nanoseconds). In contrast to the P cylinder the M cylinder splits the C39 U40 and makes contact transiently stacking the 3 nucleotides of the bulge. (G–L) Entry of P enantiomer: (G) starting position of P cylinder on 1anr1. (H) Cylinder splits the CU nucleotides at the non-bulged strand and (J) pushes the AU base pair (order of microseconds). (K) The bp opens and the cylinder aligns parallel to the GC base pair (order of nanoseconds) and (L) after the AU closes the P cylinder is in the centre of the bulge. (M) Transition timescales for the M cylinder between states.*

As the simulation proceeds, the sodium cations leave and the bulge opens. U25 and C24 are flipped out and stack with each other. The cylinder remains stacked in the V-shaped cleft afforded by U40 (or U40–A22) and C39–G26. The cylinder starts to slide around placing its pyridyls into the bulge; these pyridyls initially encounter the sugar of U23. U25 and C24 swing back and forth with U25 also encountering the pyridyls and transiently stacking with pyridyls as does A22. The crucial point of insertion involves the cylinder stacked with G26–C39, twisting around and inserting through the centre of the bulge (Figure 4.6E). It does so facilitated by transient stacking interactions with U25, A22 and C41 which help to guide it into the cavity. With the cylinder now in the cavity, U40–A22 stack onto a pair of the cylinder phenyls, and so (re-)form the V-shaped cleft (now U40–A22; C39–G26) that is similar to two sides of the 3WJ structure. This process has been replicated in 5 independent simulations.

The MSM analysis of this entry process shows just two principal states; once the cylinder has moved from its location just outside and starts to open and enter the bulge, the energy landscape drops rapidly down into the final position where the cylinder is fully inserted and where it remains (Figure 4.6M, S147 and ESI Table 5‡).

4.13 Entry mechanism for P enantiomer (Movies S3, S4; ‡Figure 6G–L)

In the case of the P enantiomer from the same starting position (Figure 4.6G), the entry mechanism is different but has similar features. The cylinder splits the U25 G26 bulge nucleotides and still stacks the G26–C39 base pair while stabilizing it (Figure 4.6H and J). On the other side, the cylinder pyridyls press upon the A22:U40 base pair (Figure 4.6K). Within 3 ns the base pair opens, the cylinder stacking aligns to G26–C39 and the A22–U40 base pair re-forms, now enclosing the cylinder in the bulge pocket (Figure 4.6L). For the rest of the simulation the cylinder resides in the familiar triangle only this time it is splitting nucleotides U25 and G26 as opposed to C39 and U40 with the M enantiomer. U40 now plays a supportive role in stacking the cylinder phenyls and its base pairing with A22 becomes transient. This mechanism has been replicated in

4 independent simulations.

It is instructive that both cylinder enantiomers slide into the cleft down the RNA bases and locate in the V-shaped cleft of the bulge which is similar to that in the 3WJ (Figure 4.5). The longer range effect of the insertion is that the helicity of the second stem is disturbed which is consistent with the experimentally observed increased cutting of the C30:U31 by RNAase A[43].

This bulge insertion is a fascinating illustration of how a three dimensional nano-size agent might target the interior of an RNA structural feature, not by hydrogen bonding to the bases but rather by using its external pi-surfaces to recognize the surfaces inside the structure. To that extent the structure resembles a three-dimensional version of intercalation, and in that context it is notable that Barton has shown that the 'light-switch' intercalator $[Ru(bpy)_2(dppz)]^{2+}$, which doesn't intercalate into duplex RNA, can bind at RNA mismatch sites[76], where it is proposed to do so by insertion, with extrusion of the mispaired bases. The organic intercalator ethidium has been proposed to bind one-base bulges in RNA[77], and metal complexes bearing a 'phi' intercalator suggested to bind near the TAR bulge from cleavage experiments, though that is not yet well understood at a structural level[78][79][80]. This insertion of a three-dimensional structure represents a unique and exciting approach to target RNA structures.

4.14 Considerations regarding free energy landscape of RNA-cylinder complex

The simulations suggest that the binding interaction between the cylinder and the TAR-RNA should be characterised as an "induced fit" interaction, meaning that the cylinder does not recognise the bulge cavity in the traditional lock-key manner but rather it induces the precise conformation of the RNA. This complicates the free energy landscape estimation. Although we do get an idea of the landscape using TiCA and PCA we do not believe that the space is sufficiently sampled and therefore MSM probabilities only reflect the sampled space. Mmpbsa techniques cannot be used as removing the cylinder from the final complex exposes a large hydrophobic cavity and an RNA structure that is not in a minimum. Therefore, in this paper we have focused on the kinetics and mechanics of the binding process and not on the free energy estimation of the binding. However, in other systems, metadynamics and transition path sampling (TPS) have previously been applied to study the interaction of metal complexes with nucleic acids and proteins[81][82].

4.15 Methods

4.15.1 DFT of cylinders

Density functional theory optimisation of the two cylinders were performed in Nwchem 6.8.1 (ref. 83) with SSB-D[83] becke97-d[84], and TPSSh[85] with D3 dispersion correction[86] for the first two and D3BJ for the last with Def2-SVP basis set. The optimisation was performed under tight driver criteria and increased grid to xfine settings for convergence. Partial charge distribution on atomic positions was calculated with the ESP module under overall restrain of charge. Visualisation of the charge distribution at the surface was done in VMD 1.9.2 (ref. 88) on surface after converting the nwchem output .molden and .esp files to mol2.

4.15.2 Docking

Autodock vina[69] was used to create pdbqt files for all solutions of pdb 1ANR as well as the first solutions of coxackievirus stem loop and HIV-2. The cylinder structure after DFT optimisation was entered as a ligand – the searching box was big enough to contain the entire molecule and the cylinder (at least 20 Å away from the biomolecule). Exhaustiveness was set to 1000. Additional docking to just the terminal bases, specifying the docking box to the first 3 base pairs showed that only 2 out of the 20 solutions allowed for capping-mode docking.

4.15.3 Molecular dynamics simulations

Parametrization of supramolecular cylinder: already DFT optimised geometries of the cylinders were split into 5 residues (3 ligands and 2 metal ions) that were fed to MCPB.py[87] that generated parameters for the metal centres at the wB97XD9/6-31G*[85] level of theory using Gaussian09[88] as well as partial charges using RESP. The coordinate and parameter files were converted to gromacs using ParmEd (http://parmed.github.io/ParmEd/html/index.html).

Preparation of parameters with AMBER99SB was achieved with pdb2gmx program of GROMACS 2019.2[89] whereas for the ROC forcefield[50] it was achieved using tleap program of Amber18[90] and the files provided in ref. 50 The parameters and coordinates were then converted to gromacs using Parmed.

In all systems, unless otherwise stated, the RNA was put in a dodecahedral box with edges at least 1.5 nm from the solute filled with TIP3P water. Initial minimisation was carried to at least 500 kJ mol^{-1} nm^{-1} or 50 000 steps followed by heating and NVT equilibration for 1000 ps using V-rescale modified Berendsen thermostat, coupling the cylinder with the RNA at 310 K. All simulations use 2 fs time step and Parrinello–Rahman pressure coupling and PME electrostatics at 1.0 nm cut-off. Attempts to run the simulation with a 4 fs time step led quickly to blow up of the system, although 3 fs time step was more stable.

After completion the compressed trajectories (.xtc) were analysed to remove periodic boundary conditions and rotations using gromacs' trjconv program. After removing the water the trajectories were analysed with pyemma2.5.6 and pyemma 2.5.7[60], barnaba[62]. Free energy calculations used g_mmpbsa.[91]

I also explored simulations for the ruthenium cylinder (total 17.3 μs) in place of the iron cylinder. The ruthenium cylinder behaved analogously in its binding, though its movement was slower due to the increased molecular mass.

4.15.4 Simulation analysis

To analyse the simulations and identify different micro-states on the energy landscape of each run, I followed the Pyemma workflow[60]. The workflow involves principal component analysis, time dependent component analysis, and Markov state modeling and Perron cluster cluster analysis.

To identify the best features to apply the workflow to, I explored a variety of potential different features to see which best captured the kinetic variance that occurred during the simulations:

1. Position of centre of mass (COM) of each residue is a low dimensional and relatively efficient way to capture different states, including simulations that involve one or more cylinders.
2. Taking advantage of the fact that each residue has an atom named N3, which is away from the backbone, I created a matrix of distances between these N3 atoms, which although high in dimensionality captures nearly all the kinetic variance. For the cylinder simulations, we also added the distances of the metal ions (Fe or Ru) and the resulting matrix can capture adequately the kinetics of the system during the simulation.
3. The distances between the phosphorus atoms in the backbone.

Of these approaches 2 proved the most useful and was applied to all the simulations.

For each simulation, Principal Component Analysis (PCA) was carried and the projections between the first 4 PCs are plotted, followed by time-lagged independent component analysis (TICA) for lag times 1 to 5000 steps. The lag time for which the fewer number of TICA dimensions were necessary to capture 95% of the kinetic variance was chosen for further analysis. The number of clusters was chosen by examining the convergence with regards to VAMP2 as described the original paper and http://www.emma-project.org/latest/index.html. Lag times for MSM model were chosen from the convergence at timescales of identified processes. Only models that used all of the states and could pass the Chapman–Kolmogorov test were continued to Perron-cluster cluster analysis (PCCA) which led to extraction of states with certain probability and structure in pdb format. Not all simulations were long enough to produce an appropriate Markov state model, and it should be noted that the Markov state models as used here are meant to describe or sum up the particular simulations and not the whole system.

The extracted state and the full length of the simulation were analysed with Barnaba:[62] all long production molecular dynamics runs, as well as states identified by PCCA, were analysed using barnaba resulting in 2D Leontis/Westhof classification[61] of base interactions as well as E-RMSD as defined by barnaba software, RMSD and J-couplings.

4.16 Conclusions

This study provides an unprecedented platform to inform design of agents that target different important RNA structural motifs found in nucleic acid nanoscience and biology, such as this bulge cavity present in the UTR of many different viruses. We show that MD simulations, in conjunction with Markov state modeling, allow the dynamic conformational landscape of RNA to be probed and thus different and more relevant binding modes and capabilities of a potential drug to be identified; by contrast, docking to rigid RNA structures is not sufficient to guide such drug designs. The simulations provide crucial new information, not readily accessible by experiment: they show insertion of the cylinders into the cavity of the RNA bulge in a similar binding to that seen for RNA 3-way junctions; they not only provide insight into the ultimate bound structure but also its wider effect on RNA conformation reducing the RNA conformational flexibility once the cavity is bound; and, for the first time, they provide insight about the molecular mechanism through which a drug might enter a cavity in the RNA UTR, involving stacking on and sliding down bases and base pairs. Together these new molecular insights and the combined modelling and analysis approaches that have enabled them and can be more widely applied, will transform understanding of how to create supramolecular drugs that insert effectively into RNA cavities and can guide new designs against a spectrum of critical RNA viruses that threaten human well-being.

4.17 References

[1] S. P. Velagapudi, M. D. Cameron, C. L. Haga, L. H. Rosenberg, M. Lafitte, D. R. Duckett, D. G. Phinney, M. D. Disney, *Proc. Natl. Acad. Sci.* **2016**, *113*, 5898–5903.

[2] N. F. Rizvi, G. F. Smith, *Bioorganic Med. Chem. Lett.* **2017**, *27*, 5083–5088.

[3] M. D. Disney, A. J. Angelbello, *Acc. Chem. Res.* **2016**, *49*, 2698–2704.

[4] M. D. Disney, B. G. Dwyer, J. L. Childs-Disney, *Cold Spring Harb. Perspect. Biol.* **2018**, *10*, DOI 10.1101/cshperspect.a034769.

[5] G. J. R. Zaman, P. J. A. Michiels, C. A. A. Van Boeckel, *Drug Discov. Today* **2003**, *8*, 297–306.

[6] C. H. Li, Y. Chen, *Int. J. Biochem. Cell Biol.* **2013**, *45*, 1895–1910.

[7] V. V. Smirnova, I. M. Terenin, A. A. Khutornenko, D. E. Andreev, S. E. Dmitriev, I. N. Shatsky, *Biochimie* **2016**, *121*, 228–237.

[8] M. De Nova-Ocampo, M. C. Soliman, W. Espinosa-Hernández, C. Velez-del Valle, J. Salas-Benito, J. Valdés-Flores, L. García-Morales, *Mol. Biol. Rep.* **2019**, *46*, 1413–1424.

[9] J. Gilmore, K. Deguchi, K. Takeyasu, *Microsc. imaging Sci. Pract. approaches to Appl. Res. Educ.* **2017**, 300–306.

[10] R. Comandur, E. D. Olson, K. Musier-Forsyth, *Rna* **2017**, *23*, 1850–1859.

[11] S. M. Villordo, C. V. Filomatori, I. Sánchez-Vargas, C. D. Blair, A. V. Gamarnik, *PLoS Pathog.* **2015**, *11*, 1–22.

[12] R. W. Fulton, J. F. Ridpath, L. J. Burge, *Vaccine* **2013**, *31*, 886–892.

[13] E. J. Kelly, E. M. Hadac, S. Greiner, S. J. Russell, *Nat. Med.* **2008**, *14*, 1278–1283.

[14] C. K. Damgaard, E. S. Andersen, B. Knudsen, J. Gorodkin, J. Kjems, *J. Mol. Biol.* **2004**, *336*, 369–379.

[15] I. Boeras, B. Seufzer, S. Brady, A. Rendahl, X. Heng, K. Boris-Lawrie, *Sci. Rep.* **2017**, *7*, 1–10.

[16] D. Yang, J. L. Leibowitz, *Virus Res.* **2015**, *206*, 120–133.

[17] B. Hsue, P. S. Masters, *J. Virol.* **1997**, *71*, 7567–7578.

[18] L. Li, H. Kang, P. Liu, N. Makkinje, S. T. Williamson, J. L. Leibowitz, D. P. Giedroc, *J. Mol. Biol.* **2008**, *377*, 790–803.

[19] D. E. Alvarez, A. L. De Lella Ezcurra, S. Fucito, A. V. Gamarnik, *Virology* **2005**, *339*, 200–212.

[20] K. C. Liao, V. Chuo, W. C. Ng, S. P. Neo, J. Pompon, J. Gunaratne, E. E. Ooi, M. A. Garcia-Blanco, *Rna* **2018**, *24*, 803–814.

[21] A. M. Fleming, Y. Ding, A. Alenko, C. J. Burrows, *ACS Infect. Dis.* **2016**, *2*, 674–681.

[22] W. C. Ng, R. Soto-Acosta, S. S. Bradrick, M. A. Garcia-Blanco, E. E. Ooi, *Viruses* **2017**, *9*, 1–14.

[23] L. Sethaphong, A. Singh, A. E. Marlowe, Y. G. Yingling, *J. Phys. Chem. C* **2010**, *114*, 5506–5512.

[24] T. Kulinski, M. Olejniczak, H. Huthoff, L. Bielecki, K. Pachulska-Wieczorek, A. T. Das, B.

Berkhout, R. W. Adamiak, *J. Biol. Chem.* **2003**, *278*, 38892–38901.

[25] T. N. Do, E. Ippoliti, P. Carloni, G. Varani, M. Parrinello, *J. Chem. Theory Comput.* **2012**, *8*, 688–694.

[26] L. Pascale, S. Azoulay, A. Di Giorgio, L. Zenacker, M. Gaysinski, P. Clayette, N. Patino, *Nucleic Acids Res.* **2013**, *41*, 5851–5863.

[27] D. Maity, S. Kumar, F. Curreli, A. K. Debnath, A. D. Hamilton, *Chem. - A Eur. J.* **2019**, *25*, 7265–7269.

[28] M. J. Selby, E. S. Bain, P. A. Luciw, B. M. Peterlin, *Genes Dev.* **1989**, *3*, 547–558.

[29] R. Nifosi, *Nucleic Acids Res.* **2000**, *28*, 4944–4955.

[30] F. Musiani, G. Rossetti, L. Capece, T. M. Gerger, C. Micheletti, G. Varani, P. Carloni, *J. Am. Chem. Soc.* **2014**, *136*, 15631–15637.

[31] D. M. Krüger, J. Bergs, S. Kazemi, H. Gohlke, *ACS Med. Chem. Lett.* **2011**, *2*, 489–493.

[32] E. Ennifar, J. C. Paillart, A. Bodlenner, P. Walter, J. M. Weibel, A. M. Aubertin, P. Pale, P. Dumas, R. Marquet, *Nucleic Acids Res.* **2006**, *34*, 2328–2339.

[33] H. Dong, D. Ray, S. Ren, B. Zhang, F. Puig-Basagoiti, Y. Takagi, C. K. Ho, H. Li, P.-Y. Shi, *J. Virol.* **2007**, *81*, 4412–4421.

[34] H. Ling, M. Fabbri, G. A. Calin, *Nat. Rev. Drug Discov.* **2013**, *12*, 847–865.

[35] L. Cardo, I. Nawroth, P. J. Cail, J. A. McKeating, M. J. Hannon, *Sci. Rep.* **2018**, *8*, 13342.

[36] H. Sepehrpour, W. Fu, Y. Sun, P. J. Stang, *J. Am. Chem. Soc.* **2019**, *141*, 14005–14020.

[37] A. Casini, B. Woods, M. Wenzel, *Inorg. Chem.* **2017**, *56*, 14715–14729.

[38] A. Pöthig, A. Casini, *Theranostics* **2019**, *9*, 3150–3169.

[39] B. Woods, R. D. M. Silva, C. Schmidt, D. Wragg, M. Cavaco, V. Neves, V. F. C. Ferreira, L. Gano, T. S. Morais, F. Mendes, J. D. G. Correia, A. Casini, *Bioconjug. Chem.* **2021**, DOI 10.1021/acs.bioconjchem.0c00659.

[40] J. Han, A. F. B. Räder, F. Reichart, B. Aikman, M. N. Wenzel, B. Woods, M. Weinmüller, B. S. Ludwig, S. Stürup, G. M. M. Groothuis, H. P. Permentier, R. Bischoff, H. Kessler, P. Horvatovich, A. Casini, *Bioconjug. Chem.* **2018**, *29*, 3856–3865.

[41] A. Oleksi, A. G. Blanco, R. Boer, I. Usón, J. Aymamí, A. Rodger, M. J. Hannon, M. Coll, *Angew. Chemie - Int. Ed.* **2006**, *45*, 1227–1231.

[42] S. Phongtongpasuk, S. Paulus, J. Schnabl, R. K. O. Sigel, B. Spingler, M. J. Hannon, E. Freisinger, *Angew. Chemie - Int. Ed.* **2013**, *52*, 11513–11516.

[43] J. Malina, M. J. Hannon, V. Brabec, *Chem. - A Eur. J.* **2015**, *21*, 11189–11195.

[44] J. M. C. A. Kerckhoffs, J. C. Peberdy, I. Meistermann, L. J. Childs, C. J. Isaac, C. R. Pearmund, V. Reudegger, S. Khalid, N. W. Alcock, M. J. Hannon, A. Rodger, *J. Chem. Soc. Dalt. Trans.* **2006**, *2*, 734–742.

[45] G. I. Pascu, A. C. G. Hotze, C. Sanchez-cano, B. M. Kariuki, M. J. Hannon, *Angew. Chemie - Int. Ed.* **2007**, *46*, 4374–4378.

[46] J. Šponer, G. Bussi, M. Krepl, P. Banáš, S. Bottaro, R. A. Cunha, A. Gil-Ley, G. Pinamonti, S.

Poblete, P. Jurečka, N. G. Walter, M. Otyepka, *Chem. Rev.* **2018**, acs.chemrev.7b00427.

[47] A. Cesari, S. Bottaro, K. Lindorff-Larsen, P. Banáš, J. Šponer, G. Bussi, *J. Chem. Theory Comput.* **2019**, *15*, 3425–3431.

[48] D. Tan, S. Piana, R. M. Dirks, D. E. Shaw, *Proc. Natl. Acad. Sci. U. S. A.* **2018**, *115*, E1346–E1355.

[49] S. Vangaveti, S. V. Ranganathan, A. A. Chen, *Wiley Interdiscip. Rev. RNA* **2017**, *8*, DOI 10.1002/wrna.1396.

[50] A. H. Aytenfisu, A. Spasic, A. Grossfield, H. A. Stern, D. H. Mathews, *J. Chem. Theory Comput.* **2017**, *13*, 900–915.

[51] A. J. Angelbello, R. I. Benhamou, S. G. Rzuczek, S. Choudhary, Z. Tang, J. L. Chen, M. Roy, K. W. Wang, I. Yildirim, A. S. Jun, C. A. Thornton, M. D. Disney, *Cell Chem. Biol.* **2021**, *28*, 34-45.e6.

[52] D. H. Mathews, *Methods* **2019**, *162–163*, 60–67.

[53] P. D. Dans, D. Gallego, A. Balaceanu, L. Darré, H. Gómez, M. Orozco, *Chem* **2019**, *5*, 51–73.

[54] M. Zgarbová, M. Otyepka, J. Šponer, F. Lankaš, P. Jurečka, *J. Chem. Theory Comput.* **2014**, *10*, 3177–3189.

[55] N. Gresh, J. E. Sponer, M. Devereux, K. Gkionis, B. De Courcy, J. P. Piquemal, J. Sponer, *J. Phys. Chem. B* **2015**, *119*, 9477–9495.

[56] M. Thesis, J. Wang, Y. Xiao, M. Usman Mirza, S. Rafique, A. Ali, M. Munir, N. Ikram, A. Manan, O. M. H. Salo-Ahen, M. Idrees, L.-Z. Sun, D. Zhang, S.-J. Chen, R. F. Garmann, M.

Comas-Garcia, M. S. T. Koay, J. J. L. M. Cornelissen, C. M. Knobler, W. M. Gelbart, K. Zhang, S. C. Keane, Z. Su, R. N. Irobalieva, M. Chen, V. Van, C. A. Sciandra, J. Marchant, X. Heng, M. F. Schmid, D. A. Case, S. J. Ludtke, M. F. Summers, W. Chiu, B. Shu, P. Gong, G. Palermo, Y. Miao, R. C. Walker, M. Jinek, J. A. McCammon, P. Sarkies, E. A. Miska, V. K. Korboukh, C. A. Lee, A. Acevedo, M. Vignuzzi, Y. Xiao, J. Arnold, S. Hemperly, J. D. Graci, A. August, R. Andino, E. Craig, D. Yamane, D. R. McGivern, E. Wauthier, M. Yi, V. J. Madden, C. Welsch, I. Antes, Y. Wen, P. E. Chugh, C. E. McGee, D. G. Widman, I. Misumi, S. Bandyopadhyay, S. Kim, T. Shimakami, T. Oikawa, J. K. Whitmire, M. T. Heise, D. P. Dittmer, C. C. Kao, S. M. Pitson, A. H. Merrill, L. M. Reid, S. M. Lemon, S. Manjula, S. Magudeeswaran, K. Poomani, J. A. Lemkul, A. D. MacKerell, H. Yamamoto, M. Collier, J. Loerke, J. Ismer, A. Schmidt, T. Hilal, T. Sprink, K. Yamamoto, T. Mielke, J. Bürger, T. R. Shaikh, M. Dabrowski, P. W. Hildebrand, P. Scheerer, C. M. T. Spahn, I. Liko, T. M. Allison, J. T. Hopper, C. V. Robinson, M. V. Schrodt, C. T. Andrews, A. H. Elcock, E. O. Freed, G. Lichinchi, S. Gao, Y. Saletore, G. M. Gonzalez, V. Bansal, Y. Wang, C. E. Mason, T. M. Rana, B. S. Zhao, Y. Wu, Z. Lu, Y. Qin, C. He, T. M. Rana, G. M. Pocock, L. L. Zimdars, M. Yuan, K. W. Eliceiri, P. Ahlquist, N. M. Sherer, A. Savelyev, A. D. MacKerell, S. Kirmizialtin, S. P. Hennelly, A. Schug, J. N. Onuchic, K. Y. Sanbonmatsu, M. Zgarbová, M. Otyepka, J. Šponer, F. Lankaš, P. Jurečka, M. Meli, M. Gasset, G. Colombo, B. Artegiani, A. Lyubimova, M. Muraro, J. H. van Es, A. van Oudenaarden, H. Clevers, L. Casalino, G. Palermo, N. Abdurakhmonova, U. Rothlisberger, A. Magistrato, *J. Chem. Theory Comput.* **2016**, *13*, 927–935.

[57] G. Pinamonti, J. Zhao, D. E. Condon, F. Paul, F. Noè, D. H. Turner, G. Bussi, *J. Chem. Theory Comput.* **2017**, *13*, 926–934.

[58] H. Shi, A. Rangadurai, H. Abou Assi, R. Roy, D. A. Case, D. Herschlag, J. D. Yesselman, H. M. Al-Hashimi, *Nat. Commun.* **2020**, *11*, DOI 10.1038/s41467-020-19371-y.

[59] J. Copperman, D. Zuckerman, **2019**, 3–9.

[60] M. K. Scherer, B. Trendelkamp-Schroer, F. Paul, G. Pérez-Hernández, M. Hoffmann, N. Plattner, C. Wehmeyer, J. H. Prinz, F. Noé, *J. Chem. Theory Comput.* **2015**, *11*, 5525–5542.

[61] N. B. Leontis, *Nucleic Acids Res.* **2002**, *30*, 3497–3531.

[62] S. Bottaro, G. Bussi, G. Pinamonti, S. Reiber, W. Boomsma, K. Lindorff-Larsen, *Rna* **2019**, *25*, 219–231.

[63] J. A. N. Zoll, M. Tessari, F. J. M. Van Kuppeveld, W. J. G. Melchers, H. A. Heus, *Rna* **2007**, *13*, 781–792.

[64] O. Ohlenschläger, J. Wöhnert, E. Bucci, S. Seitz, S. Häfner, R. Ramachandran, R. Zell, M. Görlach, *Structure* **2004**, *12*, 237–248.

[65] K. T. Dayie, A. S. Brodsky, J. R. Williamson, *J. Mol. Biol.* **2002**, *317*, 263–278.

[66] F. Aboul-ela, J. Karn, G. Varani, *Nucleic Acids Res.* **1996**, *24*, 3974–3981.

[67] N. I. Orlovsky, H. M. Al-Hashimi, T. G. Oas, *J. Am. Chem. Soc.* **2020**, *142*, 907–921.

[68] B. Davis, M. Afshar, G. Varani, A. I. H. Murchie, J. Karn, G. Lentzen, M. Drysdale, J. Bower, A. J. Potter, I. D. Starkey, T. Swarbrick, F. Aboul-Ela, *J. Mol. Biol.* **2004**, *336*, 343–356.

[69] A. J. Trott,O., Olson, *J. Comput. Chem.* **2019**, *31*, 455–461.

[70] M. A. Ditzler, M. Otyepka, J. Šponer, N. G. Walter, *Acc. Chem. Res.* **2010**, *43*, 40–47.

[71] G. A. Bermejo, G. M. Clore, C. D. Schwieters, *Structure* **2016**, *24*, 806–815.

[72] D. P. Buck, C. B. Spillane, J. G. Collins, F. R. Keene, *Mol. Biosyst.* **2008**, *4*, 851–854.

[73] C. B. Spillane, J. A. Smith, D. P. Buck, J. G. Collins, F. R. Keene, *Dalt. Trans.* **2007**, *2*, 5290–5296.

[74] J. Malina, P. Scott, V. Brabec, *Chem. - A Eur. J.* **2020**, *26*, 8435–8442.

[75] J. Malina, M. J. Hannon, V. Brabec, *Sci. Rep.* **2016**, *6*, 29674.

[76] A. J. McConnell, H. Song, J. K. Barton, *Inorg. Chem.* **2013**, *52*, 10131–10136.

[77] W. D. W. Lynda S Ratmeyer, Ravi Vinayak, Gerald Zon, **1992**, *35*, 966–968.

[78] E. Alberti, M. Zampakou, D. Donghi, *J. Inorg. Biochem.* **2016**, *163*, 278–291.

[79] H. R. Neenhold, T. M. Rana, *Biochemistry* **1995**, *34*, 6303–6309.

[80] P. J. Carter, C. C. Cheng, H. H. Thorp, *J. Am. Chem. Soc.* **1998**, *120*, 632–642.

[81] R. C. Bernardi, M. C. R. Melo, K. Schulten, *Biochim. Biophys. Acta - Gen. Subj.* **2015**, *1850*, 872–877.

[82] D. Wragg, A. de Almeida, R. Bonsignore, F. E. Kühn, S. Leoni, A. Casini, *Angew. Chemie - Int. Ed.* **2018**, *57*, 14524–14528.

[83] M. Swart, M. Solà, F. M. Bickelhaupt, *J. Chem. Phys.* **2009**, *131*, DOI 10.1063/1.3213193.

[84] H. P. Varbanov, M. A. Jakupec, A. Roller, F. Jensen, M. Galanski, B. K. Keppler, *J. Med. Chem.* **2013**, *56*, 330–344.

[85] K. P. Kepp, *Inorg. Chem.* **2016**, *55*, 2717–2727.

[86] S. Grimme, J. Antony, T. Schwabe, C. Mück-Lichtenfeld, *Org. Biomol. Chem.* **2007**, *5*, 741–758.

[87] P. Li, K. M. Merz, *J. Chem. Inf. Model.* **2016**, *56*, 599–604.

[88] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. a. Robb, J. R. Cheeseman, J. a. Montgomery, T. Vreven, K. N. Kudin, J. C. Burant, J. M. Millam, S. S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G. a. Petersson, H. Nakatsuji, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, H. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X. Li, J. E. Knox, H. P. Hratchian, J. B. Cross, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. Ochterski, P. Y. Ayala, K. Morokuma, G. a. Voth, P. Salvador, J. J. Dannenberg, V. G. Zakrzewski, S. Dapprich, A. D. Daniels, M. C. Strain, O. Farkas, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. V. Ortiz, Q. Cui, A. G. Baboul, S. Clifford, J. Cioslowski, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. L. Martin, D. J. Fox, T. Keith, M. a. Al-Laham, C. Y. Peng, A. Nanayakkara, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, C. Gonzalez, J. a. Pople, *an Introd. To Comput. Chem. Using G09W Avogadro Softw.* **2009**, 34.

[89] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, E. Lindah, *SoftwareX* **2015**, *1–2*, 19–25.

[90] R. Salomon-Ferrer, D. A. Case, R. C. Walker, *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2013**, *3*, 198–210.

[91] R. Kumari, R. Kumar, A. Lynn, *J. Chem. Inf. Model.* **2014**, *54*, 1951–1962.

Chapter 5

Informative statement regarding the chapter

Contribution to this work published in Angewandte Chemie on 29 April 2021 https://doi.org/10.1002/ange.202104179 and previous version of the manuscript was submitted to bioRxiv doi: https://doi.org/10.1101/2021.03.30.437757

Although this chapter describes the efforts to use the cylinder as an antiviral strategy against SARS-COV2 only, the fundamental principles and applications used here were originally conceived for general viral outbreak preparedness.

The fundamental framework is the pipeline; given the sequence of a virus, classification can be made based on phylogenetic distances and therefore identification of the genomic regions that have play roles beyond the coding purpose of a sequence, ie. obtain structure that facilitates the viral replication cycle. Having the structures, one can intelligently modified existing drugs to increase specificity to the novel virus.

In the case of most (+)ssRNA viruses, some of those are in the 5’ and 3’ so the exploration of SARS-COV2 started with the 5’ UTR, applying the most recent or most citied secondary structure prediction software against the novel sequence, employing a segmentation strategy with step of 1 nucleic acid.

FARFAR2 was used to predict potential tertiary structure clusters which then were used as starting configuation for molecular dynamics, allowing us first to capture the metastable dynamics of the RNA (SL5, SL3 and SL6). The simulations suggested greater flexibility of the bulged regions therefor the use of modified cylinder with caps was suggested.

In parallel to the simulations, I co-designed electrophoresis experiments that were carried out by Dr Pawel Grzechnik, communicating results from simulations to the experimental design. Other members of the Hannon group (James Craig, Aditya Farai, Catherine Hooper, Ross Egan) provided us with the modified cylinders.

Then, having access to the SARS-COV2 virus in Dr Zania Stamataki's laboratory, we proceeded with in cellulo characterisation of the impact of the cylinders to the replication. I prepared the samples and Harriet Hill and Scott Davies cultured and dosed the cells with the virus and the cylinders. MTT assays on the cylinders were carried at the Hodges lab by Nic Coltman and me. I contacted the final image and statistical analysis of the high content imaging platform for identification of percentage of infected cells.

Finally, I located all the single nucleotide polymorphisms (SNPs) that were observed up to the date of publication and assessed them in terms of impact to the stability of the RNA structures identified.

Supramolecular Cylinders Target Bulge Structures in the 5′ UTR of the RNA Genome of SARS-CoV-2 and Inhibit Viral Replication

5.1 Abstract

The untranslated regions (UTRs) of viral genomes contain a variety of conserved yet dynamic structures crucial for viral replication, providing drug targets for the development of broad spectrum anti-virals. We combine in vitro RNA analysis with molecular dynamics simulations to build the first 3D models of the structure and dynamics of key regions of the 5′ UTR of the SARS-CoV-2 genome. Furthermore, we determine the binding of metallo-supramolecular helicates (cylinders) to this RNA structure. These nano-size agents are uniquely able to thread through RNA junctions and we identify their binding to a 3-base bulge and the central cross 4-way junction located in stem loop 5. Finally, we show these RNA-binding cylinders suppress SARS-CoV-2 replication, highlighting their potential as novel anti-viral agents.

5.2 Introduction

SARS-CoV-2 is a novel coronavirus that causes COVID-19 and as of 1st March 2021 there have been 113 267 303 recorded cases and 2 520 550 deaths worldwide.[1] Emerging so soon after other major coronavirus outbreaks (SARS, MERS), this global pandemic has highlighted the need for greater preparedness to tackle newly emergent viruses that may spread with lethal consequences. Fundamental understanding of viral processes needs to be coupled to the development of a variety of broad-acting antiviral strategies to interfere with these processes, in order to maximise the armory of drugs that we have available to treat novel pathogens. To date, antiviral drug designs have largely targeted viral proteins[2, 3] especially those with enzymic functions such as proteases and polymerases.[4, 5] An alternative antiviral approach is to target viral nucleic acid structures that are essential for replication. With current advances in sequencing technology, the sequence of a new virus can be identified within the first weeks of an outbreak, identifying both the protein coding regions and the untranslated regions (UTRs). The role of the UTRs is not completely understood for many viral families, but their conserved structures underline their functional importance. Where UTRs have been studied to determine function (retrovirus HIV-1,[6, 7] flavivirus,[8-11] to a lesser extent coronavirus[12-14]) they have been shown to have dynamic structures crucial for the viral replication.[15, 16]

These non-coding RNA regions are highly structured with multiple stem loops, bulges, crosses, and pseudo-knots, with common structural elements seen in many viral UTRs. These structures play a role in RNA-RNA interactions (both within the viral genome and with host machinery) and in protein binding for the initiation of mRNA production, translation, and viral replication. Moreover, these RNA structures may act as trans acting elements or mediate translational frameshifting, a common feature in viruses with plus-strand RNA genomes.

Nucleic acid sensors mediate the early detection and host response to virus infections, and recognise either viral nucleic acids or “unusual” cellular nucleic acids present upon infection.[17] Sensors from the RIG-I-Like Receptor (RLR) family are key pattern recognition receptors for coronaviruses[18, 19] which detect RNAs with specific structures such as 5′-triphosphate or 5′-diphosphate ends.[20, 21] Therefore UTR structures within double-stranded viral RNA provide attractive drug targets, both for direct inhibition of viral replication[13] and induction of host innate immune responses.

Compared to protein- and DNA-recognition, RNA-recognition by drugs has been much less explored. Nucleic acid recognition often focuses on sequence recognition but for RNA, which folds into complex shapes, its structure provides an opportunity for specific targeting; indeed, it is the structure of the UTR that is conserved for function, rather than sequence. Small molecule libraries have been screened for RNA binding (analogous to protein drug screens)[22-24] and agents targeting RNA structures include small molecules that hydrogen bond within the heart of trinucleotide DNA/RNA repeats,[25] and planar RNA quadruplex binders.[26-31]

We have explored nano-size metallo-supramolecular cylinders (Figure 5.**1**) as RNA-binding agents.[32] They are larger than traditional small molecules, with extensive aromatic surfaces to stack with the RNA bases (Figure **1 b**) and cationic charge (4+) that ensure strong binding and excellent shape-fit for RNA cavities. We have characterized the binding of cylinders in an RNA 3-way junction[32] by crystallography (Figure **1 c**) and showed analogous binding in an RNA bulge structure.[33, 34] Furthermore, we demonstrated cylinder binding to an RNA 3-base bulge in the TAR region of the HIV-1 genome (located in its UTR), that prevented HIV-1 replication.[34]

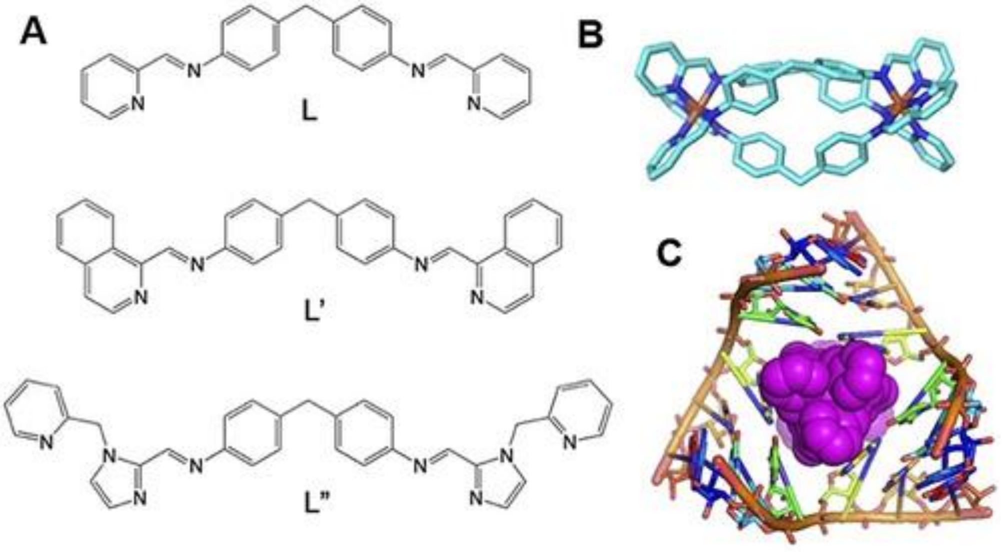


***Figure 5.1** A) Structure of the ligands used in this study. B) Structure of the $[Ni_2L_3]^{4+}$ cylinder of ligand L.[35] L',[36] and L''[37] form analogous cylinders that bear further aryl rings on their external surfaces. C) View of the crystal structure of a cylinder bound in an RNA 3-way junction cavity from PDB 4JIY[32] showing its unique binding.*

Given this anti-viral activity against HIV-1, we were interested to assess whether these cylinders would bind structures in the UTR of SARS-CoV-2. We report now combined modelling and biophysical approaches to define the 3D structures of the SARS-CoV-2 5′ UTR, and demonstrate cylinder binding to specific bulge structures in the 5′ UTR. Furthermore, we show that cylinders inhibit SARS-CoV-2 viral replication in cells.

5.3 Results and Discussion

To create a 3D dynamic model of the 5′ UTR from the published genome sequence[38] (original Wuhan strain, NC_045512), our approach was to predict the secondary structures in silico, obtain experimental evidence to verify these structures, and then model the tertiary structure and its dynamic behavior, again with experimental validation. RNA secondary structure prediction has improved dramatically over the last decade, with free energy approximations and machine learning algorithms available (adding to the attraction of the RNA as a rapid-response drug target). However, there are significant challenges with longer RNA sequences that can yield multiple distinct structures that occupy a small space in the energy landscape. We compared ≈10 folding prediction algorithms (see Supplementary Information) with many failing to cope well with the large size of the SARS-CoV-2 5′ UTR. Three representative predictions are shown in Figure **2**. The free energy RNAfold[39] and Mxfold2[40] algorithms gave similar predictions, both akin to the known UTR structures of related coronaviruses,[16, 41] while the machine learning based VFold[42] gave a quite distinct structure.

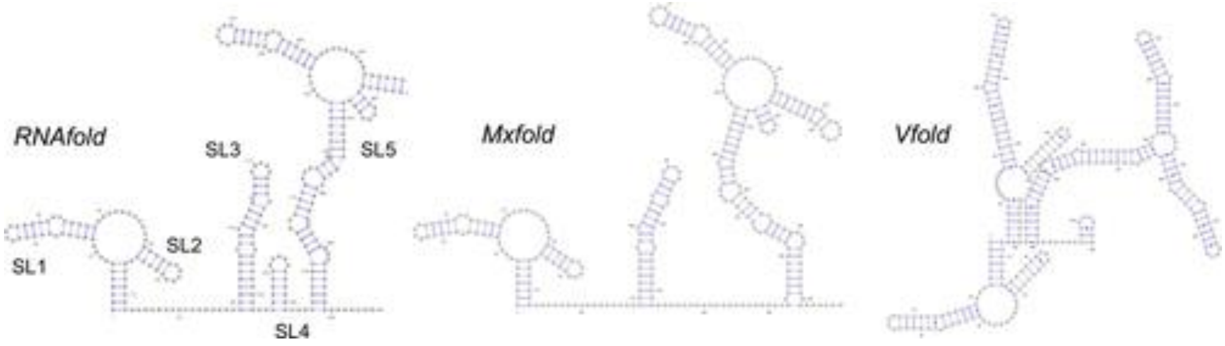


***Figure 5.2** Secondary structure predictions of the UTR of SARS-CoV-2 using three different* algorithms.

To experimentally probe the UTR, we used SHAPE, (Selective 2′-Hydroxyl Acylation Analysed by Primer Extension Sequencing) analysis where the 5′ UTR RNA sequence was first folded in vitro and the open strand (non-duplex) RNA sites (e.g. single stranded, bulges, hairpins) acylated with 1-methyl-7-nitroisatoic anhydride (1M7). These sites were then identified through a reverse transcription reaction that generates DNA fragments which end at the 1M7 tagged sites and were readily analysed by gel electrophoresis (Figure 5.**3 A**). Two primers (RT1 and RT2) conjugated with fluorescent IRDye700 were used to cover the whole 5′ UTR sequence. RT1 mapped the UTR from position +1 to +140, and RT2 the distal region of the UTR (+141 to +300).

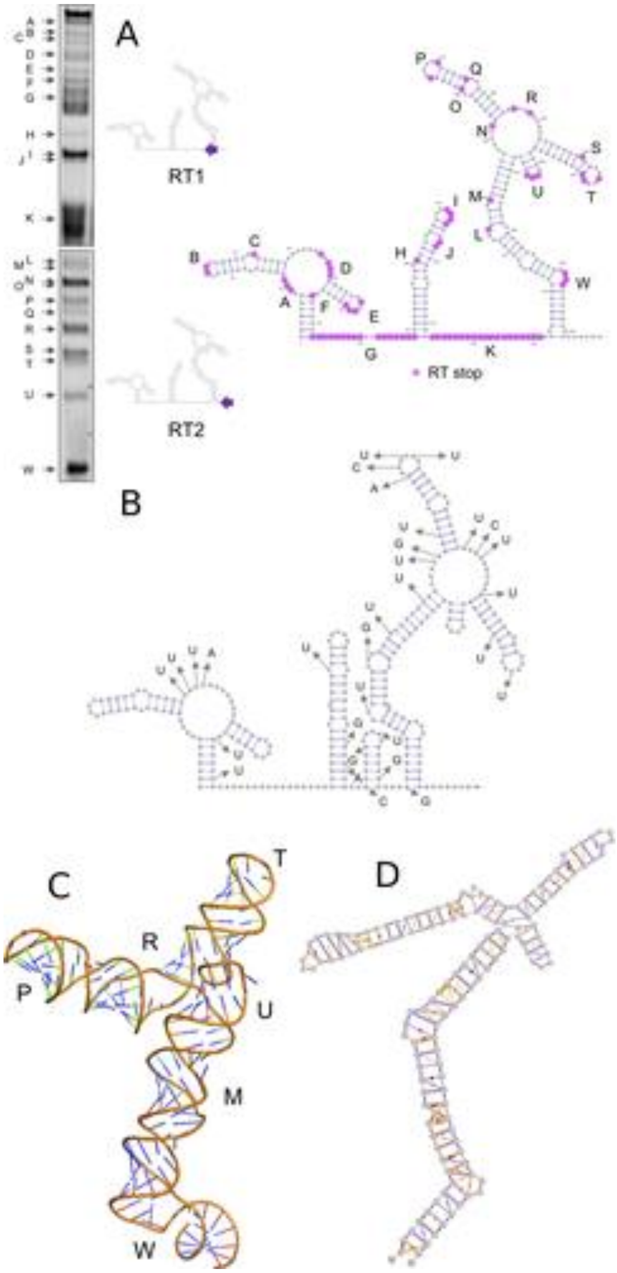


Figure 5.3 *The structure of the SARS-CoV-2 5′ UTR. A) RNA SHAPE gel results. Diagrams are included showing positions of the two IRD700 reverse transcription (RT) primers used; RT2 primer maps the whole sequence; however, longer molecules are not very well separated by electrophoresis, so RT1 was used to map the 5′ region in more detail. B) SARS-CoV-2 5′ UTR secondary structure showing the acylated nucleotides revealed by RT stops as purple dots. Open*

structures are labelled A–W. C) Positions of SNPs observed in SARS-CoV-2 viral sequences up until 7 Jan 2021. See also Figure S6 for overlay of Figure 3 B and 3 C. D) Snapshot of the dynamic three-dimensional structure of the SL5 RNA from MD simulations. E) Leontis Westhoff diagrams highlighting the dynamic base-pairing within the structure.

The results (summarized as a diagram in Figure 5.**3 B**) demonstrate that the RNAfold/Mxfold predicted structures best represent that formed in vitro. In particular, the long run of acylation around position G confirms that the Vfold prediction does not adequately describe the experimental data. The additional stem-loop (SL4) predicted by RNAfold but not Mxfold is acylated (region K) which suggests that if such a stem loop forms it may be transient. Recent studies of the whole RNA viral genome in cellulo by Miska[43] (COMRADES assay) and Pyle[44, 45] (long amplicons with SHAPE-MaP) show dynamic folding and interaction between the 5′ UTR and the 3′ UTR, but that these key stem-loop structures (SL1, 2, 3, 5 depicted in Figure **2**) are retained, affording further support and confidence that our in vitro findings are physiologically relevant.

The extensive whole-genome sequencing of SARS-CoV-2 affords the opportunity to monitor the single nucleotide polymorphism (SNPs) mutations in the 5′ UTR. We examined the available sequences in the gisaid[46] that were deposited before 7 January 2021 that contained complete 5′ UTRs. Interestingly the positions of SNPs within the UTR (Figure 5.**3 C**) often occur near the acylated positions in our SHAPE experiment (Figures 5.**3 B**, S6), suggesting that positions where the nucleotide has greater flexibility and hence less structural importance for the UTR are more likely to be substituted. Although not corrected for frequency, it is interesting to note that around 60 % (19/31) of the SNP sites identified to date involve replacement with a U residue, with the largest subset (11/31) being a C-U mutation (Figure S6). These mutations do not affect the key structures of the 5′ UTR.

After identifying the distinct stems loops (SLn) that were conserved throughout the results from the secondary structure prediction, we attempted the more challenging step of creating a 3-dimensional representation of the structure. We focused on SL3 and SL5 as they have a variety of different structural features including bulges and loops. Although the exact structure/function of

SL5 is not yet determined (to our knowledge), it contains the initiation codon and it is similar to the SL5 of SARS-CoV-1[12, 13] suggesting a functional role. Understanding the tertiary structure and behaviour from the sequence, is more complicated than predicting the secondary sequence since RNA is an inherently flexible molecule and a single static conformation will not be sufficient to understand the binding properties. Recent advances in molecular dynamics parameterization of RNA and wider availability of high-performance computer facilities can provide new insights into the dynamic structure of the RNA and show the key regions of flexibility—usually bulges and junctions, where both the secondary and tertiary structure is highly dynamic. After creating initial models using the short list of open-source software available, the ROSETTA platform (FAR-FAR2)[41, 47] gave a starting structure most consistent with the SHAPE analysis (notably the SL5 junction point having nucleotide interactions rather than being very open). We explored the dynamics around this central structure.

We employed the recent RNA-force field developed by Mathews,[48, 49] which retains NMR characteristics of RNA structures even in non-minimum starting conformations, and coupled it with Markov state modeling[50] to analyse the conformational space accessed across different simulations. We started with 3 independent 1 microsecond molecular dynamics simulations of the SL5 alone, and then performed additional 1 microsecond simulations with both enantiomers of the cylinder (three runs of at least 1 microsecond each; with parent cylinder and both enantiomers) to identify RNA regions that can be recognised by the cylinder. The simulations total 9 µs. Additionally, Markov state modelling revealed micro states where the cylinder can be positioned within the RNA helix in the bulge regions. We also performed simulations on the SL3, comprising overall 4 µs. Just as for the secondary structure predictions, the observations in the molecular dynamics of SL5 were verified experimentally by the SHAPE results, and by using these two techniques in concert we gain a molecular level understanding of the three dimensional structure and dynamic behaviour of the RNA (Figure 5.**3 C**, E), and of how the cylinder binds.

Considering the SL5 RNA in absence of cylinder, molecular dynamics reveal the following features of the stem: a) There is a bulge at G138-U140 which is highly flexible with a lot of transient stacking between its bases (region W in Figure 5.**3**). G138 base pairing with C10 elongates the bulge forcing U139-U141 to point outwards of the helical axis. This is seen experimentally in SHAPE. This sharp twist of the backbone often creates a bend to the stem. b) There is a mismatch at C15 (halfway between regions L and W) however there are many transient non-Watson-Crick base pairings between A14-A16 and C133 and those nucleotides did not produce a SHAPE signal; that is, there is no significant bulge or base flipping outwards and the helix is contiguous. c) The next bulge (U21-U25; region L) is different. Relative stability is provided by three G:C base pairs (G20:C128, C24:G126, C26:G124), causing flagging out of A23 as seen on SHAPE (region M). d) At the 4-way junction (regions N, R) the base pairings ("CUG"36-37 and "CAG"78-80) hold throughout the simulation (3 µs) creating an additional 7 nucleotide bulge on SL5a (G72-A79) where on the opposite strand there are only C38 A39. Although C38 remains stacked to G37 and transiently binds nucleotides of the opposite strand A39 lacks both strong stacking or base pairing, therefore it can be seen on SHAPE. The junction is less open (i.e. contains more pairing) than the secondary structure prediction and this is reflected in the SHAPE experiment where there is only limited acylation. e) Higher up on the SL5a CG Watson–Crick (WC) pairs create rigidity which stops on the U47, which stacks strongly on C46 allowing stable non WC base paring with U67 but leaves U48 randomly pairing U66 and G66 (region O,Q). U48 and G66 are both identified by SHAPE. The stem closes with strong CG pairings and a short loop (region P), whose bending exposes U91 and U96 and they are identified by SHAPE. f) On SL5b five CG pairs add rigidity allowing/stabilising non WC pairings. However, between C86:G100 and G89:C98 (region S) there is an additional base and as U87 and G99 strongly stack on the C86:G100 A88 is exposed and tagged by SHAPE. On the loop (region T) stacking continues strongly up to U92 and G95 creating a tight bend exposing U93. g) The short SL5c is also stabilised by 2 CG pairs and all three A residues are stacked together but point outwards of the stem (region U).

These combined simulation/experimental pictures of the RNA dynamics were then complemented by analogous SHAPE experiments and MD simulations of the SL5 RNA in the presence of the $[Fe_2L_3]^{4+}$ cylinder (Figure 5.**4**). Four batches of simulations were carried out in the presence of cylinder; for each enantiomer of the cylinder and with the cylinders positioned either away from the RNA or inside the bulges. Importantly, the MD simulations locate the cylinder binding sites on SL5 at the same positions that are affected experimentally in the SHAPE analysis, and not at the other areas of SL5 that are unaffected in SHAPE. As seen in free SL5, the bulges serve as dynamic hinges giving flexibility to the surrounding stems. In the simulations where the cylinders started away from the RNA, they quickly localized ON those hinges, reducing flexibility of the hinge drastically (in regions W, L, N, R). From studies with three base bulges (on HIV TAR) we know that such hinges can open and from such a binding position the cylinder can reorient and insert, though this can take very long on the time scales of simulations;[51] we can model this by pre-positioning the cylinder at or close to this position. The cylinders bind strongly to these structures.[32-34, 51] Once the cylinder is in the SL5 bulge (Figure 5.**4 A**, cylinder D), the simulations show that the helical structure of the surrounding stems is disturbed, opening up the stem nucleotides to attack from 1M7, and this is confirmed experimentally in SHAPE leading to an increase in the signal in these regions (around L and M and towards W, close to the RT primer).

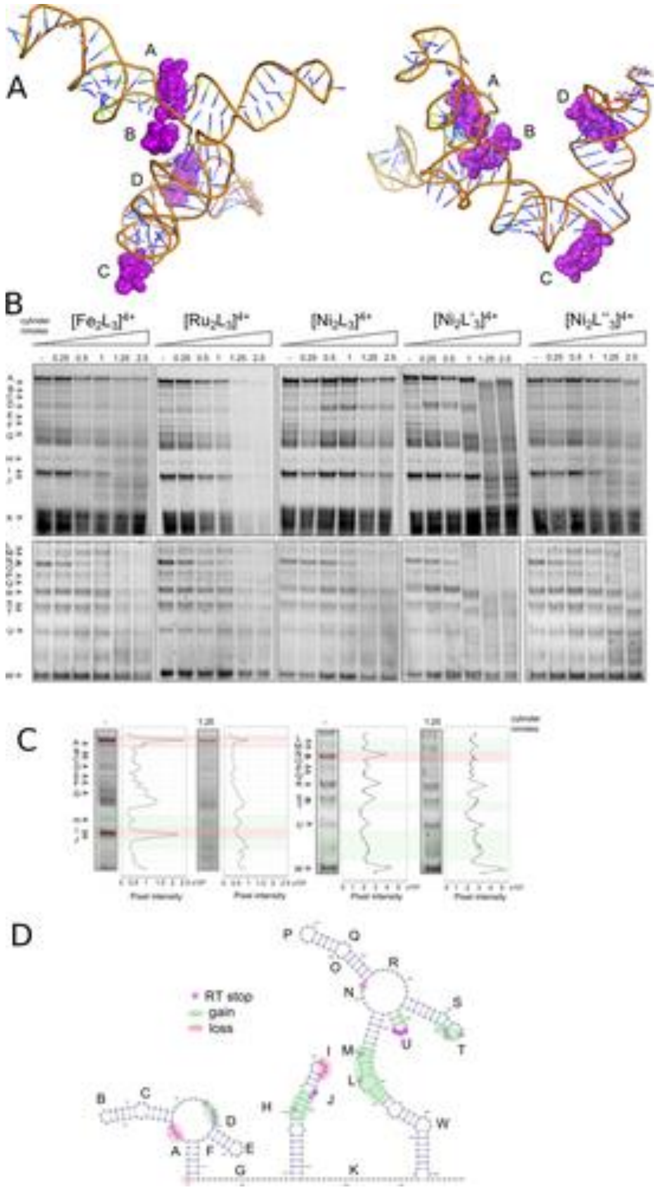


Figure 5.4 *A) View from two angles of a representative snapshot of a simulation of 4 cylinders on the SL5 RNA, revealing the same interaction points as indicated experimentally by SHAPE. Cylinder A is threaded through the central cross (4-way junction) with cylinder D threaded through the 3-base bulge at W. Cylinder B is at position N and cylinder C at position L. B) SARS-CoV-2 5′ UTR folding in the absence (lane 1) and at increasing concentrations (lanes 2–6) of five different cylinders. Cylinders were incubated with the viral 5′ UTR (0.05 nmoles) followed by SHAPE (acylation, reverse transcription, and electrophoresis). C) Band intensity of lanes 1*

(without cylinder) and 5 (with) of the [Fe_2L_3]$^{4+}$ gel. D) SARS-CoV-2 5′ UTR diagram showing the RNA regions where the folding was affected by the presence of cylinder, as indicated by SHAPE.

In addition to the bulge as a site of binding, in the simulations the cylinder can also insert into the cavity at the central cross (4-way junction) (Figure 5.**4**, cylinder A), protecting A193. This cavity is larger than the 3-base bulge and thus although the binding site may not offer as good a structural fit, it will be kinetically quite accessible. The binding also to this site was confirmed experimentally by the disappearance of this SHAPE signal (A193, RNA position N) at increased concentration of cylinder. At the loading of cylinder used in the simulation, interaction with the stems containing regions U and T was not observed. The SHAPE results suggest that these regions are also affected as the loading increases.

In SL3 there are no large bulges similar to that found in SL5, however mismatched pairs create a distortion on the helical structure that can lead to exposure of nucleotides to IM7. Specifically, molecular dynamics simulations (Figure 5.**5**) on the free RNA (no cylinder) revealed short lived pairings of different types from G96:C126 to A102:U120. Furthermore, higher up the stem U104:A118 to G106:G115 is also a region of multiple cross strand pairings. Equally important for understanding the SHAPE results is the transient stacking between this stem's nucleotides revealed in the 3D model.

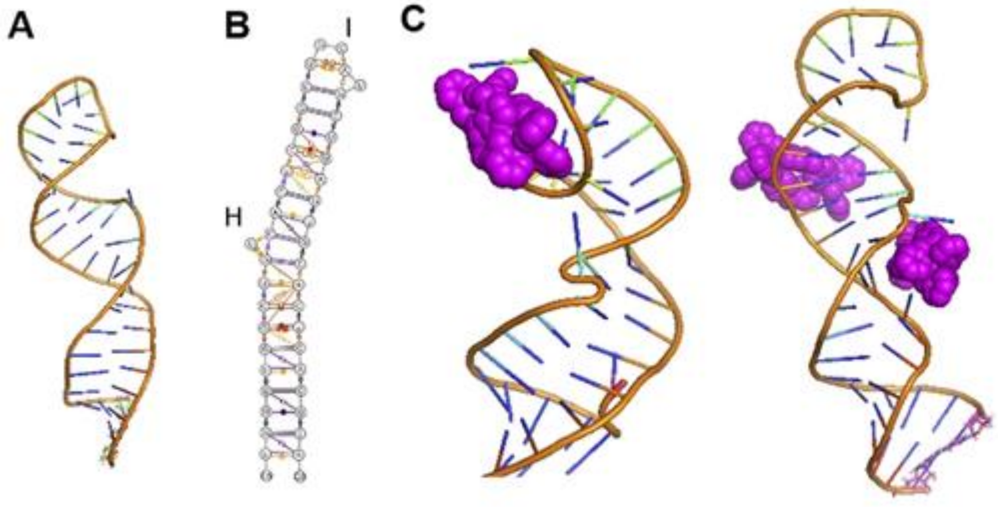


Figure 5.5 *A) Snapshot of the dynamic three-dimensional structure of the SL3 RNA from MD simulations together with a Leontis Westhoff diagram (B) highlighting the dynamic base-pairing within the structure. C) View of representative snapshots of simulations of cylinders on the SL3 RNA, showing binding at the stem loop and on the stem as also revealed by the SHAPE analysis.*

In the presence of cylinders, we observed that the cylinder is attached to the stem loop (Figure 5.**5 C**) in a stable manner, decreasing the flexibility of those residues and thus protecting the loop nucleotides from acylation, where we saw a reduced signal in SHAPE (Figure 5.**4** region I). Cylinders can also bind lower on the stem (region H/J) and this leads to an enhancement of acylation as seen on the stem of SL5.

Alongside the SHAPE experiments with the $[M_2L_3]^{4+}$ iron(II) cylinder (M=Fe), we also compared the analogous nickel(II) and ruthenium(II) cylinders (M=Ni, Ru; Figure 5.**4**). Changing the metal does not affect the overall cylinder structure or charge, and analogous patterns/effects are seen in the SHAPE mapping confirming that they bind the RNA at the same locations and it is the cylinder shape/charge that is responsible for the binding not the choice of metal. High cylinder excess (two last conditions, 1.25 and 2.5 nmoles corresponding to 25 and 50 cylinders per UTR) in most cases severely affected RNA structures and so SHAPE bands become less well defined indicating more random RT stops. In PCR experiments the $[Ru_2L_3]^{4+}$ cylinder is stable to the heat cycles and can inhibit polymerase amplification;[52] the reverse transcription efficiency seems similarly affected at the highest concentrations of this cylinder. Some small gel shifts are also observed at high cylinder loading, possibly suggesting some cylinder-binding to the DNA transcript.

We also tested the effect of two substituted cylinders based on ligands L′ and L′′, to confirm the key binding area of the cylinder design (Figure 5.**4 B**). These cylinders bear additional aryl rings at their ends while the central regions of the cylinder (which insert into the junctions/bulges) are unchanged. Both show similar patterns in the SHAPE analysis to the cylinders of ligand L, but while $[Ni_2L''_3]^{4+}$ had very a similar impact on folding, the isoquinoline cylinder $[Ni_2L'_3]^{4+}$ caused some changes in the SHAPE pattern even at the lowest cylinder concentrations. The results suggest that it may be possible to modify the cylinder structure to modulate the affinity for the binding sites.

Having established that the cylinder can bind to and modify the structure and reactivity of the SARS-CoV-2 5′ UTR in vitro, we explored their potential to inhibit viral replication in cellulo. Simian Vero cells were infected with SARS-CoV-2 virus England 2 (Wuhan strain; identical 5′ UTR to reference sequence) in the presence and absence of the Ru and Ni cylinders, $[M_2L_3]^{4+}$ (M=Ru, Ni), and the frequency of cells expressing the viral encoded spike glycoprotein quantified (Figure 5.**6**). Both cylinders reduced spike-protein-expressing cells in a dose responsive manner, with the ruthenium cylinder being more effective and reducing the frequency of infected cells to <5 % at the highest doses tested (75 µM). MTT cell metabolic activity/viability assays confirmed that the cylinder is not cytotoxic to Vero cells in the timeframe of these experiments (See Supplementary Information).

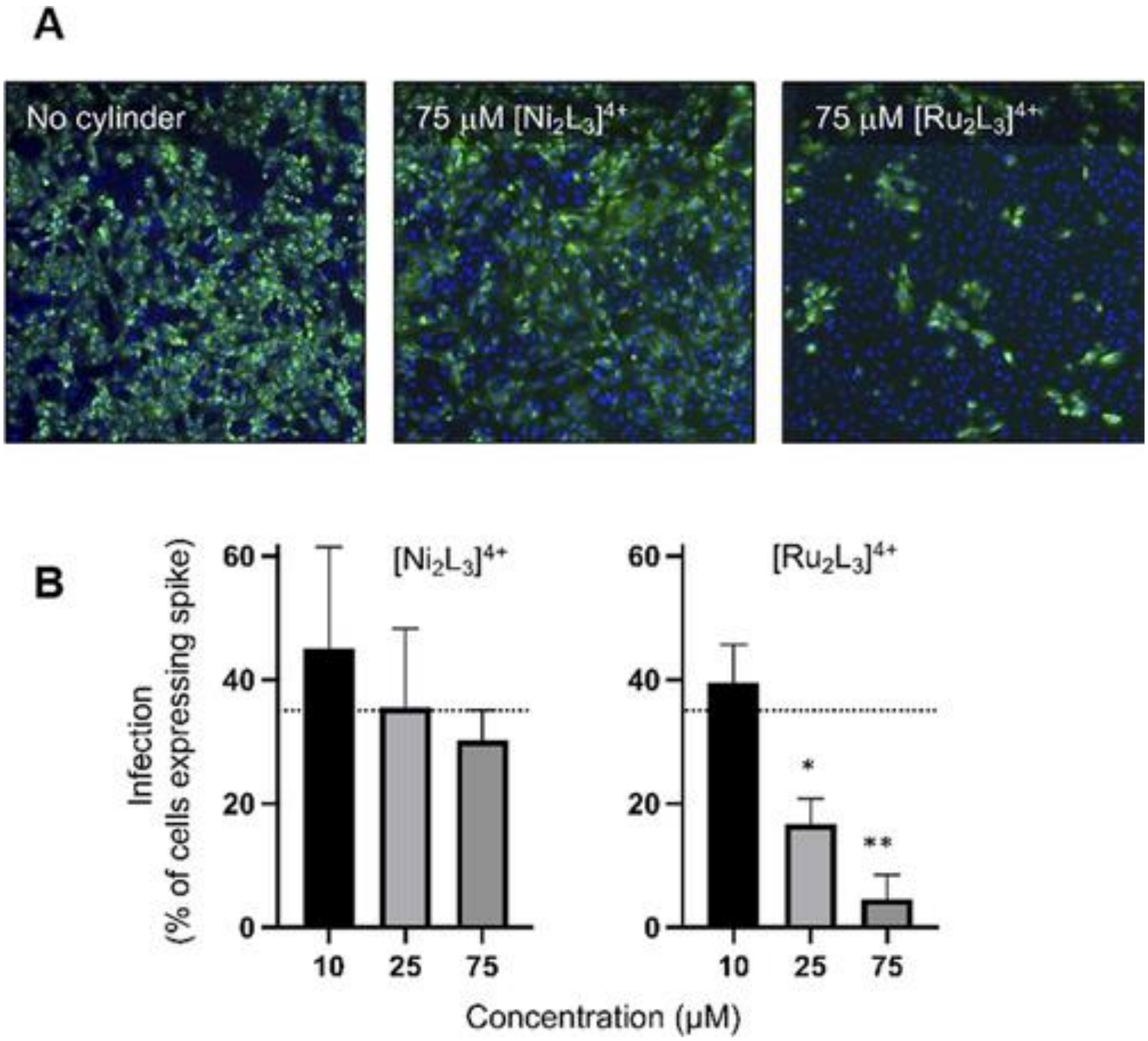


Figure 5.6 *Effects of the [M_2L_3]$^{4+}$ (M=Ru, Ni) cylinders on SARS-CoV-2 infection of Vero cells. Cells were infected with SARS-CoV-2 (MOI=0.04) in the presence or absence of cylinders and fixed at 48 hours post-infection and spike-protein expression quantified by rabbit anti-spike-protein monoclonal antibody (CR3022) and mouse anti-rabbit Alexa 555 (green). Cell nuclei were visualised with Hoechst 33342 (blue). Total cell numbers and percentage of spike-protein-expressing cells were enumerated by high content imaging at x10 magnification using a CellInsight CX5 high content microscope (Thermo Fisher Scientific). A) Representative images of untreated or 75 µM [Ni_2L_3]$^{4+}$ or [Ru_2L_3]$^{4+}$ treated cells. B) Data represents the mean from three independent experiments and the error bars show standard deviations. Statistical analyses show*

*Student's t tests with Welch's correction compared to no cylinder (dotted line), * p=0.0168 and ** p=0.0037.*

5.4 Conclusion

We have shown that by combining experimental SHAPE results with molecular dynamics simulations we can create 3D models of the structure and dynamics of key individual stems that make up the 5′ UTR of SARS-CoV-2. These stems contain a number of intriguing structural motifs also found in the UTRs of other viruses, and which offer the possibility of developing new anti-viral agents that act against a broad spectrum of diseases. The unique nucleic acid binding activity of the supramolecular cylinders is ideally suited to target these types of structures and we show that the cylinders can bind non-covalently to an RNA bulge in stem loop 5, as well as the central cross (4-way junction) of that loop. The ability to bind at different crucial RNA structural sites that are essential for virus replication limits the opportunity for the virus to mutate and to evade drug action. In line with their RNA binding, these nanosized supramolecular helicates inhibit infection at concentrations where they have negligible cellular toxicity.

These helicate cylinders are currently the only metallo-supramolecular architectures that have been demonstrated to thread through RNA bulge and junction structures, but there is a growing interest in metallo-supramolecular designs to bind nucleic acid structures.[53, 54] While the SHAPE experiments provide further demonstrations of cylinder selectivity for junctions and bulges over other nucleic acid structures, an exciting possibility is that cylinders might also be able to bind host-cell RNA structures, machinery on which the virus depends for replication, causing a dual anti-proliferation effect. The results herein suggest that nucleic acid binding metallo-supramolecular architectures, and the cylinder designs in particular, merit further exploration as anti-viral agents.

5.5 Acknowledgements

This work was funded by the EPSRC Physical Sciences for Health Centre (L.M., T.C., J.S.C., M.J.H.: EP/L016346/1), BBSRC MIBTP (N.J.C., C.A.J.H., BB/M01116X/1) with Sygnature Drug Discovery (N.J.C., CASE 1940003), and an EU Marie Curie Fellowship (A.G., H2020-MSCA-IF-2018-844145). P.G. and K.W. were supported by a Sir Henry Dale Fellowship from the Wellcome Trust and the Royal Society (200473/Z/16/Z) and Z.S. by a Medical Research Foundation intermediate career fellowship (MRF-169-0001-F-STAM-C0826). J.A.M. is funded by a Wellcome Investigator Award (IA) 200838/Z/16/Z, UK Medical Research Council (MRC) project grant MR/R022011/1 and Chinese Academy of Medical Sciences (CAMS) Innovation Fund for Medical Science (CIFMS), China (grant number: 2018-I2M-2-002). We thank Henri Huppert for expert assistance with developing CellInsight quantification algorithms to measure infection in the microneutralisation assay and Jack Dismorr for synthetic assistance. Simulations used the Bluebear and Castles HPC facility (U. Birmingham).[55]

5.6 References

1 World Health Organisation data for 1 March **2021**.

2 S. Yuan, R. Wang, J. F. W. Chan, A. J. Zhang, T. Cheng, K. K. H. Chik, Z. W. Ye, S. Wang, A. C. Y. Lee, L. Jin, H. Li, D. Y. Jin, K. Y. Yuen, H. Sun, *Nat. Microbiol.* 2020, **5**, 1439– 1448.

3 W. Dai, B. Zhang, H. Su, J. Li, Y. Zhao, X. Xie, Z. Jin, F. Liu, C. Li, Y. Li, F. Bai, H. Wang, X. Cheng, X. Cen, S. Hu, X. Yang, J. Wang, X. Liu, G. Xiao, H. Jiang, Z. Rao, L.-K. Zhang, Y. Xu, H. Yang, H. Liu, *Science* 2020, **368**, 1331- 1335.

4 A. Shannon, N. T. T. Le, B. Selisko, C. Eydoux, K. Alvarez, J. C. Guillemot, E. Decroly, O. Peersen, F. Ferron, B. Canard, *Antiviral Res.* 2020, **178**, 104793.

5 Y. Furuta, B. B. Gowen, K. Takahashi, K. Shiraki, F. Donald, D. L. Barnard, *Antiviral Res.* 2013, **100**, 446– 454.

6 C. K. Damgaard, E. S. Andersen, B. Knudsen, J. Gorodkin, J. Kjems, *J. Mol. Biol.* 2004, **336**, 369– 379.

7 B. S. Brigham, J. P. Kitzrow, J. P. C. Reyes, K. Musier-Forsyth, J. B. Munro, *Proc. Natl. Acad. Sci. USA* 2019, **116**, 10372– 10381.

8 D. E. Alvarez, A. L. De Lella Ezcurra, S. Fucito, A. V. Gamarnik, *Virology* 2005, **339**, 200– 212.

9 X. Liu, Y. Liu, Q. Zhang, B. Zhang, H. Xia, Z. Yuan, *Virology* 2018, **524**, 114– 126.

10 M. G. De Castro, F. B. De Nogueira, R. M. R. Nogueira, R. Lourenço-De-Oliveira, F. B. Dos Santos, *Virol. J.* 2013, **10**, 3.

11 R. Ochsenreiter, I. L. Hofacker, M. T. Wolfinger, *Viruses* 2019, **11**, 298.

12 K. Sharma, M. Surjit, N. Satija, B. Liu, V. T. K. Chow, S. K. Lai, *Biochemistry* 2007, **46**, 6488– 6499.

13 D. Yang, J. L. Leibowitz, *Virus Res.* 2015, **206**, 120– 133.

14 L. Li, H. Kang, P. Liu, N. Makkinje, S. T. Williamson, J. L. Leibowitz, D. P. Giedroc, *J. Mol. Biol.* 2008, **377**, 790– 803.

15 I. Manfredonia, D. Incarnato, *Biochem. Soc. Trans.* 2020, **48**, 1– 12.

16 I. Manfredonia, C. Nithin, A. Ponce-Salvatierra, P. Ghosh, T. K. Wirecki, T. Marinus, N. S. Ogando, E. J. Snijder, M. J. van Hemert, J. M. Bujnicki, D. Incarnato, *Nucleic Acids Res.* 2020, **48**, 12436– 12452.

17 E. Bartok, G. Hartmann, *Immunity* 2020, **53**, 54– 77.

18 E. de Wit, N. van Doremalen, D. Falzarano, V. J. Munster, *Nat. Rev. Microbiol.* 2016, **14**, 523– 534.

19 A. Park, A. Iwasaki, *Cell Host Microbe* 2020, **27**, 870– 878.

20 J. Rehwinkel, M. U. Gack, *Nat. Rev. Immunol.* 2020, **20**, 537– 551.

21 A. G. Dias, Jr., N. G. Sampaio, J. Rehwinkel, *Trends Microbiol.* 2019, **27**, 75– 85.

22 T. Hermann, *Wiley Interdiscip. Rev. RNA* 2016, **7**, 726– 743.

23 M. D. Disney, A. J. Angelbello, *Acc. Chem. Res.* 2016, **49**, 2698– 2704.

24 M. D. Disney, B. G. Dwyer, J. L. Childs-Disney, *Cold Spring Harbor Perspect. Biol.* 2018, **10**, a034769.

25 L. Nguyen, L. M. Luu, S. Peng, J. F. Serrano, H. Y. E. Chan, S. C. Zimmerman, *J. Am. Chem. Soc.* 2015, **137**, 14180– 14189.

26 A. M. Fleming, Y. Ding, A. Alenko, C. J. Burrows, *ACS Infect. Dis.* 2016, **2**, 674– 681.

27 Y. Zhang, S. Liu, H. Jiang, H. Deng, C. Dong, W. Shen, H. Chen, C. Gao, S. Xiao, Z. F. Liu, D. Wei, *RNA Biol.* 2020, **17**, 816– 827.

28 C. Zhao, G. Qin, J. Niu, Z. Wang, C. Wang, J. Ren, X. Qu, *Angew. Chem. Int. Ed.* 2021, **60**, 432– 438;

29 P. Krafčíková, E. Demkovičová, V. Víglaský, *Biochim. Biophys. Acta Gen. Subj.* 2017, **1861**, 1321– 1328.

30 G. Biffi, M. Di Antonio, D. Tannahill, S. Balasubramanian, *Nat. Chem.* 2014, **6**, 75– 80.

31 C. K. Kwok, S. Balasubramanian, *Angew. Chem. Int. Ed.* 2015, **54**, 6751– 6754;

32 S. Phongtongpasuk, S. Paulus, J. Schnabl, R. K. O. Sigel, B. Spingler, M. J. Hannon, E. Freisinger, *Angew. Chem. Int. Ed.* 2013, **52**, 11513– 11516;

33 J. Malina, M. J. Hannon, V. Brabec, *Sci. Rep.* 2016, **6**, 29674.

34 L. Cardo, I. Nawroth, P. J. Cail, J. A. McKeating, M. J. Hannon, *Sci. Rep.* 2018, **8**, 13342.

35 G. I. Pascu, A. C. G. Hotze, C. Sanchez-Cano, B. M. Kariuki, M. J. Hannon, *Angew. Chem. Int. Ed.* 2007, **46**, 4374– 4378;

36 M. Pascu, G. J. Clarkson, B. M. Kariuki, M. J. Hannon, *Dalton Trans.* 2006, 2635– 2642.

37 C. A. J. Hooper, L. Cardo, J. S. Craig, L. Melidis, A. Garai, R. T. Egan, V. Sadovnikova, F. Burkert, L. Male, N. J. Hodges, D. F. Browning, R. Rosas, F. Liu, F. V. Rocha, M. A. Lima, S. Liu, D. Bardelang, M. J. Hannon, *J. Am. Chem. Soc.* 2020, **142**, 20651– 20660.

38 F. Wu, S. Zhao, B. Yu, Y. M. Chen, W. Wang, Z. G. Song, Y. Hu, Z. W. Tao, J. H. Tian, Y. Y. Pei, M. L. Yuan, Y. L. Zhang, F. H. Dai, Y. Liu, Q. M. Wang, J. J. Zheng, L. Xu, E. C. Holmes, Y. Z. Zhang, *Nature* 2020, **579**, 265– 269.

39 a A. R. Gruber, R. Lorenz, S. H. Bernhart, R. Neuböck, I. L. Hofacker, *Nucleic Acids Res.* 2008, **36**, 70– 74;39bR. A. C. Oliveira, R. V. M. Almeida, M. D. A. Dantas, F. N. Castro, J. P. M. S. Lima, D. C. F. Lanza, *BMC Bioinformatics* 2014, **15**, 1– 14.

40 K. Sato, M. Akiyama, Y. Sakakibara, *Nat. Commun.* 2021, **12**, 1– 9.

41 R. Rangan, A. M. Watkins, W.Kladwang, R.Das, *bioRxiv* **2020**, https://doi.org/10.1101/2020.04.14.041962.

42 X. Xu, P. Zhao, S. J. Chen, *PLoS One* 2014, **9**, e107504.

43 O. Ziv, J. Price, L. Shalamova, T. Kamenova, I. Goodfellow, F. Weber, E. A. Miska, *Mol. Cell* 2020, **80**, 1067– 1077.e5.

44 R. de Cesaris Araujo Tavares, G. Mahadeshwar, H. Wan, N. C. Huston, A. M. Pyle, *J. Virol.* 2020, **95**, e02190- 20.

45 N. C. Huston, H. Wan, M. S. Strine, R. de Cesaris Araujo Tavares, C. B. Wilen, A. M. Pyle, *Mol. Cell* 2021, 81, 584– 598.e5.

46 S. Elbe, G. Buckland-Merrett, *Glob. Challenges* 2017, **1**, 33– 46.

47 A. M. Watkins, R. Rangan, R. Das, *Structure* 2020, **28**, 963– 976.e6.

48 A. H. Aytenfisu, A. Spasic, A. Grossfield, H. A. Stern, D. H. Mathews, *J. Chem. Theory Comput.* 2017, **13**, 900– 915.

49 L. G. Smith, J. Zhao, D. H. Mathews, D. H. Turner, *Wiley Interdiscip. Rev. RNA* 2017, **8**, e1422.

50 M. K. Scherer, B. Trendelkamp-Schroer, F. Paul, G. Pérez-Hernández, M. Hoffmann, N. Plattner, C. Wehmeyer, J. H. Prinz, F. Noé, *J. Chem. Theory Comput.* 2015, **11**, 5525– 5542.

51 L. Melidis, I. B. Styles, M. J. Hannon, *Chem. Sci.* 2021, https://doi.org/10.1039/D1SC00933H.

52 C. Ducani, A. Leczkowska, N. J. Hodges, M. J. Hannon, *Angew. Chem. Int. Ed.* 2010, **49**, 8942– 8945;

53 53aFor structurally characterised binding of a metallo-intercalator inside a DNA 4-way junction, see: V. H. S. van Rixel, A. Busemann, M. F. Wissingh, S. L. Hopkins, B. Siewert, C. van de Griend, M. A. Siegler, T. Marzo, F. Papi, M. Ferraroni, P. Gratteri, C. Bazzicalupi, L. Messori, S. A. Bonnet, *Angew. Chem. Int. Ed.* 2019, **58**, 9378– 9382; 53b A. Oleksi, A. G. Blanco, R. Boer, I. Usón, J. Aymamí, A. Rodger, M. J. Hannon, M. Coll, *Angew. Chem. Int. Ed.* 2006, **45**, 1227– 1231;53cD. R. Boer, J. M. C. A. Kerckhoffs, Y. Parajo, M. Pascu, I. Uson, P. Lincoln, M. J. Hannon, M. Coll, *Angew. Chem. Int. Ed.* 2010, **49**, 2336– 2339;

54 A J. Gómez-González, Y. Pérez, G. Sciortino, L. Roldan-Martín, J. Martínez-Costas, J.-D. Maréchal, I. Alfonso, M. Vázquez López, M. E. Vázquez, *Angew. Chem. Int. Ed.* 2021, **60**, 8859– 8866; 54bK. Duskova, P. Lejault, É. Benchimol, R. Guillot, S. Britton, A. Granzhan, D. Monchaud, *J. Am. Chem. Soc.* 2020, **142**, 424; 54cL. Guyon, M. Pirrotta, K. Duskova, A. Granzhan, M.-P. Teulade-Fichou, D. Monchaud, *Nucleic Acids Res.* 2018, **46**, e16; 54dX. Li, J. G. Wu, L. Wang, C. He, L. Chen, Y. Jiao, C. Duan, *Angew. Chem. Int. Ed.* 2020, **59**, 6420– 6427; e M. Galindo, D. Olea, M. Romero, J. Gómez, P. del Castillo, M. Hannon, A. Rodger, F. Zamora, J. Navarro, *Chem. Eur. J.* 2007, **13**, 5075– 5081; 54fJ. Zhu, C. J. E. Haynes, M. Kieffer, J. L. Greenfield, R. D. Greenhalgh, J. R. Nitschke, U. F. Keyser, *J. Am. Chem. Soc.* 2019, **141**, 11358– 11362; 54gS.-F. Xi, L.-Y. Bao, Z.-L. Xu, Y.-X. Wang, Z.-D. Ding, Z.-G. Gu, *Eur. J. Inorg. Chem.* 2017, 3533– 3541; 54hA. Terenzi, C. Ducani, V. Blanco, L. Zerzankova, A. F. Westendorf, C. Peinador, J. M. Quintela, P. J. Bednarski, G. Barone, M. J. Hannon, *Chem. Eur. J.* 2012, **18**, 10983– 10990;54iL. Cardo, M. J. Hannon, *Met. Ions Life Sci.* 2018, **18**, 303– 324; 54jH. Crlikova, J. Malina, V. Novohradsky, H. Kostrhunova, R. A. S. Vasdev, J. D. Crowley, J. Kasparkova, V.

Brabec, *Organometallics* 2020, **39**, 1448– 1455; 54kC. Zhao, H. Song, P. Scott, A. Zhao, H. Tateishi-Karimata, N. Sugimoto, J. Ren, X. Qu, *Angew. Chem. Int. Ed.* 2018, **57**, 15723– 15727;54lH. Song, M. Postings, P. Scott, N. J. Rogers, *Chem. Sci.* 2021, **12**, 1620– 1631; 54mD. Wragg, S. Leoni, A. Casini, *RSC Chem. Biol.* 2020, **1**, 390– 394;54nA. Pöthig, A. Casini, *Theranostics* 2019, **9**, 3150– 3169.

55 S. J. Thompson, S. E. M. Thompson, J. Cazier, **2019**, https://doi.org/10.5281/zenodo.3250616

Chapter 6

Interactions of metal complexes with G-quadruplex.

6.1 Introduction

G quadruplexes are a class of nucleic acid tertiary structure characterized by the formation of planar guanine tetrads, stabilized by Hoogstein base pairing between them. Stacks of tetrads are further stabilized by monovalent ions within or between the tetrads as well as the different topologies of the backbone. Sequences that form G-quadruplexes have been found across all kingdoms of life[1] and the human genome[2].

The function of G quadruplexes has been extensively studied over the last 20 years[3] [4], with efforts initially focused on telomeric sequences[5] and later on their role in cancer[6] and cell differentiation[7].

Compared to the previously discussed RNA structures, the relatively rigid core of G-quadruplexes make easier structural characterisation of simple sequences either with XRD[8], [9] or NMR[10]. Specifically, structures solved by solution NMR contain multiple conformations that can be used for parallel MD exploring the conformation and energy landscape. Nevertheless, beyond the parallel and antiparallel loop conformation that have been well characterised, the vast majority of G4 forming sequences are predicted to form hybrid forms[11] where the dynamics of the system are a lot more complicated and thus drug targeting more challenging. High throughput methods can be a useful tool for optimising drug design against these targets[12].

In this chapter we employ MD to examine the interaction of metal complexes with G4s. The chapter starts with planar Pt and Pd complexes that have been previously synthesised in the Hannon group, and later focuses on the interaction of the cylinder with G4s, examining the impact of chirality, as it has been previously published[13]. Finally, the majority of simulations attempt to map the interaction between cylinder and the newly (2018) solved structure of the HIV-1 LTR G4[10], as it is unique among solved G4 structures, having loops that form a stem and can possibly be an interesting target for antivirals and G4 related retro elements on host genome[14].

6.2 Simulations with Pd and Pt complexes on MYC (5w77)

Previously, the Hannon group created planar G4 targeting ligands, as discussed in chapter 3. As most experimental characterisation in the previous work uses MYC DNA G-quadruplexes, here MYC is also used to identify the binding interaction, with two 1-microsecond long simulations for each molecule (Pd and Pt). Parameters for the metal centre were created following the MCPB.py[15] pipeline, although visualisation of the coordination bonds is limited by PyMOL[16] bond visualisation parameters.

Simulations always start with the ligand placed randomly away from myc-g4 and proceeds under the same conditions as previously described but with KCl instead of NaCl ionic atmosphere. K atoms from crystal structure are retained in their coordinates and remain there throughout the simulation without artificially enforcing it. Unsurprisingly, in this time scale both molecules showed identical localisation on the top open quartet (one containing the 5' end), with in this case the tailing additional DNA residue contributing to the binding (Figure 6.1).

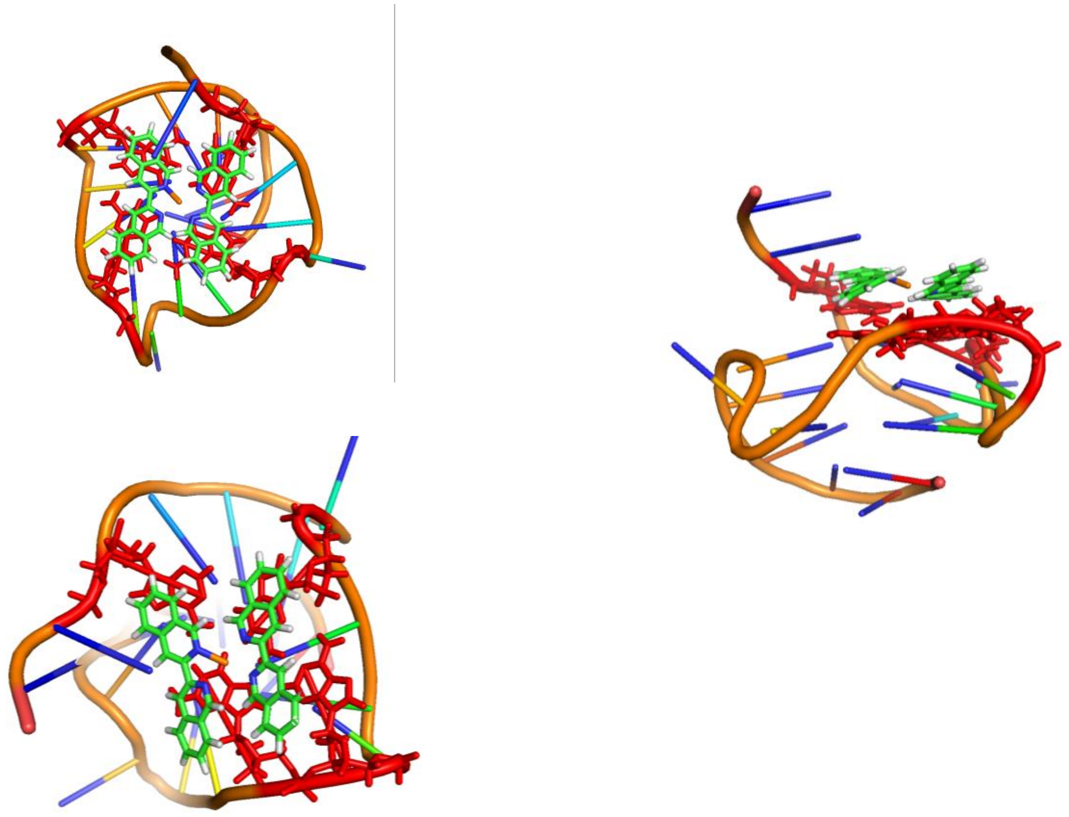


Figure 6.1 *(a)Pd binding to top of MYC. (b) Pt Binding to top of MYC, (c) interacting residues on MYC are shown in red c) side view of dominant binding mode.*

Interestingly throughout the simulations different metastable/transient binding positions were observed which were captured by PCA and TiCA analysis of the trajectories, the strongest such binding was captured in a simulation of Pd where it transiently bound to the bottom side. This is not however on a G tetrad rather the 3’ end of the oligomer (TAA). Although the simulations do not contain enough data to produce MSM maps, it is useful to represent the simulation run through PCA and TiCA , Figure 6.2.

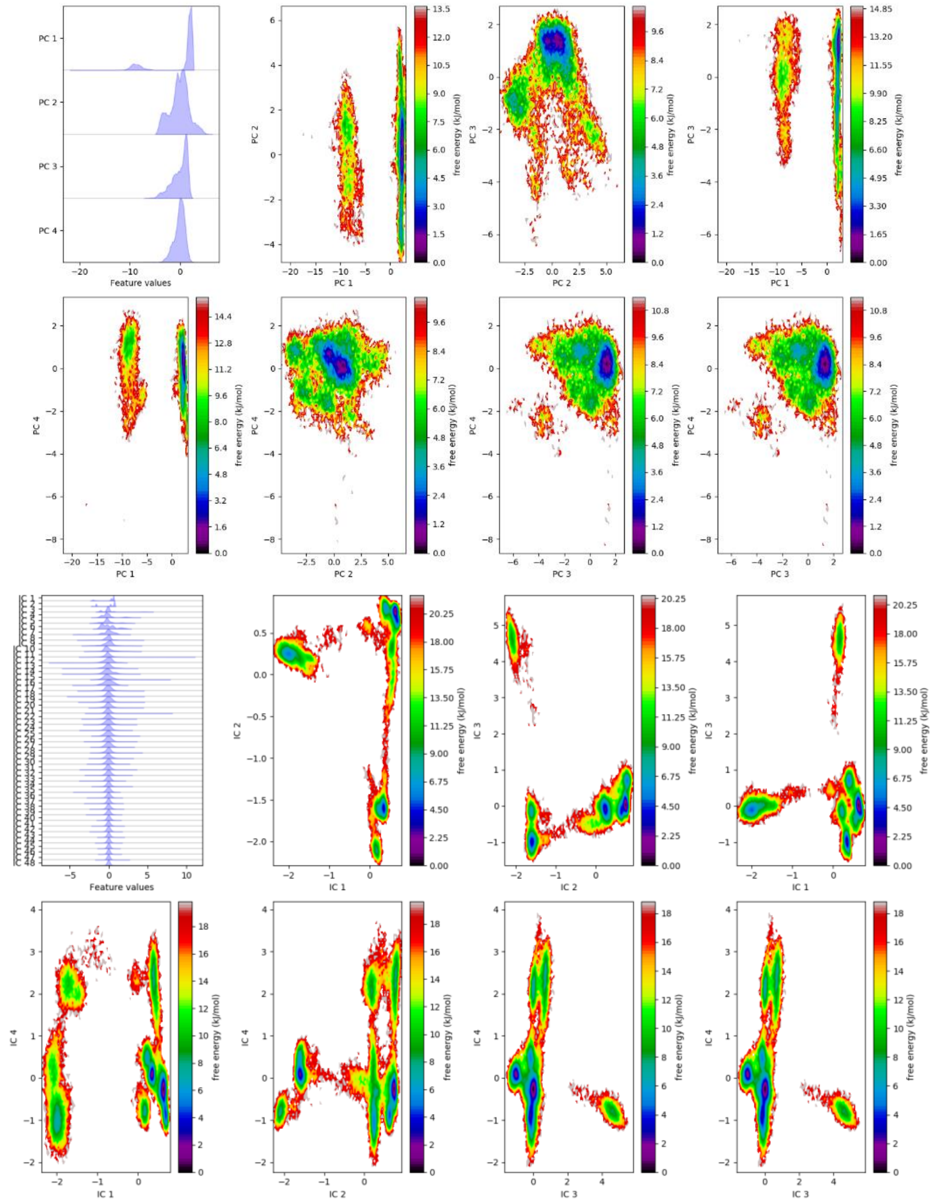


__Figure 6.2__ PCA and TiCA of simulations with Pd complex.

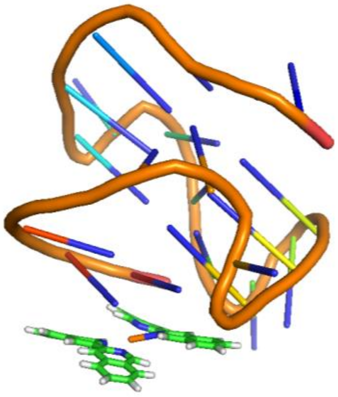


Figure 6.3 *Transient Pt binding to the 3' end.*

Pt's ability to create coordination bonds with N7 and N1 of guanines creates additional interest for this compound, for potential use in Chem-SEQ experiments, providing modifications only to G that are in G4 conformation.

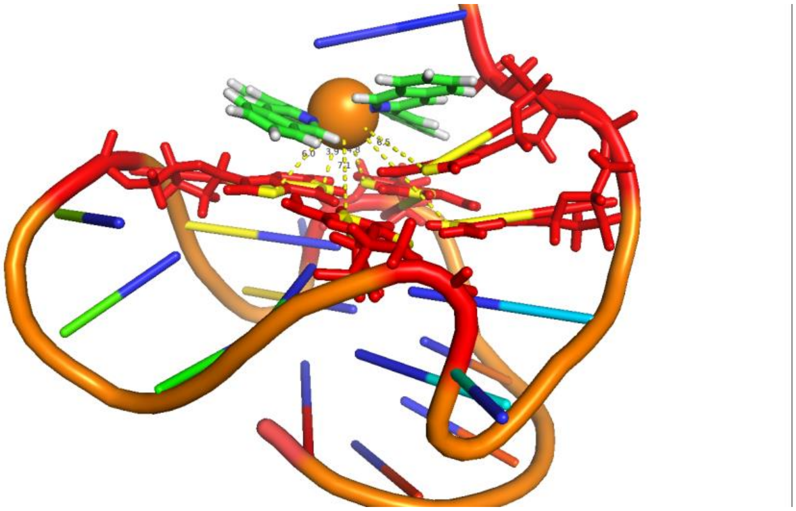


Figure 6.4 *Pt complex proximity of Pt to N1 and N7 atoms of Gs.*

Concluding remarks

There are over 3000 compounds that have been linked to G4 binding in the recently updated G4LDB 2.2 database [17], although structural characterisation has been almost universally ignored, partially due to the ill defined use of "G-quadruplex structure". These compounds and especially the Pt metal centre has the potential to create coordination bonds with the guanines and therefore creates a new class of targeting G quadrats, one that can result in chemical modification of Gs only in G4 conformation which can then be used in sequencing pipelines (ChemMAP-unpublished).

6.3 Cylinder binding on MYC G4 DNA

Previous experimental studies showed different binding constants for different chiral enantiomers of the cylinder. Using the same G4 (MYC;5w77) MD reveals the difference in the preferred binding mode resulted from simulations under 10 microseconds. Again, in the case of cylinders the 5' end of the oligomer is also stabilising the interaction, especially in the case of the P enantiomer (Figure 4). The most common binding interaction between MYC G4 and each cylinder has been indeed different, with the two enantiomers showing different orientation with respect to the central axis of symmetry of the G4. This results in more residues participating in the interaction in the case of P cylinder and enables the flagging 5' sequence to transiently stabilise the structure.

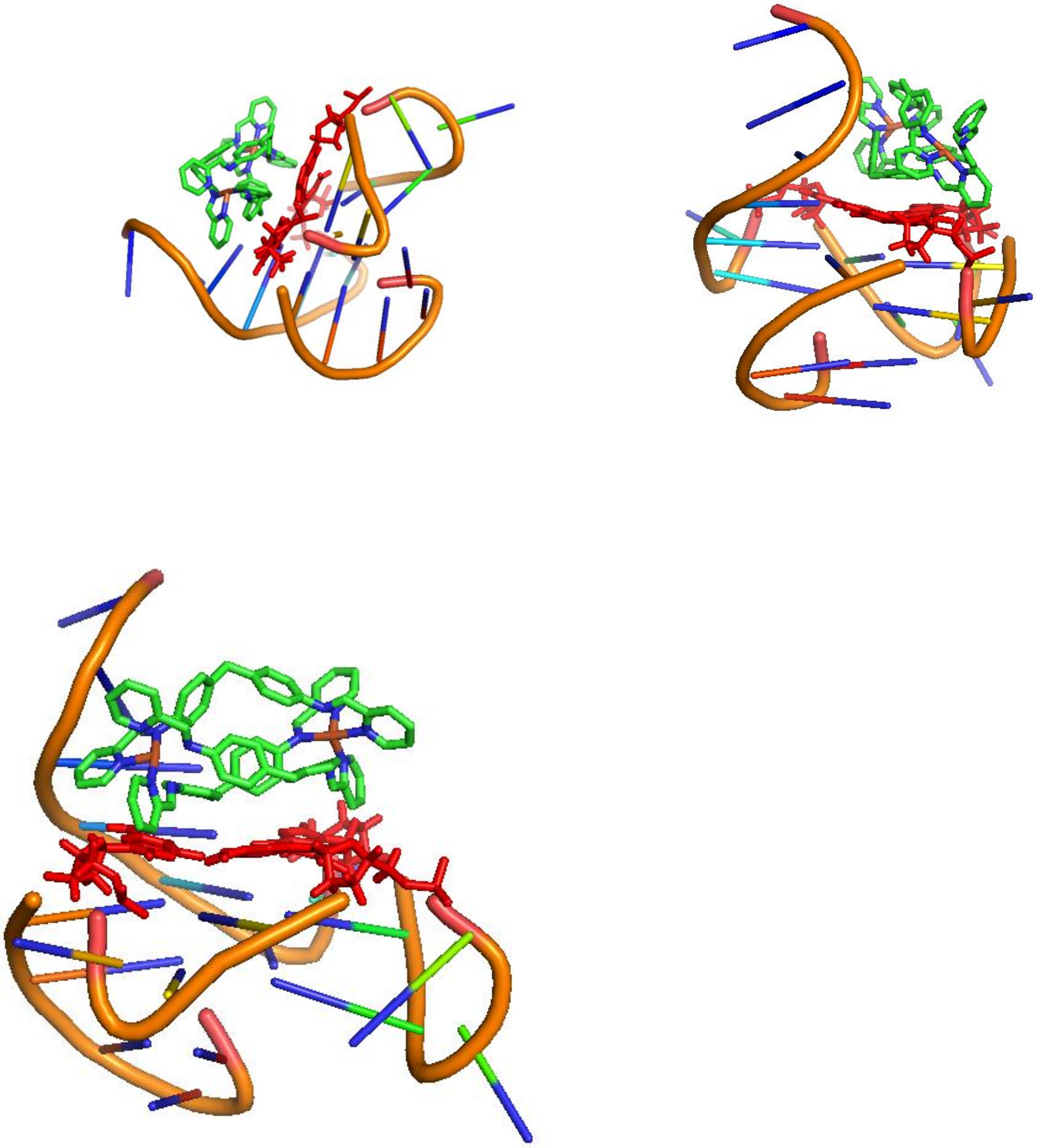


Figure 6.5 *Snapshots of dominant mode of interaction with P enantiomer of the cylinder.*

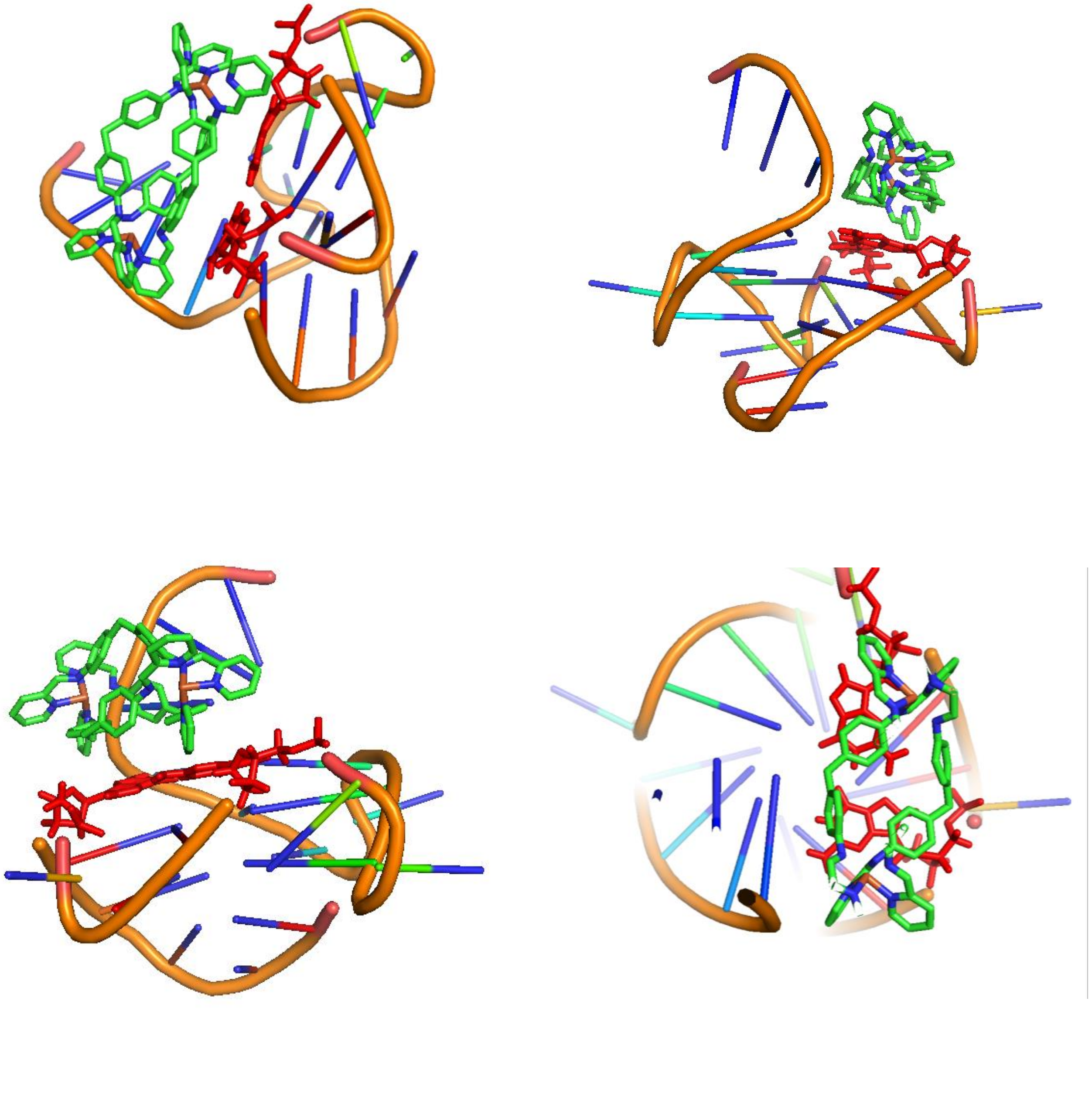


Figure 6.6 *Snapshots of dominant mode of interaction with M enantiomer of the cylindrer.*

6.4 Cylinder binding to HIV-1 LTR G quadruplex

A more interesting G4 target is the HIV-1 LTR G4 DNA. Published by the Phan group[10] in 2018, the PDB entry contains 10 conformations (Figure 6). Interestingly solutions 3 and 7 (Figure 7) suggest there are stable conformations that have a cavity big enough to accommodate the cylinder between the top quartet and the stem. This observation started a purely theoretical/computational journey on the interaction between the cylinder and this exotic G4 structure. The sequence of the oligo is GGGAGGCGTGGCCTGGGCGGGACTGGGG.

Similarly to the other, published, computational explorations in this thesis, the first step has been the MD characterisation of the system, through 3 simulations runs of 2 microseconds each of the free HIV-1 LTR G4 structure (pdb;6h1k) aiming assess its stability and overall dynamics. In order to explore as much as possible of the landscape both of the latest Amber DNA forcefields (OL15 and BSC1) have been used, as well as both Na and K ionic atmospheres of 0.05M (similar to the NMR experiment) on top of the required number to neutralise the system.

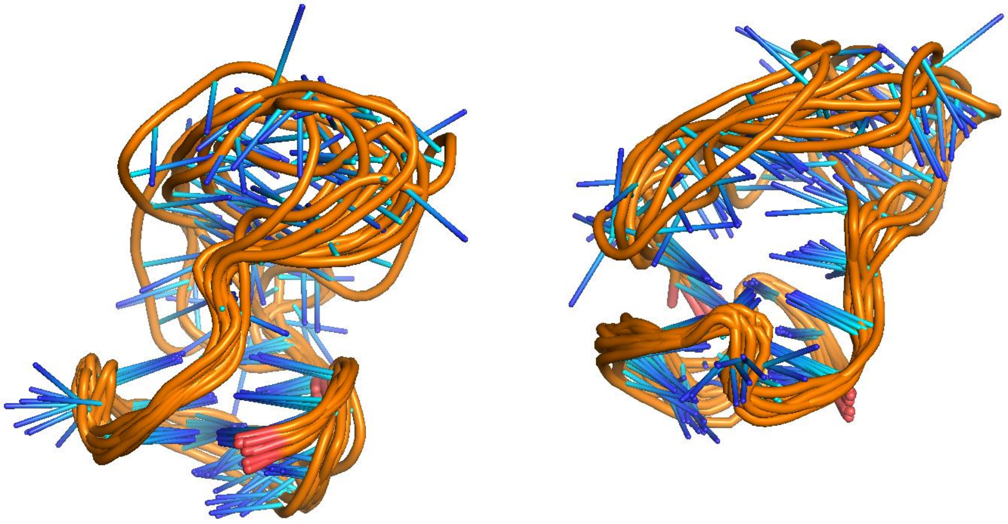


Figure 6.7 *6H1K overlap of the 10 published solutions.*

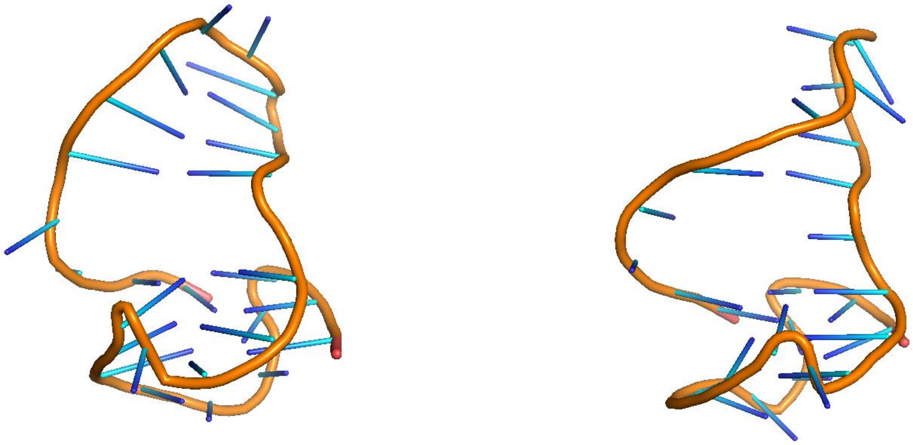


Figure 6.8 *Solution 3 (left) and solution 7 (right) of 6H1K.*

As the starting structures are NMR solutions, positions of the counter ions are not included, but K atoms were trapped in the expected positions, on the central axis of the G4 domain even during the stabilisation of each run.

Throughout the simulations of free DNA, the G4 region remained stable but the loop of the stem shows some expected flexibility and folding back to the stem/G4, at it also observed in NMR solution 5. PCA and TiCA maps of the combined simulations did not fully integrate into a continuous space and therefore a Markov State Model is not meaningful.

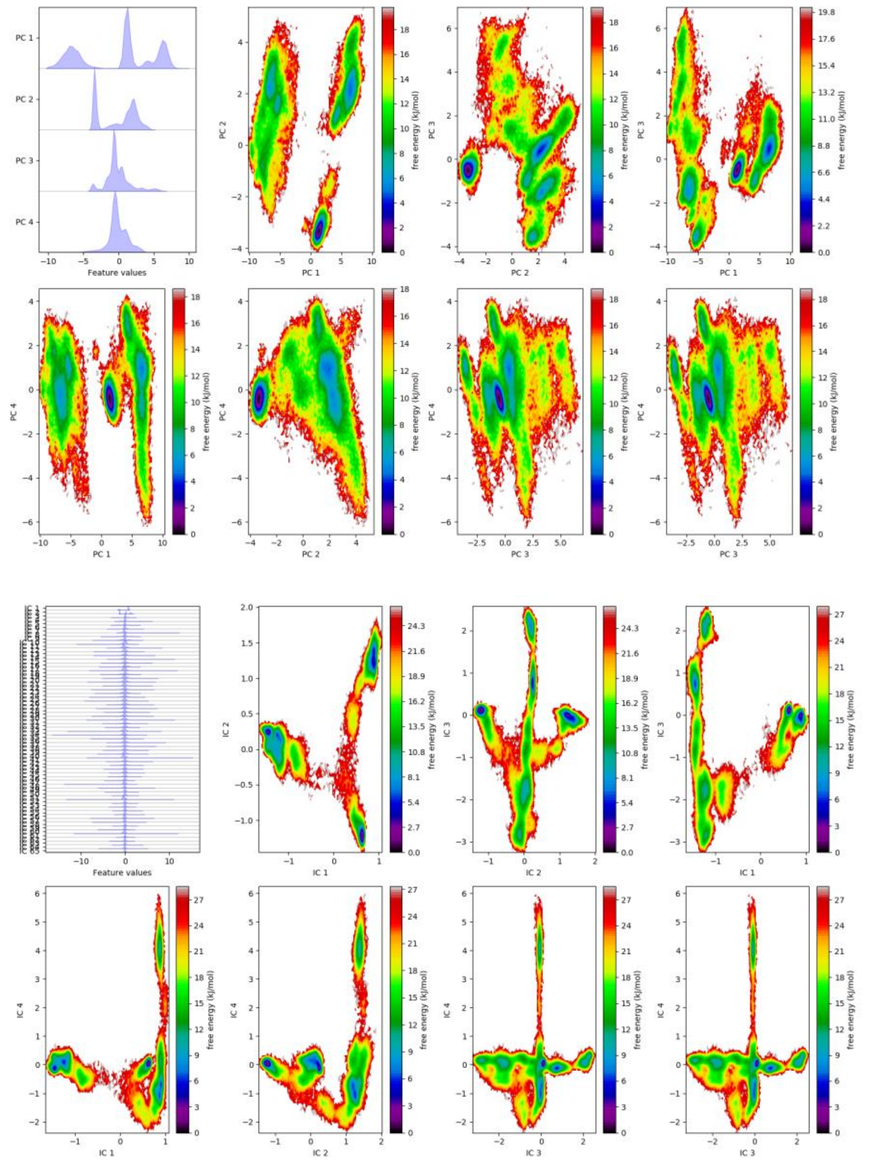


Figure 6.9 *PCA (A) and TiCA (B) projections of the first 4 eigenvectors of free 6H1K from 3 independent simulations with combined simulation time of 10μs using OL15 forcefield.*

However, aiming to address the interaction with the cylinder, 12 independent multi-microsecond (3-15μs) have been performed, allowing to map the encounter. Finally, a novel binding mode, with the cylinder occupying the space between the quartet and the stem is being proposed and evaluated.

Combining 60μs of simulations into one dataset resulted n a continues PCA and TiCA landscapes (Figure 6.9) where MSM model can be constructed.

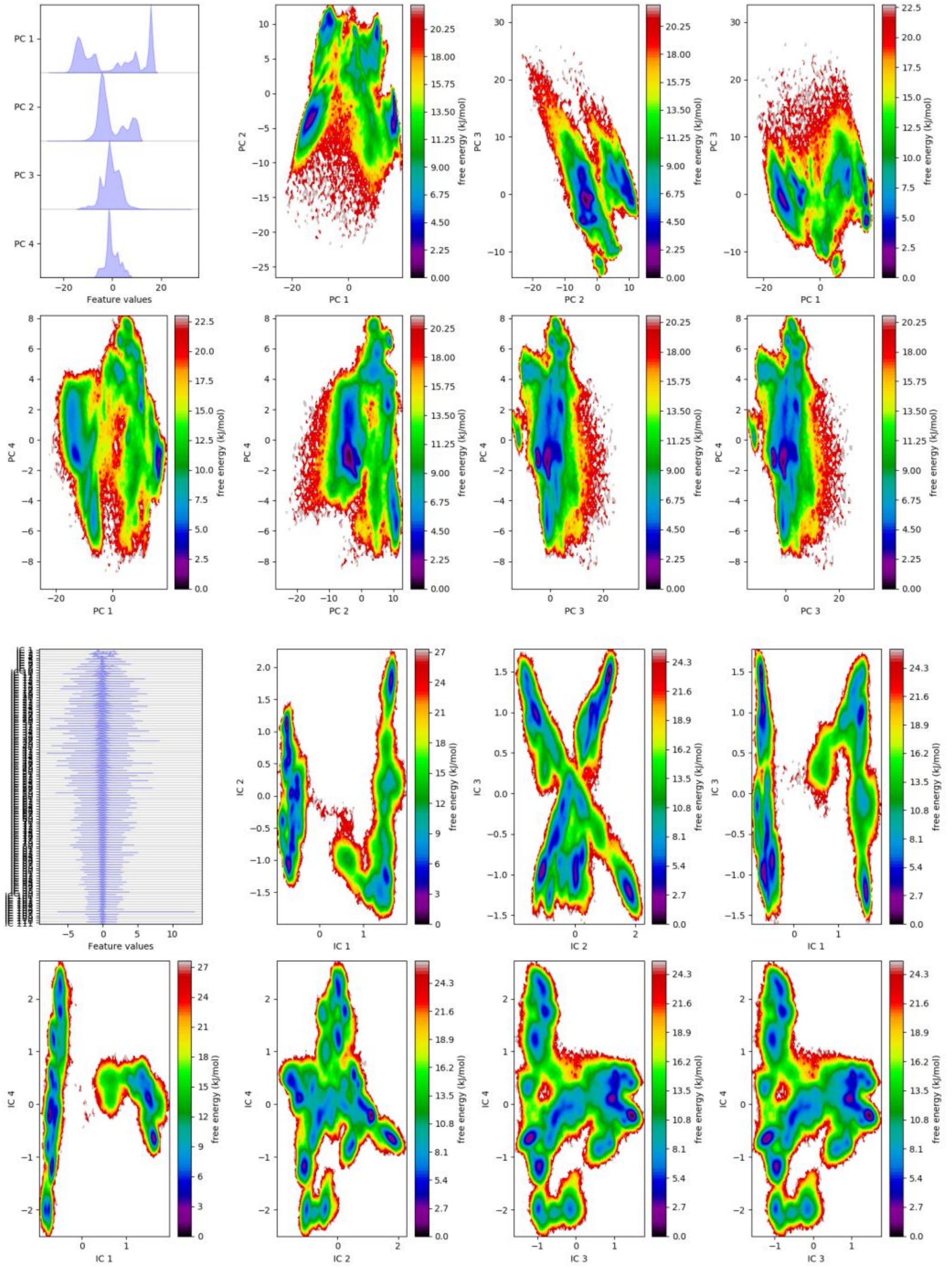


Figure 6.10 *PCA (top) and TiCA (bottom) of combined simulations with free DNA and cylinder.*

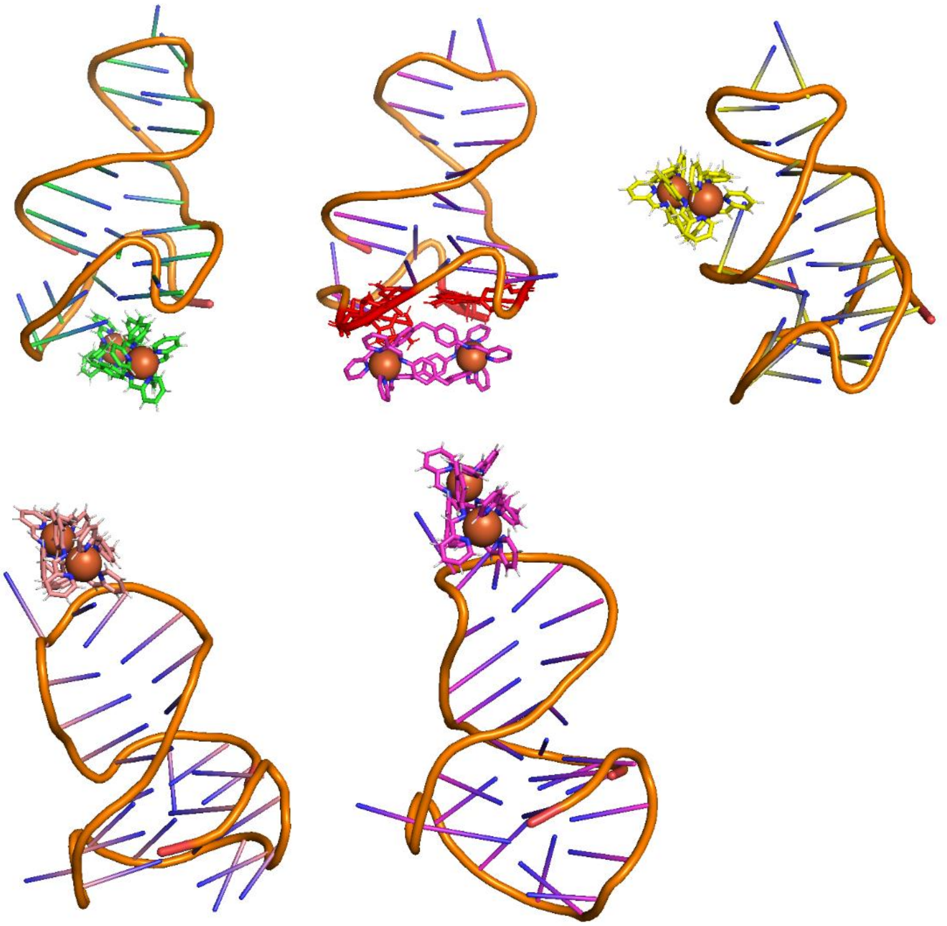


Figure 6.11 *1-5 MSM states in descending order of pseudo-free-energy.*

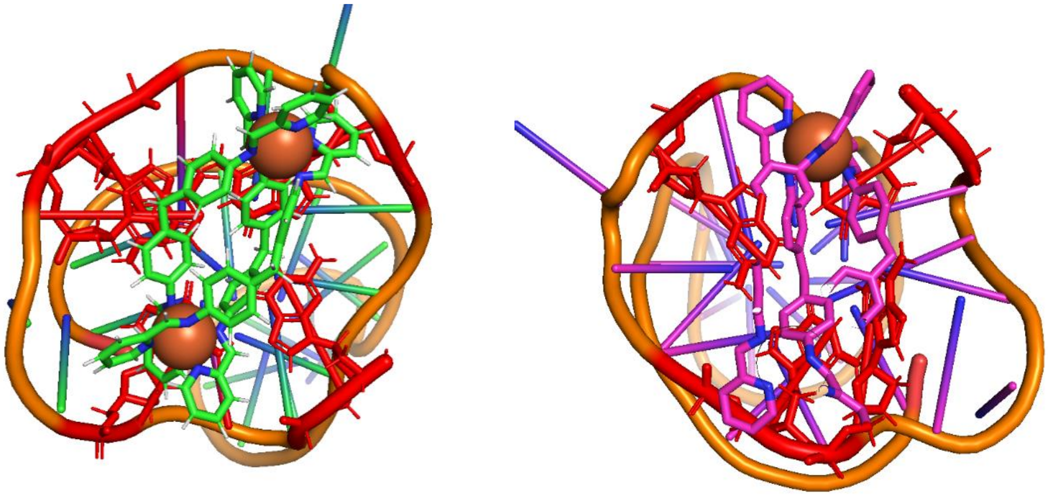


Figure 6.12 *The two most dominant MSM states showing the two different binding modes to the open G4 quartet.*

There are 3 metastable sites of interaction, 1. the open G4 quartet, 2. the grove between the stem and the top G4 quartet (A4, G5, C13) and 3. the stem's loop C7-G10. The dominating binding site in this data set is the open quartet. This binding site contains multiple local minima, identified as 2 minima close together in PCA. These two, are expanded into multiple minima close together in TiCA space.

Proposed binding site

The first data set explored the interaction of this G4 with the cylinder starting with the cylinder away from the DNA. This approach only explores the interaction close to the most dominant conformation of the DNA. However, one can explore the wider landscape by initiating the interaction with a conformation other than the dominant species. In this case, taking advantage of solution 3 of the 6H1K dataset, the cylinder is manually positioned between the G4 and the stem. 4 multi-microsecond simulations have been performed, 4 with Na and 2 with K as counterions, combining 30μs.

The cylinder remained in the site throughout all the simulations adopting multiple combinations of stable binding modes that have be observed in TAR RNA in chapter 4. In detail, there are 8 nucleotides that contribute to binding;

1. The cylinder caps the stable G5:C13 base pair base pair in a similar way to the base pairs around the bulge in TAR (chapter 4). Neighbouring bases A4 and T14, which normally form a transient base pair, open towards the same direction with respect to the central axis of the stem, both positive, and the familiar splitting of consecutive bases can be observed on the A4 site (between A4 and G5). On the opposite strand (downstream), T14 is positioned parallel to the cylinder's axis, again similarly to the bulged nucleotides of TAR.
2. Additionally, the single bulged nucleotide that separates the stem from the G4; G3 further encapsulates the cylinder as it caps the cylinder from the opposite direction of G5, contributing to the overall stability.
3. Finally, G2 and G15, are the two nucleotides which are part of the G quadruplex structure that contribute most to the binding. It is worth noting that these are the two on the same strand as the stem, neighbouring in sequence with the previously mentioned residues.

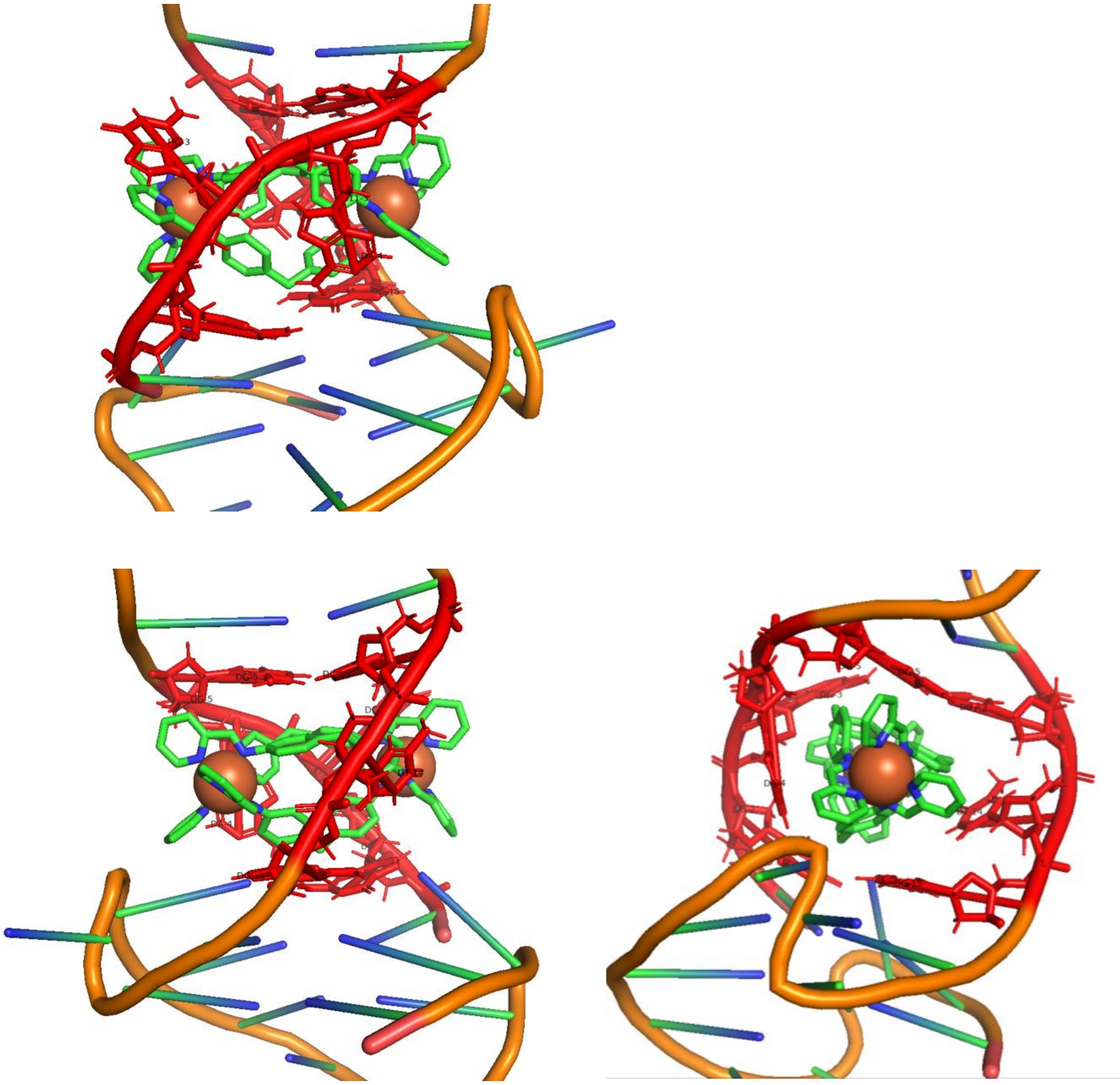


Figure 6.13 *Interaction of the proposed binding mode.*

Due to the very stable binding in this case, PCA and TiCA maps (Figure 10) are dominated by the conformational changes of the DNA rather than the movement of the cylinder and a continuous space could be completed identifying 5 MSM states (Figure 11). These maps also show that the changes on the DNA are small and transient as the maps are dominated by one minimum, and contains a few shallow ones. In only one of the 5 MSM states (MSM 5) identified the binding mode is slightly different, with G3 pointing the other direction.

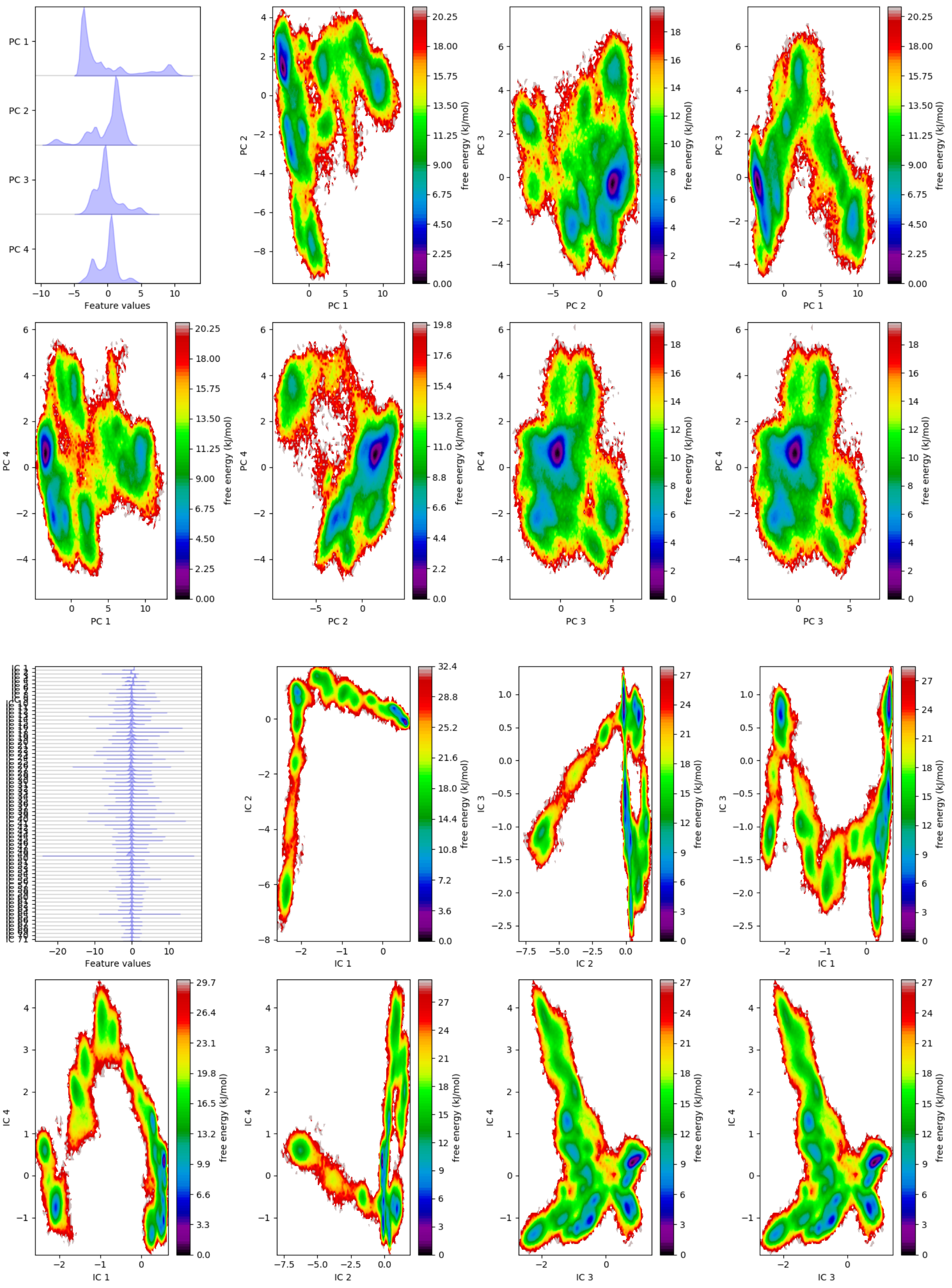


Figure 6.14 *PCA (top) and TiCA(bottom) maps of simulations of the P cylinder between G4 and stem.*

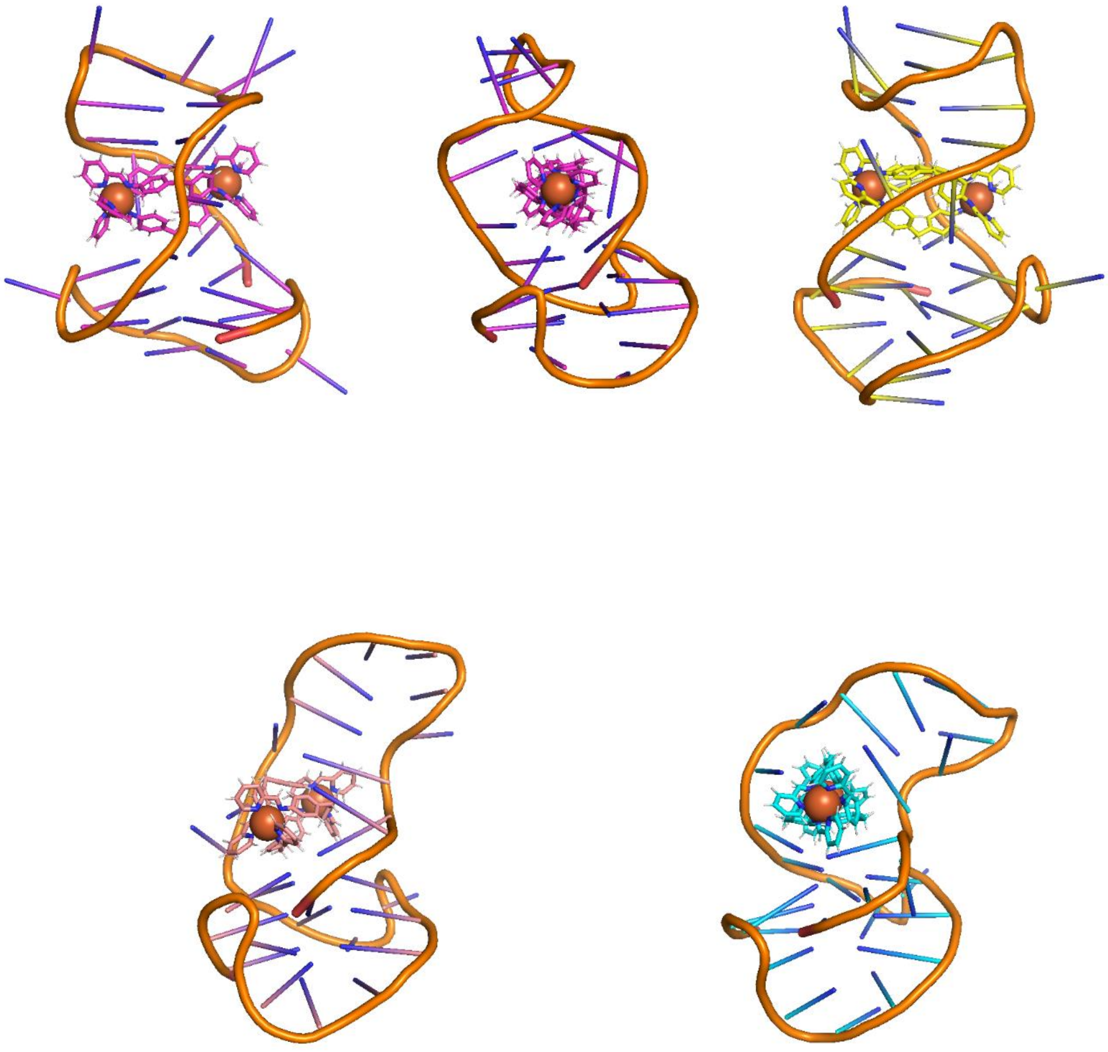


Figure 6.15 *1 to 5 MSM states identified in decreasing abundance within the simulation.*

To further examine this binding mode and potentially quantify the binding energy LiGaMD has been employed, as described in chapter 2. However none of the trial parameters resulted in dissociation of the complex and thus no conclusion can be drawn.

6.5 Conclusions

1. The Pt complex shows promising binding to the open quartet, and distances to N1 and N7 atoms of guanines suggest potential coordination
2. Cylinders with different chirality have shown to bind differently to the MYC G4 potentially explaining some experimental results. Here we show through simulation the role chirality can play on the binding orientation of the cylinder.
3. HIV1 LTR G4, is a unique and exciting structure to study not only because it is found in HIV1 and is potentially an antiviral target, but also it is a structure that expands the landscape of sequences which could form G4 beyond the PQS (potential quadruplex-forming sequences) resulted by all G4 predicting software. Additionally, it's position in the genome, on the crossover of the integrating site, and the known correlation of G4 forming capable sequences and retro elements in host DNA creates new avenues for investigation on the DNA G4 stability and function within the nucleus.
4. Although the data collected for this chapter on the interaction of the cylinder with HIV1 LTR G4 has been extensive in comparison to other G4 MD studies, further simulations need to be undertaken evaluating the stability of the free structure. Mapping of the free interaction has been extensive, showing the open quartet binding as the most dominant potential binding mode, but in light of the binding mode proposed in this chapter could be irrelevant for the dominant binding in solution. The proposed binding has been the strongest that has been captured throughout this thesis although a potential path to that binding mode has not captured.

6.6 Future work

1. Pt-complex binding interaction should be verified experimentally with Mass spectroscopy experiments, after which, QM/MD can be performed in order to capture and optimise the transition of coordinating from ligand to G4. Modifications to the ligand can include asymmetric units, which could ease the dissociation of one ligand.

2. In cylinder chirality specific binding of the MYC G4 NMR experiments should focus on identifying potential NOEs between the cylinder and the identified residues. If these NOEs are identified, titration experiments are expected to reveal the rising population of the secondary (open quartet) binding mode.
3. Finally, the HIV-1 LTR G4 study here only used the P form of the cylinder so preliminary binding assays should compare the two forms for potential differences before embarking on expensive computational comparison between the two enantiomers.

6.7 References

[1] G. Marsico, V. S. Chambers, A. B. Sahakyan, P. McCauley, J. M. Boutell, M. Di Antonio, S. Balasubramanian, *Nucleic Acids Res.* **2019**, *47*, 3862–3874.

[2] V. S. Chambers, G. Marsico, J. M. Boutell, M. Di Antonio, G. P. Smith, S. Balasubramanian, *Nat. Biotechnol.* **2015**, *33*, 877–881.

[3] D. Varshney, J. Spiegel, K. Zyner, D. Tannahill, S. Balasubramanian, *Nat. Rev. Mol. Cell Biol.* **2020**, *21*, 459–474.

[4] K. G. Zyner, D. S. Mulhearn, S. Adhikari, S. M. Cuesta, M. Di Antonio, N. Erard, G. J. Hannon, D. Tannahill, S. Balasubramanian, *Elife* **2019**, *8*, 1–40.

[5] S. Burge, G. N. Parkinson, P. Hazel, A. K. Todd, S. Neidle, *Nucleic Acids Res.* **2006**, *34*, 5402–5415.

[6] S. A. Ohnmacht, C. Marchetti, M. Gunaratnam, R. J. Besser, S. M. Haider, G. Di Vita, H. L. Lowe, M. Mellinas-Gomez, S. Diocou, M. Robson, J. Šponer, B. Islam, R. Barbara Pedley, J. A. Hartley, S. Neidle, *Sci. Rep.* **2015**, *5*, 1–11.

[7] K. G. Zyner, A. Simeone, S. M. Flynn, C. Doyle, G. Marsico, S. Adhikari, G. Portella, D. Tannahill, S. Balasubramanian, *Nat. Commun.* **2022**, *13*, 142 DOI 10.1038/s41467-021-27719-1.

[8] Y. Geng, C. Liu, B. Zhou, Q. Cai, H. Miao, X. Shi, N. Xu, Y. You, C. P. Fung, R. U. Din, G. Zhu, *Nucleic Acids Res.* **2019**, *47*, 5395–5404.

[9] G. N. Parkinson, M. P. H. Lee, S. Neidle, *Nature* **2002**, *417*, 876–880.

[10] E. Butovskaya, B. Heddi, B. Bakalar, S. N. Richter, A. T. Phan, *J. Am. Chem. Soc.* **2018**, *140*, 13654–13662.

[11] S. Shi, H. L. Huang, X. Gao, J. L. Yao, C. Y. Lv, J. Zhao, W. L. Sun, T. M. Yao, L. N. Ji, *J. Inorg.*

Biochem. **2013**, *121*, 19–27.

[12] G. Wu, D. Tillo, S. Ray, T. C. Chang, J. S. Schneekloth, C. Vinson, D. Yang, *Molecules* **2020**, *25*, 15 ,3465

[13] H. Yu, X. Wang, M. Fu, J. Ren, X. Qu, *Nucleic Acids Res.* **2008**, *36*, 5695–5703.

[14] A. B. Sahakyan, P. Murat, C. Mayer, S. Balasubramanian, *Nat. Struct. Mol. Biol.* **2017**, *24*, 243–247.

[15] P. Li, K. M. Merz, *J. Chem. Inf. Model.* **2016**, *56*, 599–604.

[16] M. A. Lill, M. L. Danielson, *J. Comput. Aided. Mol. Des.* **2011**, *25*, 13–19.

[17] Y. H. Wang, Q. F. Yang, X. Lin, D. Chen, Z. Y. Wang, B. Chen, H. Y. Han, H. Di Chen, K. C. Cai, Q. Li, S. Yang, Y. L. Tang, F. Li, *Nucleic Acids Res.* **2022**, *50*, D150–D160.

Chapter 7

Summary and future perspectives

The aim of this thesis has been to lay the theoretical foundation for broad spectrum antiviral drug design, focusing on using coordination compounds for viral RNA targeting.

The introduction explores fundamental viral replication processes across the different viral groups of the Baltimore scheme, in an effort to explore common bottlenecks where nucleic acid structure plays a functional roll in the replication process. These structures are metastable in nature as they change conformation during the different stages of replication, and therefore arresting the structure to one state, or influencing the exchange rate between these metastable states can provide means to dysregulate the replication process.

The introduction also explores the state of the art in molecular modelling across scales, starting with electronic representation of molecules, with DFT techniques and then how the results in one scale can be used to parametarize molecules to be used in bigger scale in both time and space, using molecular dynamics.

The general underlying theme that has emerged through this thesis has been the identification of metastable states and their effect in the equilibrium of highly dynamic multi-layered systems.

Chapter 2 sets the pipelines of computational tools that have been used throughout the thesis.

Chapter 3 investigates the dynamics of coordination compounds and especially different variants of the cylinder. Traditionally in chemistry, out of global minimum states have been explored in terms of transition states, as means of describing reactions. However, here the potential diversity of a molecule around the global minimum has been studied, including different molecular spin states and their effect to the molecular structure. This can be used to assess the stability of the global minimum in solution during synthesis as well as the stability in the biological world where the molecule would encounter multiple different interactions and possible reactions. The case of the Ni cylinder is especially interesting as the parent cylinder is very stable in solution and in biophysical experiments but loses biological activity compared to the very stable Ru cylinder. At the same time, changing the pyridines to imidazoles, changes the electronic landscape to a more stable ground state which retains the activity in biological experiments. Further to the study of electronic equilibrium, the interaction of cylinders with CB10 has also been studied. In this case, it is the energy of solvation that creates metastable states in the interaction between two molecules. This equilibrium has been controlled by creating rotaxanes, published in JACS[1] and the efficiency of partially capped cylinders has been studied, in a manuscript that is close to be finalised.

Chapter 4 has been published in Chemical Sciences [2]. In this chapter and publication, HIV-1's TAR RNA has been used as a model example of well-studied dynamic RNA structure with antiviral target potential. Here, for the first time molecular dynamics have been used to demonstrate the conformational changes on the target molecule (TAR RNA) upon interaction with a coordination compound. Molecular dynamics could de novo predict and describe the binding interaction along with the path of conformational changes, filling in the knowledge gaps between experiments, which can be used for further optimisation of the targeting compound. Along with HIV-1 TAR, stem-loops of 5' UTR of other viruses have been studied (Polio, HIV-2, coxsackievirus).

Chapter 5 has been published in Angewandte Chemie[3], and describes a pipeline for targeted drug design response to a novel pathogenic virus. Within days after the publication of SARS-CoV2 sequence, models of secondary and tertiary structure of the 5' UTR were developed and molecular dynamics protocols described in the previous chapter revealed the regions where the cylinder can be used as a template for further optimisation. The use of 2 more molecules, meth-pyridine cylinder (previously used for the rotaxane paper) and isoquinoline Nickel cylinder have been suggested to have higher affinity to these structures which was verified when tested biophysically in experiments designed by the author of this thesis and performed in the Grezchnik lab. Finally, in cellulo activity was quantified using high throughput imaging in collaboration with the Stamataki lab. The imaging platform needs to be further optimised for use with other molecules, specifically regarding the choice of fluorophores and levels of initial viral load. Most importantly this work demonstrates a robust pipeline for secondary structure prediction of RNA, including the metastable steps during folding (junction of stem-loop 5 and linear fragment between SL4 and SL5).

Chapter 6 studies non B DNA structures specifically G quadruplexes. Initially the interaction of metal cantered planar compounds are used in molecular dynamics revealing a potential for coordination bonds forming between the metal centre and the exposed guanines. Although this is a first step for the analysis, further computational studies need to be undertaken allowing for electronic interaction between Pt and G with QM/MD in coordination with mass spectrometry experiments that would create a better understanding of the interaction. Finally, the chapter closes with one of the most interesting G quadruplex structures found in biology, HIV-1's LTR G4 whose solution NMR structure was published during the PhD. Only one chiral enantiomer was studied and two binding modes were proposed. Although some CD experiments have been performed further characterisation is needed to coordinate computational and experimental methods before publication.

To summarise, this has been a journey of metastable states in various length and time scales and aims to show how this formulation of the phenomena can aid broad-spectrum antiviral drug design and nucleic acid targeting in general. At the electronic structure level, we explore the potential alternative conformations of coordination compounds allowing different spin states which explores the potential paths of molecular degradation in a complex chemical environment. The use of molecular dynamics to describe metastable states also observed in previous NMR studies has been demonstrated in chapter 4 and their application in novel virus de novo predictions of RNA structure has been used in chapters 5 and 6.

Fundamentally, this work lays the foreground in incorporating thermodynamics in drug design targeting highly flexible structures, but also exposes the ability to transform a dynamic landscape by incremental stabilisation of one of the metastable states. This thinking can be applied far beyond the chemical space and the same representation of data can produce informative yet lower dimensional representation of any dynamic system, including the viral replication and viral stability in multiple scales. Specifically for chemical space, the field of targeting nucleic acid structures has been increasing rapidly, with new structural diversity being achieved after expanding the nucleic acid alphabet with endogenous base modifications. This type of modifications and dysregulation of their abundance has been linked to multiple cancer types and other diseases of epigenetic origin, nevertheless their contribution to structural diversity of nucleic acids is only now being realised. The computational framework developed here enables a first principle fundamental understanding of these new landscapes and can enable accurate drug design.

[1] C. A. J. Hooper, L. Cardo, J. S. Craig, L. Melidis, A. Garai, R. T. Egan, V. Sadovnikova, F. Burkert, L. Male, N. J. Hodges, D. F. Browning, R. Rosas, F. Liu, F. V. Rocha, M. A. Lima, S. Liu, D. Bardelang, M. J. Hannon, *J. Am. Chem. Soc.* **2020**, *142*, 20651–20660.

[2] L. Melidis, I. B. Styles, M. J. Hannon, *Chem. Sci.* **2021,** 12, 7174-7184.

[3] L. Melidis, H. Hill, N. Coltman, S. Davies, K. Winczura, T. Chauhan, J. Craig, A. Garai, C. Hooper, R. Egan, J. McKeating, N. Hodges, Z. Stamataki, P. Grzechnik, M. Hannon, *Angew. Chemie Int. Ed.* **2021**, 133, 18292-18299.

8. Appendix I and Supplementary information

This has been taken verbatim from the supplementary information of the Chemical science publication in chapter 4. Supplementary information found at 7174-7184. doi: 10.1039/d1sc00933h, (Amended by APR 08/08/23)