Spring 2021

Molecular Simulation Studies of Dynamics and Interactions in Nucleic Acids

Lev Levintov
University of New Hampshire, Durham

MOLECULAR SIMULATION STUDIES OF DYNAMICS AND INTERACTIONS IN NUCLEIC ACIDS

BY

Lev Levintov

BS, Chemical Engineering, University of New Hampshire, 2016

DISSERTATION

Submitted to the University of New Hampshire
in Partial Fulfillment of
the Requirements for the Degree of

Doctor of Philosophy
in
Chemical Engineering

May, 2021

This dissertation has been examined and approved in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Chemical Engineering by:

Dissertation Director, Dr. Harish Vashisth

Associate Professor, Department of Chemical Engineering

Dr. Nivedita Gupta

Professor, Department of Chemical Engineering

Dr. Kang Wu

Associate Professor, Department of Chemical Engineering

Dr. Krisztina Varga

Associate Professor, Department of Molecular, Cellular, and Biomedical Sciences

Dr. Paul Robustelli

Assistant Professor, Department of Chemistry (Dartmouth College)

On [Date of Defense]

Original approval signatures are on file with the University of New Hampshire Graduate School.

DEDICATION

I dedicate this to my family.

ACKNOWLEDGEMENTS

First, I want to acknowledge my advisor, Prof. Harish Vashisth, who continuously guided and supported me throughout my PhD study. His dedication and professional attitude to work always inspired me to work harder. Moreover, I want to thank him for all the professional skills that I have learned from him during my PhD.

I also want to thank Professors Nivedita Gupta, Kang Wu, Krisztina Varga, and Paul Robustelli, for serving on my thesis committee.

I am very grateful to all my collaborators, Dr. Shambhavi Tannir (University of Wyoming), Dr. Krisztina Varga (University of New Hampshire), Dr. Mark Townley (University of New Hampshire), Dr. Milan Balaz (Yonsei University), Dr. Brian Leonard (University of Wyoming), and Dr. Jan Kubelka (University of Wyoming) for a collaborative work on the porphyrin/DNA system. Specifically, I want to thank Dr. Krisztina Varga and Dr. Milan Balaz for insightful conversations on the self-assembly of the porphyrin/DNA systems and for the opportunity to study these systems.

I also thank Dr. Sanjib Paul (New York University) for a collaborative work on the base flipping mechanism in dsRNA. Moreover, I want to thank him for his patience in teaching me the foundations of the transition path sampling methods and for sharing his codes and scripts with me.

I thank the past and present lab members for all the questions and the constructive criticism that I have received from them during our group meetings. These conversations were always helpful and improved my critical thinking.

I am forever grateful to all my family who have always supported me and believed in my dreams. No matter how far they are, I can always feel their love and support for me. I greatly thank my grandfather, Boris Levintov, who nurtured love for science in me when I was a small kid. He taught me to always ask questions about nature around us and gifted me a lot of books about biology.

Lastly, but importantly, I thank my girlfriend Eivet for all her love, support, care, and patience.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

ARM — Arginine-rich Motif.

BC — Boundary Conditions.

BIC — Bayesian Information Criterion.

BSA — Buried Surface Area.

CV — Collective Variable.

DNA — Deoxyribonucleic Acid.

dsRNA — Double-stranded Ribonucleic Acid.

HIV — Human Immunodeficiency Virus.

MD — Molecular Dynamics.

NAMD — Nanoscale Molecular Dynamics.

NMR — Nuclear Magnetic Resonance.

OP — Order Parameter.

PBC — Periodic Boundary Conditions.

PDB — Protein Data Bank.

PMF — Potential of Mean Force.

RC — Reaction Coordinate.

RNA — Ribonucleic Acid.

RMSD — Root Mean Squared Deviation.

RMSF — Root Mean Squared Fluctuation.

RRE — Rev Response Element.

SASA — Solvent Accessible Surface Area.

SMD — Steered Molecular Dynamics.

TAR — Transactivation Response Element.

TPS — Transition Path Sampling.

VMD — Visual Molecular Dynamics.

WC — Watson-Crick.

ABSTRACT

MOLECULAR SIMULATION STUDIES OF DYNAMICS AND INTERACTIONS IN NUCLEIC ACIDS

by

Lev Levintov

University of New Hampshire, May, 2021

In my thesis work, I conducted molecular simulation studies to explore dynamics and interactions in nucleic acids. I began my work by applying conventional molecular dynamics (MD) simulations to study the local and global dynamics of the transactivation response (TAR) element from the type-1 human immunodeficiency virus (HIV-1) and the effect of binding of ligands on the dynamics of TAR RNA. I determined that the TAR RNA structure was stabilized on binding of ligands due to the decreased flexibility in helices that comprise TAR RNA. This rigidity of the TAR RNA structure was coupled with the decreased flipping of bulge nucleotides. I also observed that different initial conformations of TAR RNA converged to similar conformations in the course of MD simulations. Finally, I observed the formation of binding pockets in unliganded TAR structures that could accommodate ligands of various sizes.

After comprehensively exploring the dynamics of TAR RNA with and without ligands, I conducted more specific studies on the interactions that were formed or broken during the (un)binding process of two ligands, a small molecule inhibitor and a helical peptide, from

the viral RNA molecules using non-equilibrium simulations. Firstly, I observed that the dissociation of a small molecule is coupled with a base flipping event which I described using physical variables and thermodynamic properties. Secondly, I observed that the dissociation process of a helical peptide is facilitated by a network of hydrogen bonding and salt bridging interactions which are formed across four distinct dissociation pathways. I also resolved the free-energy profiles for each pathway which revealed metastable states and dissociation barriers. Based on the free-energy profiles, I proposed a preferred dissociation pathway and identified one arginine amino acid that plays an important role in the recognition of the peptide by the viral RNA.

Next, I focused on studying a more complex reaction coordinate (RC) that could describe a base flipping mechanism in a double-stranded RNA (dsRNA) molecule using transition path sampling (TPS) methods. Additionally, I used the likelihood maximization method to determine a refined RC based on an ensemble of 1000 transition trajectories created by the path sampling algorithm. The refined RC consisted of two collective variables (CVs), a distance and a dihedral angle between the neighboring nucleotides and the flipping base. I also projected a free-energy profile along the refined RC which revealed three free-energy minima. I proposed that one of the free-energy minima represented a wobbled conformation of the flipping nucleobase. I also analyzed the reactive trajectories which showed that the base flipping is coupled with global conformational changes in a stem-loop of dsRNA.

Outside of studies involving RNA, I conducted conventional MD simulations to study the dynamics of a porphyrin/DNA nanoassembly which revealed the overall left-handed orientation of the nanoassembly. I characterized the resulting porphyrin/DNA system using various physical variables. Overall, my thesis revealed the local and global dynamics of RNA as well as DNA systems, and perturbations to dynamics originating in binding of ligands of various sizes.

CHAPTER 1
INTRODUCTION

1.1 Brief History of Nucleic Acids Research

The research work on nucleic acids has witnessed major breakthroughs during the 20th century [1]. In 1944, it was discovered that deoxyribonucleic acid (DNA) is the carrier of genetic information and not the proteins, as believed previously [2]. In 1952, Erwin Chargaff showed that the amount of guanine in DNA equaled the amount of cytosine and the amount of adenine equaled the amount of thymine [3]. Later in 1953, Rosalind Franklin and Maurice Wilkins, Francis Crick and James Watson solved the structure of B-DNA (Figure 1.1A) using X-ray crystallography which was the first accurate molecular structure of DNA [4, 5]. Finally, in 1961, deciphering of the genetic code began after Marshall Nirenberg and Heinrich Matthaei discovered the first codon that got translated into a specific amino acid [6].

In 1965, the first complete nucleotide sequence of a ribonucleic acid (RNA) molecule was reported by Holley *et al.* [7]. For decades, according to the central dogma of biology, RNA was considered only as a passive carrier of genetic information from DNA to proteins [8]. In particular, a messenger RNA (mRNA) was thought to act as an intermediate carrier of genetic information and a transfer RNA (tRNA) was responsible for the transport of amino acids to the translation machine of the cell (Figure 1.1B,C) [9]. Only in the early 1980s it was discovered that RNA can catalyze certain chemical reactions by breaking and reforming phosphodiester bonds [10], and its catalytic function can achieve acceleration rates that are comparable to protein enzymes [11]. The next crucial step in the RNA research was the crystallization of the ribosome [12], a large cellular machine that is responsible for protein

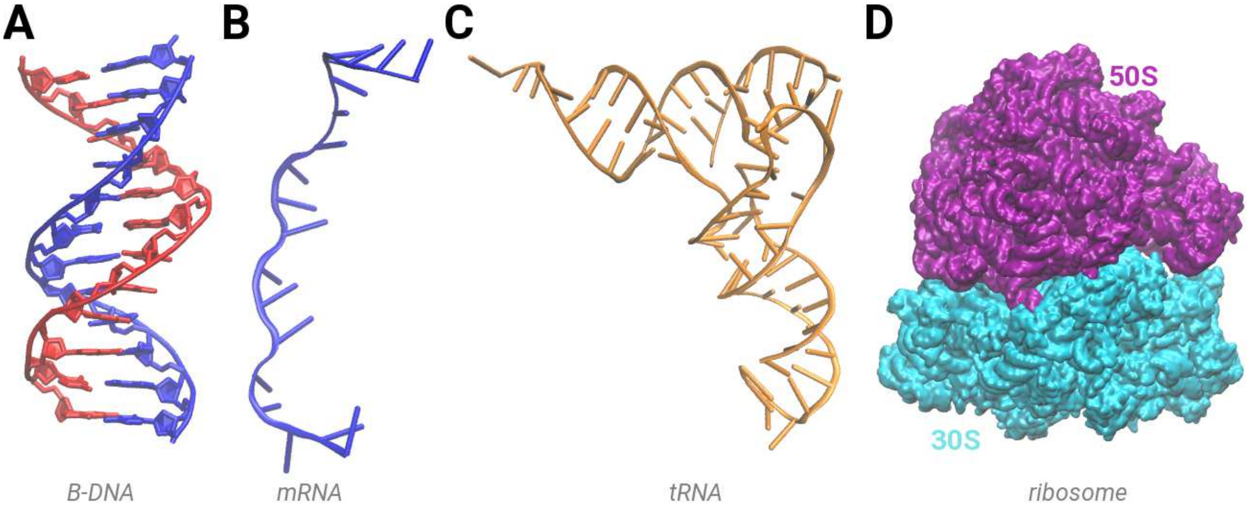

Figure 1.1: Shown are the snapshots of different types of nucleic acids: (A) a double-stranded DNA molecule with each strand drawn in a unique color (PDB: 1BNA); (B) a cartoon representation of an mRNA molecule (PDB: 6DNC); (C) a cartoon representation of a tRNA molecule (PDB: 4V42); (D) a representation of the ribosome structure with the 30S and the 50S subunits highlighted in cyan and purple, respectively (PDB: 6DNC).

synthesis. The structure and further research on the ribosome [13–24] revealed that the ribosome structure is mostly consisted of large RNA molecules which are combined with small proteins (Figure 1.1D). In recent years, it has been shown that the RNA molecules are involved in other cellular processes, including gene silencing, regulation, and processing of genetic information [25–28]. Moreover, RNA molecules are implicated in the development of various diseases, including cancers [29,30], neurological disorders [31], cardiovascular diseases [32], as well as in the replication and survival mechanisms of many viruses and bacteria [33–35].

These discoveries have tremendously expanded our knowledge about the molecular biology by demonstrating the significance of RNA molecules in performing various functions in cells. However, despite the increased understanding of the RNA structure-dynamics-function relationships in cells, we still need to better characterize the dynamics of RNA and the interactions between RNA and its binding partners at the atomic level [36].

1.2 Structures of Nucleic Acids

Nucleic acids are either single-stranded (RNA; Figure 1.1B) or double-stranded (DNA; Figure 1.1A) biopolymers that are composed of individual building blocks called nucleotides. Each nucleotide consists of three components: a phosphate group, a 5-carbon sugar moiety, and a nitrogenous base. If the sugar group is a deoxyribose, the biopolymer is a DNA molecule and if the sugar group is a ribose, the biopolymer is an RNA molecule. The sugar group is linked to the 5′-phosphate group and to the aromatic base through the N-glycosidic bond (Figure 1.2A). DNA is comprised of the following four nucleobases (bases): adenine (A), cytosine (C), guanine (G), and thymine (T). Adenine, guanine, and cytosine also occur in RNA, but in RNA thymine is substituted by uracil (U). All five bases are shown in Figure 1.2B.

Each base can form hydrogen bonding interactions with another base and thus their combination forms a base-pair. Based on that, J. Watson and F. Crick proposed the complementary pairing of bases (also known as Watson-Crick base pairing) which states that an adenine base forms a pair with a thymine base (or with a uracil base in RNA) and a guanine base forms a pair with a cytosine base (Figure 1.2C) [5]. These base-pairing interactions as well as stacking of base pairs, which results from van der Waals and electrostatic interactions between bases that stack on top of one another, stabilize the double-helical structure of DNA molecules (Figure 1.2D) and of short helical segments in RNA molecules that often fold upon themselves to form WC-complementary base pairs. Each nucleic acid chain of nucleotides is assembled through a sequential phosphodiester linkage mechanism, in which a phosphate group links the 3′-carbon of each sugar to the 5′-carbon of the next sugar moiety, leaving an unlinked 5′-carbon at one end of the strand (termed as the "5′-end") and an unlinked 3′-position at the other end of the strand (termed as the "3′-end"). Therefore, the strands are asymmetric and can form base-pairing interactions. The 5′-end is considered to be the beginning of the strand because the synthesis of nucleic acids is initiated at that end.

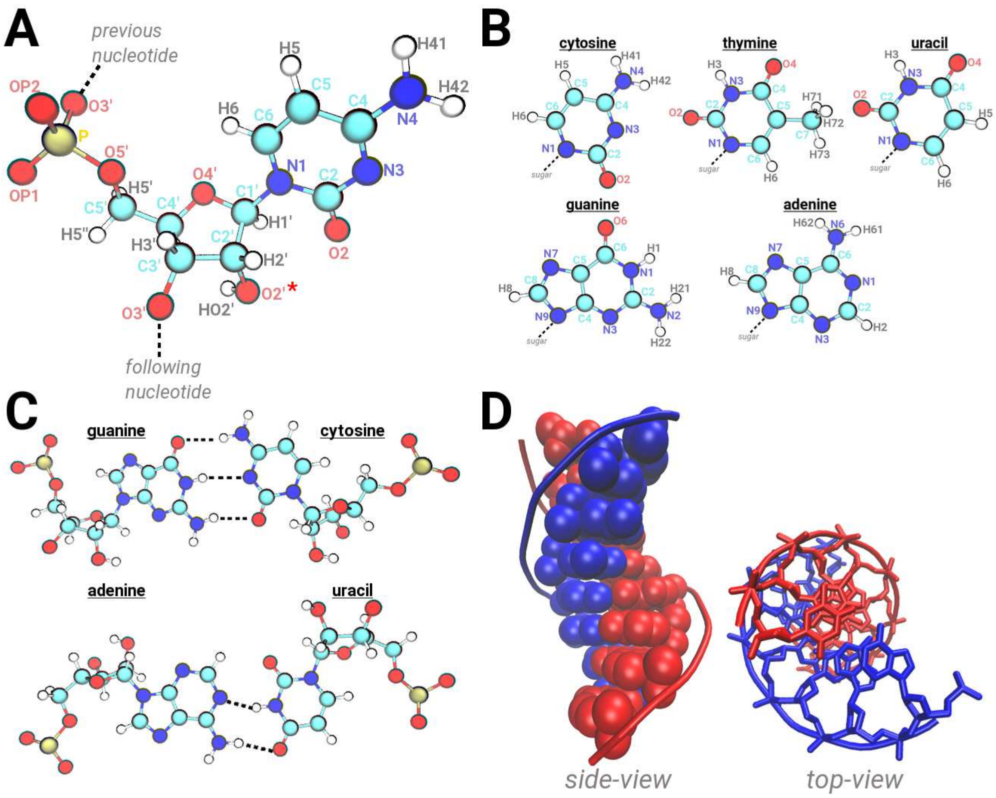

Figure 1.2: **Structures of nucleic acids.** (A) An example of a single nucleotide within a strand. Each atom is highlighted and labeled in a unique color. A red asterisk marks an oxygen atom that is not present in a DNA nucleotide. (B) The five aromatic bases that are present in nucleic acids. Each atom is highlighted and labeled in a unique color. (C) The Watson-Crick base pairs. The dashed lines indicate hydrogen bonds. (D) Two snapshots of a double-helical DNA structure highlighting base-stacking interactions (side-view) and base-pairing interactions (top-view). Each strand is represented in a unique color.

Nucleic acids can also be analyzed and presented by their primary, secondary (2D), and tertiary (3D) structures. The primary structure describes a sequence of nucleotides, while the secondary structure defines the base-pairing interactions in nucleic acids. For RNA, the secondary structure can be divided into the following motifs (elements): helices, internal loops, hairpins, bulges, and junctions [37]. The tertiary structure describes the three-dimensional shape of nucleic acids and results from interactions between specific secondary structural motifs [37]. The tertiary structure depends on the formation of numerous van der Waals contacts and hydrogen bonds which result from the base-pairing interactions between secondary structural motifs. Thus, the tertiary structure of RNA molecules largely depends on the secondary structure interactions [38,39]. However, RNA structures are not static in the solution rather are highly flexible and dynamic.

1.3 RNA Dynamics

As stated previously, RNA molecules play an essential role in various cellular processes, such as translation and transcription [40], regulation of gene expression [28], and protein synthesis [41]. The variability in functions of RNA is associated with its ability to undergo conformational changes since a single RNA molecule can adopt multiple complex three-dimensional shapes in response to external stimuli while also undergoing local conformational changes at the level of base pairs [42,43]. These dynamics can be classified into distinct modes that range over various timescales, for example, base-pairing rearrangements at the ns-μs timescales or complex interhelical motions that lead to global transitions in the three-dimensional shape of the RNA molecule at the μs timescales [44,45]. However, despite the tremendous amount of information on RNA dynamics that the experimental techniques provide us, the characterization of all possible parameters that are required to describe the RNA dynamics at the atomic level of detail is still challenging [36,46–48]. Therefore, computational methods can be utilized to simulate the dynamics of RNA molecules to support experimental findings and to provide further insights into the RNA dynamics as well as to understand the recognition mechanisms between the RNA molecules and various ligands [36,49].

1.4 RNA-Ligands Interactions

The misregulation of the activity of RNA molecules can lead to the development of various diseases, including cancer, neurological disorders, and cardiovascular diseases [50–52]. Moreover, RNA molecules play a crucial role in various processes of viral and bacterial life cycles, such as replication and survival mechanisms [35,53–55]. Therefore, RNA molecules serve as compelling targets for novel therapeutic agents [31,35,55–57]. In this thesis, I studied three types of ligands the examples of which are provided in Figure 1.3: (*i*) small molecules that

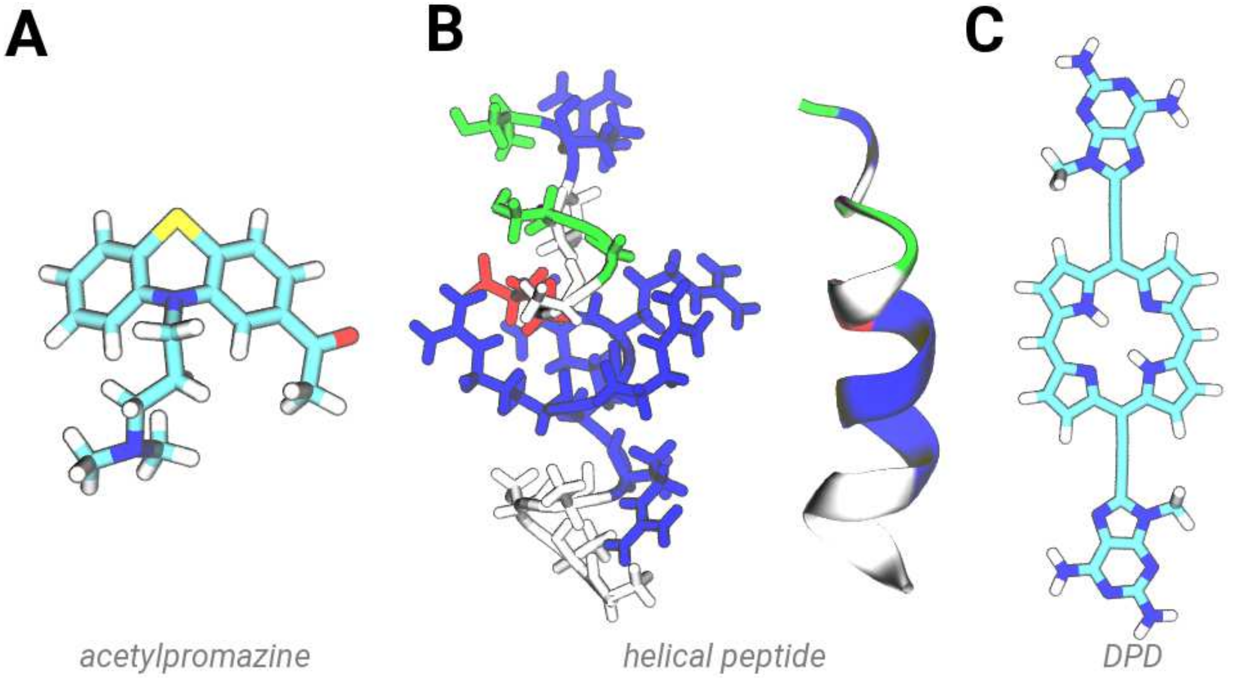

Figure 1.3: Snapshots of representative ligands that were studied in this work: (A) the structure of a small molecule, acetylpromazine, with each atom highlighted in a unique color (PDB: 1LVJ); (B) a stick (left) and a cartoon (right) representations of a helical peptide based on the type of residue (PDB: 1G70); (C) the structure of a porphyrin-diaminopurine (DPD) molecule, with each atom highlighted in a unique color.

inhibit viral replication in human immunodeficiency virus type 1 (HIV-1); (*ii*) helical and cyclic peptides that inhibit viral replication in HIV-1; (*iii*) modified porphyrin molecules that interact with DNA strands to form a supramolecular nanoassembly.

1.5 Background of Systems Studied

In this thesis, I studied three classes of nucleic-acid systems: (i) viral RNA molecules, (ii) a double-stranded RNA (dsRNA) molecule, and (iii) single-stranded DNA (ssDNA) molecules. Among viral RNA molecules, I studied the HIV-1 transactivation response element (TAR) RNA and the HIV-1 Rev response element (RRE) RNA which participate in the viral replication process.

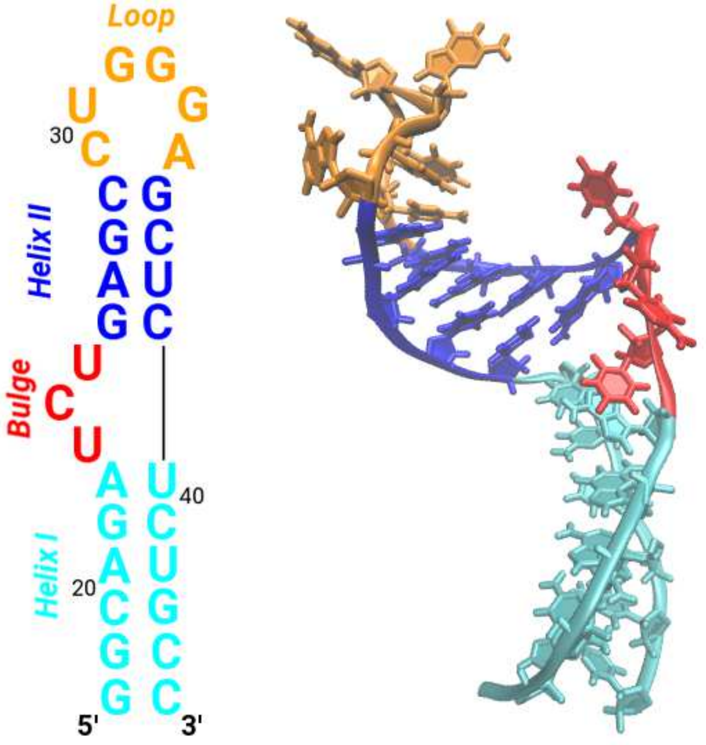

Figure 1.4: Secondary structure and a snapshot of the three-dimensional structure of HIV-1 TAR RNA (PDB: 1ANR) are shown. Various structural motifs are uniquely colored and labeled.

1.5.1 Viral RNA Molecules

Due to its ability to adapt multiple states, a key model for studying RNA dynamics is TAR RNA (Figure 1.4) from HIV-1 [48,58]. TAR RNA is located at the 5′ end of HIV-1 transcripts where it interacts with the viral transactivator (Tat) protein and the host cofactor cyclin T1 to promote efficient transcription of the downstream genome and is therefore considered to be an important drug target. TAR RNA has been studied using nuclear magnetic resonance (NMR) spectroscopy [58–68], coarse-grained MD simulations [69], electron paramagnetic resonance (EPR) [70], gel mobility [71], combinations of NMR and MD methods [62, 72], and combinations of NMR and structure prediction software [73]. Collectively, these studies have shown that TAR RNA undergoes complex dynamics by sampling different interhelical conformations around the bulge junction, thus forming various conformational ensembles [48, 61]. Several studies have revealed that a bulge motif in TAR RNA (red; Figure 1.4) is

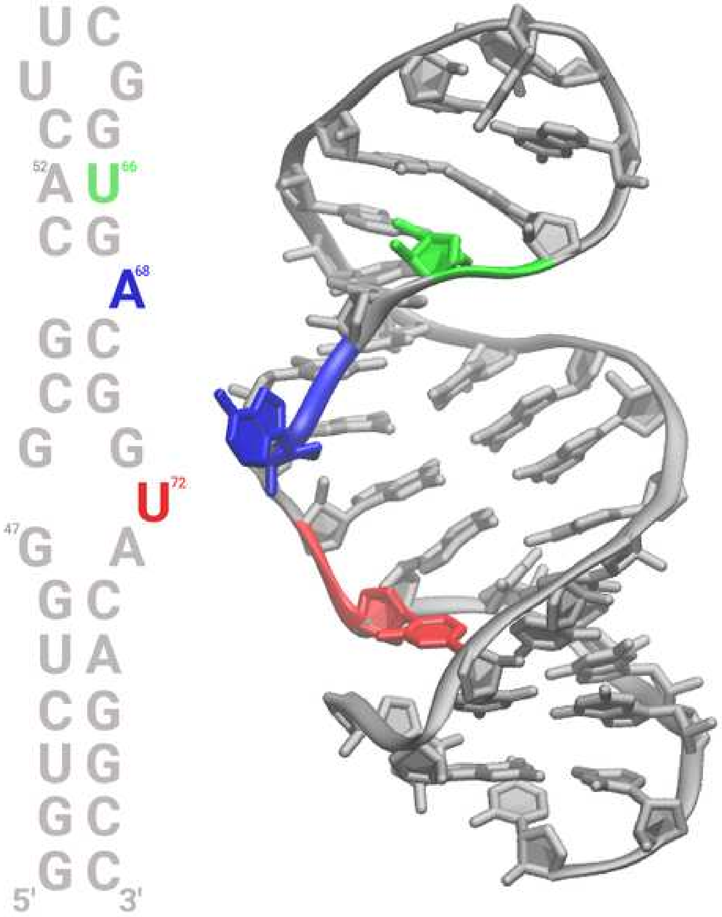

Figure 1.5: Secondary structure and a snapshot of the three-dimensional structure of HIV-1 RRE RNA (PDB: 1G70) are shown. Key nucleotides are uniquely colored and labeled.

especially critical for its recognition by the Tat protein [59, 60, 74]. Several previous studies have shown that TAR RNA can bind to peptide mimics [59,75–79], to small molecules [80–83], to proteins [84–87], and to divalent cations [88]. While detailed dynamics in TAR RNA have not been characterized in the presence of all known ligands, it has been suggested that ligands can potentially induce structural transitions in TAR by stabilizing pre-existing conformers or an ensemble of states in the *apo* TAR RNA structure [66, 89–91].

Another viral RNA system that I studied in my thesis work, was the conserved HIV-1 RRE RNA segment which is located in the *env* coding region and plays an essential role in viral replication [92] (Figure 1.5). In particular, I studied the dissociation process of a helical peptide, arginine-rich peptide (RSG-1.2) [93], from the RRE RNA. This peptide

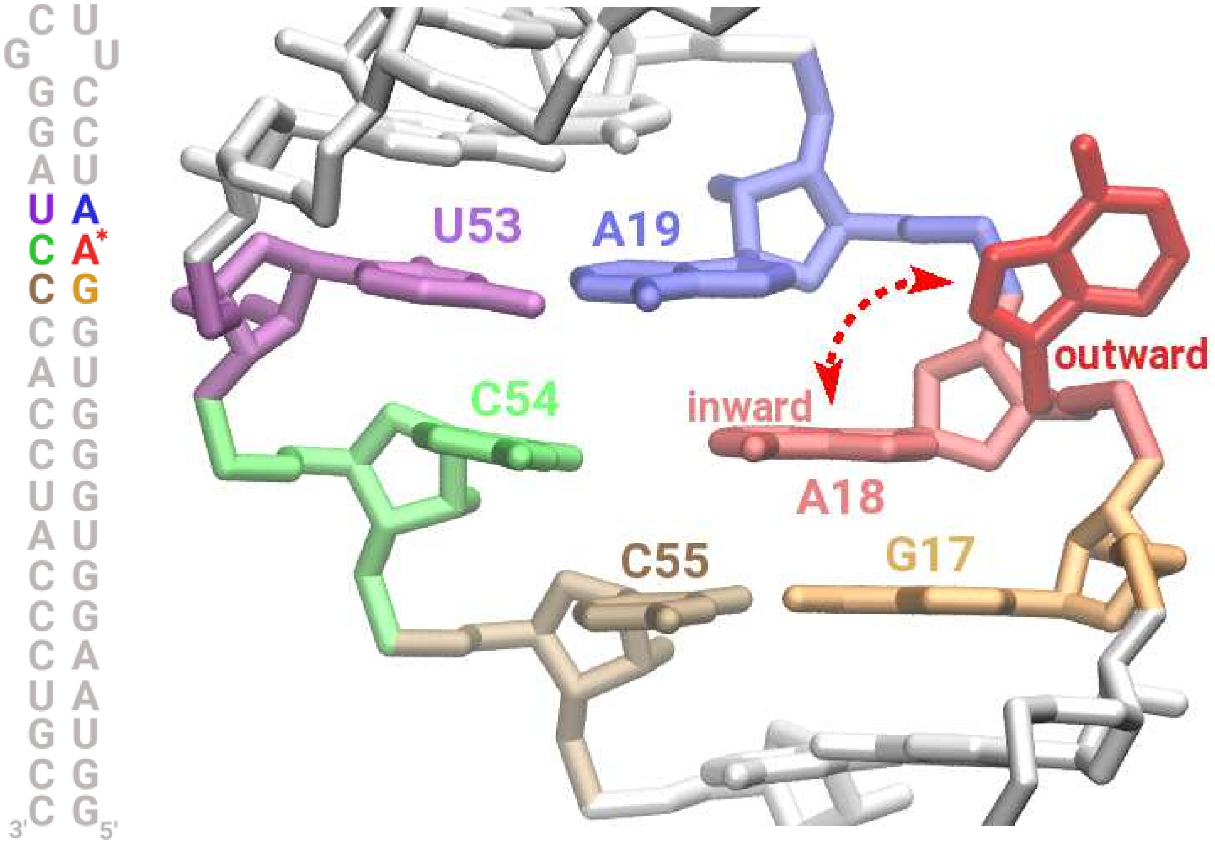

Figure 1.6: (*left*) Secondary structure of dsRNA with key nucleotides highlighted. The nucleobase studied in this work is marked with an asterisk. (*right*) A side-view of the three-dimensional structure of dsRNA where each key nucleotide is highlighted in a unique color and labeled. Specifically, A18 is shown both in the flipped in and out conformations.

binds the RRE RNA with a higher binding affinity and specificity than the Rev protein and displaces it to inhibit the viral replication process [93, 94]. The RRE RNA has been studied using X-ray crystallography [95, 96], electron microscopy [97], single-molecule fluorescence spectroscopy [98], circular dichroism [94], and MD simulations [99]. These studies collectively revealed how the Rev protein binds the RRE RNA, the overall assembly of the Rev domain and the RRE RNA, and proposed the importance of hydrogen bonding and salt bridging interactions between the Rev/RSG-1.2 peptides and the RRE RNA. However, a detailed mechanism of the binding/unbinding of the RSG-1.2 peptide has not been investigated. In my work I revealed the sequence of events that underlie the binding/unbinding mechanism and proposed the pathway with the smallest free energy barrier of dissociation.

1.5.2 Double-stranded RNA (dsRNA)

I also investigated the base flipping mechanism of a nucleobase in a dsRNA which is known to flip out and get chemically modified by an enzyme (Figure 1.6) [100]. Studying spontaneous base flipping is a challenging process both for experimental and computational methods due to a lower likelihood of observation of flipping in a single nucleobase in otherwise stable structures of nucleic acids. NMR methods have been applied to study base flipping [100–102], along with other experimental techniques including X-ray crystallography [103], fluorescence-based assays [104, 105], melting point studies [106], and combined approaches [107, 108]. In DNA, NMR studies have shown that the lifetime of the extrahelical state of a base can be on the order of μs, and that of the intrahelical state in the range of ms depending on the stability of individual bases [109, 110]. Additionally, several studies revealed that the base becomes accessible to the solvent for NMR detection when the base pair opens to a pseudo-dihedral angle of at least 30°, thereby indicating that the bases are still within the cutoff of a hydrogen bond formation [111, 112]. Therefore, the fluctuations measured by NMR may need to be reassigned to base wobbling as opposed to flipping and the mechanistic understanding may not be directly applicable to a base flipping process [108]. Thus, despite key mechanistic information emerging from the application of NMR methods, there remains the need for additional analyses at the atomic level for obtaining further insights into this molecular mechanism.

On the computational side, due to limitations in conformational sampling by conventional molecular dynamics (MD) simulations, enhanced sampling methods have been applied to probe this event [108, 111–119]. Among previous studies of base flipping, some have used external forces to induce base flipping transitions [111, 115], which likely leads to a loss of critical information on key variables that may contribute to base flipping. Enhanced sampling methods also rely on the definition of an appropriate reaction coordinate (RC) which is a single variable to discriminate between a given pair of stable states and using which the key thermodynamic (e.g. free energy) properties can be computed. Although establishing an appropriate RC is challenging [120], once it is identified the multidimensional free energy surface can be reduced to a one-dimensional profile along the RC to obtain crucial mechanistic insights into the transition mechanism. Many computational methods have been applied to study nucleobase stacking/unstacking in nucleic acids [113,114,117,121–125]. Several significant studies have been conducted to study the base flipping process in DNA in association with protein binding [126,127]. Several of these studies explored simplified systems that consisted only up to three base pairs and may be limited in describing the dynamics in a larger RNA system with many base pairs. [121–123] Additionally, several previous studies were reported over a decade ago and the force-fields for nucleic acids have significantly improved in recent years. [36,128] Moreover, the candidate variables that potentially contribute to RC have not been examined systematically. Therefore, the application of simulation methods that permit systematic testing of a suitable RC is needed to improve our understanding of the mechanism of base flipping in nucleic acids.

1.5.3 Porphyrin/DNA System

Nucleic acids are commonly used as templates to prepare supramolecular nanoassemblies due to their ability to form stacking and hydrogen bonding interactions [129,130]. Achiral porphyrins (Figure 1.3C) are perfect building blocks for nanoassemblies because they have the ability to self-stack and form hydrogen bonding interactions with ssDNA molecules. Highly ordered left-handed and right-handed supramolecular porphyrin nanostructures that were templated between two DNA strands can be prepared under different experimental conditions which control the orientation (handedness) of the overall assembly. However, despite knowing the orientation of each nanoassembly under specific experimental conditions, the detailed mechanism of self-assembly was not known. Therefore, I conducted MD simulations of an achiral stack of porphyrin derivatives (DPD molecules) with and without DNA strands (Figure 1.7) to understand the assembly mechanism and to identify the handedness of the nanoassembly.

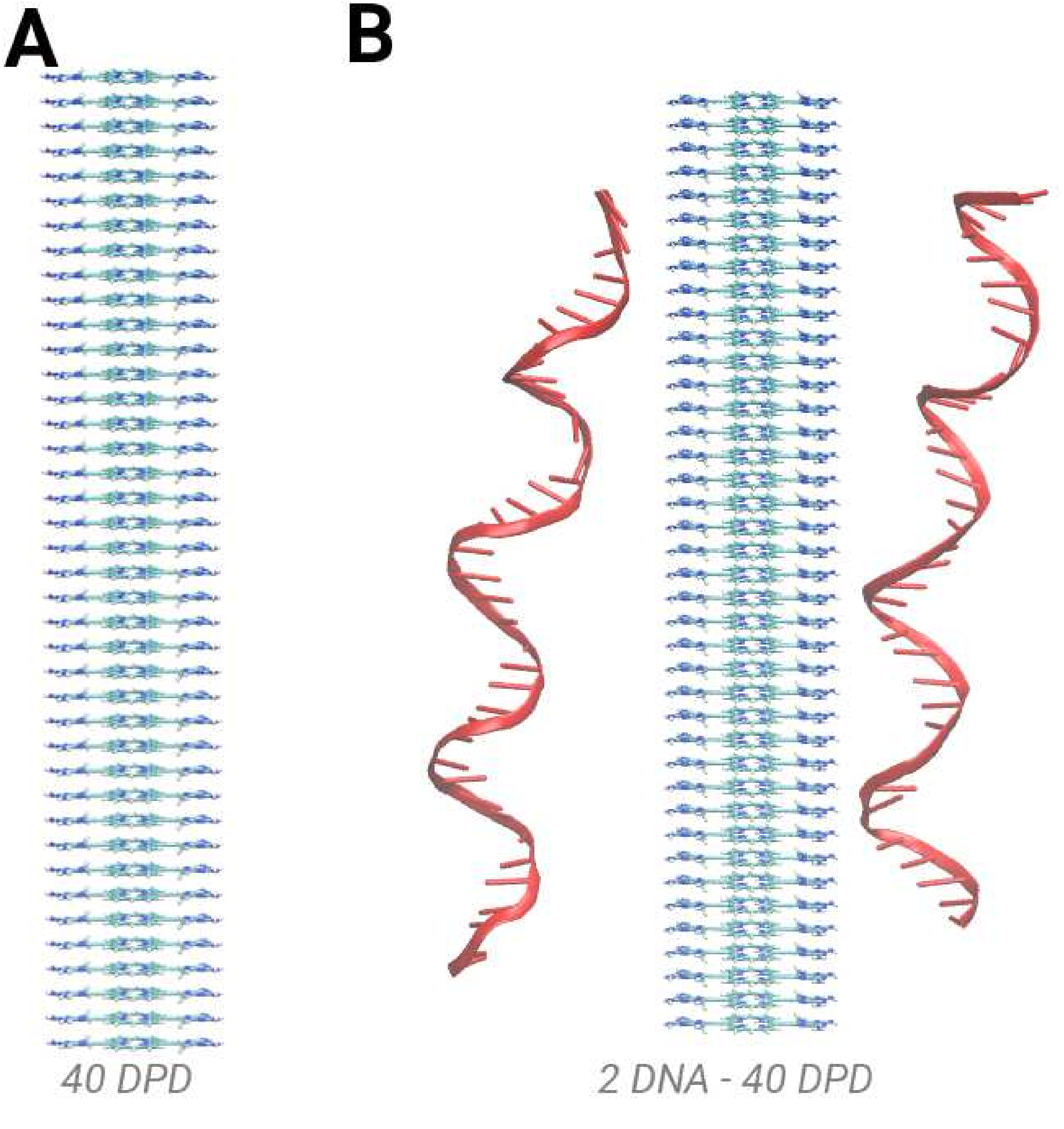

Figure 1.7: (A) A stack of 40 DPD molecules. Each molecule is shown in sticks representation. (B) A stack of 40 DPD molecules with 2 DNA strands are shown in sticks and cartoon representations, respectively.

1.6 Specific Aims

The main goal of my thesis is to investigate the dynamics and thermodynamics of nucleic acid molecules as well as to probe the interfacial interactions of nucleic acid molecules with various ligands. In this section, I introduce specific aims of my thesis.

1.6.1 Specific Aim 1: Probe the conformational dynamics in HIV-1 TAR RNA with/without ligands

RNA molecules are known to undergo conformational changes in response to environmental changes, but the dynamics in RNA molecules, with and without ligands, have been studied only to a limited extent. I conducted explicit-solvent MD simulations to study the dynamics in a model RNA system, the HIV-1 TAR RNA, that were initiated from 13 different initial conformations of the TAR RNA with ligands, and from 14 conformations without these ligands. By utilizing 27 different initial systems, I aimed to obtain a broader sampling of TAR RNA dynamics. These studies are reported in chapter 3.

1.6.2 Specific Aim 2: Characterize conformational transitions associated with recognition of a small molecule inhibitor by the HIV-1 TAR RNA

RNA has become an important target for developing novel therapeutic agents, however the conformational transitions that are coupled with ligand binding/unbinding are poorly understood. In this aim, I used non-equilibrium simulations to study the dissociation pathway of a small molecule inhibitor with low toxicity and high binding affinity from a binding pocket in TAR RNA. The study revealed several local conformational transitions in the nucleotides which constitute the binding pocket, specifically a base flipping event and a rotation of a base around its glycosidic bond in bulge nucleotides. Additionally, I have reported the free energy profile and the corresponding dissociation constant that describe the small molecule dissociation which are in reasonable agreement with the experimental values. These studies are reported in chapter 4.

1.6.3 Specific Aim 3: Probe the recognition mechanism of a helical peptide by the HIV-1 RRE RNA

In this aim, I conducted non-equilibrium simulations to probe the dissociation process of an arginine-rich helical peptide from the HIV-1 RRE RNA along four distinct pathways. These simulations revealed key interactions that were formed in a step-wise ordered pattern between specific amino acids of the peptide and specific nucleotides of the RNA molecule in each pathway. These interactions often occurred simultaneously, thus forming a network of salt bridging and hydrogen bonding interactions which are critical for the recognition of the peptide by the RRE RNA. Moreover, the analysis of the free energy profiles indicated the preferred pathway and the mechanism of peptide recognition. These studies are reported in chapter 5.

1.6.4 Specific Aim 4: Perform a systematic examination of collective variables to describe the base flipping mechanism in RNA

Base flipping is a critical biophysical event involved in recognition of various ligands by nucleic acids. The mechanism of base flipping in nucleic acids has been explored using various experimental and computational techniques. However, our understanding of molecular scale details of this mechanism still remains limited, specifically which interactions contribute the most to this event and which variables best characterize it. In this aim, I performed a systematic examination of collective variables (CVs) using transition path sampling methods in combination with likelihood maximization method to describe the base flipping mechanism. I have reported which CVs are key components of the base flipping mechanism and how they can be combined into a one-dimensional reaction coordinate (RC). I also report the free energy surface and transition dynamics which are projected along the determined RC. These studies are reported in chapter 6.

1.6.5 Specific Aim 5: Characterize the self-assembly and dynamics of porphyrin/DNA systems

In this aim, I conducted explicit-solvent MD simulations of several variations of porphyrin/DNA systems with different molecular compositions. These simulations revealed the dynamics and the preferred orientation of each porphyrin/DNA system. I characterized the resulting nanoassemblies using various physical variables. These studies are reported in chapter 7.

1.7 Thesis Outline

In chapter 2, I provide details on the computational methods, software tools, and mathematical models that I used in my thesis work. In chapter 3, I describe the results of a study on the conformational dynamics of the HIV-1 TAR RNA with and without ligands. In chapter 4, I describe the dissociation process of a small molecule inhibitor from the HIV-1 TAR RNA. In chapter 5, I describe the dissociation process of a helical peptide from the HIV-1 RRE RNA. In chapter 6, I present the results of a study on the local dynamics of bases in RNA, specifically on the base flipping mechanism in a dsRNA molecule. In chapter 7, I discuss a study on the self-assembly of porphyrin/DNA systems. In chapter 8, I share my thoughts on future work.

Appendices A and B provide supporting information, scripts, and analysis codes for a study presented in chapter 3. Appendices C and D provide supporting information, scripts, and analysis codes for a study presented in chapter 4. Appendices E and F provide supporting information, scripts, and analysis codes for a study presented in chapter 5. Appendices G and H provide supporting information, scripts, and analysis codes for a study presented in chapter 6. Appendix I provides parameter files for small molecules and porphyrins that I have generated in studies presented in chapters 3, 5, and 7. Appendix J provides links to media sources which have highlighted the study presented in chapter 4. Appendix K concludes my thesis with my curriculum vitae.

CHAPTER 2
MODELS AND METHODS

2.1 Introduction

One of the key hypotheses in molecular biology is that biomolecules are flexible and dynamic molecules and the atoms that constitute these biomolecules are in constant motion. Therefore, it is critical to understand the underlying atomic interactions and the dynamics of biomolecules which regulate their overall structure and function. Computer simulations have emerged as a tool to probe the dynamics of different many-particle systems, by capturing the motions of atoms and their interactions [131]. Therefore, computer simulations can be used to answer specific questions about properties and functions of various biomolecular systems. Specifically, conventional MD simulation is one of the most common computational technique to probe the equilibrium and transport properties of many-particle systems. In 1977, the first study on the dynamics of a folded protein was published which showed that MD simulations could capture the dynamic properties of proteins [131, 132]. During the following 40 years, the impact and variability of computer simulation methods in predicting various molecular motions and interactions have expanded dramatically. For example, computer simulations have been applied to study the protein structure and dynamics [133–136], the dynamics in nucleic acids [137–139], the protein-ligand interactions [140, 141], the ion channels [142], and the ligand binding [143, 144].

2.2 Molecular Dynamics (MD) Simulations

In my work, MD simulations were the primary method to explore the dynamics in nucleic acids and their interactions with ligands. MD simulations can capture different properties of molecular systems and provide crucial insights on the atomic details that underlie biomolecular processes. However, not all properties and quantities can be directly calculated in an MD simulation and vice versa certain quantities that can be directly estimated in a simulation cannot be tracked in an experiment [145]. A representative example is a simulation of liquid water in which we can measure the coordinates and velocities of each molecule (microscopic properties) at any instance of time [145]. However, there is no experimental method that can produce this kind of information, but rather it will provide us with the averaged properties across a large number of molecules (macroscopic properties) [145]. Statistical mechanics is used to connect the microscopic measurements from computer simulations with the macroscopic properties using laws of thermodynamics and Newton's laws of motions.

In a conventional MD simulation, Newton's second law is used to update atomic positions in time

$$\vec{F}_i = m_i \vec{a}_i = -\frac{\partial U}{\partial r_i} \tag{2.1}$$

where $\vec{F}_i$ is the force on a particle i with a mass m_i and an acceleration $\vec{a}_i$, U is the interatomic potential energy and r_i represents the Cartesian set of coordinates of a particle i. The potential energy term (U) is described in section 2.2.6.

2.2.1 Ensembles

The macroscopic state of a system is defined using macroscopic properties, including temperature (T), pressure (P), and volume (V). Other thermodynamic properties can be computed using equations of state or other fundamental equations of thermodynamics. However, in the microscopic state (microstate) we can obtain the coordinates and velocities of each particle in the system. To connect the dynamics of particles which are defined by their microscopic properties to the overall macroscopic properties of the system, a concept of an *ensemble* is defined. An ensemble is a collection of weighted microstates that have an identical macrostate [146]. In other words, a single macrostate corresponds to many microstates. Different types of ensembles exist with specific properties controlled and held fixed, including T, P, V, total number of particles (N), total energy (E), or chemical potential (μ). In my work, all MD simulations were conducted using either the NVT ensemble with fixed variables N, V, and T or the NPT ensemble with fixed variables N, P, and T. These ensembles are commonly used in the MD simulations since they consistently represent the experimental conditions.

2.2.2 Langevin Dynamics

A key requirement for any MD simulation method is to generate the correct ensemble at a specified temperature, pressure or volume. For this purpose, the Newtonian equation of motion of a particle (equation 2.1) is modified by adding a friction term which improves the stability of the system. For that matter, the Langevin equation is implemented in all MD software that I used in my work as follows

$$m\dot{v} = F(r) - m\gamma v - m\gamma\sqrt{\frac{2k_BT}{m}}R(t) \tag{2.2}$$

where m is the mass of a particle, $\dot{v}$ is the acceleration, $F(r)$ is the force, r is the position vector, γ is the friction coefficient, $v = \dot{r}$ is the velocity, k_B is the Boltzmann constant, T is the temperature, $R(t)$ is a univariate Gaussian random process. It is often advantageous to use smaller values of γ around 1 ps^{-1}, 2 ps^{-1} or 5 ps^{-1} to improve sampling [147,148] or stability of integration [149]. The equation 2.1 is modified by adding the dissipative ($-m\gamma v$) and the fluctuating (the last term) forces in order to mimic the viscosity of a solvent and the molecular collisions which are present in the realistic experimental systems.

2.2.3 Initial Conditions

In an MD simulation, the initial coordinates and velocities of each atom in the system must be specified before the simulation is launched. The initial coordinates for atoms can be obtained from the Research Collaboratory for Structural Bioinformatics (RCSB) website[1] which provides the Protein Data Bank (PDB) files for experimentally resolved structures of biomolecules. These structures have been resolved using various experimental techniques, such as X-ray crystallography, NMR spectroscopy, and cryo-electron microscopy. The initial velocities for atoms in the system are randomly assigned from a Maxwell-Boltzmann distribution:

$$P(v) = \sqrt{(\frac{m}{2\pi k_B T})^3} 4\pi e^{-\frac{mv^2}{2k_B T}} \tag{2.3}$$

where m is the mass of the particle, v is the velocity, k_B is the Boltzmann constant, and T is the temperature.

2.2.4 Numerical Integration

After defining all the initial positions and velocities of the particles, we need to perform numerical integration of the Newton's equations of motion. Various integration algorithms have been designed to perform this task and each algorithm has its specific advantages and disadvantages. However, the complexity of the physical and chemical systems which consist of thousands of particles as well as the stochastic nature of the evolution of these systems imply that the convergence of any integration algorithm is a challenging task [145]. Moreover, an integration algorithm is required to maintain the accuracy of system properties, such as the temperature and the pressure must be kept constant around the specified values in the NPT ensemble. In this respect, one of the simplest yet efficient algorithms is the Verlet algorithm and its variations [145]. A variation of the Verlet algorithm is the

[1] http://www.rcsb.org/

Brünger—Brooks—Karplus (BBK) method [150], which is a common algorithm to integrate the Langevin equation 2.2:

$$r_{n+1} = r_n + \frac{1 - \gamma\Delta t/2}{1 + \gamma\Delta t/2}(r_n - r_{n+1}) + \frac{1}{1 + \gamma\Delta t/2}\Delta t^2[m^{-1}F(r_n) + \sqrt{\frac{2\gamma k_B T}{\Delta m}}Z_n] \tag{2.4}$$

where Z_n is a set of Gaussian random variables of zero mean and variance of one. In the BBK method, only one random variable is needed for each degree of freedom. This method has a global error proportional to Δt^2 [150]. Another common integration algorithm is the velocity Verlet algorithm in which the velocity and position are calculated at the same timestep:

$$r_{n+1} = r_n + v\Delta t + \frac{F_n}{2m}\Delta t^2 \tag{2.5}$$

$$v_{n+1} = v_n + \frac{F_{n+1} + F_n}{2m}\Delta t \tag{2.6}$$

In the velocity Verlet algorithm, we need to compute the new positions first and only after that we can compute the new velocities which can be further used to compute the forces [145]. This method has a global error proportional to Δt^2 [145].

2.2.5 Integration Timestep (Δt)

It is crucial to set a proper timestep (Δt) for the numerical integration since it will determine the accuracy and convergence in MD simulations. A small timestep increases the accuracy of the simulations but simultaneously results in the increased computational costs. A larger timestep leads to increased sampling of the conformational space but causes instabilities in the simulation. Therefore, it is critical to select the timestep to achieve accuracy and convergence in the simulations. One of the requirements for the numerical integrators is that the timestep should be small enough with respect to the most rapid component of the motion [151]. Thus, a recommended timestep for the MD simulations of biomolecules, where the most rapid component is the motion of hydrogen atoms, is either 1 fs, if the bonds to hydrogen atoms are flexible or 2 fs, if the bonds to hydrogen atoms are rigid [152].

2.2.6 Potential Energy (U)

One of the most crucial components of any MD simulation is solving Newton's equations of motion which in return requires the calculation of the potential energy function (U). There are various models of the potential energy function that are commonly implemented in MD simulations. Each model is based on different assumptions and on different parameterization schemes that are used for determining the parameters of the corresponding model that is commonly referred to as a force-field. In my work, I implemented the Amber force-field which has been shown to be the most promising force-field for conducting MD simulations of nucleic acids [36, 153]. The functional form of the Amber force-field describes the bonded interactions (the first three terms in equation 2.7) and non-bonded interactions (the last two terms in equation 2.7) [154].

$$\begin{aligned} U(\mathbf{r}) = \sum_{bonds} K_b(b-b_0) + \sum_{angles} K_\theta(\theta-\theta_0) + \sum_{dihedrals} (\frac{V_n}{2})(1+\cos[n\phi-\delta]) \\ + \sum_{i<j} \epsilon_{ij}[(\frac{R^0_{ij}}{R_{ij}})^{12} - 2(\frac{R^0_{ij}}{R_{ij}})^6] + \sum_{i<j} \frac{q_i q_j}{R_{ij}} \end{aligned} \tag{2.7}$$

The first term in the potential energy function accounts for the bond oscillations around the equilibrium bond length of b_0 with the specified bond force constant of K_b. The second term describes the angle oscillations around the equilibrium angle of θ_0 with the specified angle force constant of K_θ. The third term accounts for the dihedral angles or torsional rotations where V_n is the amplitude, ϕ is the dihedral angle, n is the periodicity, and δ is the phase. Interatomic interactions between pairs of atoms (labeled as i, j in equation 2.7) are approximated by the 12-6 Lennard-Jones potential which is the fourth term. The Lennard-Jones potential represents the attractive and repulsive forces between pairs of atoms with the equilibrium interatomic van der Waals (VDW) distance of R_{ij}^0 and the potential well depth of ϵ_{ij}. The last term is the Coulombic potential which describes the electrostatic interactions between pairs of atoms which are represented as point charges (q_i or q_j in equation 2.7). Hydrogen bonding interactions are taken into account through the Lennard-Jones and Coulombic potentials. In order to conduct an MD simulation, all the above parameters should be specified. These parameters are computed using quantum-mechanical calculations and then compared against experimental data [36].

I used the Amber force-field for nucleic acids [155–158], TIP3P model for water [159], ions [160], peptides [161], small molecules, and porphyrins. The force-field for porphyrins and small molecules was designed using the general Amber force-field (GAFF) with the AM1-BCC charge method [162, 163]. The derived parameters for the small molecules are provided in Appendix I.2 and for the DPD molecule in Appendix I.3. In the next section, I provide a brief overview on the history of the Amber force-field for nucleic acids.

2.2.7 History of the Amber Force-Field for Nucleic Acids

The main factor that is responsible for all the interatomic interactions in the system is the potential energy function and the associated force-field. The force-field defines the functional form and the parameter set (e.g. K_b, b_0, K_θ, and etc.) for the potential energy function. Until now, the majority of nucleic acids simulations are performed using non-polarizable force-fields whose form was based on the work by Cornell *et al.* [164]. This force-field is often abbreviated as parm94 or as ff94, which was a modification of the Weiner *et al.* force-field [165], and became the first Amber force-field for simulations of proteins, nucleic acids, and organic molecules. The Cornell *et al.* nucleic acids force-field was considered to be a great success due to the choice of the scheme for fitting the atomic charges which led to a good description of the hydrogen bonding and stacking interactions [36]. This force-field was later modified by adjusting the pucker and the χ-dihedral profiles to yield the parm98 [166] and parm99 [167] force-field, which are alternatively abbreviated as ff98 and ff99.

After that, the development of Amber parameters for nucleic acids progressed along two primary pathways. One pathway is led by the Orozco group and is named after the Barcelona Supercomputing Center (BSC). In 2007, the bsc0 version of the Amber force-field for nucleic acids was released which improved upon parm99 by updating the α and γ dihedral angles in the backbone of nucleic acids [155]. These modifications prevented sampling of non-native γ-trans backbone dihedral states which led to the collapse of B-DNA and RNA structures. In 2016, this group released a new modification of this force-field, bsc1, which includes additional modifications to the sugar pucker, the χ glycosidic dihedral angle, and the ϵ and ζ dihedral angles [158]. Bsc1 force-field is considered to be a general force-field for simulating DNA systems [168] and I used it to simulate DNA strands in the study presented in chapter 7.

The other pathway is the collective research performed by various groups in Czech Republic from the city of Olomouc which gave the name to this series of force-fields, "OL". In 2010, this group released RNA-specific correction χ_{OL3} which reparameterized the χ glycosidic dihedral angle [156, 169]. As a result of this correction, the *anti* to high-*anti* χ shifts in RNA molecules which caused irreversible transitions into untwisted ladder-like structures, were suppressed. This modification also improved the description of the *syn* region and the *syn/anti* balance. It is a well-tested force-field and is on e of the recommended force-fields to use for simulating RNA molecules [36]. This group also released several separate modifications to the DNA force-field, specifically, χ_{OL4} modification to improve the χ glycosidic dihedral angle in DNA nucleotides [170], ϵ/ζ_{OL1} modification to improve the ϵ and ζ dihedral angles in the DNA backbone [171], and β_{OL1} modification to improve the β dihedral angle in the DNA backbone [172]. Combinations of these modifications to the DNA force-field are often referred to as OL15 and is a good alternative to the bsc1 force-field [168]. The group also developed a general pair potential (HBfix) which can tune particular non-bonded terms responsible for hydrogen bonding interactions in base pairs of RNA molecules [173]. This potential can be combined with other Amber force-fields and it does not affect any other interactions. The group further released a new version of HBfix which refines the simulations of RNA tetranucleotides (tHBfix) [174].

The Amber community has also benefited from the studies conducted by many other research groups. Specifically, the χ reparameterization by Yildirim *et al.* (χ_{YIL}) corrected the *syn/anti* balance in RNA structures on the basis of NMR data and the resulting effect was similar to the χ_{OL3} force-field [36]. An alternative set of torsions for RNA has been reported by the Rochester group ("ROC") which fit five backbone and four glycosidic dihedral parameters [157]. Another RNA force-field was developed by Shaw *et al.* ("Shaw") in which the charges and VDW parameters of the nucleobase atoms and the χ, γ, and ζ dihedral angles were modified to improve stacking and base-pairing interactions [175]. The level of accuracy of this force-field corresponds to the most promising protein force-fields which was tested by conducting 30 μs - 180 μs MD simulations of various RNA structures [175]. Cesari *et al.* performed a refinement of all dihedral angle potentials in the Amber RNA force-field using the data from solution NMR [176]. As the authors claim, simulations of RNA tetraloops using the corrected dihedral angle potentials showed good agreement with the experimental results, however additional testing of the derived parameters is required on larger RNA systems [176]. New OPLS-AA/M force field has been developed by the Yale group which optimized torsional potentials of the α and γ backbone dihedral angles [177]. In my work, I used the ff99+χ_{OL3} or the RNA.ROC force-fields for simulating RNA molecules.

Overall, a large number of Amber force-fields for nucleic acids as well as modifications or corrections have been released in the past decade. Each set of parameters has its own advantages and disadvantages, specifically several force fields require additional testing for a variety of RNA systems [176, 177]. The user needs to make a choice of which force-field or which set of modifications to apply to the system of study.

2.2.8 Boundary Conditions

There are three types of boundary conditions (BC) which can be specified in MD simulations: (*i*) vacuum, (*ii*) a reflecting wall, and (*iii*) periodic boundary conditions (PBC). The vacuum is the simplest BC which mimics a gas-phase environment but the dynamics of the global system properties will not reproduce the condensed phase [178]. The reflecting wall BC means that the particle is immediately reflected when it crosses the wall. PBC means that the system is placed in a simulation box and is considered to have infinitely many images in space [179]. Each simulation domain has 26 nearest neighbors in three dimensions. When the particle crosses the boundary of the simulation box on one side, an image of the particle enters the simulation box from the opposite side, and thus the overall number of particles in the system is conserved. In my work, I used PBC in all MD simulations.

2.2.9 Minimization

Even when initial structures are obtained from the experimental work, there may still be missing hydrogen atoms or other atoms in residues. Therefore, energy minimization of the initial coordinates is required to remove any potential steric clashes between atoms if the missing atoms are added during initial structure preparation. In my work, I used either steepest descents or conjugate gradient schemes prior to conducting MD simulations.

2.2.10 Temperature and Pressure Control

In all MD simulations reported in this thesis, I used the Langevin thermostat, where additional damping and random forces are introduced to the system. The temperature control is implemented through a frequent adjustment of momenta of all atoms in the system. The pressure is controlled using the Nose-Hoover barostat algorithm in all MD simulations in my thesis work [180–182].

2.3 Software Packages

2.3.1 MD Simulation Software

I used the version 18 of Amber software package (Amber 18) to conduct MD simulations for studies presented in chapters 3 and 6. Amber is a suite of biomolecular simulation programs that can be used to setup, perform, and analyze MD simulations [154, 183]. Specifically, the Amber software suit includes Ambertools which is a collection of freely available programs for analysis and set up of simulations [154]. The Amber code supports serial as well as parallel CPU and GPU simulations. Amber also refers to a set of molecular mechanical force-fields for the simulation of biomolecules which are available to the practitioners of biomolecular simulations.

I also used the NAMD software package (NAMD 2.12) to conduct MD simulations for studies presented in chapters 4, 5, and 7. NAMD is a parallel MD software designed for highperformance simulation of large biomolecular systems [152]. NAMD can be scaled to use hundreds of processors for conducting MD simulations and is compatible with Amber and CHARMM potential functions, parameters, and file formats [152]. Advanced techniques such as steered molecular dynamics (SMD) are also implemented in NAMD. This software is free for academic purposes and has an open source code which can be modified by users.

2.3.2 Conducting MD Simulations

As stated previously, I conducted MD simulations using the Amber and NAMD software suites with the Amber force-field in my work. Therefore, I used programs in the Ambertools package in combination with Visual Molecular Dynamics (VMD) to set up and analyze the systems. The resulting files were compatible with both NAMD and Amber software suits. Here, I provide a brief overview on conducting MD simulations.

(*i*) *Preparation of input files:* Before conducting MD simulations, I prepared coordinate files (PDB or CRD files), parameter/topology files (PARM files), and configuration files with simulation settings. The information on these files can be found on the Amber website[2]. During this step, solvent and ions are added to the simulation domain using LEaP program in Ambertools [154] or VMD [184]. At this step, I also used the Antechamber program to generate force-field files for small molecules and DPD molecules. I provide these files in Appendix I.

(*ii*) *Conducting MD simulations:* Energy minimization is the first step in MD simulations. After that, the simulations are continued either in the NPT or NVT ensembles. Frequently, a short (∼0.5 ns) MD simulation is conducted in the NPT ensemble to equilibrate the water box prior to conducting simulations in the NVT ensemble. All simulations were conducted on supercomputing resources at UNH or on Comet (San Diego Supercomputer Center). In the studies presented in chapters 4 and 5, I applied weak restraints to the phosphorus atoms in the RNA backbone while conducting enhanced sampling simulations.

(*iii*) *Data analysis:* The resulting trajectory and log files contain information on the atomic coordinates as well as other information depending on the type of simulation that is being conducted (e.g. force data in SMD simulations). In the following section, I describe the software that I used in my work to perform data analysis.

2.3.3 Modeling and Analysis

I used the software tool VMD 1.9 [184] to visualize and analyze trajectories generated by NAMD and Amber. VMD can also be utilized to perform solvation and ionization of the system. I used the Tk console available in VMD to execute Tcl scripts to analyze the following metrics in my trajectories: root mean squared deviation (RMSD), root mean squared fluctuation (RMSF), buried surface area (BSA), distance between atom pairs, hydrogen bond distances, salt bridge distances, etc. I also used the CPPTRAJ program [185] in Ambertools package to perform cluster analysis and average structure analysis and to compute various dihedral angles, and angles between bases.

[2] https://ambermd.org/tutorials/BuildingSystems.php

I also used MATLAB (ver. R2019a) to perform additional calculations, data analysis, and to create plots. I used the MDpocket tool [186] to analyze trajectories in the study presented in chapter 3. The MDpocket tool is an open-access pocket detection tool for MD trajectories.

All simulation preparations and data analyses were performed using the Linux operating systems (openSUSE). Therefore, I also wrote simple bash shell scripts (e.g. AWK and SED) for analyses. Gnuplot was another graphing utility that I used to create plots. Finally, I used various plugins in VMD for my simulation set up or analysis, including *SSRestraints*, *CatDCD*, and *NAMDEnergy*. All the scripts that I have generated in my studies are included in Appendices B, D, F, and H.

2.4 Enhanced Sampling Methods

Conventional MD simulations often cannot explore the entire conformational space due to a large number of degrees of freedom. Moreover, a system in an MD simulation frequently gets trapped in an energy minimum with high energy barriers to transition into a different state resulting in insufficient sampling. A good example is a ligand dissociation process which often cannot be fully captured using conventional MD simulations. Therefore, a variety of enhanced sampling methods have been developed to overcome these limitations of conventional MD simulations [187].

These methods often rely on the definition of one or more CVs, which can reduce the number of degrees of freedom and apply bias to the dynamics of the system in a controlled manner. Various variables could be used as CVs, for example the distances between atoms or groups of atoms, the angles between atoms or groups of atoms, the RMSD of the system, the secondary structure of the system, and so on[3]. The CV is also often termed as an "order parameter" or a "reaction coordinate". However, in my work, I have different definitions for the order parameter and the reaction coordinate which are further discussed in section 2.4.2

[3] https://www.ks.uiuc.edu/Research/namd/2.9/ug/node53.html

and these terms are not used interchangeably.

In my thesis, I used two enhanced sampling techniques, steered molecular dynamics (SMD) and transition path sampling (TPS) method. I applied SMD simulations to study the dissociation process of a small molecule and a helical peptide from viral RNA elements (chapters 4 and 5) and TPS method to study the base flipping mechanism in the dsRNA (chapter 6).

2.4.1 Steered Molecular Dynamics (SMD)

In SMD simulations, an external force is applied to an atom or a group of atoms, termed as SMD atom(s), to enhance conformational sampling in biophysical processes (e.g. ligand dissociation) that are difficult to observe in conventional MD simulations. Specifically, in my thesis work, I used constant velocity SMD (cv-SMD) simulation in which pulling is performed at a constant velocity (Figure 2.1). For simplicity, I refer to cv-SMD as SMD in my thesis. This method was first introduced in 1997 by Klaus Schulten *et al.* [188] and was inspired by the atomic force microscopy (AFM) experiment, in which a mechanical probe is used to obtain the force-extension data of various structures, including biomolecules. While the AFM experiments provided macroscopic insights into structurefunction relationships of various systems, the mechanism of these events at the atomic level and the underlying interactions were not fully understood [189]. SMD simulations later proved to be a reliable method which provided crucial details on the structurefunction relationship of macromolecular complexes involving proteins and ligands and complemented experimental data [189, 190].

In SMD simulations[4], the SMD atom is attached to a dummy atom via a virtual spring. This dummy atom is moved at a constant velocity and the applied force between both is measured using

[4] https://www.ks.uiuc.edu/Training/Tutorials/namd/namd-tutorial-html/node18.html

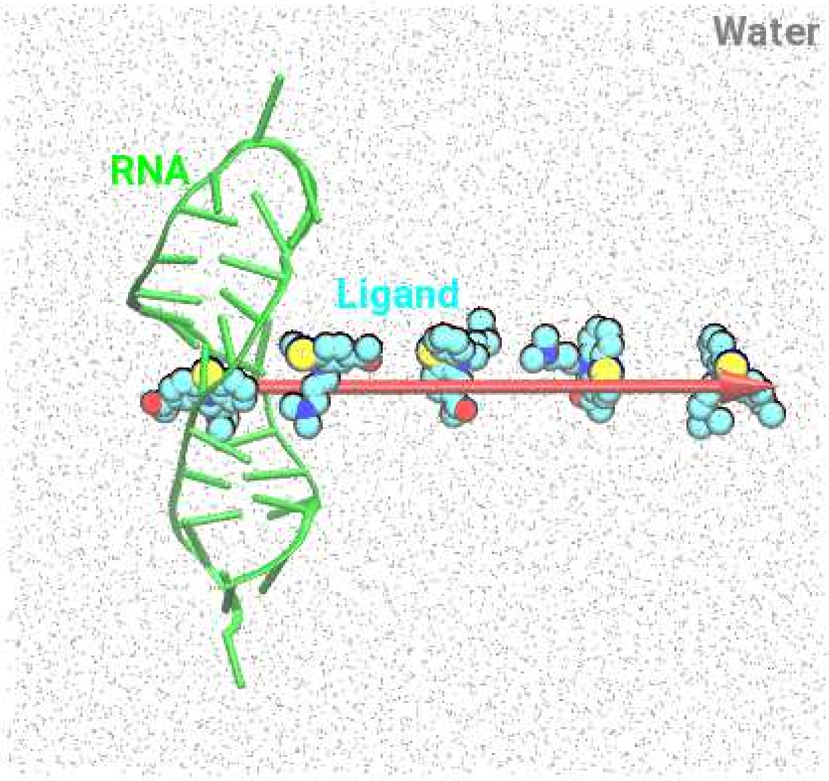

Figure 2.1: A side-view of the simulation domain during an SMD simulation: RNA, green cartoon; water molecules, gray points; and ligand, space-filling. The ligand is presented at various time points to show how it is being pulled along the reaction coordinate (depicted by a red arrow) which indicates the direction of pulling.

$$\vec{F} = -\nabla U \tag{2.8}$$

$$U = \frac{1}{2}k[vt - (\vec{r} - \vec{r_0}) \cdot \vec{n}]^2 \tag{2.9}$$

where U is the potential energy, k is the spring constant, v is the pulling velocity, t is time, $\vec{r}$ is the actual position of the SMD atom, $\vec{r_0}$ is the initial position of the SMD atom, $\vec{n}$ is the direction of pulling. The external work performed for a trajectory can be estimated by

$$W_{0\to t} = -kv\int_0^t (r - (r_0 + vt))dt \tag{2.10}$$

The second law of thermodynamics states that the average work done on the system is greater than the free energy difference between the initial and the final states of the system. The overbar denotes an average over an ensemble of measurements of W:

$$\Delta G \leq \overline{W} \tag{2.11}$$

The equality holds only when the process is carried out at an infinitely slow rate. However, in 1997, Christopher Jarzynski discovered an equality that relates a set of non-equilibrium processes between the two states to an equilibrium free energy difference between the same two states [191]. The resulting expression is

$$\overline{\exp(-\beta W)} = \exp(-\beta \Delta G) \tag{2.12}$$

or, equivalently,

$$\Delta G = -\beta^{-1} \ln \overline{\exp(-\beta W)} \tag{2.13}$$

where $\beta = 1/k_B T$. It is also possible to compute the free energy difference (ΔG) using the second-order cumulant expansion of the Jarzynski's equality which is computed as follows:

$$\Delta G = \langle \overline{W} \rangle - \frac{1}{2}\beta(\langle \overline{W}^2 \rangle - \langle \overline{W} \rangle^2) \tag{2.14}$$

where overbars represent averages over defined time windows, and angle brackets denote ensemble averages over independent cv-SMD simulations.

Thus, using the Jarzynski's equality we can estimate the equilibrium free energy difference from an ensemble of non-equilibrium processes. In my thesis work, I followed the protocol developed by Jensen *et al.* [192] who described how to relate the work values extracted from SMD simulations (equation 2.10) to the free energy difference computed per Jarzynski's equality (equations 2.12, 2.13, and 2.14) [193, 194].

2.4.2 Transition Path Sampling (TPS)

In my thesis, I used the TPS methods to study the base flipping mechanism in a dsRNA (chapter 6). Frequently, one wants to investigate a rarely occurring dynamical process that connects two metastable states of a system. This rare transition occurs rapidly, but observing it using conventional simulation methods is a challenging problem due to high free-energy barriers that separate the two metastable states and limit the sampling of the system to one of these states [195]. If this transition is captured and characterized, key mechanistic details on the transition can be identified [196].

Many enhanced sampling methods assume a prespecified CV that is intuitively selected for the process being studied, e.g SMD [190], metadynamics [197], adaptive biasing force method [198], umbrella sampling [199], and other methods [200–202]. However, for many reactions in complex systems with a lot of degrees of freedom it is a non-trivial task to identify an accurate RC because the reaction involves simultaneous changes in many degrees of freedom [120,203]. A reaction coordinate is a single variable that characterizes the dynamical mechanism of the transition and can discriminate between a given pair of stable states. Thus, the determined RC can be utilized in various ways: it is possible to monitor the events occurring during the transition along the resulting RC; the free energy can be projected along the resulting RC; rate constants of the transition can be computed using the RC [203, 204].

A technique that has been successfully applied to study rare events is the TPS method which generalizes basic Monte Carlo procedures to construct the transition path ensemble (TPE) [196, 205]. Specifically, TPS generates an ensemble of transition paths that connect a pair of initial (reactant) and final (product) states that are separated by a free energy barrier and, importantly, TPS does not require *a priori* knowledge of the reaction coordinate [196, 205, 206]. Now, I provide a more detailed overview on the individual steps that are performed during the implementation of the TPS used in my work.

(a) *List of Appropriate CVs:* Before conducting simulations, it is recommended to identify a list of all possible CVs ($\{X_k\}, k = 1, 2, \cdots, N_{tot}$) that can characterize the transition between the reactant (A) and the product (B) states. These CVs can be identified from intuitively observing the experimental structure, from conducting a biased simulation and exploring the resulting trajectory or from reading the existing studies on that system [207].

(b) *Definition of Stable States:* The next step of the TPS method is to define the reactant and the product states in terms of a single order parameter (X_p) [196, 208]. An order parameter is defined as a collective variable that can unambiguously discriminate between states A and B [196]. Each state is classified by a basin (region) of attraction, since the system experiences equilibrium fluctuations. Thus, the ranges of X_p should not only be large enough for each state to accommodate equilibrium fluctuations of the system, but they should also not overlap to prevent harvesting of the wrong basin [195, 196]. For example, I can define the range for state A as $X_p < c_1$ and the range for state B as $X_p > c_2$, then the region defined by $c_1 < X_p < c_2$ corresponds to a shooting region, which is located close to the transition region [203].

(c) *Initial Reactive Trajectory:* An important step in TPS is to obtain an initial reactive trajectory (seed trajectory) that connects states (basins) A and B. The seed trajectory can be generated by any tools available, as long as it sequentially connects states A and B. Several possible ways are conducting a long conventional MD simulation, conducting an MD simulation at an elevated temperature or conducting a biased MD simulation [195,209,210]. The resulting seed trajectory does not have to be a true dynamical pathway since it will eventually reach the TPE by successive sampling as per TPS algorithm [196].

(d) *Transition Path Ensemble (TPE):* After generating an initial reactive trajectory, an ensemble of unbiased MD trajectories (shooting trajectories) is generated between states A and B. The seed trajectory is used to select configurations (shooting points) that are located in the shooting region. The shooting points are then altered by sampling momenta afresh for each atom in the system from the Boltzmann distribution [203, 204]. The total energy of the system, as well as the total linear and angular momenta of the system are conserved. This is followed by the aimless shooting algorithm, or in other words, a number of short

MD simulations are conducted. The shooting trajectories have a high probability of rapidly relaxing to one of the stable basins because the aimless shooting algorithm that I used in my work generates shooting points near the barrier region [203, 204]. If the shooting trajectory terminates at either of the stable basins, it is accepted as a new transition path and used as a new seed trajectory and new shooting points are generated [203]. Otherwise, the trajectory is rejected.

(e) *Orthogonal Collective Variables:* The CVs that are identified in part (a), should be computed at the terminal and shooting points of each shooting trajectory. Each of the CVs at the shooting point is then normalized according to the following expression:

$$q_k = \frac{1}{\sigma_k}(X_k - \langle X_k \rangle), \tag{2.15}$$

where $\langle X_k \rangle$ and σ_k respectively represent the mean and standard deviation of X_k (CV), for all shooting points. The normalized variable has a mean of 0 and a standard deviation of 1. Thus, $\mathbf{q}(\{q_k\})$ represents a set of CVs to be tested in construction of the RC.

(f) *Determination of the Reaction Coordinate:* The RC is defined as a linear combination of the identified and normalized CVs as

$$r(\{\mathbf{q}\}) = a_0 + \sum_{k=1}^{m} a_k q_k, \tag{2.16}$$

where m is the number of OPs and is less than or equal to the total number of identified CVs, a_k's are adjustable parameters. For example, if the final RC consists of two CVs, then it will have a form of $r = a_0 + a_1 q_1 + a_2 q_2$. I applied the likelihood maximization method [203, 204] to find the best set of CVs and associated a_k's that are chosen to maximize the likelihood $\ln(L)$ and have the committor function p_O defined as the probability that a transition path, initiated from a shooting point, commits to the product state, O. Per aimless shooting algorithm and likelihood maximization methods, the committor is modeled as

$$p_O(r) = \frac{1}{2}[1 + tanh(r)], \tag{2.17}$$

and L is defined as

$$L = \prod_{x_k \to O} p_O(r(\mathbf{q})) \prod_{x_k \to I} [1 - p_O(r(\mathbf{q}))], \tag{2.18}$$

The products over $x_k \to O$ and $x_k \to I$ represent the product over all shooting points x_k committed to state O and I. By varying m in equation 2.16, different models of the RC can be investigated. For each model, a_k's are determined for each combination of q_k's by maximizing $\ln(L)$. The parameter a_0 is adjusted so that the transition between states I and O appears at $r = 0$.

The models of the RC with the same number of OPs ($m = n$) are then compared against each other using the maximum likelihood scores to pick the best combination of CVs. The best model with n parameters is then compared against the best model with n+1 parameters and the significance of the addition of an extra CV is evaluated using the Bayesian information criterion (BIC) [204], which determines when additional complexity of the model shows no further improvement or increased significance because an extra parameter in the model is significant only if the likelihood increases by a value larger than the value set by the BIC [204]. The BIC is applied using the following expression:

$$\text{BIC} = \frac{1}{2}\ln(N_{shoot}), \tag{2.19}$$

where N_{shoot} is the total number of shooting points that is generated. For example, if 1000 shooting trajectories are generated then by using $N_{shoot} = 1000$, the BIC is equal to 3.45 using equation 2.19. If $\ln(L)$ does not change more than this number on increasing model complexity, adding another parameter to RC is not considered significant.

(g) *Free Energy Profile along the RC:* The potential of mean force (PMF)/free energy profile is obtained along the RC using

$$G(r) = -k_B T \ln P(r), \tag{2.20}$$

where k_B is the Boltzmann constant, T is the temperature, and *P(r)* is the histogrammed population. Per Peters *et al.*, [204] if an RC is optimized using shooting points from TPS simulations, then the resulting RC based on the transition path ensemble is also a good RC in the equilibrium ensemble, thereby permitting equation 2.20 for obtaining the PMF.

2.5 Summary

I used conventional MD simulations in studies presented in chapters 3 and 7. I used SMD simulations in studies presented in chapters 4 and 5. I used TPS method in a study presented in chapter 6.

CHAPTER 3
STUDY ON THE ROLE OF CONFORMATIONAL HETEROGENEITY IN LIGAND RECOGNITION BY VIRAL RNA MOLECULES

3.1 Abstract

RNA molecules are known to undergo conformational changes in response to various environmental stimuli including temperature, pH, and ligands. In particular, viral RNA molecules are a key example of conformationally adapting molecules that have evolved to switch between many functional conformations. The TAR RNA from the HIV-1 is a viral RNA molecule that is being increasingly explored as a potential therapeutic target due to its role in the viral replication process. For work described in this chapter, I have studied the dynamics in TAR RNA in apo and liganded states by performing explicit-solvent MD simulations initiated with 27 distinct structures. I determined that the TAR RNA structure is significantly stabilized on ligand binding with especially decreased fluctuations in its two helices. This rigidity is further coupled with the decreased flipping of bulge nucleotides, which were observed to flip more frequently in the absence of ligands. I found that initially-distinct structures of TAR RNA converged to similar conformations on removing ligands. I also report that conformational dynamics in unliganded TAR structures leads to the formation of binding pockets capable of accommodating ligands of various sizes.

3.2 Significance

In the studies presented in this chapter, I reveal how global and local dynamics in viral RNA molecules is influenced by non-covalent ligand binding. To the best of my knowledge, this is the first work which utilized a large number of various initial conformations of the viral RNA molecule with/without ligands. I determined that ligand binding stabilizes the viral RNA structure which is characterized by decreased fluctuations of the two helices that comprise the RNA molecule and by decreased flipping of the bulge nucleotides. Additionally, I observed that initially diverse conformations of the viral RNA molecule became more similar in the course of MD simulations. These results enhance our understanding of the dynamics in viral RNA molecules and the role of ligand binding. Finally, I discovered the formation of binding pockets at various segments of the RNA molecule which could be potentially targeted with new therapeutic agents.

3.3 Background

RNA molecules have long been considered primarily as passive carriers of genetic information but this conception has changed in recent years due to enhanced understanding of the roles of RNA in different cellular processes including translation and transcription [40], regulation of gene expression [28], and protein synthesis [41]. RNA is also implicated in various diseases, including cancers, neurological disorders, and viral infections [29, 51, 211, 212]. This involvement of RNA presents an opportunity to target RNA by small-molecules for influencing the progression of various diseases [213]. Viral RNA molecules are a compelling target for small-molecule therapeutics since many viruses have RNA genomes, for example, HIV, hepatitis C virus (HCV), influenza virus, and severe acute respiratory syndrome coronavirus (SARS CoV/CoV2) [35, 53, 214].

The variability in functions of RNA is rooted in its ability to undergo conformational changes that often lead to complex three-dimensional folds and an ensemble of structures which determine the function of RNA [43, 48]. Conformational changes in RNA may be due to changes in physiological conditions or due to binding of ligands (proteins, small ligands, and ions) [90, 215, 216]. The conformational flexibility in RNA can also lead to formation of transient binding pockets that can be exploited for drug design [35, 55, 56]. While viral genomes encode for a limited number of protein targets, often considered "undruggable" [213], conserved and structured RNA motifs in viral genomes are viable targets to discover binding pockets for small molecules [35, 213].

Although the understanding of RNA dynamics has increased via different experimental techniques [49, 57, 213, 217], especially those of RNA structure determination, the role of dynamics of all structural motifs in RNA and their coupling to ligand binding has not been fully explored to date [218]. Several experimental techniques including X-ray crystallography and NMR spectroscopy can provide information on the dynamical properties of nucleic acids [58, 72, 219], but even these methods are often limited in probing all possible parameters that can fully describe the dynamics in RNA molecules [90, 220]. However, computational tools can further enhance the understanding of RNA dynamics by providing additional insights at the atomic level. Moreover, these tools can potentially assist in identifying binding pockets that can be explored in drug design.

TAR RNA (Figure 3.1A) from HIV-1 is a key model system to study conformational transitions in RNA molecules due to its ability to adapt multiple states [48, 58]. The TAR RNA is located at the 5′ end of HIV-1 transcripts where it interacts with the viral trans-activator (Tat) protein and the host cofactor cyclin T1 to promote efficient transcription of the downstream genome and is therefore considered to be an important drug target. Several previous studies have shown that TAR RNA can bind to peptide mimics [59, 75–79], to small molecules [80–83], to proteins [84–87], and to divalent cations [88] (examples are shown in Table A.1, Figure A.1). TAR RNA has been studied using NMR methods [58–68, 221], coarse-grained MD simulations [69], electron paramagnetic resonance (EPR) [70], gel mobility [71], combinations of NMR and MD methods [62, 72], and combinations of NMR and structure prediction software [73]. Collectively, these studies have shown that TAR RNA undergoes complex dynamics by sampling different interhelical conformations around the bulge junction, thus forming various conformational ensembles [48, 61]. While detailed dynamics

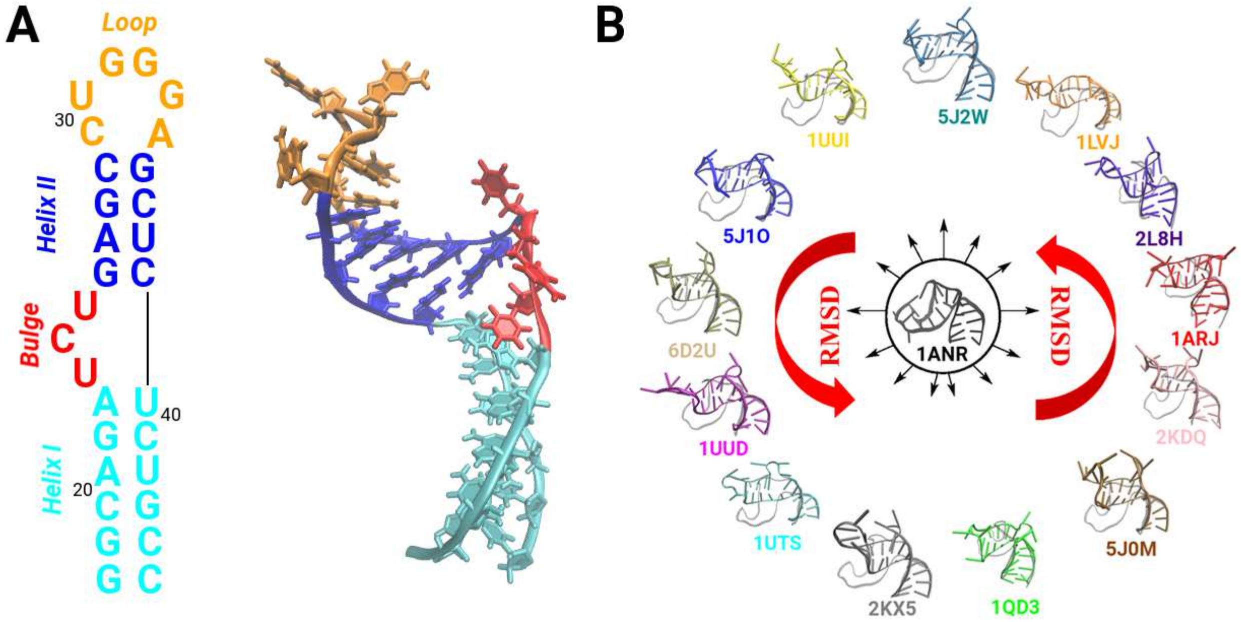

Figure 3.1: **Sequence and structural details of HIV-1 TAR RNA.** (A) Shown is the secondary structure and a snapshot of the three-dimensional structure of HIV-1 TAR RNA (PDB code 1ANR). Various structural motifs (Bulge, Helix I, Helix II, and Loop) are uniquely colored and labeled. (B) Shown are the snapshots of the initial states of TAR RNA (cartoon representation) in unliganded simulations. The *apo* conformation of TAR (PDB code 1ANR) is shown at the center (black cartoon) and superimposed onto other TAR RNA initial states. The initial conformations are placed in a circle such that the RMSD of the initial state relative to the *apo* conformation increases counterclockwise, with the 5J2W structure having the least RMSD and the 1LVJ structure having the highest RMSD.

in TAR RNA have not been characterized in the presence of all known ligands, it has been suggested that ligands can potentially induce structural transitions in TAR by stabilizing pre-existing conformers or an ensemble of states in the *apo* TAR RNA structure [66,89–91].

Several studies have revealed that a bulge motif in TAR RNA is especially critical for its recognition by the Tat protein [59,60,74]. Therefore, the bulge motif has been exploited in the design of inhibitors to disrupt the TAR/Tat interaction (Figure A.2A) [79,222]. As shown in Figure A.2A, peptide ligands mostly interact with the apical loop (orange in Figure 3.1A), helix II (blue in Figure 3.1A), and the bulge motif (red in Figure 3.1A), while small molecules are scattered between helices I and II (cyan and blue in Figure 3.1A). The ligands differ in size and the charge value, which results in an increased buried surface area (BSA) as the size of the ligand increases but the structural changes in TAR RNA are not correlated with the ligand size or BSA (Figure A.2B).

I studied dynamics in TAR RNA by conducting long time-scale MD simulations that were initiated from 13 different initial conformations of TAR with ligands, and from 14 conformations without these ligands, 13 conformations after removing ligands and one conformation based on the experimental *apo* structure (Table A.2; Figures 3.1B and A.3). By utilizing several different initial structures, I aimed to obtain a broader conformational mapping of TAR RNA which has not been carried out yet. Moreover, I studied the effect of ligand binding on the dynamics in TAR RNA by comparing unliganded and liganded simulations.

3.4 Methods

3.4.1 System Preparation

I have studied the dynamics in HIV-1 TAR RNA using 27 different initial conformations with/without ligands (Figures 3.1B and A.3). The initial coordinates for these conformations were obtained from the Protein Data Bank (PDB codes: 1ANR, 1ARJ, 1LVJ, 1QD3, 1UTS, 1UUD, 1UUI, 2KDQ, 2KX5, 2L8H, 5J0M, 5J1O, 5J2W, 6D2U) [59,60,75–83]. Several of these structures had either a different type of nucleotide or a different number of atoms in the deposited structure files. I selected the 1ANR conformation as the standard set of nucleotide sequence and mutated or removed those nucleotides or atoms in the other 13 systems that were different from the *apo* structure (PDB code 1ANR) which resulted in each TAR RNA system consistently having 29 nucleotides and 931 atoms (Table A.2). Each unliganded and liganded TAR structure was solvated in a periodic simulation domain of TIP3P water molecules where the overall number of atoms in various systems ranged between 19443 and 25647 (Table A.2). The overall charge in simulations of unliganded systems was neutralized with 29 Na^+ ions while the number of Na^+ ions in the liganded systems varied depending on the charge of the ligand.

3.4.2 Simulation Details

All MD simulations were carried out and analyzed using software packages Amber, CPPTRAJ and VMD [154, 184, 185] combined with the Amber force-field for RNA (ff99+χ_{OL3}) [155, 156] and for peptides (ff14sb) [161]. For solvent, TIP3P water model [159] and for ions the Li/Merz parameters were used [160]. The Antechamber package was used to design force-fields for small molecules by using the general Amber force-field (GAFF) with the AM1-BCC charge method (see Appendix I) [162, 163]. The temperature and pressure were maintained at 300 K and 1 atm using the Langevin thermostat and the Berendsen barostat. The steepest descent minimization was performed for 1000 steps followed by 100-500 steps of conjugate gradient minimization. The periodic boundary conditions were used with a cutoff of 9.0 Å for nonbonded interactions. Each of the 27 systems was subjected to a 2 μs long MD simulation in the NPT ensemble with a 2 fs timestep, which resulted in the overall 54 μs dataset and the frames in each trajectory were saved every 20 ps.

3.4.3 Conformational Metrics

Torsional Flexibility: The overall torsional flexibility of each TAR RNA structure was investigated by computing (from MD simulation data) all backbone dihedral angles and the

χ dihedral angle (which describes the relative position of a nucleobase relative to the sugar group). The dihedral angles were computed for each nucleotide in each system across the entire trajectory (100,000 values per MD trajectory). The resulting values of dihedral angles were then used to compute their normalized distributions. These distributions were further compared against the typical ranges of values of dihedral angles known from experimentally determined structures of nucleic acids that were extracted from the Protein Data Bank and reported in the textbook by Tamar Schlick (Figures 3.2A and A.4) [179].

Buried surface area (BSA): I calculated the BSA for each ligand in the liganded simulations. The BSA was computed using the following expression:

$$\mathrm{BSA} = \mathrm{SASA_R} + \mathrm{SASA_L} - \mathrm{SASA_{RL}}$$

where $\mathrm{SASA_R}$ represents the solvent accessible surface area (SASA) of the RNA, $\mathrm{SASA_L}$ represents the SASA of the ligand, and $\mathrm{SASA_{RL}}$ represents the combined SASA of the RNA/ligand complex. The BSA values indicate the area of contact between a ligand and the TAR RNA conformation.

Root mean squared deviation (RMSD): I calculated the all-atom RMSD for each system and for nucleotides in the bulge motif (U23, C24, and U25) to understand the effect of ligand binding on the overall TAR RNA structure. The alignment of each structure was performed against all non-hydrogen atoms in the initial state. The RMSD values indicate changes in the TAR RNA structure relative to a reference state. I used the initial conformations as well as the average structures computed from each simulation as the reference states.

Root mean squared fluctuation (RMSF): I computed the backbone phosphorous (P) atom based RMSF per residue to further study the flexibility of each nucleotide and ΔRMSF to compare the differences in dynamics between unliganded and liganded simulations. A negative value of ΔRMSF signifies decreased fluctuations in the presence of ligands or increased fluctuations in the absence of ligands.

Average Structure: I computed the average structures of TAR RNA from each unliganded and liganded simulation using the CPPTRAJ [185] program in the Amber software. The global rotational and translational motions were removed prior to computing the average structure. I then cross-compared all average structures using the RMSD as a comparison metric.

Clustering Analysis: To determine the population of similar TAR RNA conformations in MD simulations, I performed a clustering analysis using the CPPTRAJ [185] tool with DBSCAN [223] clustering algorithm. I used the P-atom based RMSD in each TAR RNA conformation as a distance metric and a minimum of 25 conformations (RMSD within ∼1.0-1.5 Å) were required to form a cluster. In addition to estimating clusters in individual trajectories, I performed combined cluster analysis (CCA) on the entire dataset of unliganded simulations and also on the entire dataset of liganded simulations by combining all trajectories. I used DBSCAN clustering algorithm with the minimum number of points set to 50 and the RMSD was set as a distance metric (RMSD within 1.9Å). To conserve memory, I used every second frame of each trajectory resulting in 700,000 and 650,000 frames for combined clusters of unliganded and liganded simulations, respectively. The initial "sieve" value was set to 40 to form initial clusters, which means that every 40$^{\text{th}}$ frame was used to generate an initial cluster, resulting in 17,500 and 16,250 initial frames from unliganded and liganded simulations.

Helical Dynamics: The TAR RNA structure consists of two helices (termed Helix I and II) that are linked by a flexible three nucleotide bulge motif (Figure 3.1A). The dynamics in these helices were characterized using the γ_1 and γ_2 angles, which describe the twist of each helix around the helical axis, and by the ϕ angle which describes the relative position of the helices. The ϕ angle was defined between the centers of mass of the two helices. For the calculation of angles, Helix I was defined by the base pairs G17-C45, G18-C44, C19-G43, A20-U42, G21-C41, and A22-U40 (cyan in Figure 3.1A), while Helix II was defined by the base pairs G26-C39, A27-U38, G28-C37, and C29-G36 (blue in Figure 3.1A). The axes in helices were defined as passing through the centers of mass of the bottom and top base pairs in respective helices and the CPPTRAJ [185] program was used to compute the twist angles.

Nucleotide Flipping: I characterized the flipping of nucleotides in the bulge motif (highlighted in red; Figure 3.1A) using the pseudo-dihedral angle computed between the centers of mass (COM) of four groups of atoms: the nitrogenous bases of U40 and A22, or C39 and C26, sugar moiety attached to A22 or C26, sugar moiety attached to U23, C24, or U25, and the nitrogenous base of U23, C24, or U25. The definition of the flipping angle was adopted from previous studies [124, 224]. In all systems, the inward (flipped-in) state of a nucleotide corresponds to pseudo-dihedral angle values between -60° and 60° but other values of the angle characterize an outward (flipped-out) state.

Binding Pocket Analysis: To characterize the binding pockets in TAR RNA conformations based on unliganded MD simulations, I used the MDpocket tool, which is an open-access pocket detection tool for MD trajectories [186]. Before analyzing each frame from MD simulations for pocket analysis, each trajectory was aligned to the initial structure based on the backbone P-atoms.

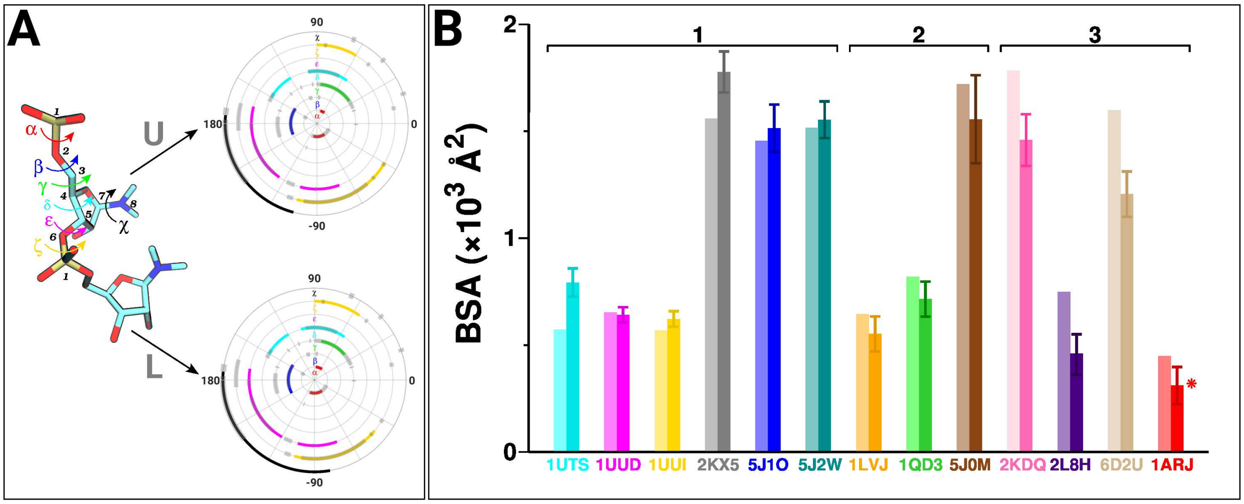

Figure 3.2: **Conformational metrics of torsional flexibility and BSA.** (panel A; *left*) All dihedral angles are shown by an arrow and labeled on a snapshot of the polynucleotide chain. The atoms in the chain are labeled as follows: 1, P; 2, O5'; 3, C5'; 4, C4'; 5, C3'; 6, O3'; 7, C1'; and 8, N9/N1. (panel A; *right, top and bottom*) The normalized distributions of each RNA backbone dihedral angle (α, β, γ, δ, ϵ, ζ) and the glycosidic dihedral angle (χ) for unliganded (U) and liganded (L) simulations. The transparent and thicker gray lines represent expected ranges of dihedral angles based on experimentally known RNA structures. See also Figure A.4. (B) The histograms of mean BSA values based on liganded MD simulations (darker shades) and the initial liganded structures (lighter shades) are shown. The error bars (vertical lines marked on histograms) were computed based on each liganded simulation. The BSA histograms are organized into three groups (labeled 1, 2, and 3; marked by overbars). A red asterisk highlights a system (PDB code 1ARJ) which exhibited a partial dissociation of the ligand. See also Figure A.6.

3.5 Results

3.5.1 Assessment of Torsional Flexibility and Ligand Stability

I first assessed the overall torsional flexibility of TAR by computing all backbone or glycosidic dihedral angles from unliganded and liganded simulations (Figures 3.2A and A.4). Specifically, the distributions of these dihedral angles (Figure 3.2A) were computed from the combined data of 14 unliganded simulations and similarly from the combined data of 13 liganded simulations. I found that the values spanned by these dihedral angles are consistent with the known ranges of dihedral angles in experimental structures of nucleic acids (marked by transparent gray lines on angle distributions in Figure 3.2A), thereby highlighting that the TAR RNA conformations generated with the AMBER force-field are consistent with the expected dynamics in the backbone of RNA structures.

I also assessed whether the ligands remained associated with each TAR RNA structure during MD simulations. In Figure 3.2B, I show the histograms of the mean values of BSA from liganded MD simulations (darker shades) along with the initial BSA values of ligands in various TAR structures (lighter shades). These BSA data are organized into three groups based on the distribution of BSA values in MD simulations in comparison to initial BSA values. I observed that most systems exhibited similar or higher ligand BSA values (e.g. group 1 systems with PDB codes 1UTS, 1UUD, 1UUI, 2KX5, 5J1O, and 5J2W) in comparison to the initial BSA due to conformational rearrangements of ligands in the binding pocket that led to a deeper burial of some ligands in the binding pocket (e.g., see Figure A.5). However, some systems exhibited fluctuations in nucleotides which allowed ligands to conformationally rearrange in the binding pocket and partially move out of the initial pocket, as indicated by the decreased BSA values (e.g. group 2 systems with PDB codes 1LVJ, 1QD3, and 5J0M; Figure 3.2B).

A larger decrease in BSA of the ligand was observed in systems organized in group 3 (e.g. PDB codes 2KDQ, 2L8H, 6D2U, and 1ARJ) indicating that the ligands in these systems

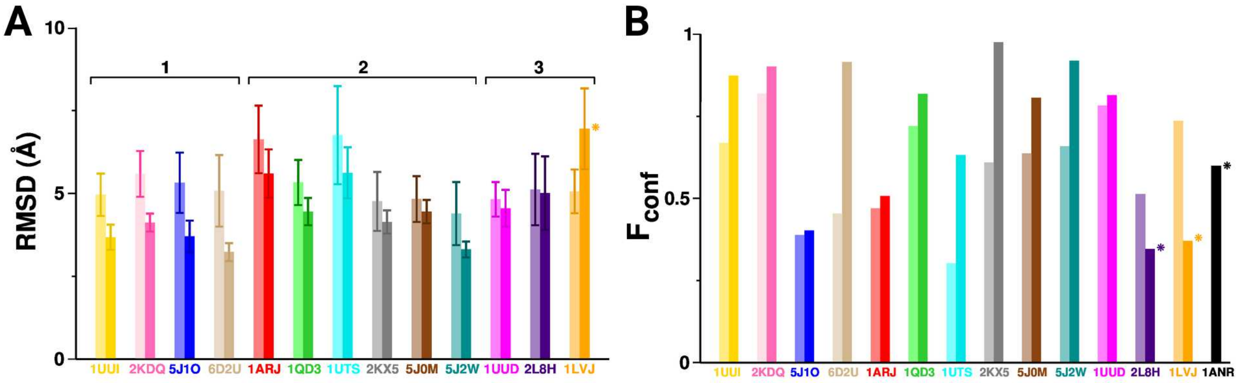

Figure 3.3: **RMSD and clustering analyses.** (A) Shown are the histograms (with error bars) of mean values of RMSD for all unliganded (lighter shades) and liganded (darker shades) simulations. The RMSD histograms are organized into three groups (labeled 1, 2, and 3; marked by overbars). An orange asterisk marks a system (PDB code 1LVJ) which showed a different behavior in comparison to other systems. (B) The fraction of conformations (F_{conf}) from a given simulation (100,000 conformations per simulation) that constitute the most populated cluster for each system in unliganded (lighter shades) and liganded (darker shades) simulations. Each bar corresponds to a unique system. The purple and orange asterisks indicate those systems in which F_{conf} was higher in unliganded simulations than in corresponding liganded simulations. A black asterisk marks the experimental *apo* TAR structure (PDB code 1ANR).

exhibited increased rearrangements in the binding pocket. For example, the ligand arginine amide in one of the TAR RNA structures (PDB code 1ARJ; marked by a red asterisk in Figure 3.2B) exhibited brief dissociation for about ∼18 ns before binding again in the initial binding pocket (Figure A.6). This observation is consistent with the largest dissociation constant for this ligand [59] and the smallest size of this ligand among all ligands studied (Table A.1). However, I did not observe full dissociation of any ligand during MD simulations.

3.5.2 Ligands Rigidify TAR RNA

To assess the differences in conformations of TAR RNA in unliganded and liganded states, I first computed the RMSD values with respect to the initial structure in each simulation (Figure 3.3A; lighter and darker shades for unliganded and liganded simulations, respectively). I observed that the unliganded systems diverged from their respective initial states on average by 5.28 ± 1.12 Å and the liganded systems by 4.52 ± 1.19 Å. For several systems, I observed a significant decrease in mean RMSD values and no overlap in error bars in liganded states compared to unliganded states (group 1; Figure 3.3A). About half of liganded systems fluctuated more and thus had less distinct RMSD distributions in comparison to respective unliganded systems, as characterized by overlapping error bars (group 2; Figure 3.3A). Despite having less distinct distributions, these liganded simulations still showed lower mean RMSD values in comparison to unliganded simulations (group 2; Figure 3.3A). Two systems (PDB codes 1UUD and 2L8H) had almost no difference in their RMSD distributions in liganded and unliganded states (group 3; Figure 3.3A). These data suggest that the TAR RNA structures became conformationally more rigid when ligands were present since liganded structures deviated to a smaller extent from their initial conformations in comparison to unliganded simulations.

However, one system (marked by an orange asterisk in group 3; Figure 3.3A) showed increased fluctuations and a higher mean RMSD in the liganded state in comparison to the unliganded state. On further probing this structure (PDB code 1LVJ), I found that the structural deviation is likely a result of ligand rearrangements in the binding pocket that conformationally altered the TAR RNA structure from an initially bent conformation to a relaxed conformation with a higher RMSD value (Figure A.7). The fact that the presence of a ligand caused higher perturbations in the TAR RNA structure that resulted in an unexpected conformational behavior needs to be considered in designing new inhibitors that target TAR RNA.

I further report the ΔRMSF values per residue based upon unliganded and liganded simulations to understand the effect of ligand binding on the flexibility of a particular residue or a motif (Figure A.8). In Figure A.8, I show the difference between the RMSF values (ΔRMSF) of the unliganded and liganded simulations where a negative value corresponds to an increased flexibility on ligand removal. I observed that in eight systems, all of the residues became more flexible when ligands were removed (e.g. systems with PDB codes 1QD3, 1UTS, 1UUI, 2KDQ, 2KX5, 5J0M, 5J1O, and 6D2U). In four systems, I observed larger flexibility in a portion of residues in the liganded simulations (e.g. systems with PDB codes 1ARJ, 1UUD, 2L8H, and 5J1O). Importantly, one of those systems (PDB code 5J1O) had only one residue (U25) with increased flexibility in the liganded state due to the flipping out of that residue during the simulation. Finally, in one system (PDB code 1LVJ) all of the residues became more rigid after ligands were removed which was consistent with the aforementioned observation that the fluctuations and mean RMSD for this specific system was higher in the liganded conformation (Figure 3.3A).

Due to the significance of the bulge motif in ligand recognition [59, 74], I also separately computed the RMSD values for the bulge motif (Figure A.9). Overall, the bulge motifs in unliganded and liganded systems deviated by 7.82 ± 1.83 Å and 5.6 ± 1.73 Å, respectively. In particular, I observed that the majority of systems had higher mean RMSD values in the unliganded states with the exception of the system with the PDB code 1LVJ that had a higher mean RMSD value of the bulge motif in the liganded state (Figure A.9). This observation is also consistent with ΔRMSF data that showed increased flexibility of the bulge nucleotides in the system with the PDB code 1LVJ (Figure A.8). The behavior of the bulge motif in this system is coupled with the ligand movement in the binding pocket and the local rearrangements of bulge nucleotides that result in the overall change in the conformation of TAR RNA (Figure A.7). These data suggest that the TAR RNA structures and the bulge nucleotides in liganded systems deviated to a smaller extent from their initial conformations compared to in the unliganded systems.

3.5.3 Comparison of Average Structures and Formation of Conformational Clusters

I further investigated the conformational variability of the data by comparing the average structures from each simulation and by performing a cluster analysis. Using RMSD as a conformational metric, I cross-compared all initial TAR RNA structures before simulations were initiated (Figure 3.4A) as well as by obtaining the average structures of TAR RNA from

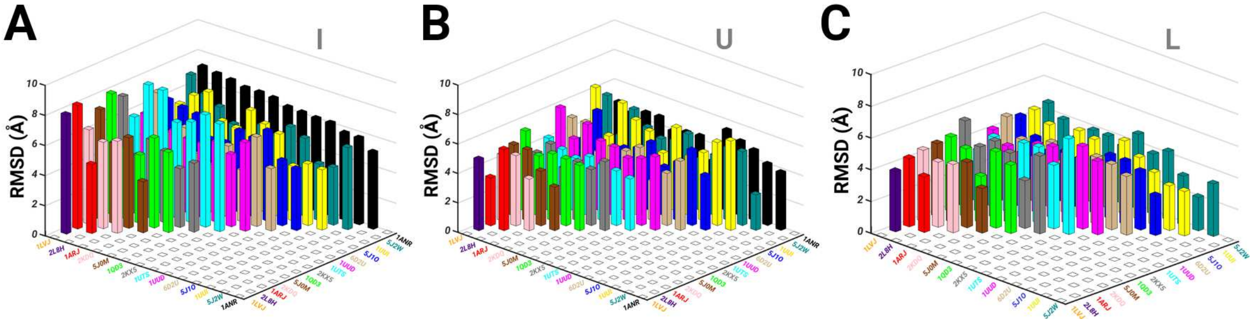

Figure 3.4: **Cross-comparison of initial, unliganded, and liganded TAR structures.** A cross-comparison of TAR RNA conformations via RMSD is highlighted for all structures in the initial states (panel A; labeled I) and based upon average structures derived from unliganded (panel B; labeled U) and liganded (panel C; labeled L) MD simulations.

each unliganded (Figure 3.4B) and liganded simulation (Figure 3.4C). This cross-comparison showed that the RMSD between a pair of structures decreased in unliganded simulations (Figure 3.4B), thereby indicating that the TAR RNA structures became on average more similar to each other in unliganded simulations. In liganded simulations, the RMSD between a pair of structures also decreased on average, but to a smaller extent in comparison to the unliganded simulations and the structures were still reasonably distinct (Figure 3.4C). For example, the initial RMSD between the structures with the PDB codes 1LVJ (orange) and 2L8H (purple) was 7.91 Å (dark purple bar in Figure 3.4A). After unliganded and liganded simulations, the RMSD between these systems for average structures was 3.8 Å and 4.8 Å, respectively (dark purple bars in Figures 3.4B,C). I also observed that the average structures of all systems obtained from unliganded MD simulations adopt conformations similar to the average structure obtained from an MD simulation of the experimental *apo* system (PDB code 1ANR) (Figure A.10).

To further assess the fluctuations and the flexibility in TAR RNA, I computed the RMSD values in the course of each simulation with respect to the average structure of the corresponding simulation (Figure 3.5). I observed that the unliganded systems diverged from their respective average structures by 3.07 ± 0.97 Å and the liganded systems by 2.31 ± 0.57 Å. The majority of the systems had a decrease in mean RMSD values and smaller magnitude

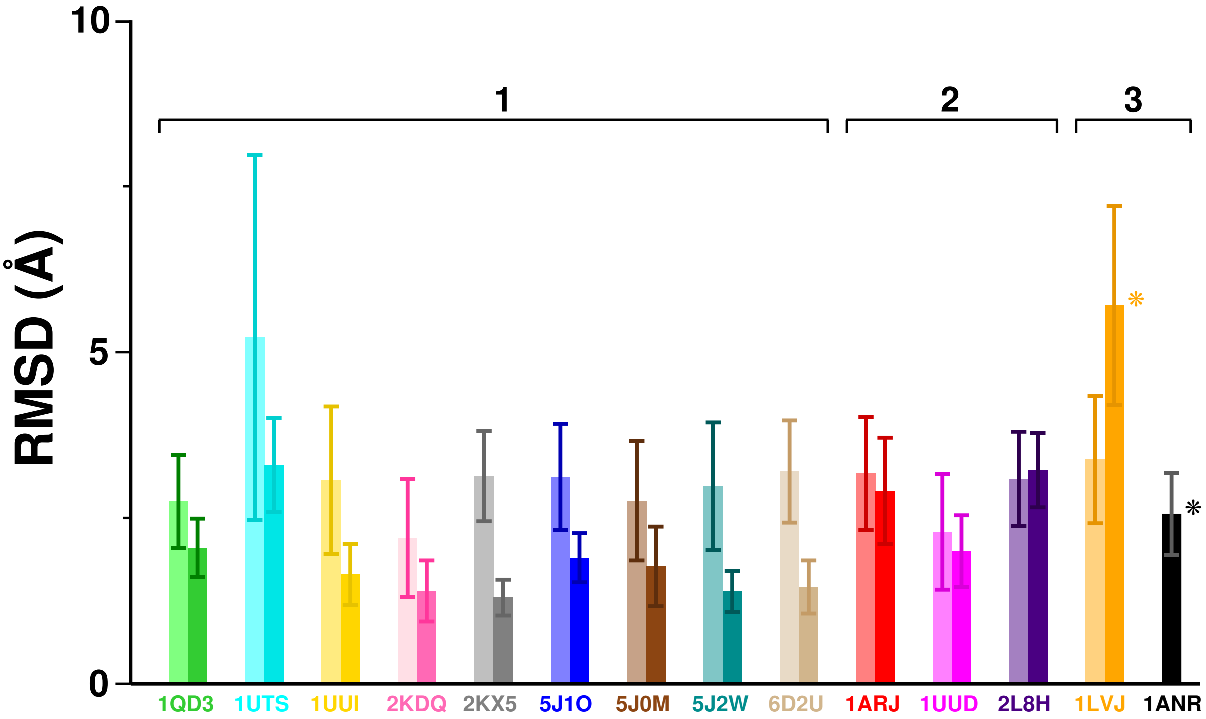

Figure 3.5: Shown are the histograms (with error bars) of mean values of RMSD computed with respect to the average structure for all unliganded (*lighter shades*) and liganded (*darker shades*) simulations. The RMSD histograms are organized into three groups (labeled 1, 2, and 3; marked by overbars). An orange asterisk marks a system (PDB code 1LVJ) which showed a different behavior in comparison to other systems. A black asterisk marks the experimental *apo* TAR structure (PDB code 1ANR).

of fluctuations in the liganded states compared to the unliganded states (group 1; Figure 3.5). Three systems exhibited very similar RMSD distributions in unliganded and liganded states (group 2; Figure 3.5), however, two of them still showed decreased mean RMSD values in the liganded state (PDB codes 1ARJ and 1UUD). Finally, one system (PDB code 1LVJ; Figure 3.5) showed increased fluctuations and a higher mean RMSD value in the liganded state in comparison to the unliganded state which is consistent with the observations described earlier. Overall, this metric showed that the fluctuations in the TAR RNA structures decreased in the presence of ligands.

In addition to comparing the average structures from each simulation, I performed clustering analysis to detect similarities among structures within each simulation and to understand the effect of presence of ligands on conformational variability in TAR RNA. In Figure 3.3B, I present the fraction of conformations in the most populated clusters derived from each unliganded and liganded simulation along with more comprehensive details on the distributions of clusters in Figures A.11 and A.12. I observed a larger variation in conformational clusters in unliganded simulations in comparison to liganded simulations. For example, only two unliganded systems (PDB codes 1UUD and 2KDQ) had a cluster that contained at least 75% of structures while eight liganded simulations (PDB codes 1QD3, 1UUD, 1UUI, 2KDQ, 2KX5, 5J0M, 5J2W, and 6D2U) had a cluster of this type (Figure 3.3B). The only exceptions were the unliganded systems with the initial structures based on PDBs 1LVJ (orange histograms in Figure 3.3B) and 2L8H (purple histograms in Figure 3.3B) that contained clusters with a higher fraction of conformations in the most populated cluster than in corresponding liganded simulations. Importantly, all liganded systems with peptide ligands (PDB codes 2KDQ, 2KX5, 5J0M, 5J1O, 5J2W, and 6D2U), except for the initial structure with the PDB code 5J1O, had the most populated cluster containing the majority of conformations (Figure 3.3B). This analysis further supports the observation that ligands in general rigidify the TAR RNA structure by restricting its motion within an ensemble of structures that constitute the most populated cluster.

I also performed the combined cluster analysis (CCA) to further investigate conformational clusters in TAR simulations that were initiated with distinct initial structures. In Figure A.13, I show F_{conf} for the top clusters from the datasets of unliganded and liganded simulations. The CCA of unliganded simulations revealed three clusters that contain more than 5% of the total number of configurations each and are composed of multiple systems (Figure A.13A). The CCA of liganded simulations, that was performed at the same value of the RMSD metric for constructing clusters as for the CCA of the unliganded simulations, revealed only one cluster which contains most of the systems (Figure A.13B). These observations show that a number of simulations, that were initiated from distinct initial structures, have conformations that are similar to each other.

3.5.4 Ligands Alter Helical Dynamics in TAR RNA

The TAR RNA structure is comprised of two helices (termed Helix I and II in Figure 3.1A) and the dynamics in these helices are described using the angles, γ_1 and γ_2 (Figure 3.6A,B), which describe the twist of each helix around the helical axis and by the interhelical bending angle (ϕ) which describes the relative positioning of helices (Figure 3.6C). I observed that the initial γ_1 values for all systems were between 29° and 34° except for one system (PDB code 2L8H) where the angle was 66.5°. Similarly, the γ_2 values were between 27° and 38° for all systems except one system (PDB code 1QD3) where it was 16°.

Based on the angle distributions from MD simulations, I observed that γ_1 was mostly confined between -90° and 90° for the unliganded and liganded systems (Figure 3.6A). Most of the liganded systems showed a decrease in the width of populated angles in comparison to the analogous systems in the unliganded form with the exception of the system with the PBD code 2L8H (depicted in purple color in Figure 3.6) which exhibited a higher twisting in the Helix I with γ_1 angles between 70° and 115°. The system with the PBD code 1ARJ in the liganded form showed values similar to the unliganded conformation given conformational rearrangements in the ligand and a decreased BSA (Figures 3.2, A.6). The average γ_1 values for the unliganded and liganded systems were estimated to be 19 ± 40° and 20 ± 33°, respectively. Overall, the presence of ligands decreased the standard deviation in γ_1 by 7°, thereby leading to narrower γ_1 distributions.

I also observed that Helix II showed more flexibility compared to Helix I in both liganded and unliganded simulations. In the unliganded simulations, γ_2 was mostly distributed between -90° and 105° with the exception of the structures with PDB codes 5J0M, 2KDQ, and 5J2W that spanned additional conformations between -105° and -140°, between -120° and -130°, and between -105° and -145°, respectively. The γ_2 angle in the liganded systems was mostly confined between -105° and 120° but even though the width of distributions were

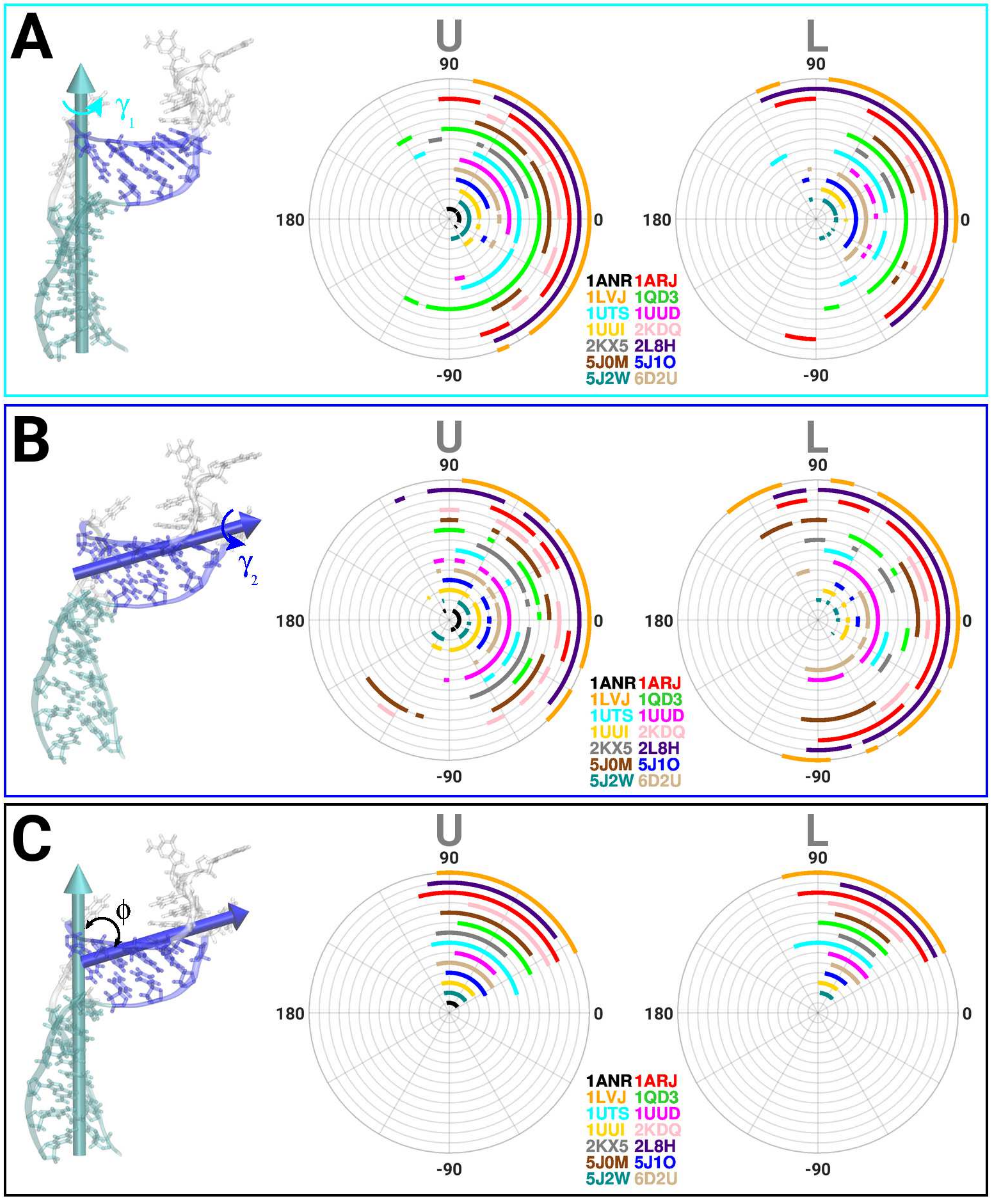

Figure 3.6: **Intrahelical and interhelical dynamics in TAR RNA.** (A; *leftmost panel*) A snapshot of the TAR RNA structure depicting the intrahelical angle γ_1, which describes the rotation of Helix I. The reference axis for the rotation of Helix I is marked by a cyan arrow. (A; *middle and rightmost panels*) The distributions of γ_1 are shown for the unliganded (U) and liganded (L) simulations of each structure. (B, C) Data similar to panel A are shown for the intrahelical angle γ_2, which describes the rotation of Helix II (panel B), and ϕ, the interhelical angle between Helix I and II (panel C). The reference axis for the rotation of Helix II is marked by a blue arrow. The color scheme in histograms is same as the PDB label.

similar, the number of states that were populated decreased. The average γ_2 values for the unliganded and liganded systems were estimated to be 12 ± 51° and 11 ± 46°. Overall, the presence of ligands decreased the standard deviation in γ_2 by 5°.

I observed that ϕ is mostly confined between 25° and 105° for the unliganded systems, with the exceptions of structures with the PDB codes 1QD3 and 2KDQ, which occupied states with angles between 25° and 80°, and 1UUD which occupied states between 45° and 85°. The distributions of ϕ in the liganded systems became narrower (ranging between 35° and 75°) with the exception of the structures with the PDB codes 1LVJ, 1ARJ, and 1UTS which spanned angles between 25° and 105°, 25° and 100°, and 45° and 110°, respectively. Overall, the structural bending in TAR RNA decreased in the liganded systems, except for the systems with initial states based on the PDB codes 1UUD, 1UTS, 1LVJ and 1ARJ which have distributions similar to their unliganded systems. The average ϕ for the unliganded and liganded systems was estimated to be 70 ± 14° and 60 ± 11°, respectively. While previous NMR analysis has suggested high amplitude bending and twisting motions in TAR RNA helices [72], these data further suggest that the conformations of helices in TAR RNA are altered and stabilized on ligand binding.

3.5.5 Ligands Stabilize Nucleotide Flipping in TAR RNA

Beyond global motions in TAR structures, I further probed local motions in key motifs such as the bulge region, which is considered important for the viral replication process because the rearrangements in nucleotides in this region (U23, C24, and U25) determine the orientation of helical motifs (Helix I and II) in TAR [59, 60, 74], Specifically, the outward flipped conformations of nucleotides C24 and U25 facilitate coaxial stacking of Helices I and II, thereby rigidifying the TAR structure. In Figure 3.7, I show the nucleotides used in defining the flipping angle (θ) and the time-traces of θ for three bulge-nucleotides, as obtained from unliganded and liganded simulations. The inward flipping of a nucleotide is characterized by θ values between -60° and 60°, and the outward flipping for all other θ

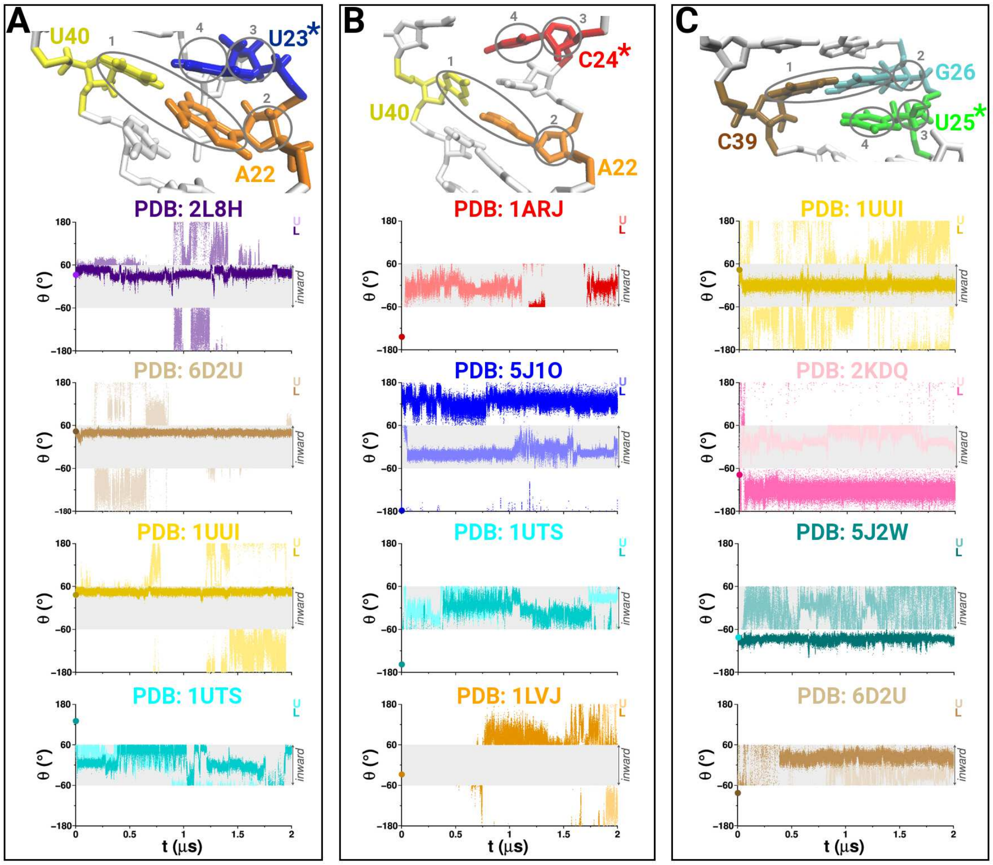

Figure 3.7: **Conformational transitions in bulge nucleotides.** (A) A snapshot of TAR RNA nucleotides (stick representation) used in defining the flipping angle (θ) for U23 (blue sticks and marked by an asterisk) and the traces of θ vs. simulation time (t) are shown for four systems in which a conformational transition was observed either in the unliganded state (U; lighter shade) or in the liganded state (L; darker shade) or in both. The initial value of θ for U23 in each system is marked on the y-axis by a filled circle in the same color as traces. The inward flipped state is characterized by θ values between -60° and +60° (labeled and shown by a transparent gray rectangle). All other values of θ indicate an outward flipped state. For those unliganded and liganded simulations where a transition occurred in both simulations, only those values of θ are plotted where a transition was observed. In case the transition was observed only in an unliganded simulation (or vice versa in a liganded simulation), in addition to plotting θ values in the unliganded simulation where the transition occurred, all values of θ are shown for the corresponding liganded simulation (or vice versa corresponding unliganded simulation) where the transition was not observed. (B and C) Data similar to panel A are shown for the flipping of nucleotides C24 (red sticks and marked by an asterisk; panel B) and U25 (green sticks and marked by an asterisk; panel C). The flipping angle of a nucleotide is defined by the center of mass of each of the following four groups: the nitrogenous bases of base-paired nucleotides (labeled 1) neighboring the flipping base, sugar moiety (labeled 2) attached to the base that is stacked with the flipping base, sugar moiety (labeled 3) attached to the flipping base, and the nitrogenous base (labeled 4) of the flipping nucleotide. See also Figure A.14.

values.

The first bulge-nucleotide U23 (Figure 3.7A) was initially in a flipped-in state in most structures, except in two structures (PDB codes 1QD3 and 1UTS) where it was in a flipped-out state (as marked by a filled circle on the y-axis at $t = 0$ in time-traces of θ in Figure 3.7A). I observed that U23 flipped out during five unliganded simulations (PDB codes 1ARJ, 1UUI, 2L8H, 5J2W, and 6D2U) in which U23 was initially in a flipped-in state (traces with lighter shades in Figures 3.7A and A.14). In most of these systems U23 eventually returned to its initial position after briefly transitioning to a flipped-out conformation. For example, in one of the unliganded simulations (PDB code 2L8H; light purple time-traces in Figure 3.7A) U23 flipped outward at $t = {\sim}0.9$ μs, maintaining the flipped out state for ${\sim}0.5$ μs, and then flipping inward, resuming its initial position.

I observed a similar conformational behavior in systems with PDB codes 6D2U (light brown time-traces in Figure 3.7A) and 5J2W (light cyan time-traces in Figure A.14A). In another unliganded system (PDB code 1UUI), U23 flipped out at $t = {\sim}1.25$ μs and remained in a flipped-out state until the end of the simulation (light yellow time-traces in Figure 3.7A). However, in one system (PDB code 1UTS) I observed conformational transitions in both unliganded and liganded simulations (cyan time-traces in Figure 3.7A), where U23 flipped inward from an initially outward conformation with $\theta = 130°$, retaining the inward position for almost the entirely of liganded simulation and flipping outward after ${\sim}1$ μs in the unliganded simulation. In all liganded simulations, U23 transiently flipped outward only in one system (PDB code 1ARJ; Figure A.14A). Overall, I observed that the presence of ligands significantly decreased conformational transitions in U23. This conformational behavior is consistent with observations from experiments, where a smaller pool of ligands (arginine amide and a linear as well as a cyclic peptide) were tested [225]. However, I consistently observed conformational stabilization of U23 for several ligands with different binding affinities.

The second bulge-nucleotide C24 (Figure 3.7B) was initially in a flipped-out conformation in all systems, except in three systems (PDB codes 1ANR, 1LVJ, and 1QD3). I observed that C24 flipped inward in three unliganded simulations (PDB codes 1ARJ, 5J1O, and 1UTS) (lighter shade time-traces shown in red, blue, and cyan in Figure 3.7B). For example, C24 flipped inward in one of the unliganded systems (PDB code 1ARJ; Figure 3.7B) and remained in the inward conformation until ∼1.15 μs. However, in other unliganded systems (PDB codes 5J1O and 1UTS), C24 flipped inward at the beginning of simulations where it either remained flipped inward during the entire simulation (PDB code 5J1O; Figure 3.7B) or flipped outward and then flipped back inward toward the end of the simulation (PDB code 1UTS; Figure 3.7B). For two unliganded systems where C24 was initially in an inward flipped conformation (PDB codes 1LVJ and 1QD3), it flipped outward toward the end of simulations (Figures 3.7B, A.14B). In liganded simulations, I observed that C24 flipped inward during two simulations (PDB codes 1ARJ and 1UTS; darker red and cyan time-traces in Figure 3.7B) and flipped outward during one simulation (PDB code 1LVJ; darker brown time-trace in Figure 3.7B). Overall, I observed that C24 showed conformational transitions both in unliganded and liganded simulations, but less frequently in liganded simulations.

The third bulge-nucleotide U25 (Figure 3.7C) was initially in a flipped-in conformation in most structures except in three structures (PDB codes 2KDQ, 5J2W, and 6D2U). I observed U25 to be significantly flexible in both unliganded and liganded simulations since it flipped inward or outward in most of the systems. For example, U25 flipped outward from an initially inward conformation in several unliganded simulations (PDB codes 1UUI, 1UTS, 2KX5, 2L8H, and 5J1O; Figures 3.7C and A.14C). It also flipped inward from an initially outward conformation during unliganded simulations of several systems (PDB codes 2KDQ, 5J2W, and 6D2U) in which it remained in the inward conformation until the end of each simulation (Figure 3.7C). In several liganded simulations (PDB codes 1ARJ, 1UTS, 2L8H, and 5J1O), U25 flipped outward from an initially inward conformation (Figure A.14C), while in other systems (e.g. PDB code 6D2U) it flipped inward from an initially outward conformation (Figure 3.7C). Overall, I observed that all three bulge nucleotides (U23, C24, and U25) can

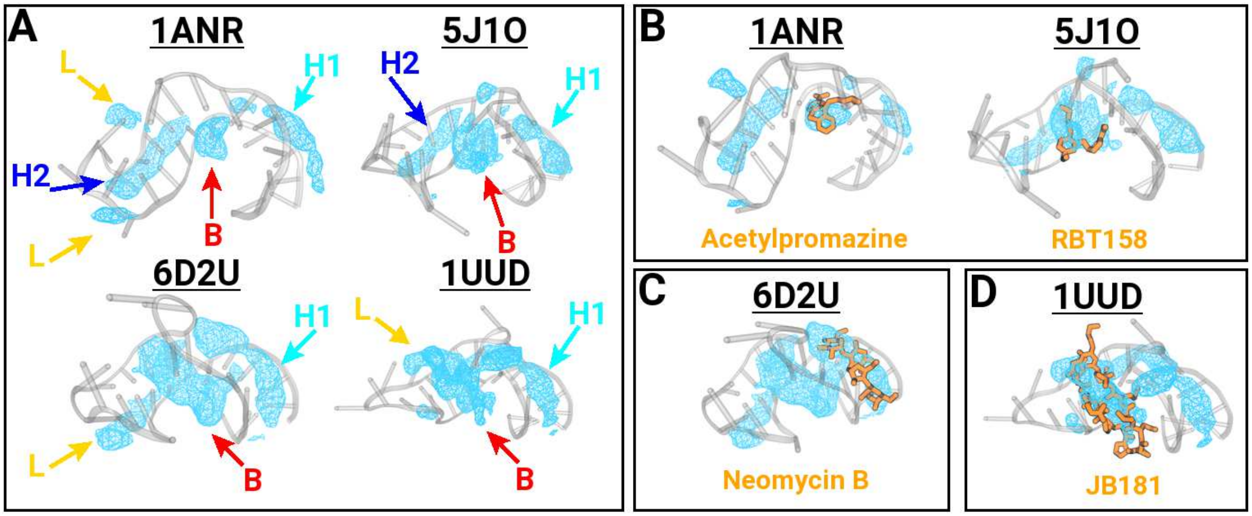

Figure 3.8: **Predicted binding pockets in unliganded TAR structures.** (A) Predicted binding pockets (cyan surfaces) are shown overlaid on each TAR RNA structure (transparent white cartoon). (B-D) Snapshots of the overlay of each ligand (orange sticks) on the predicted binding pocket where the ligand is known to bind in each structure. See also Figures A.15 and A.16.

conformationally transition between inward and outward states although ligands decrease the frequency of these transitions.

3.5.6 Conformational Dynamics in TAR RNA Reveal Ligand Binding Pockets

Given that the knowledge of binding pockets is useful in developing novel inhibitors [213], I probed all unliganded simulations for the presence of binding pockets that may form as a result of conformational dynamics. In Figures 3.8A and A.15, I show several binding pockets (depicted as cyan surfaces overlaid on the initial structure) that appear in various regions of each unliganded TAR RNA structure (labeled B, L, H1, and H2 for the bulge region, loop region, and helices I and II, respectively; see also Figure 3.1A).

I also assessed whether the density of pockets observed in specific regions in a structure could accommodate ligands in conformations observed in liganded TAR RNA structures. I found that the observed pockets were sufficiently large in size to encapsulate the ligand known to bind to that specific TAR RNA structure. For example, TAR RNA is known to bind to small-molecule ligands (acetylpromazine and RBT158) in the bulge region, where I observed binding pockets (labeled B for PDBs 1ANR and 5J1O in Figure 3.8A) large enough to accommodate each ligand (Figure 3.8B). Furthermore, TAR RNA is also known to bind other ligands (neomycin B and JB181) in the Helix I and the apical-loop/bulge regions, where I observed binding pockets (labeled H1 for PDB 6D2U and L/B for PDB 1UUD in Figure 3.8A) large enough to accommodate respective ligands (Figure 3.8C, D). I observed that unliganded simulations with different initial structures showed several similar binding pockets as well as previously unknown binding pockets (in the apical loop or Helices I and II) that accommodated ligands known to experimentally bind to other conformations (Figure A.16).

3.6 Discussion

In this study, I have carried out long time-scale MD simulations (totaling 54 μs) of the HIV-1 TAR RNA structure in unliganded and liganded states. Specifically, these simulations were conducted with initial coordinates derived from the experimentally resolved *apo* structure of TAR RNA (PDB code 1ANR) as well as 13 other structures of TAR RNA that were bound to a variety of ligands including small-molecules and peptides. To increase the pool of unliganded simulations, I also conducted simulations of 13 liganded TAR RNA structures by removing ligands and retaining the initial coordinates for RNA atoms. I aimed to probe conformational heterogeneity in ensembles of TAR RNA structures in unliganded and liganded states to understand the predisposition of the unliganded TAR conformations to ligand binding and the effect thereafter.

I initially assessed the overall torsional flexibility of TAR RNA by computing distributions of all backbone dihedral angles from unliganded and liganded simulations (Figures 3.2A and A.4) and found these distributions to be consistent with the range of values from experimentally known structures of nucleic acids. This observation supports the ability of the interatomic potential in adequately capturing the dynamics in TAR RNA structures. By computing the buried surface area (BSA) of each ligand, I also assessed the stability of ligands during long time-scale MD simulations and found that in most TAR RNA structures ligands remained bound throughout simulations except in a few cases where ligands conformationally rearranged and/or partially dissociated.

I then probed the global dynamics in TAR RNA by comparing all unliganded and liganded conformations from MD simulations using global RMSD, ΔRMSF, and clustering analyses (Figures 3.3, A.8, and 3.5). The primary observation from the RMSD analysis was that the mean RMSD of unliganded conformations was higher than the mean RMSD of liganded conformations, thereby suggesting decreased conformational fluctuations in TAR RNA on ligand binding. The analysis of ΔRMSF further supported this observation since the magnitude of fluctuations in nucleotides was smaller in the liganded simulations than in the unliganded simulations. These observations are consistent with the notion that RNA molecules are stabilized by binding of ligands because liganded TAR RNA structures were conformationally more rigid compared to unliganded structures.

However, one of the small molecules, acetylpromazine, resulted in distinct perturbations in the overall structure of the liganded TAR RNA in comparison to other liganded systems and in comparison to the corresponding unliganded simulation. In the presence of acetylpromazine, the TAR RNA structure transitioned between two distinct (bent and stretched) conformations (Figure A.7). I have previously also shown that the (un)binding process of acetylpromazine is associated with the flipping of nucleotides in the binding pocket [226]. The main structural difference between acetylpromazine and other small molecules is the presence of a sulfur moiety in one of the benzoic rings (Figure A.1), which could be an important design feature for future development of inhibitory compounds.

The bulge motif which connects two helices in the TAR RNA structure (Figure 3.1A) also became more rigid in the presence of ligands, which decreased the twisting and bending fluctuations in TAR RNA helices. The clustering analysis further supported these observations by showing a higher fraction of similar conformations in liganded structures in comparison to unliganded structures. In fact some liganded structures with peptide ligands exhibited a single cluster containing more than 90% of the RNA conformations, implying that these liganded simulations exhibited small conformational variability in the presence of peptides since the TAR RNA conformations were similar to each other (e.g. PDB code 2KX5; Figure. 3.3B and A.12). I also observed that various simulations have similar conformations that form combined clusters by performing combined cluster analysis (Figure A.13) and by comparing average structures from each simulation (Figures 3.4, A.10). These observations further support the previously proposed hypothesis [48, 61, 220] that despite being a highly flexible molecule, TAR RNA potentially adopts a set of conformations forming an ensemble of structures that can recognize various ligands.

I further probed the local dynamics in bulge nucleotides (U23, C24, and U25), the conformational flipping motions in which facilitate ligand binding [59, 60], as well as alter global dynamics in TAR RNA. As opposed to the notion that the binding of ligands may prevent conformational transitions in nucleotides, I observed that the bulge nucleotides can transition between inward and outward conformations in both unliganded and liganded states although the frequency of transitions significantly decreases in the presence of ligands. Overall, I found bulge nucleotides C24 and U25 to be more flexible than U23.

Importantly, as a result of conformational heterogeneity in TAR RNA structures and coupling between local and global dynamics, I observed the formation of ligand binding pockets near several structural motifs (bulge region and helices I/II). This observation is consistent with the suggestion that TAR RNA may adopt conformations with pre-existing binding pockets where ligands can fit [48]. I observed that these binding pockets form consistently in all unliganded simulations with enough volume to accommodate different ligands (Figures 3.8, A.15, and A.16), including larger ligands (e.g. peptides). Moreover, I observed the formation of binding pockets in other structural motifs (e.g. the apical loop) in TAR RNA which are potentially useful to future inhibitor design. As an example, Patwardhan et al. [222] showed that the amiloride ligands can bind to nucleotides in the apical loop where I observed several binding pockets.

3.7 Conclusions

Although RNA molecules are known to undergo conformational changes during various cellular processes, the conformational dynamics in RNA molecules, with and without ligands, have been studied only to a limited extent. I used explicit-solvent MD simulations to study the dynamics in a model RNA system, the HIV-1 TAR RNA, which is known to recognize several types of ligands including small-molecules and peptides. I observed that the ligands rigidified TAR RNA structures by interacting with the bulge nucleotides and decreased the overall bending and twisting motions in helical motifs in TAR RNA. Therefore, I found that ligands overall decreased conformational heterogeneity in TAR structures. While RNA is considered a highly flexible molecule, I observed that TAR RNA structures on average became more similar to each other in the unliganded and liganded simulations compared to their initial conformations. I also observed that the conformational transitions leading to flipping of nucleotides in RNA molecules likely occur irrespective of the presence of ligands although the frequency of these transitions decreases on ligand binding. As a result of conformational heterogeneity, I also showed that unliganded RNA molecules possess ligand binding pockets that may be amenable to targeting by novel inhibitory molecules.

3.8 Supporting Information

Additional data and figures are shown in Appendix A. I have performed preliminary analysis of the principal components which is also presented in Appendix A. In Appendix B, I also provide example scripts that I used to set up, conduct, and analyze my simulations. I have also included in Appendix B the scripts for creating figures. For sharing with the scientific community, I have further made the simulation data available via the Zenodo platform (`https://doi.org/10.5281/zenodo.4521164`).

3.9 Publication

The work described in this chapter is reproduced from Ref. [227], with permission from the Royal Society of Chemistry. The citation is as follows:

Levintov, L., and Vashisth, H. (2021). Role of Conformational Heterogeneity in Ligand Recognition by Viral RNA Molecules. *Phys. Chem. Chem. Phys.* doi: 10.1039/D1CP00679G.

CHAPTER 4
STUDY ON THE BINDING/UNBINDING PROCESS OF A SMALL MOLECULE FROM A VIRAL RNA MOLECULE

4.1 Abstract

RNAs are conformationally flexible molecules that fold into three-dimensional structures and play an important role in different cellular processes as well as in the development of many diseases. RNA has therefore become an important target for developing novel therapeutic approaches. The biophysical processes underlying RNA function are often associated with rare structural transitions that play a key role in ligand recognition. In this chapter, I describe studies where I probed these rarely occurring transitions using nonequilibrium simulations by characterizing the dissociation of a ligand molecule from an HIV-1 viral RNA element. Specifically, I observed base-flipping rare events that are coupled with ligand binding/unbinding and also provided mechanistic details underlying these transitions.

4.2 Significance

In the studies presented in this chapter, I reveal the key interactions that are required to be created or ruptured during the dissociation process of a small molecule with inhibitory properties from a viral RNA molecule. Specifically, I observed base-flipping rare events which are involved in the recognition mechanism of the small inhibitor by the viral RNA molecule. Additionally, I determined that these transitions contribute to a sequence of events relating five nucleotides which have not been observed previously. These results enhance our understanding of the recognition mechanisms of small molecules by viral RNAs and knowledge of these transitions can be potentially useful for designing new inhibitory molecules for targeting viral RNA molecules.

4.3 Background

RNA molecules were considered only as passive carriers of genetic information until RNA was implicated in diverse cellular processes (translation and transcription [40], regulation of gene expression [28, 228], and protein synthesis [41]). Many RNAs are also involved in progression of various diseases including neurological disorders, cancers, and cardiovascular diseases [29, 31, 229]. Moreover, RNAs play a critical role in the replication and survival mechanisms of many viruses and bacteria [34,230,231]. Thus, it is promising to target RNA molecules for developing therapeutic modalities because RNA lies upstream of proteins and its activity can be modulated before or during its synthesis [213].

Particularly, viral genomes do not provide a large number of protein targets due to the lack of well-defined binding pockets for small molecules [35]. However, conserved and structured RNA motifs of viral genomes are flexible and fold into complex three-dimensional structures that may provide transient binding pockets for small molecules, and thereby activities of "undruggable" proteins could be modulated before they are synthesized [35,213]. For example, new amiloride derivatives were shown to interact with several HIV-1 RNAs and inhibit the replication process of the virus [222, 232].

However, it is more challenging to target RNAs than proteins due to the highly charged nature of the RNA backbone, conformational flexibility of RNA, and a relatively low abundance of cellular RNAs in comparison with the ribosomal RNA [217]. In addition, designing new ligands to target RNA is limited by a poor understanding of the recognition mechanisms between RNA and its binding partners. These mechanisms are important for the function of RNA and the knowledge of the conformational dynamics of binding, as well as their thermodynamic and kinetic properties, will be useful in the drug discovery process [49].

Experimental techniques including X-ray crystallography and NMR spectroscopy provide crucial insights into the dynamics of RNA and its interactions with ligands [219,233]. AFM is another technique to study interactions between ligands and receptors or unfolding processes by obtaining force-extension data [189]. However, characterization of all possible atomic details of large and complex biomolecular systems continues to be a challenging process for experimental techniques. The number of parameters that need to be measured exceeds the number of parameters that can be tracked in experiments, even with the advanced NMR methods [46–48].

However, computational methods, such as MD simulations, are becoming increasingly important in characterizing the dynamics of biomolecules and their interactions with ligands by providing additional insights at the atomic level. Although many biophysical processes occur on time-scales challenging to probe using conventional MD simulations, non-equilibrium techniques, such as SMD simulations, that enhance conformational sampling are useful in probing critical ligand recognition events. During these processes, interactions that are important for the overall stability of the system are perturbed to reveal key structural motifs involved in ligand binding/unbinding. SMD has been successfully applied to study unfolding of RNA/DNA [234, 235], unbinding mechanisms of protein/ligand [236, 237], RNA/ligand [238, 239] complexes, and to study other systems [240, 241].

For work described in this chapter I applied MD and SMD simulation methods to study the TAR RNA from the HIV-1 (Figure 4.1A) that is located at the $5'$ end of the viral RNA genome. It is a key model system to study RNA dynamics and has been shown to transition between multiple conformations (e.g. bent and coaxially stacked configurations) along with other less populated states [48, 58]. It also has an important function in the viral replication mechanism because it interacts with the viral Tat protein and the host cofactor cyclin T1 to promote efficient transcription of the downstream genome [242]. Therefore, it has been targeted with molecules of various types and sizes and has become a primary drug target in the HIV-1 genome.

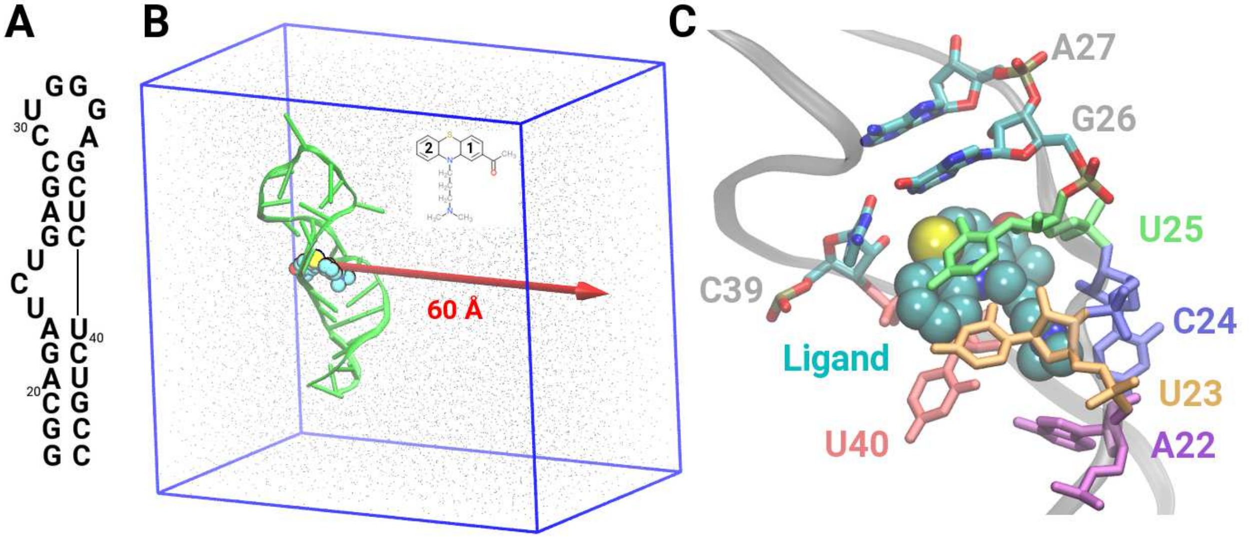

Figure 4.1: **System setup and structural details.** (A) Secondary structure of HIV-1 TAR RNA. (B) A side-view of the simulation domain: RNA, green cartoon; water molecules, gray points; ligand, space-filling; and the bounding box, blue. A red arrow indicates the direction of pulling. The chemical structure of the ligand is also shown with labeled aromatic rings (inset). (C) A side view of the binding pocket: ligand is shown in a space-filling representation and each key nucleotide is highlighted in a unique color and labeled.

Specifically, I conducted a long time-scale MD simulation spanning 2 μs and 300 non-equilibrium SMD simulations (see sections 4.4.1 and 4.4.2) to study the dissociation pathway of a small molecule, acetylpromazine (*inset* in Figure 4.1B) [83], which represents a compound with low toxicity and high binding affinity with interactions (Figure 4.1C) in the common binding pocket in TAR-RNA.

4.4 Methods

4.4.1 System Preparation and Simulation Details

Software and Force-Field: For work described in this chapter I focused on studying the unbinding process of acetylpromazine from the HIV-1 TAR RNA using conventional MD and SMD simulations. The initial coordinates for the system were obtained from the NMR structure deposited in the Protein Data Bank (PDB code 1LVJ) [83]. A 2 μs long classical MD simulation of the RNA/ligand complex was conducted using the Amber force-field (ff99+χ_{OL3}) [155, 156] using the Amber software [154]. All SMD simulations were also carried out using the Amber force-field (ff99+χ_{OL3}) [155, 156] but using the NAMD software [152]. The analyses of all trajectories were carried out using the CPPTRAJ and VMD software [152, 184, 185]. For acetylpromazine, the Antechamber program [243, 244] in Amber was used to develop the force field parameters with atomic charges using the AM1-BCC charge method (see Appendix I) [162].

MD: The system was solvated in a 57 Å × 84 Å × 55 Å periodic box of TIP3P water molecules and the total number of atoms was 22053. The overall charge of the system was neutralized with 29 Na^+ ions. No constraints were imposed on the RNA/ligand complex. The temperature was maintained using the Langevin thermostat at 310K, consistent with experimental conditions [83], and the pressure was maintained at 1 atm using the Berendsen barostat. The steepest descent minimization was initially performed for 1000 steps followed by 100-500 steps of conjugate gradient minimization. The system was subjected to a 2 μs long MD simulation in the NPT ensemble with a 2 fs timestep. The configurations were saved every 20 ps. Data from this simulation are shown in Figures C.1 and C.2.

SMD: The system was solvated in a 54 Å × 90 Å × 90 Å periodic box of TIP3P water molecules and the total number of atoms was 33936. The overall system was charge neutralized with 29 Na^+ ions and was energy-minimized via 500 cycles of conjugate-gradient optimization. To equilibrate the box volume, a 500 ps MD simulation with a 2 fs timestep was initially conducted. The coordinates from the end of this MD simulation were used as initial conditions for subsequent 5 ns long SMD simulations in the NPT ensemble, conducted using a 2 fs timestep. Even though the ligand dissociated at 25 Å, I continued SMD simulations up to a distance of 60 Å. The temperature and pressure were maintained at 310 K and 1 atm using the Langevin thermostat and the Nose-Hoover barostat. Periodic boundary conditions were used in all simulations, electrostatics were computed every time step using the particle mesh Ewald method, and the van der Waals interactions were cut-off at 10 Å with switching initiated at 8 Å. In these simulations, phosphorus atoms in the RNA backbone were weakly restrained to prevent the overall rotation and translation of the RNA molecule. Configurations were saved every picosecond and SMD output was saved every 20 fs.

4.4.2 SMD Simulations and the Potential of Mean Force (PMF) Calculation

The cv-SMD simulations were implemented by applying a harmonic external force using a spring with a spring constant of k = 7 kcal mol^{-1} Å^{-2} that was attached to the center of mass of the ligand and was pulled at a constant velocity of 0.0125 Å/ps along the reaction coordinate r. The force constant value was chosen per stiff-spring approximation [194] to closely follow the reaction coordinate for ligand dissociation. The potential of mean force (PMF) was calculated using the exponential averaging and the second order cumulant expansion of the Jarzynski's equality presented in section 2.4.1 (equations 2.13 and 2.14).

4.4.3 Buried Surface Area (BSA)

I calculated the BSA for the ligand acetylpromazine from a 2 μs classical MD simulation. The BSA was computed using the following equation:

$$\mathrm{BSA} = \mathrm{SASA}_{\mathrm{RNA}} + \mathrm{SASA}_{\mathrm{Ligand}} - \mathrm{SASA}_{\mathrm{Complex}}$$

where $\mathrm{SASA}_{\mathrm{RNA}}$ represents the solvent accessible surface area (SASA) of RNA, $\mathrm{SASA}_{\mathrm{Ligand}}$ represents the SASA of ligand, and $\mathrm{SASA}_{\mathrm{Complex}}$ represents the SASA of the RNA/ligand complex. The BSA value indicates the area of contact between the ligand and RNA. Data are shown in Figure C.2.

4.5 Results and Discussion

4.5.1 Thermodynamics of Ligand Dissociation

The studies of ligand dissociation from bound conformations are most suitably done using non-equilibrium simulations because the system is trapped in an energy minimum with high energy barriers to dissociation where the ligand is stabilized by interactions in the binding pocket. Conventional MD simulations are often non-ergodic due to incomplete sampling and as a result systems usually remain trapped in energy minima. In the work presented in this chapter, I did not observe a spontaneous dissociation of acetylpromazine in a conventional and long time-scale (2 μs) MD simulation. As seen in snapshots from the MD trajectory (Figure C.1), the ligand remained stably bound to RNA. The BSA, which represents the interface area of contact between the RNA and the ligand, supports this observation since the average BSA is 552 ± 82 Å^2 (Figure C.2) with an initial value of 645 Å^2. Thus, observing spontaneous dissociation is a non-trivial task even in μs-long MD simulations and non-equilibrium enhanced sampling methods (e.g. SMD) are needed. I used cv-SMD simulations for studying the dissociation process of acetylpromazine and for computing the non-equilibrium work of ligand dissociation. The non-equilibrium work values were then used to compute the unbinding free-energy (ΔG) using the exponential averaging as well as the second-order cumulant expansion of the Jarzynskis equality [191–194]. An SMD simulation with the lowest work value will have the highest contribution to the free-energy computed via Jarzynski's equality and therefore provides the most valuable information about key interactions that have to be broken or created during the dissociation process since the system requires the least amount of work to overcome those interactions. In contrast, simulations with higher work values provide a less than optimal pathway for ligand dissociation. Thus, the comparison between simulations requiring the lowest and highest work values can reveal the salient features of the binding/unbinding process of ligands. Specifically, I performed 300 cv-SMD simulations, each 5 ns long, where a harmonic spring with a spring constant $k = 7$ kcal mol^{-1} Å^{-2} was

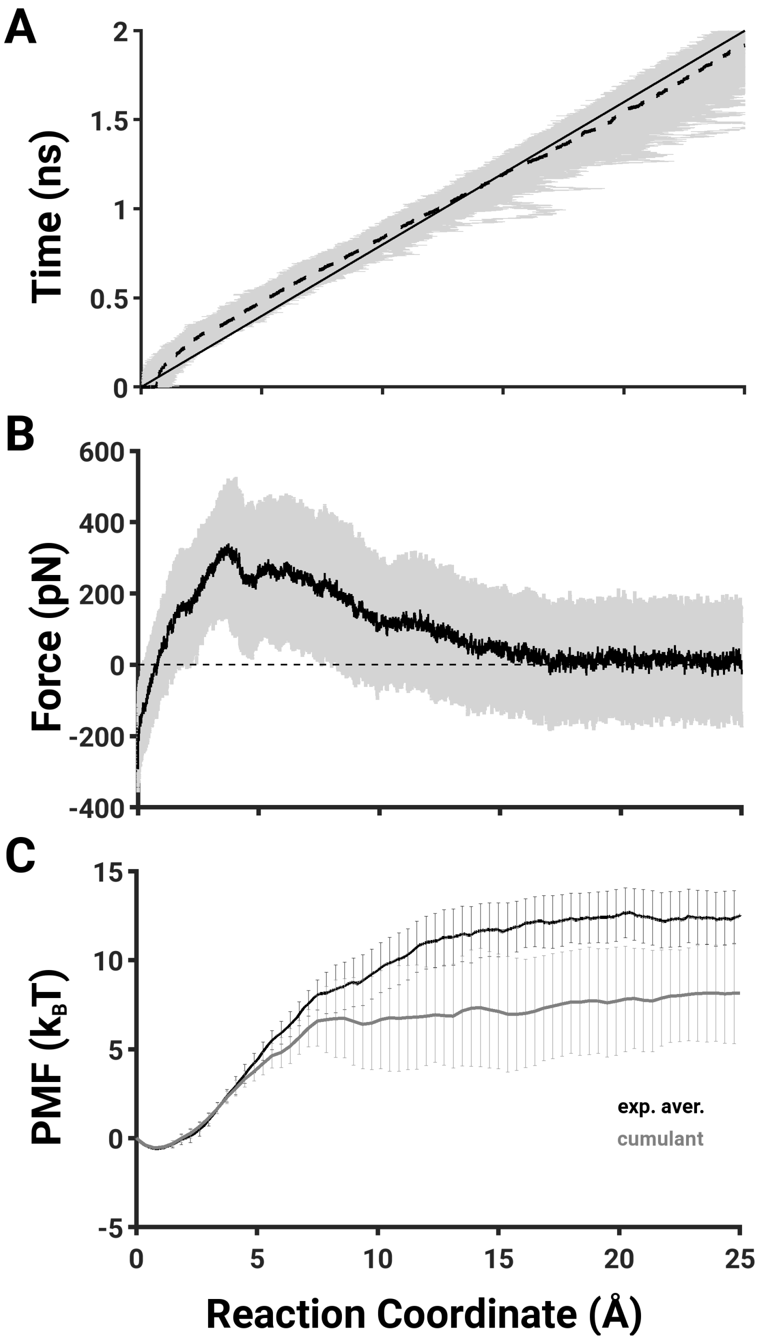

Figure 4.2: **Reaction coordinate, unbinding force, and free-energy from SMD simulations**. (A) The COM trajectory of the ligand. Black solid line represents the actual RC, black dotted line represents the average trace across 102 trajectories, and gray lines represent all SMD trajectories. (B) Unbinding force with the mean force (black solid line) and standard deviation profiles (gray) from all SMD simulations are shown. (C) Potential of Mean Force *vs.* RC, as computed using the exponential averaging (black line) and using the second-order cumulant expansion (gray line) with error bars.

attached to the COM of acetylpromazine and pulled with a velocity of 0.0125 Å/ps along the z-direction. The external work, W (Figure C.3), of ligand dissociation from the RNA pocket was computed from 102 trajectories out of all SMD simulations that consistently followed the reaction coordinate (Figures 4.2A and C.4A).

The unbinding cv-SMD force profile (Figure 4.2B) starts with the ligand in the bound state with no external force applied. Negative forces at the beginning indicate the dominance of system forces over the external force. As the external force values started to increase, overcoming the system forces restricting the ligand to its original conformation, acetylpromazine dissociation begins. The continued increase in the mean force until reaching a maximum value represents the displacements of various nucleotides in the binding pocket and perturbations in stacking interactions between the benzene rings of acetylpromazine and nucleotides. The maximum force corresponds to the point where ligand displaced all nucleotides leading to an open dissociation pathway. A small decrease in the force profile (between 4.3 Å and 4.9 Å) corresponds to a state where both of the benzene rings moved out of the binding pocket. The unbinding forces then decreased as the ligand moved away from the binding pocket. The fluctuations in force were measured after 17.5 Å to ascertain that the average force converged to zero indicating full dissociation of the ligand with no interactions to RNA (Figure C.4B).

The free-energy profiles computed using the exponential averaging and second-order cumulant expansion of Jarzynski's equality (Figure 4.2C) show an energy minimum corresponding to the bound state and converged free-energy values for the unbound state. The free-energy difference between the bound and unbound states at 17.5 Å was calculated to be 12.5 ± 1.47 kcal/mol and 8.176 ± 2.87 kcal/mol using the exponential averaging and the second-order cumulant expansion of Jarzynskis equality, respectively. The unbinding free energies were then used to compute the dissociation constant ($K_d = e^{-\frac{\Delta G}{RT}}$, where R is the gas constant and T is the temperature) and compared against the experimentally determined values. I estimated K_d value as 1.54 nM (exponential averaging) and 1750 nM (cumulant

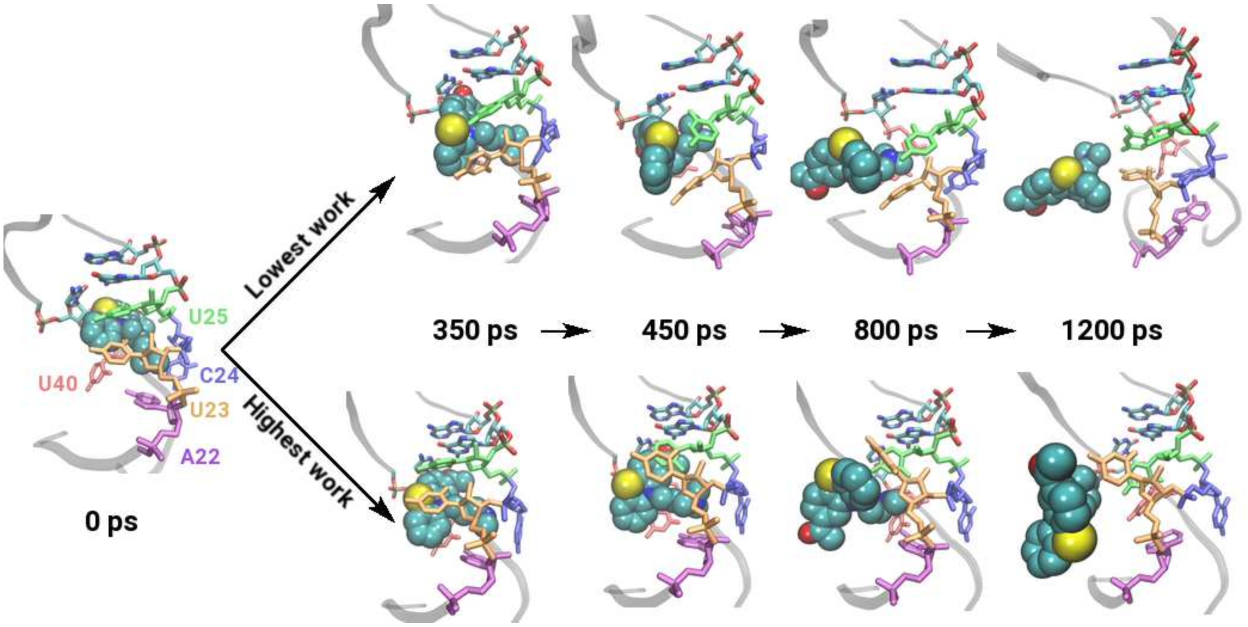

Figure 4.3: **Ligand dissociation mechanism:** Snapshots of ligand dissociation from the simulations with the lowest work (top) and the highest work (bottom) are shown. Color and labeling scheme is same as in Figure 4.1C. See also Figure C.6.

expansion). The experimental K_d value of 100 nM (corresponding to ∼9.94 kcal/mol) lies within the range of bounds predicted by our simulations.

4.5.2 Ligand Escape Pathway

Initially, the ligand was located between the base pairs G26-C39 and A22-U40 where its benzene ring **2** was inserted between U23, U25 and U40, forming stacking interactions with these bases, and the benzene ring **1** was positioned next to G26, forming an angle of ∼135° to the benzene ring **2**. The aliphatic chain of the ligand was extended along the minor groove of RNA and pushed C24 out of the stack (Figure 4.1C). I first focused on the dissociation pathway that was observed in the simulation that required to perform the least amount of work out of all SMD trajectories since that simulation has the most important details of the dissociation mechanism.

During the first 350 ps of this cv-SMD simulation, the ligand rotated counterclockwise by 90° with the sulfur atom pointing out of the binding pocket (Figure 4.3 and Figure C.5). At that time, the benzene ring **2** induced a counterclockwise rotation of U23 of the χ-dihedral by 50° and the benzene ring **2** stacked on U23, sulfur atom formed a van der Waals interaction with U25, the aliphatic chain induced a rotation of the χ-dihedral of C24 from -75° to -165°, A22 shifted by 40°, partially flipping out and providing space to C24 to rotate and flip inward, following the movement of the ligand out of the binding pocket.

At a distance of 5.6 Å (t = 450 ps), the sulfur atom continued to interact with U25 that resulted in an intramolecular conformational change in the ligand where three fused aromatic rings formed ~90° angle with the aliphatic chain (Figure 4.3 and Figure C.5). In the meantime, C24 flipped inward, occupying the free space left behind by the ligand, and formed a hydrogen bond with the oxygen atom of U40 while A22 returned to its initial position in the RNA stack. The flipping of C24 back into the RNA helix represents a rare base-flipping event in nucleic acids that occurs on a millisecond timescale and is difficult to observe both experimentally and during conventional MD simulations [108, 245].

Between 450 ps and 800 ps, U23 flipped underneath the ligand that was moving out of the pocket, thus making a pathway free of any obstacles (Figure 4.3 and Figure C.5). At a distance of 9.55 Å (800 ps), the ligand rotated again causing a minor counterclockwise rotation of U25 around the χ-dihedral by 60°. As the ligand was dissociating, U23 moved out of the binding pocket and flipped out when the ligand was at a distance of 10 Å away from U23. The ligand was free of any interactions with the RNA at $d = 17.5$ Å. Other simulations with lower work values indicated a similar mechanism of ligand dissociation (Figure C.6).

In contrast, in the simulation trajectory resulting in the highest dissociation work, the C24 nucleotide did not flip inside, despite interacting with the ligand as in the lowest work simulation (Figure 4.3 and Figure C.7). This could be potentially explained by the fact that A22 did not shift to provide additional space for C24. At 360 ps, the base part of U25 rotated around the χ-dihedral by 100° while still interacting with the sulfur atom of the ligand. In addition to that, U23 did not interact with benzene ring **2** as long as it did in the lowest work simulation and did not move closer to A22 below the ligand. Instead, when U25

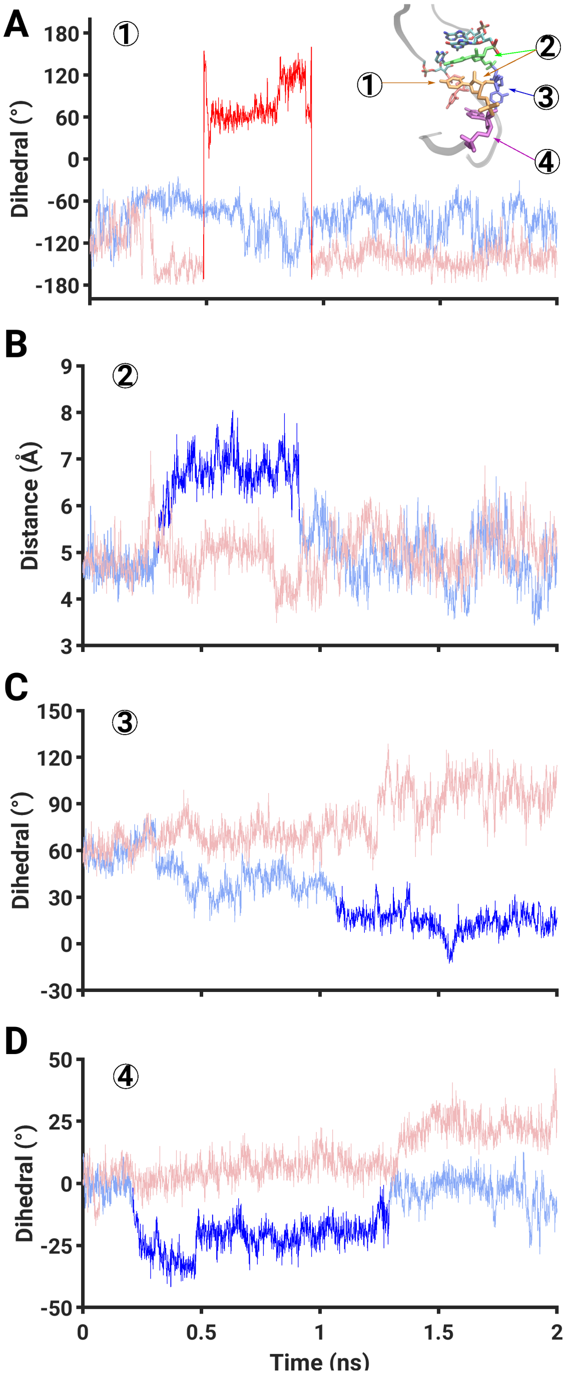

Figure 4.4: **Conformational metrics:** Shown are the traces of several conformational metrics from the lowest work (blue) and the highest work (red) simulations. Darker colors signify transition regions of interest. The numbers in each panel correspond to metrics computed for specific nucleotides (see *inset* in panel A). The conformational metrics shown are: (A) χ-dihedral of U23 nucleotide; (B) distance between the COM of U23 and U25; (C) dihedral angle that describes the flipping of C24; and (D) dihedral angle that describes the rotation of A22. See also Figure 4.5.

was rotating, U23 got shifted away from the ligand and stacked on U25 for ∼120 ps. That transition moved U23 in the outward configuration with respect to the binding pocket and above the ligand, while in the lowest work simulation U23 was below the ligand toward the binding pocket.

At 500 ps, U25 started interacting with the sulfur atom of the ligand which caused a rotation of the nucleobase in U23 around the χ-dihedral from -150° to 60° (Figure 4.4A). U23 proceeded to interact with the ligand by stacking on the benzene ring **2** between 500 ps and 950 ps which resulted in the rotation of U23 base to its original χ-dihedral value of -150° (Figure 4.4A). U23 then interacted with the aliphatic chain and remained in the flipped out state for the remainder of the simulation. These sequence of events potentially contribute to additional work required to overcome more stacking interactions between the acetylpromazine benzene ring **2** and U23/U25. Also, after the ligand moved out of the binding pocket, it continued to interact with A35 that was flipped out in the stem-loop of RNA (Figure C.8).

4.5.3 Mechanistic Details of Ligand Dissociation

To characterize the conformational rearrangements of the binding pocket nucleotides in the least work simulation, including the flipping-in of C24, and probe the reasons for not observing this flipping event in the highest work simulation, I describe a number of mechanistic details that collectively describe these events. These details can further improve our understanding of a base flipping process in the TAR-RNA and in bulge motifs of RNA in general.

I observed a sequence of conformational transitions in U23 and U25 (Figure C.9) that influenced the base flipping as well as potentially contributed to the amount of work needed to dissociate the ligand. As highlighted in earlier discussion, U23 rotated around the χ-dihedral and buried deeper in the binding pocket in the first 350 ps in the least work simulation. The movement of A22 outward was a consequence of this transition since U23 was displaced by the ligand which in turn displaced A22. Between 350 ps and 1000 ps, U23 rotated relative to A22 (Figure 4.5A) by 100° and moved away from U25 by 2 Å (Figure 4.4B) while partially filling the space that was available after A22 moved outward. At ∼830 ps, U25 rotated around the χ-dihedral by 40° counterclockwise and interacted with U23 until it (U23) flipped out at the end of the simulation. Interestingly, in the highest work simulation the same base rotated around the χ-dihedral in the opposite direction by 120°. Also, in that simulation, U23 rotated clockwise (opposite to the direction of rotation in the lowest work simulation where U23 moved inside the binding pocket) by 90° around the χ-dihedral at 350 ps which caused it to move out of the binding pocket. The difference in the directions of rotation of the χ-dihedral of U23 is a crucial detail that led to different conformational events in the binding pocket and likely influenced the final work values.

As shown in Figure 4.4C, the flipping in of C24 toward the binding pocket started after the rotation around its χ-dihedral which was observed in both the lowest and the highest work simulations. However, only in the lowest work simulation, this rotation was followed by the transition to an inward conformation. At 350 ps, A22 shifted in the outward direction by 30° counterclockwise (Figures 4.4D and C.10) providing space for C24 to move in. In the highest work simulation, on the contrary, A22 did not shift outward, remaining at its initial position and forming a base pair with U40 after the ligand dissociated.

In the lowest work simulation, the movement of A22 outward was followed by the movement of C24 inward as described by the dihedral-angle in Figure 4.4C at ∼400 ps. I observed fluctuations in C24 as it was moving in because the ligand had to first leave the binding pocket and provide space for that nucleotide. It also started to form a hydrogen bond with U40 (Figure 4.5B and C.10) and after the ligand completely dissociated, the hydrogen bond was stabilized (after 1 ns). Thus, C24 replaced the ligand which acted as a "pseudo base pair" in the initial conformation between A22 and U40. This highlights that ligands recognized by RNA likely substitute for and conformationally mimic interactions between RNA nucleobases.

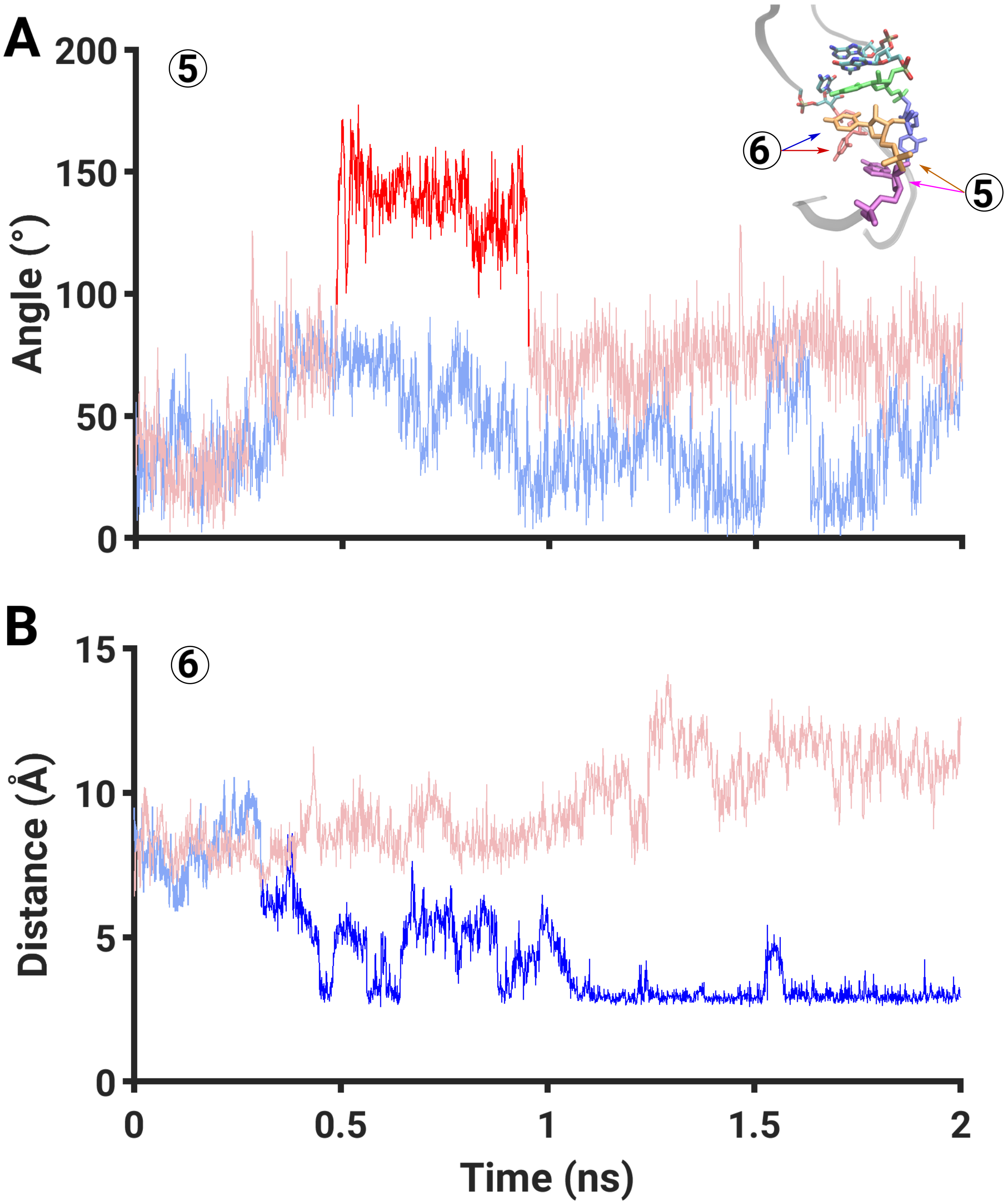

Figure 4.5: **Additional conformational metrics:** Shown are traces of additional conformational metrics from the lowest work (blue) and the highest work (red) simulations: (A) an interplane angle between A22 and U23 (marked as 5 in the *inset* in panel A and describing the relative position of U23); and (B) a hydrogen bond distance between C24 and U40 (marked as 6 in the *inset* in panel A). Darker colors signify transition regions of interest; see also Figure 4.4 for other details.

4.6 Conclusions

Binding/unbinding of ligands in RNA systems is an important biophysical process that is poorly understood. I used non-equilibrium cv-SMD and conventional MD simulations to study the dissociation pathway of acetylpromazine from TAR-RNA binding pocket to obtain key insights into the ligand binding/unbinding process. As expected, I did not observe ligand dissociation in a conventional MD simulation. On the contrary, cv-SMD simulations facilitated ligand dissociation and provided a large ensemble of trajectories to study this mechanism. In particular, I investigated in detail the lowest and the highest work simulations to identify mechanistic underpinnings of ligand dissociation. In the simulation with the lowest work value, I observed a rare base flipping event in the C24 nucleotide of TAR-RNA. This transition was a result of a sequence of complex events relating 5 nucleotides that were not observed in the highest work simulation.

Interestingly, the differences in the sequence of events between the lowest and the highest work simulations were initiated by the rotation of the χ-dihedral of U23 in opposite directions which I have identified for the first time. The counterclockwise rotation of the χ-dihedral of U23 not only decreased the amount of work but assisted in flipping-in of C24. I suggest that building a substantial ensemble of non-equilibrium trajectories is a potentially useful approach to gain insights into rare conformational transitions. These simulations, together with the Jarzynskis equality, were also able to predict the bounds on K_d within which was the experimentally measured value. Furthermore, I reported mechanistic details underlying several conformational transitions, including a dihedral-angle of C24, a hydrogen bond between C24 and U40, the χ-dihedral of U23, and an interplane angle between U23 and A22. Since the transitions in these variables exhibit two-state features, it is potentially useful to invoke rare event sampling methods to further study this mechanism in future. Specifically, transition path sampling [120, 196] along with the likelihood maximization [203, 204] is an exhaustive and accurate method to study these types of events. Its principles have been applied to study protein [246] and RNA systems [247]. Moreover, conformational transitions observed here can be potentially exploited for designing a new generation of inhibitory molecules targeting TAR-RNA.

4.7 Supporting Information

Additional data and figures are shown in Appendix C. In Appendix D, I also provide example scripts that I used to set up, conduct, and analyze my simulations. I have also included in Appendix D the scripts for creating figures.

4.8 Publication

The work described in this chapter is reproduced from Ref. [226], with permission from the American Chemical Society. The citation is as follows:
Levintov, L., and Vashisth, H. (2020). Ligand Recognition in Viral RNA Necessitates Rare Conformational Transitions. *J. Phys. Chem. Lett.* 11:5426-5432

CHAPTER 5
STUDY ON THE BINDING/UNBINDING PROCESS OF A HELICAL PEPTIDE FROM A VIRAL RNA MOLECULE

5.1 Abstract

Interactions between RNA molecules and proteins are critical to many cellular processes and are implicated in various diseases. The RNA-peptide complexes are good model systems to probe the recognition mechanism of RNA by proteins. For studies described in this chapter, I report studies on the binding/unbinding process of a helical peptide from a viral RNA element using non-equilibrium MD simulations. I explored the existence of various dissociation pathways with distinct free-energy profiles that reveal metastable states and distinct barriers to peptide dissociation. I also report the free-energy differences for each of the four pathways to be 96.47 ± 12.63 kcal/mol, 96.1 ± 10.95 kcal/mol, 91.83 ± 9.81 kcal/mol, and 92 ± 11.32 kcal/mol. Based on the free-energy analysis, I further propose the preferred pathway and the mechanism of peptide dissociation. The preferred pathway is characterized by the formation of sequential hydrogen bonding and salt bridging interactions between several key arginine amino acids and the viral RNA nucleotides. Specifically, I identified one arginine amino acid (R8) of the peptide to play a significant role in the recognition mechanism of the peptide by the viral RNA molecule.

5.2 Significance

In the studies presented in this chapter, I reveal key interactions that are involved in the recognition of a viral RNA molecule by a peptide which have not been reported previously.

Specifically, I discovered that the recognition of the peptide depends on the formation of salt bridges and hydrogen bonds that are formed between the arginine residues and the RNA backbone. I also demonstrated that these interactions formed a network of salt bridges that were spanning the major groove of RNA. These results enhance our understanding of the importance of arginine amino acids, or other basic amino acids, in the design of peptides that target viral RNA molecules.

5.3 Background

Numerous functions of RNA molecules depend on their interactions with proteins [248], which play a crucial role in various phases of the cell life cycle, including gene regulation [249, 250], transcription [251, 252], and translation [253]. Consequently, misregulation of RNA-protein interactions can lead to neurological disorders, cardiovascular problems, and oncogenic diseases [29, 50–52]. Moreover, the interactions between viral RNA molecules and cellular or viral proteins are involved in the replication and transcription processes of various viruses, for example, HIV, HCV, and SARS CoV/CoV2 [59, 254–256]. Therefore, resolving the mechanistic details of RNA-protein interactions is essential for understanding various biological and biophysical processes [29, 50–52, 59, 249–256].

Proteins and short peptides often interact with RNA molecules by adopting an α-helical or a β-sheet structure that can fit into the binding pocket of an RNA molecule [257–259, 259–263] or through the interactions with the RNA backbone [100, 248, 264]. Specifically, the RNA-peptide complexes are considered good model systems to study RNA-protein interactions and to probe the recognition mechanisms [265, 266]. A general RNA binding protein domain is the arginine-rich motif (ARM) which is found in ribosomal proteins [267], ribonucleoproteins [248, 268], and viral proteins [59, 269]. The ARMs are short peptides that have a high concentration of arginine residues and have high affinity and specificity of interaction with their targets by adopting various conformations including α-helical, β-hairpin, or extended conformations [270]. The interactions between these ARMs and RNA molecules have been investigated using NMR spectroscopy [59,93,257,271–274], CD spectroscopy [275,276], X-ray crystallography [95,96], and combinations of experimental and computational methods [270,277–279]. Several comprehensive investigations have been conducted on the nucleic acid-protein interfaces using structural and shape analyses to establish common features across known complexes [280–283]. Overall, these studies showed that the RNA-protein interactions are governed by sequence (e.g. composition of amino acids/nucleotides) or by shape (e.g. recognition of specific shapes of proteins).

However, the role of dynamics in RNA-protein interactions is still not fully understood due to challenges in capturing all the required parameters for describing a complex biomolecular system [47,49,264]. Computational methods such as MD simulations that are rooted in biophysical modeling are promising tools to enhance our knowledge of the recognition mechanism between RNA molecules and proteins by characterizing molecular motions at the atomic level [284]. Although several RNA-protein complexes [277,278,285–316] have been investigated using MD simulations, only a few studies have been conducted to investigate the interactions in *viral* RNA-protein complexes [99,317–322]. Specifically, the studies on the viral RNA-protein complexes highlighted the importance of electrostatic interactions and the interactions between water molecules and proteins. However, most of these studies [99,317–320] were reported over a decade ago and the force fields for nucleic acids and proteins have significantly improved in recent years [36]. Additionally, the time-scales of conventional MD simulations performed in these studies were limited. Thus, we still lack a full understanding of the viral RNA-protein recognition mechanisms and of specific interactions that need to be created or disrupted during the binding/unbinding process.

To address these questions, I applied non-equilibrium SMD simulations to study the binding/unbinding process of a helical arginine-rich peptide (RSG-1.2) from a conserved HIV-1 RRE RNA segment which is located in the *env* coding region and plays an essential role in viral replication (Figure 5.1A) [92]. The RSG-1.2 peptide is a mutated Rev peptide

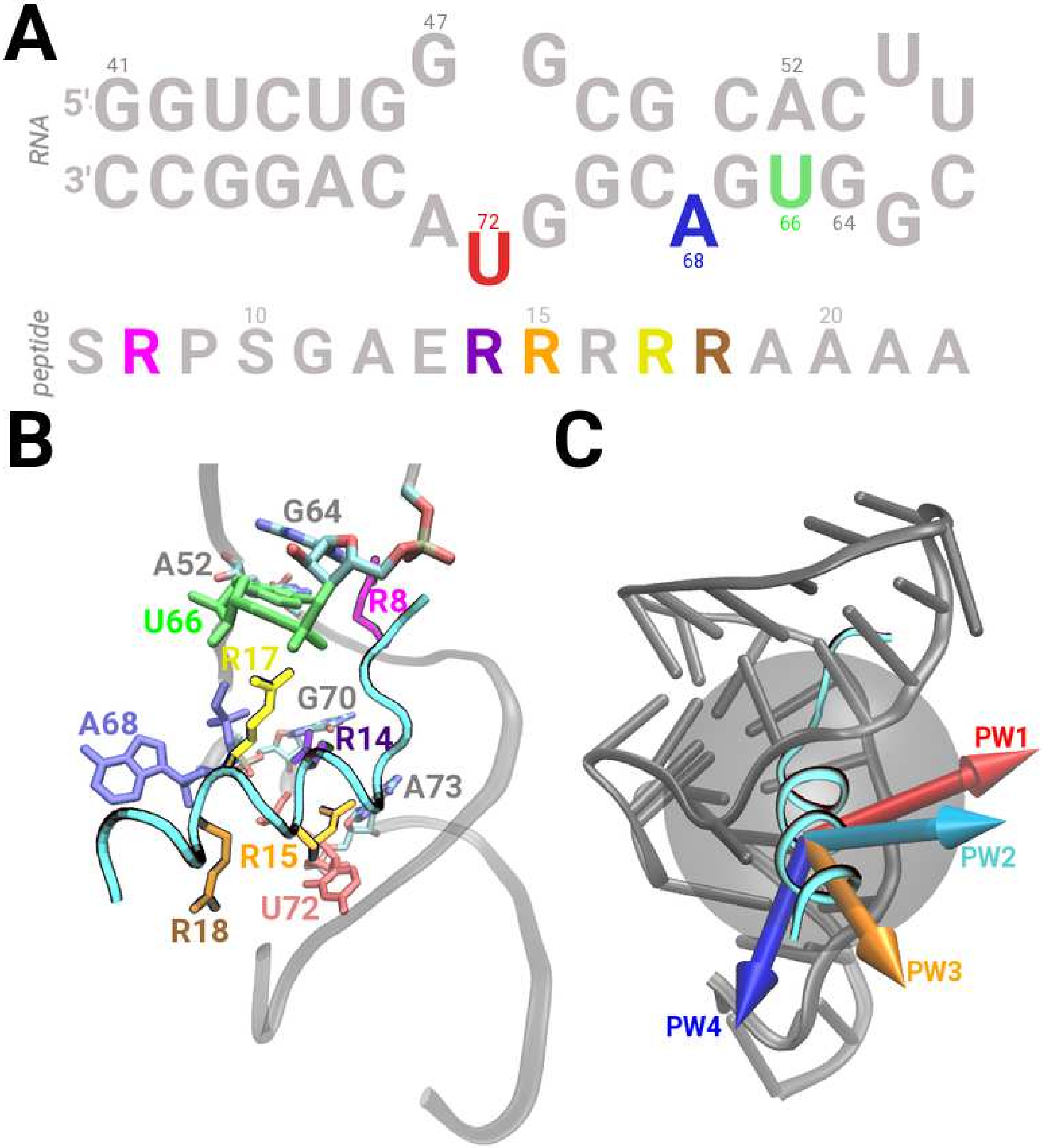

Figure 5.1: **Structural details and system setup.** (A) The sequences of the HIV-1 RRE RNA and the RSG-1.2 peptide are shown. (B) A side-view of the binding pocket is shown where the peptide is rendered as a cyan tube with the side-chains of key residues highlighted in stick representations. Each key nucleotide in the RNA and each key amino acid in the peptide are highlighted in a unique color and labeled. (C) A side-view of the RRE RNA (gray cartoon) and the peptide (cyan cartoon) complex is shown. A transparent gray sphere represents an approximate volume of the peptide binding pocket. Each arrow corresponds to the peptide dissociation coordinate/direction for one of the four pathways (PWs): PW1 (red), PW2 (cyan), PW3 (orange), and PW4 (blue).

with higher binding affinity and specificity in comparison to the canonical Rev peptide which binds RRE RNA [92] and is a good model system for studying RNA-protein interactions (Figure 5.1B) [266]. Specifically, I conducted SMD simulations along four distinct pathways (Table E.1; Figures 5.1C and E.1). In these simulations, I observed the formation of specific interactions and the sequence in which those interactions were forming or rupturing during the dissociation process of the peptide along each PW. Based on simulation results, I propose the preferred pathway as well as the mechanism of recognition of the peptide.

5.4 Methods

5.4.1 System Setup and Equilibration Details

In this work, I have studied the (un)binding process of the RSG-1.2 helical peptide from the HIV-1 RRE RNA using SMD simulations along four different pathways (Figure 5.1C). I obtained the initial coordinates for the system from the first frame of the NMR structure deposited in the Protein Data Bank (PDB code: 1G70) [93]. I centered the RNA/peptide complex at the origin and rotated to align the dissociation direction of the peptide in each pathway along the same axis (Figure E.1). I then solvated each system in a periodic simulation domain of TIP3P water molecules (Table E.1; Figure E.1). I neutralized the overall charge of the system with 27 Na^+ ions.

I energy minimized the system via the steepest descent minimization for 1000 steps that was followed by 500 cycles of conjugate-gradient minimization. To equilibrate the box volume, I conducted a 500 ps MD simulation in the NPT ensemble with a 2 fs timestep. I maintained the temperature and pressure at 310 K and 1 atm using the Langevin thermostat and the Nos-Hoover barostat in all MD and SMD simulations. I used periodic boundary conditions in all simulations and computed the electrostatic interactions using the particle mesh Ewald method. For the van der Waals interactions, I used a cut-off of 10 Å with switching initiated at 8 Å. I applied weak restraints to the phosphorous (P) atoms in the RNA backbone to prevent the overall rotation and translation of the RNA molecule. I carried out all simulations using the NAMD [152] software package combined with the Amber forcefield for RNA (RNA.ROC) [157] and for the peptide (ff14sb) [161]. I used the TIP3P water model [159] for the solvent and the Li/Merz parameters for the ions [160]. I analyzed all trajectories using the VMD and CPPTRAJ software [184, 185].

5.4.2 SMD Simulations

To study the dissociation of the peptide along each of the four pathways, I performed constant velocity SMD (cv-SMD) simulations, referred hereafter as SMD simulations. I provide additional details on the SMD method in section 2.4.1. To select the four dissociation pathways, I considered a sphere which approximated the volume of the binding pocket (gray sphere in Figure 5.1C). I then selected points on the surface of the sphere that were radially separated by ~13 Å, to prevent overlap with the RNA molecule. The arrows that are shown in Figure 5.1C pass through each of the defined points and represent unique reaction coordinates of dissociation along each of the four pathways. I used the coordinates from the end of the initial MD simulations for subsequent SMD simulations in the NPT ensemble. Specifically, for each of the four pathways, I conducted 75 SMD simulations, each of which was 13 ns long, thereby resulting in a total simulation time of 3900 ns.

Consistent with the stiff-spring approximation [194], I applied a harmonic external force with a spring constant of $k = 12$ kcal mol^{-1} Å^{-2} that was attached to the center of mass of the peptide residues Gly11 through Arg22. After testing various values, I chose a pulling velocity of 0.00625 Å/ps. I also applied a harmonic restraint to prevent the rotation of the peptide during dissociation in SMD simulations. As the reference orientation angle, I used the initial coordinates of the peptide and a force constant of 3 kcal mol^{-1}deg^{-2} for the harmonic potential. I also applied restraints to the atoms forming hydrogen bonds in the peptide residues Gly11 through Arg22 to maintain the secondary structure of the peptide during the dissociation. The configurations were saved every ps and the SMD output was saved every 20 fs.

5.4.3 PMF Calculation

I followed the protocol developed by Jensen *et al.* [192], and calculated the PMF using the exponential averaging of the Jarzynski's equality presented in section 2.4.1 (equation 2.13).

5.4.4 Interaction Energies and Salt Bridges

I also computed the non-bonded interaction energies between a specific amino acid of the peptide and a specific nucleotide of the RRE RNA. In particular, I calculated the vdW energy between all atoms in the following pairs of amino acids and nucleotides: Arg8/R8 and U66; Arg15/R15 and U72; Arg17/R17 and A68; Arg18/R18 and A68.

I also analyzed a network of hydrogen bonding and salt bridging interactions formed between a specific arginine amino acid and a specific RNA nucleotide. Hydrogen bonds were defined between a hydrogen atom of the arginine amino acid and a heavy atom (oxygen or nitrogen atom) of the RNA nucleotide. Salt bridges were defined between a nitrogen atom of the arginine amino acid and the oxygen atom of the phosphate group in the RNA backbone. The definition and the cutoff value of 3.5 Å for hydrogen bonding and salt bridging interactions were adopted from a previous study [99]. Specifically, I computed the salt bridge distances between the atoms presented in Table 5.1.

5.5 Results

5.5.1 Thermodynamics of Peptide Dissociation

Using non-equilibrium cv-SMD simulations, I studied the dissociation of the RSG-1.2 peptide from the RRE RNA along four distinct pathways (Figure 5.1C). During these SMD simulations, the peptide consistently followed the reaction coordinate (Figure E.2A). I also calculated the unbinding force profiles to ascertain that the average force converged to zero, corresponding to a fully dissociated state of the peptide and with no residual interactions with the RNA. In Figure E.2B, I show the average force profiles with error bars for each

Table 5.1: **Details on salt bridging interactions.** The details on the atom of the amino acid (*Peptide*) and the atom of the nucleotide (*RNA*) that participate in salt bridging interactions are presented for each pathway (PW).

PW	Peptide	RNA
1	NH2/R15	O1P/U45
2	NH1/R8	O1P/G48
	NH2/R14	O2P/A68
	NH2/R15	O2P/U45
	NH1/R15	O1P/C44
3	NH2/R8	O2P/A68
	NH2/R14	O2P/A68
	NH2/R15	O2P/G42
4	NH1/R8	O1P/U72
	NH2/R14	O2P/A68
	NH1/R14	O1P/C69
	NH2/R15	O1P/C44

pathway which show that the average force for the dissociation of the peptide converged to zero after ∼35-40 Å depending on the pathway. The convergence to zero is further ascertained by computing the distributions of force values after 40 Å for each pathway that reveal a mean of zero (Figure E.3). Then, I computed the non-equilibrium work required for the dissociation of the peptide from each of the 75 simulations for all four pathways (Figures E.4 and E.5). The resulting work distributions were used to estimate the free-energy/PMF profile along the reaction coordinate for each pathway (Figure 5.2B) using the Jarzynski's equality [191] that relates the non-equilibrium work to the equilibrium free-energy difference (ΔG). Since non-equilibrium trajectories with the least work have the highest contribution to the equilibrium free-energy difference estimated using the Jarzynski's equality, I provide mechanistic details from these trajectories.

The intermediate steps of the peptide dissociation in each pathway are quantitatively described using the unbinding force profiles (Figure 5.2A). At the beginning of each SMD simulation (r = 0 Å), the peptide was located in the bound state, interacting with the RNA nucleotides in the binding pocket (Figure 5.1B). In particular, the R8 amino acid was initially interacting with the U66, G64, and A52 nucleotides; the R14 amino acid was

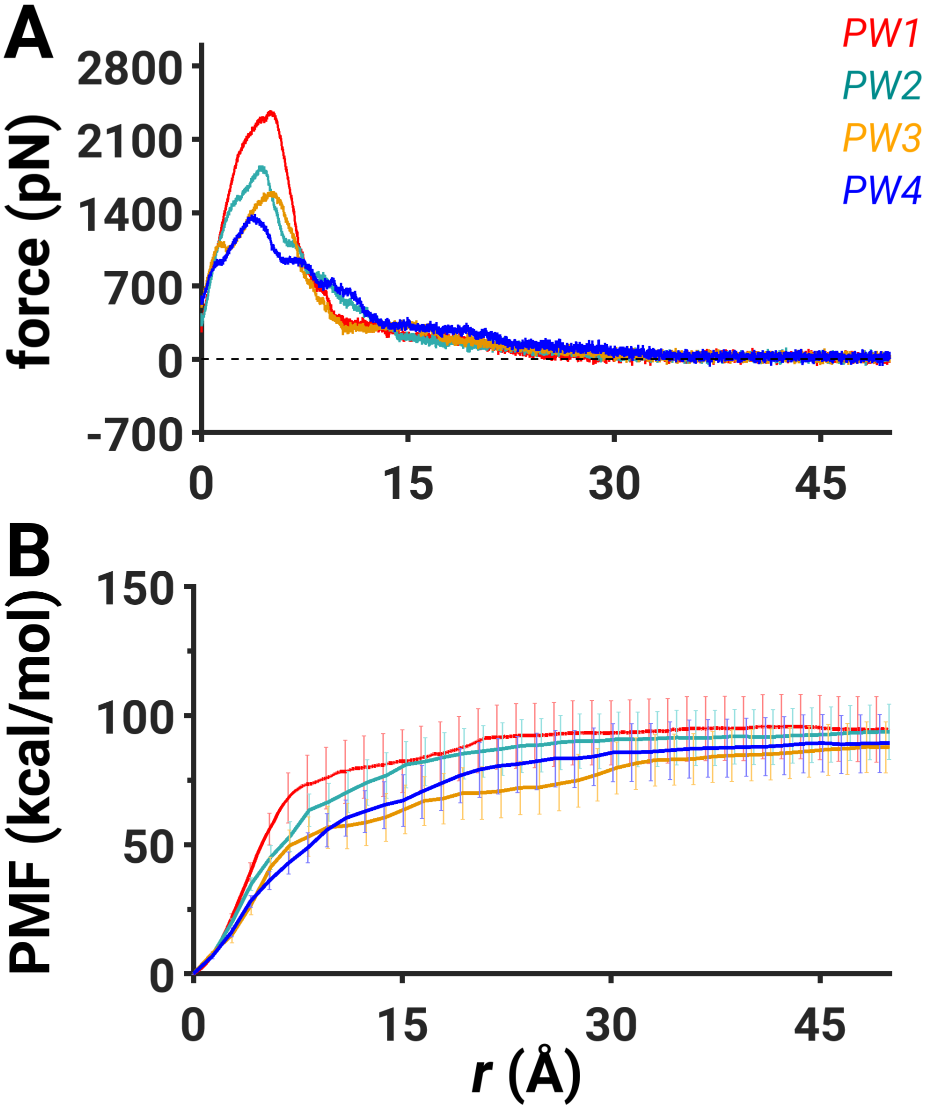

Figure 5.2: **The unbinding force and the free-energy profiles.** (A) The traces of the averaged unbinding force along each pathway are shown: PW1 (red), PW2 (cyan), PW3 (orange), and PW4 (blue). (B) The free-energy profile along each pathway is shown. See also Figures E.2 and E.6.

interacting with the G70 nucleotide; the R15 amino acid was interacting with the A73 and U72 nucleotides; and the R17 amino acid was initially interacting with the A68 nucleotide (Figure 5.1B). A gradual increase in the external force values for each pathway (Figure 5.2A) indicates that the peptide began to dissociate from the binding pocket by overcoming the interactions with the binding pocket nucleotides. The peak force values correspond to the stage when the peptide has moved out of the binding pocket by rupturing key interactions with the RNA. The external force values then decreased as the peptide was at a distance of ∼35-40 Å when the force values on average converged to zero signifying that the peptide reached the dissociated state (Figure 5.2A).

I further analyzed the unbinding force profiles which exhibited different magnitudes of the maximum force of dissociation in each pathway. Specifically, I observed that PW1 had the highest value of the maximum force of dissociation occurring at ∼5 Å which was equal to ∼2377 pN (PW1 in Figure 5.2A). The force profile in PW2 exhibited the second highest value of the maximum force of dissociation at ∼4.5 Å which was equal to ∼1850 pN (PW2 in Figure 5.2A). Additionally, I detected a smaller peak of the unbinding force at ∼7.2 Å in PW2 which was equal to ∼1090 pN. I observed that PW3 exhibited the third highest value of the maximum force of dissociation at ∼5.2 Å corresponding to a force value of ∼1600 pN (PW3 in Figure 5.2A). Moreover, I detected a smaller force peak value at 1.3 Å in PW3 corresponding to ∼1120 pN. Finally, I observed the lowest value of the maximum force of dissociation in PW4 which occurred at at ∼3.8 Å and was equal to ∼1360 pN (PW4 in Figure 5.2A). I also located smaller peaks in force at ∼1.2 Å and at ∼6.6 Å which were both equal to ∼950 pN. I detected a variability in the location of the maximum force value in the individual trajectories. The maximum force values were located between 4.6 Å and 5.4 Å in PW1, between 4.3 Å and 4.8 Å in PW2, between 4.9 Å and 5.4 Å in PW3, and between 3.6 Å and 4.5 Å in PW4 (Figure E.2B). I observed that the unbinding force profiles converged to zero at 35 Å for PW1 and PW2, and at 40 Å for PW3 and PW4 (Figures 5.2A and E.2B).

I report the free-energy profiles for each pathway (Figure 5.2B) which provide additional information on the thermodynamics of peptide dissociation, including the free-energy barriers, and the metastable states. All reported free-energy values are measured with respect to the initial state. I also show a zoomed view on the free-energy profile for r values between 0 Å and 15 Å along with the first-order derivative of the free-energy profile computed every 100 points for the same range of r values in each pathway (Figure E.6). The first-order derivative provides information on the instantaneous rate of change of the free-energy profile and I defined the wells in the first-order derivative profiles as the metastable state in the free-energy (M; Figure E.6) and the barriers separating these wells as the free-energy barriers (‡; Figure E.6). The point when the first-order derivative converges to zero corresponds to a point in the free-energy profile when there is no change and the free-energy profile plateuas.

I observed that the highest free-energy barrier of dissociation was in PW1 which was equal to 41 ± 3.67 kcal/mol at ∼4.2 Å with an additional free-energy barrier of 61.67 ± 7.41 kcal/mol at 6 Å (red ‡; Figure E.6A). I observed the second highest free-energy barrier in PW2 corresponding to 37.51 ± 2.62 kcal/mol at ∼4.4 Å with an additional free-energy barrier of 58.08 ± 5.96 kcal/mol at ∼7.5 Å (cyan ‡; Figure E.6B). In PW3, I observed several free-energy barriers at ∼1 Å and at ∼4.6 Å corresponding to the free-energy values of 4.49 ± 0.18 kcal/mol and 31.47 ± 3.77 kcal/mol, respectively (orange ‡; Figure E.6C). Finally, in PW4, I observed four free-energy barriers at ∼0.8 Å, at ∼3.8 Å, at ∼5.9 Å, and at ∼8.5 Å corresponding to the free-energy values of 3.66 ± 0.34 kcal/mol, 24.46 ± 2.08 kcal/mol, 38.46 ± 3.59 kcal/mol, and 50.48 ± 5.99 (blue ‡; Figure E.6D).

I also observed the formation of the metastable states in all pathways (labeled M in Figure E.6). I located the metastable states at ∼5.4 Å in PW1 (red M; Figure E.6A), at ∼5.4 Å in PW2 (cyan M; Figure E.6B), at 1.8 Å in PW3 (orange M; Figure E.6C), and at 1.3 Å, at 5.1 Å and at 6.9 Å in PW4 (blue M; Figure E.6D). The mechanistic details of each metastable state are provided in the following section. Finally, I observed that the free-energy differences between the initial states (r = 0 Å) and the dissociated states (r = 50 Å) were 96.47 ± 12.63 kcal/mol for PW1, 96.1 ± 10.95 kcal/mol for PW2, 91.83 ± 9.81 kcal/mol for PW3, and 92 ± 11.32 kcal/mol for PW4. Thus, the resulting free-energy differences (ΔG) have similar values, falling within the range of error bars for each pathway. Overall, I observed that PW4 has the smallest free-energy barrier for dissociation of the peptide while having additional metastable states in comparison to other pathways.

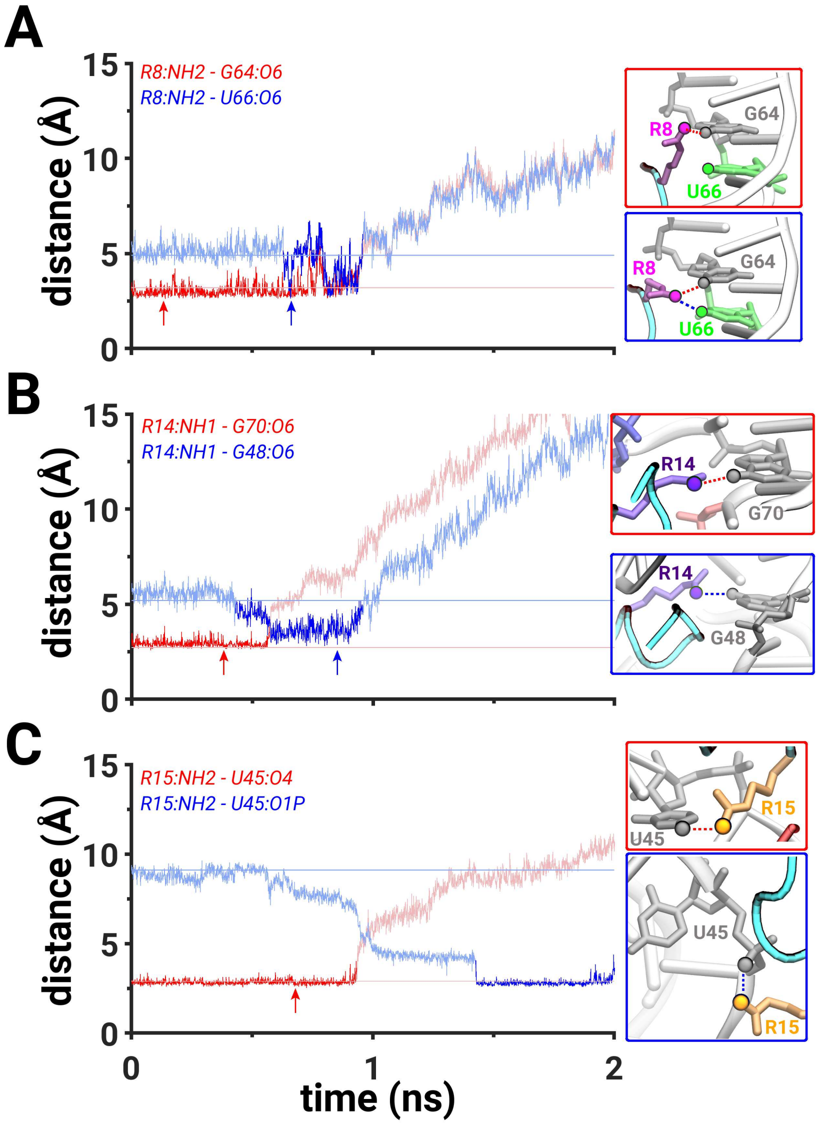

Figure 5.3: **Mechanistic details of PW1.** (A) The hydrogen bond distances between the NH2 atom of R8 and the O6 atom of G64 (red trace) and between the NH2 atom of R8 and the O6 atom of U66 (blue trace). (B) The hydrogen bond distances between the NH1 atom of R14 and the O6 atom of G70 (red trace) and between the NH1 atom of R14 and the O6 atom of G48 (blue trace). (C) The hydrogen bond distance between the NH2 atom of R15 and the O4 atom of U45 (red trace) and the salt bridge between the NH2 atom of R15 and the O1P atom of U45 (blue trace). All metrics are computed from the simulation with the lowest work value. Darker colors signify regions of interest. Lightly shaded horizontal lines indicate initial values of the corresponding distance. Each panel is accompanied with snapshots highlighting the corresponding interactions extracted from a time point marked by an arrow. Each amino acid, nucleotide, and an atom that participate in hydrogen bonding or salt bridging interactions are uniquely colored.

5.5.2 Mechanistic Details: Peptide Dissociation Pathways

In the initial conformation, the peptide is bound in the major groove of the RRE RNA between the A75-U45 and U66-A52 base pairs while largely maintaining an α-helical conformation with five residues constituting a coiled segment at the N-terminus (Figure 5.1B) [93]. The A68 and U72 nucleotides were in the flipped-out configurations, recognizing the peptide through stacking interactions with the R15 and R18 amino acids, respectively (Figure 5.1B). The Hoogsteen edge of the G70 and A73 nucleotides formed hydrogen bonding interactions with the R14 and R15 amino acids, respectively. The R8 amino acid from the coiled segment of the peptide interacts with the U66 nucleotide while the R17 and R18 amino acids also form contacts with the RNA backbone.

During the early part of the lowest-work SMD simulation in PW1, the peptide began dissociating out of the binding pocket (Figure E.7A) which was also characterized by weakening of interactions between several key amino acids and nucleotides (Figure E.8A). In particular, I observed that the van der Waals interaction energy between the R8 amino acid and the U66 nucleotide, the R15 amino acid and the U72 nucleotide, the R17 amino acid and the A68 nucleotide approached zero (Figure E.8A), indicating negligible interactions between the residues. Specifically, at t = ∼0.6 ns, the hydrogen bond between the NH2 atom of R8 amino acid and the O6 atom of G64 weakened (red trace; Figure 5.3A) and a new hydrogen bond was formed between the NH2 atom of R8 amino acid and the O6 atom of U66 (blue trace; Figure 5.3A). Additionally, at t = ∼0.6 ns, the hydrogen bond between the NH1 atom of R14 amino acid and the O6 atom of G70 broke (red trace; Figure 5.3B) which led to the formation of a hydrogen bond between the NH1 atom of R14 amino acid and the O6 atom of G48 (blue trace; Figure 5.3B). This sequence of events was a result of the peptide leaving the initial binding pocket and was coupled with the formation of new hydrogen bonding interactions between the R8 and R14 amino acids and the U66 and G48 nucleotides, respectively (Figures 5.3A,B).

At t = ∼1 ns, the peptide was located in the proximity of the backbone atoms of the

C44, U45, and G46 nucleotides that constitute the major groove of the RNA (Figure E.9) and the van der Waals interactions between the R8, R15, and R17 amino acids, and the U66, U72, and A68 nucleotides diminished (Figure E.8A). This was also characterized by the rupture of the hydrogen bonds that were previously formed at t = ∼0.6 ns between the NH2 atom of R8 amino acid and the O6 atom of U66, and between the NH1 atom of R14 amino acid and the O6 atom of G48 (blue trace; Figure 5.3A,B). The state when the peptide was located in the proximity of the backbone atoms of the C44, U45, and G46 nucleotides corresponds to a weak metastable state in the free-energy profile (red M; Figure E.6A).

At ∼1.4 ns, the peptide displaced the backbone atoms of the C44, U45, and G46 nucleotides and was located in the partially dissociated state, while the R8, R14, and R15 amino acids were still in the vicinity of the RRE RNA with the possibility to interact with the C44, U45, and G46 nucleotides (Figure E.7A). However, at t = ∼1.9 ns, I observed the formation of only one salt bridge that was formed between the NH2 atom of R15 amino acid and the O1P atom of U45 (blue trace; Figure 5.3C) which was preceded by the rupture of the hydrogen bond at t = ∼0.95 ns between the NH2 atom of R15 amino acid and the O4 atom of U45 while the peptide was still located in the binding pocket (red trace; Figure 5.3C). The peptide was free of any interactions with the RNA at a distance of 35 Å (t = 5.6 ns).

In PW2, I observed different mechanistic details underlying the dissociation process in comparison to PW1 which likely contributed to a lower free-energy barrier to dissociation (Figure 5.2B). As the peptide began dissociating out of the binding pocket (Figure E.7B), the van der Waals interactions between the R8 amino acid and the U66 nucleotide were broken at t = ∼0.1 ns (purple trace; Figure E.8B). This event occurred simultaneously with the rupture of the hydrogen bond between the NH2 atom of R8 amino acid and the O6 atom of G64 at t = ∼0.1 ns (red; Figure 5.4A). The R8 amino acid did not form any stable close contact interactions until t = ∼0.9 ns, when the NH1 atom of R8 formed a salt bridge with the O1P atom of G48. At ∼0.73 ns, the hydrogen bond between the NH1 atom of R14 and

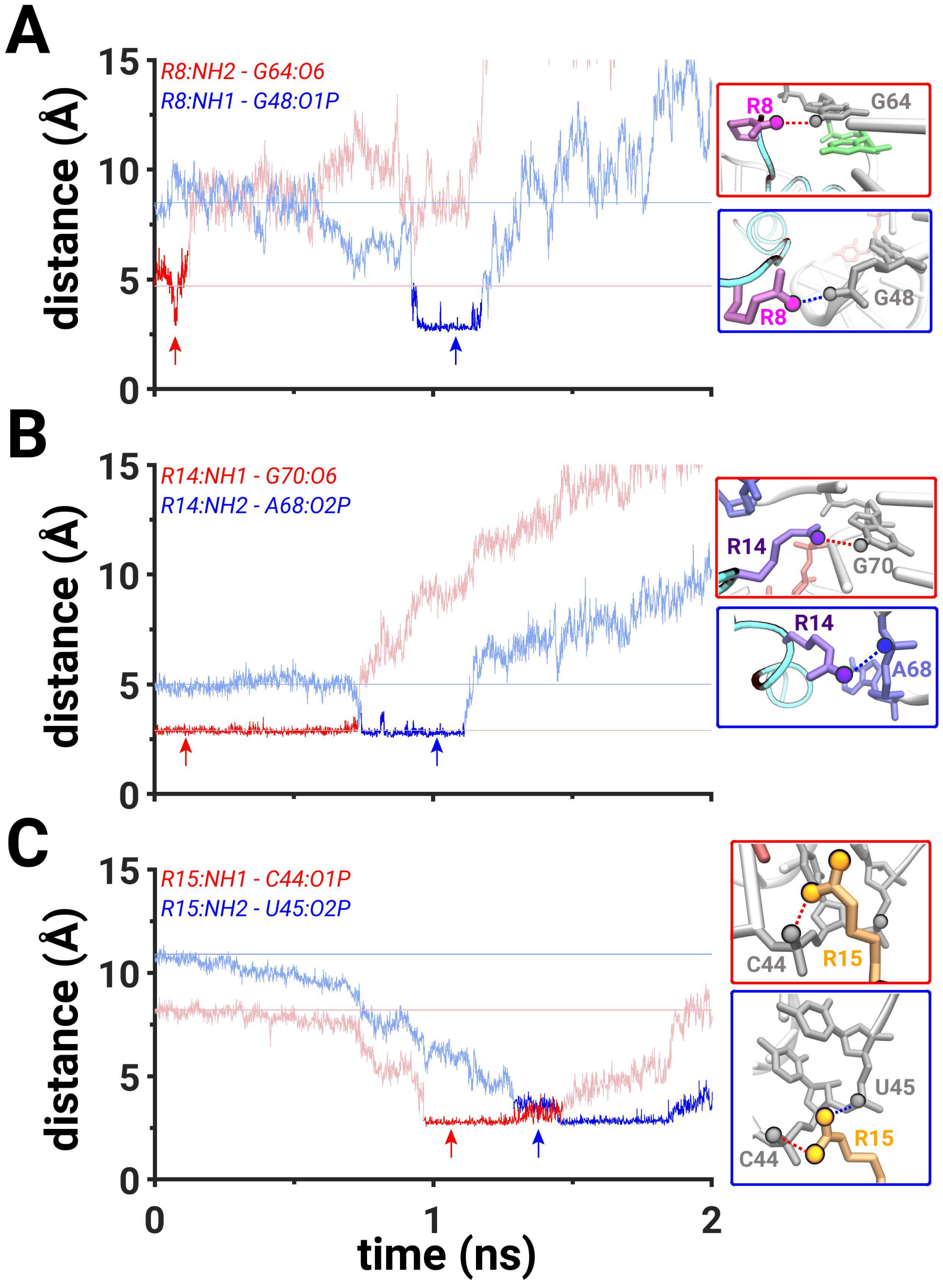

Figure 5.4: **Mechanistic details of PW2.** (A) The hydrogen bond distance between the NH2 atom of R8 and the O6 atom of G65 (red trace) and the salt bridge between the NH1 atom of R8 and the O1P atom of G55 (blue trace). (B) The hydrogen bond distance between the NH1 atom of R14 and the O6 atom of G70 (red trace) and the salt bridge between the NH2 atom of R14 and the O2P atom of A68 (blue trace). (C) The salt bridges between NH1 atom of R15 and the O1P atom of U45 (red trace) and between the NH2 atom of R15 and the O2P atom of C44 (blue trace). cf. Figure 5.3 for all other details.

the O6 atom of G70, that was preformed in the initial binding pocket, broke and the NH2 atom of R14 formed a salt bridge with the O2P atom of A68 at t = ∼0.75 ns (Figure 5.4B). Thus, two arginine amino acids, R8 and R14, formed salt bridging interactions at ∼0.9 ns, creating a network of salt bridges from the G48 nucleotide to the A68 nucleotide (Figure E.10A). This conformation also resulted in a metastable state which was highlighted in the free-energy profile at ∼6.5 Å (cyan M; Figure E.6B).

In PW2, the NH1 atom of R15 formed a salt bridge with the O1P atom of C44 (red trace; Figure 5.4C) when the peptide was in the vicinity of the backbone atoms of the C44, U45, and G46 nucleotides at t = ∼1 ns (Figure E.7B). Importantly, at t = ∼1.3 ns, the NH2 atom of R15 formed a salt bridge with the O2P atom of U45 (blue trace; Figure 5.4C). Thus, between t = ∼1.3 ns and t = ∼1.5 ns, the NH1 and NH2 atoms of R15 were fluctuating to simultaneously form two salt bridges with the O1P and O2P atoms of C44 and U45 nucleotides, respectively (Figure 5.4C). This motion was another factor that contributed to a decrease in the free-energy barrier in comparison to PW1. In addition to that, the rupture of the hydrogen bond between the NH2 atom of R8 amino acid and the O6 atom of G64 at t = ∼0.1 ns and the rupture of the van der Waals interactions between the R8 amino acid and the U66 at t = ∼0.1 ns, also contributed to a decrease in the free-energy barrier in comparison to PW1. The peptide was free of any interactions with the RNA at a distance of 35 Å (t = 5.6 ns).

In PW3, the peptide required ∼3 ns to escape the binding pocket, while in PW1 and PW2 the peptide escaped the binding pocket in ∼2 ns (Figures E.7A-C). This was in part due to the interactions of various amino acids with the A68 nucleotide in PW3 (Figure E.7C) as well as due to the interactions between the R8 amino acid and the U66 nucleotide that I characterized using the van der Waals energy (purple trace; Figure E.8C). These interactions resulted in a partial unfolding of the peptide coil between t = ∼1.8 ns and t = ∼3 ns (Figure E.7C).

During the first 0.8 ns of the simulation, the peptide disrupted interactions between the

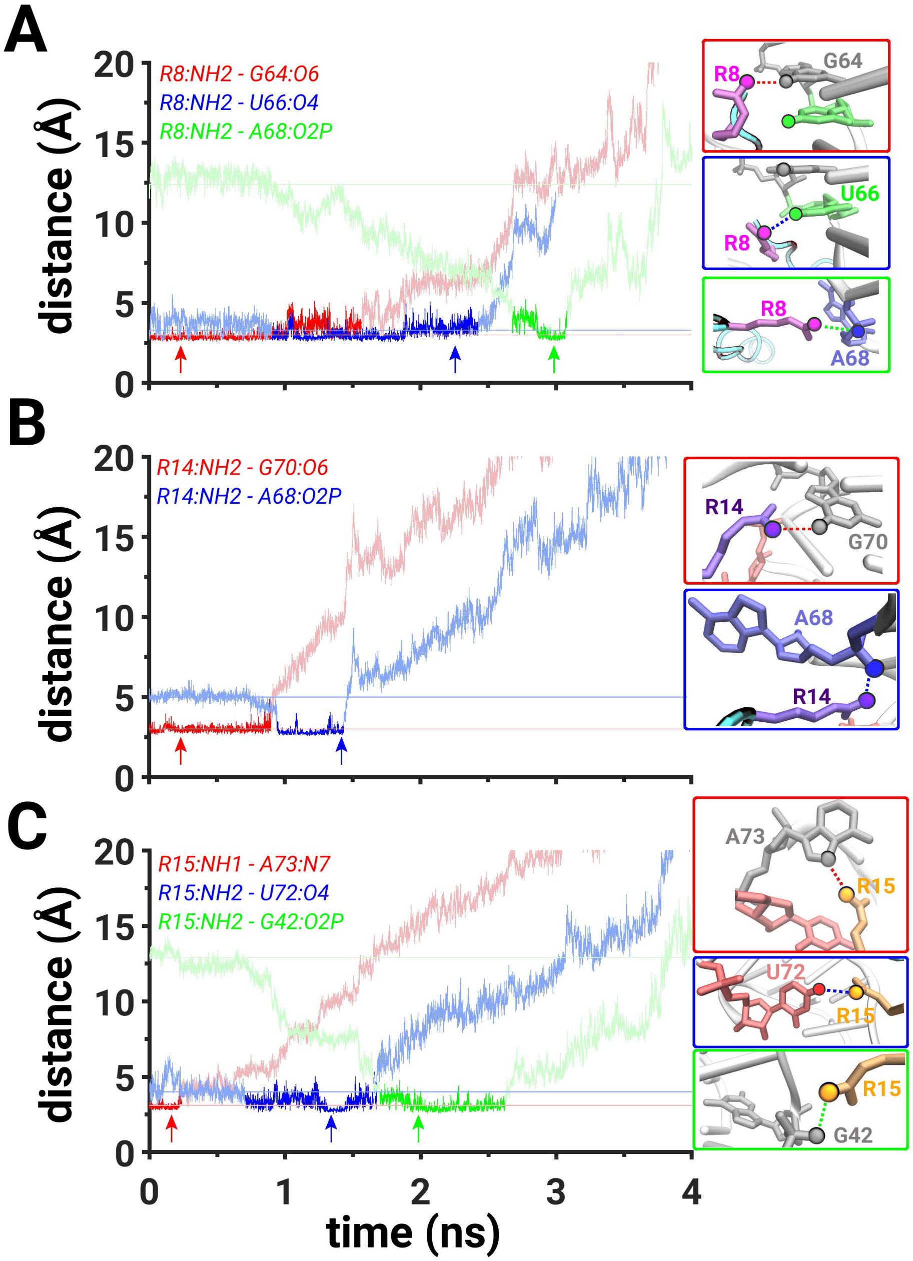

Figure 5.5: **Mechanistic details of PW3.** (A) The hydrogen bond distances between the NH3 atom of R8 and the O6 atom G64 (red trace) and between the NH2 atom R8 and the O4 atom of U66 (blue trace) and the salt bridge between the NH2 atom of R8 and the O2P atom of A68 (green trace). (B) The hydrogen bond distance between the NH2 atom of R14 and the O6 atom of G70 (red trace) and the salt bridge between the NH2 atom of R14 and the O2P atom of A68 (blue trace). (C) The hydrogen bond distances between the NH1 atom R15 and the N7 atom of A73 (red trace) and between the NH2 atom of R15 and the O4 atom of U72 (blue trace) and the salt bridge between the NH2 atom of R15 and the O2P atom of G42 (green trace). cf. Figure 5.3 for all other details.

R18 amino acid and the A68 nucleotide, as characterized by the van der Waals energy (brown trace; Figure E.8C), and started dissociating. A hydrogen bond between the NH2 atom of R8 amino acid and the O6 atom of G64 weakened at t = ∼0.9 ns (red trace; Figure 5.5A) and the NH2 atom of R8 amino acid started forming a new hydrogen bond with the O4 atom of U66 at t = ∼1 ns (blue trace; Figure 5.5A). At t = ∼0.9 ns, the hydrogen bond between the NH2 atom of R14 amino acid and the O6 atom of G70 ruptured (red trace; Figure 5.5B) and the NH2 atom of R14 amino acid formed a salt bridge with the O2P atom of A68 (blue trace; Figure 5.5B).

At t = ∼0.25 ns, the NH1 atom of R15 amino acid stopped forming the hydrogen bond with the N7 atom of A73, and the NH2 atom of R15 formed a hydrogen bond with the O4 atom of U72. Thus, the combined interactions between the NH2 atom of R8 amino acid and the O4 atom of U66, between the NH2 atom of R14 amino acid and the O2P atom of A68, between the NH2 atom of R15 amino acid and the O4 atom of U72 created a network of salt bridging and hydrogen bonding interactions at ∼1 ns and lasted for ∼0.5 ns (Figure E.10B).

At t = ∼1.7 ns, the hydrogen bond between the NH2 atom of R15 amino acid and the O4 atom of U72 ruptured (blue trace; Figure 5.5C) and a salt bridge was formed between the NH2 atom of R15 and the O2P atom of the G42 which broke at t = ∼2.7 ns (green trace; Figure 5.5C). At t = ∼2.7 ns, a salt bridge was formed between the NH2 atom of R8 amino acid and the O2P atom of A68 which lasted for ∼0.2 ns (green trace; Figure 5.5A). Thus, salt bridging interactions were forming during every step of the dissociation process in PW3. The peptide was free of any interactions with the RNA at a distance of 40 Å (t = 6.4 ns).

Finally, in PW4, which had the lowest free-energy barrier to dissociation (Figure E.6D), the mechanism of dissociation was similar to PW3 but I observed several key differences. During the first 0.8 ns of the simulation, the interactions between the R8 amino acid and the U66 nucleotide and the R18 amino acid and the A68 nucleotide weakened, as characterized by the van der Waals interaction energy (Figure E.8D). The NH2 atom of R8 amino acid

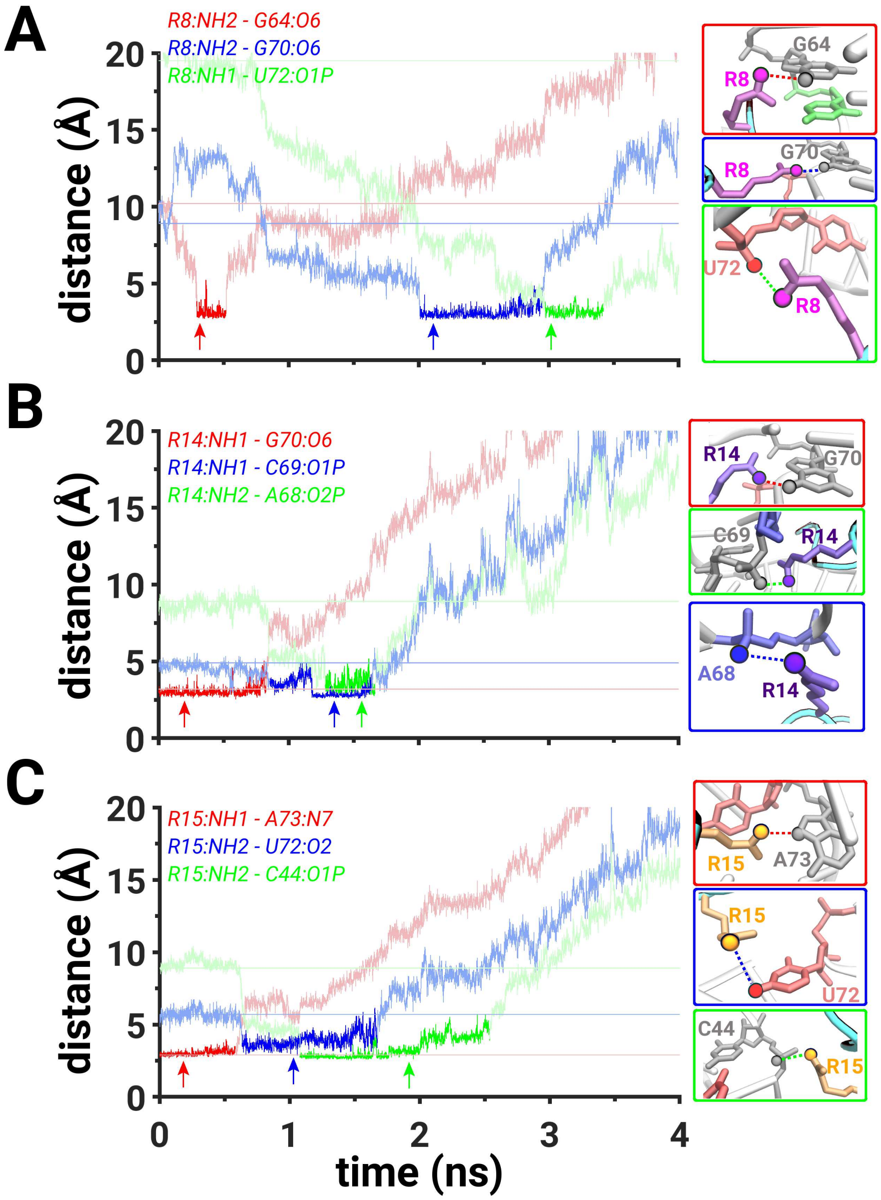

Figure 5.6: **Mechanistic details of PW4.** (A) The hydrogen bond distances between the NH2 atom of R8 and the O6 atom of G64 (red trace) and between the NH2 atom of R8 and the O6 atom of G70 (blue trace) and the salt bridge between the NH1 atom of R8 and the O1P atom of U72 (green trace). (B) The hydrogen bond distance between the NH1 atom of R14 and the O6 atom of G70 (red trace) and the salt bridges between the NH1 atom of R14 and the O1P atom of C69 (blue trace) and between the NH2 atom of R14 and the O2P atom of A68 (green trace). (C) The hydrogen bond distances between NH1 atom of R15 and the N7 atom of A73 (red trace) and between the O2 atom of R15 and the O2 atom of U72 (blue trace) and the salt bridge between the NH2 atom of R15 and the O1P atom of C44 (green trace). cf. Figure 5.3 for all other details.

formed a hydrogen bond with the O6 atom of G64 at t = ∼0.35 ns and broke it at t = ∼0.55 ns (red trace; Figure 5.6A). After that, the R8 amino acid did not form any stable interactions until t = ∼2 ns (Fig. S6D). At t = ∼0.8 ns, the hydrogen bond between the NH1 atom of R14 amino acid and the O6 atom of G70 ruptured (red trace; Figure 5.6B) and a salt bridge was formed between the NH1 atom of R14 amino acid and the O1P atom of G69 (blue trace; Figure 5.6B). At t = ∼0.65 ns, a hydrogen bond was formed between the NH2 atom of R15 amino acid and the O2 atom of U72 (blue trace; Figure 5.6C) which was preceded by the rupture of the hydrogen bond (at t = ∼0.6 ns) between the NH1 atom of R15 amino acid and the N7 atom of A73 (red trace; Figure 5.6A). The salt bridge between the NH1 atoms of R14 amino acid with the O1P atom of C69 and the hydrogen bond between the NH2 atom of R15 amino acid with the O2 atom of U72 formed a network of salt bridging and hydrogen bonding interactions at ∼0.8 ns (Figure E.10C) which corresponded to a metastable state at ∼5.1 Å (blue M; Figure E.6D).

At t = ∼1.3 ns, the NH2 atom of R14 amino acid formed another salt bridge with the O2P atom of A68 (green trace; Figure 5.6B). The NH2 atom of R15 amino acid ruptured the hydrogen bond with the O2 atom of U72 and formed a salt bridge with the O1P atom of C44 at t = ∼1.1 ns (Figure 5.6C). At t = ∼2 ns, the NH2 atom of R8 amino acid formed a hydrogen bond with the O6 atom of G70 (blue trace; Figure 5.6A) and combined with the salt bridge between the NH2 atom of R15 amino acid and the O1P atom of C44, the second network of hydrogen bonding and salt bridging interactions was created in PW4 (Figure E.10D) and corresponded to a metastable state at ∼6.9 Å (blue M; Figure E.6D). At t = ∼2.4 ns, a hydrogen bond between the NH2 atom of R8 amino acid and the O6 atom of G70 ruptured and a salt bridge was formed between the NH1 atom of R8 and the O1P atom of U72 (green trace; Figure 5.6A). The peptide was free of any interactions with the RNA at a distance of 40 Å (t = 6.4 ns).

Overall, I observed formation of unique interactions in each pathway, including the formation of salt bridging and hydrogen bonding interactions. These observations suggest that there is a network of salt bridges and hydrogen bonds that was formed in each pathway, with the exception of PW1 which had the smallest number of hydrogen bonds and salt bridges formed in comparison to other pathways.

5.6 Discussion

I have studied the dissociation mechanism of the RSG-1.2 peptide from the RRE RNA along four distinct pathways using non-equilibrium SMD simulations. Although, it has been previously proposed that the salt bridging interactions could be important for the recognition of this peptide by the RRE RNA [99], there is no study on the binding/unbinding mechanism of this peptide in the literature. I observed the formation of unique salt bridging and hydrogen bonding interactions in each pathway that form in an ordered step-wise sequence where the rupture of one interaction led to the creation of another interaction. I also estimated the free-energy profiles for each pathway using the Jarzynski's equality and observed distinct free-energy barriers in each pathway.

I observed the highest free-energy barrier of dissociation in PW1 (Figure E.6A) which was coupled with the displacement of the backbone atoms of the C44, U45, and G46 nucleotides (Figure E.9). Moreover, I observed only one salt bridge formed during dissociation in PW1 (Table 5.1; Figure 5.3C). The formation of this salt bridge between the R15 amino acid of the peptide and the U45 nucleotide of the RRE RNA was coupled with a weak recognition of the peptide by the RNA.

Even though, the overall process of dissociation in PW2 was somewhat similar to PW1, the free-energy barrier in PW2 was smaller than in PW1 (Figure E.6A,B). One of the key differences between PW1 and PW2 was the interaction between the R8 amino acid and the U66 nucleotide of the RNA that ruptured at t = ∼0.1 ns in PW2, as characterized by the van der Waals interaction energy, while the rupture of the interaction between the R8 amino acid and the U66 nucleotide only occurred at t = ∼0.8 ns in PW1 (purple traces; Figures E.8A,B). I also observed an additional salt bridge in PW2 that was formed between the NH2 atom of R15 amino acid and the O1P atom of C44 (red trace; Figure 5.4C). This interaction was formed ∼0.5 ns earlier in PW2 in comparison to a similar type of interaction between the NH2 atom of R15 amino acid and the O1P atom of U45 in PW1. The peptide passed in close proximity to the C44 and U45 nucleotides in both pathways and a faster establishment of a salt bridging interaction with an atom from one of these nucleotides is important for the recognition of the RNA backbone for the peptide if it dissociates in the direction of PW1 or PW2. The earlier rupture of the interaction between the R8 amino acid and the U66 nucleotide as well as a lack of displacement of the backbone atoms of the C44, U45, and G46 nucleotides led to a decreased free-energy barrier in PW2.

The pathways PW3 and PW4 had smaller free-energy barriers in comparison to PW1 and PW2 (Figure E.6). It should be noted that in PW3 and PW4, the peptide required a longer time to dissociate in comparison to PW1 and PW2 which was caused by additional interactions that were forming between the flipped-out A68 nucleotide and the peptide, as it was dissociating (Figures E.7C,D). These interactions were not formed in PW1 and PW2 because the peptide was dissociating in a direction away from the A68 nucleotide (Figures E.7A,B). Moreover, the dissociation reaction coordinate in PW3 and PW4 was free of any obstacles such as the atoms of the RNA backbone in the C44, U45, and G46 nucleotides that were present in PW1 and PW2. Thus, a decrease in the free-energy barriers in PW3 and PW4 was achieved by reducing any steric overlap or the displacement of the atoms in the major groove of the RNA. Therefore, these two pathways, PW3 and PW4, are preferred in comparison to PW1 and PW2 due to lower free-energy barriers for peptide dissociation (Figure E.6).

However, PW4 exhibited an even smaller free-energy barrier of dissociation by ∼7 kcal/-mol in comparison to PW3, meaning that the pathway PW4 is further preferred over PW3. This decrease in free-energy barrier in PW4 is likely a result of the behavior of the R8 amino acid which did not form any stable interactions between ∼0.5 ns and ∼2 ns in PW4 (Figure 5.6A) while it was forming stable hydrogen bonding interactions in that time range in PW3 (red and blue traces; Figure 5.5A). This behavior of the R8 amino acid was also reflected in the van der Waals interaction energies (purple traces; Figures E.8C,D) which showed that the R8 amino acid had stronger interactions with the U66 nucleotide in PW3 in comparison to PW4.

By analyzing the salt bridging and hydrogen bonding interactions in each pathway, I determined that R8, R14, and R15 were the most critical amino acids for the recognition of the peptide by the RRE RNA. Each of these amino acids were involved in a complex network of salt bridges and hydrogen bonds in PW2, PW3, and PW4 (Figures 5.4,5.5,5.6 and E.10). Moreover, these amino acids interacted with the RNA nucleotides in a step-wise pattern in which the rupture of existing interactions resulted in the formation of new interactions with other nucleotides during the dissociation process. PW3 and PW4 exhibited the formation of additional hydrogen bonds and salt bridges in comparison to PW1 and PW2 which resulted from the extended dissociation timescales.

In particular, I believe that the R8 amino acid was the most critical amino acid in the least free-energy barrier pathway PW4. Firstly, as mentioned before, the R8 amino acid had decreased interactions with the nucleotides of the RRE RNA between ∼0.5 ns and ∼2 ns which led to a decreased free energy barrier in PW4. Secondly, after the peptide dissociated from the initial binding pocket, the R8 amino acid was the only amino acid that was forming a stable interaction with the RNA nucleotide after ∼3 ns. Specifically, the NH2 atom of the R8 amino acid formed a salt bridge with the O1P atom of the U72 nucleotide between ∼3 ns and ∼3.4 ns (green trace; Figure 5.6A). Thus, I hypothesize that for the reverse process of peptide binding along PW4, the R8 amino acid will be the first amino acid to form a stable interaction with the U72 nucleotide of the RRE RNA.

Additionally, it is critical to note that the RSG-1.2 protein was synthesized by mutagenesis from the Rev peptide [323] that binds the RRE RNA during the HIV-1 replication process. One important mutation in that study, was the mutation of the arginine amino acid in the Rev protein at position 9 to proline (P9). It was hypothesized that this mutation resulted in a decrease of electrostatic contacts between the arginine amino acids in the N-terminus of the peptide and could be potentially coupled with the increased binding affinity to the RRE RNA [323]. However, it was not clear how the RSG-1.2 peptide recognized the RRE RNA during the binding process and which amino acids contributed the most to this process. In our work, I observed that the R8 amino acid, which is located next to the P9 amino acid in the polypeptide chain, formed stable hydrogen bonding and salt bridging interactions in each pathway. The R8 amino acid also was the last amino acid to interact with the RRE RNA during the dissociation and thus could be the first to interact with the RRE RNA during the binding process. Thus, the ability of the R8 amino acid to form these interactions was rooted in its flexibility that was coupled with the formation of various interactions with the RRE RNA nucleotides and resulted in the increased binding affinity and specificity with the RRE RNA in comparison to the Rev protein.

5.7 Conclusion

The (un)binding of proteins or short peptides in the RNA-protein complexes is an important biophysical process that is poorly understood. I used non-equilibrium cv-SMD simulations to study the dissociation mechanism of a helical peptide along four different pathways from the RRE RNA binding pocket to obtain key insights into the peptide binding/unbinding process and the recognition mechanism of this peptide. In particular, I investigated the mechanistic details of each pathway to identify interactions that are important for the recognition of proteins/peptides. I analyzed the resulting free-energy profiles and observed that the final free-energy differences were 96.47 ± 12.63 kcal/mol for PW1, 96.1 ± 10.95 kcal/mol for PW2, 91.83 ± 9.81 kcal/mol for PW3, and 92 ± 11.32 kcal/mol for PW4. Consistent with the similar initial (bound) and final (unbound) states of the peptide in each pathway, the resulting free-energy differences (ΔG) are consistent among different pathways. However, the free-energy profiles for each pathway exhibited different magnitudes of the free-energy barriers for dissociation of the peptide leading to the observation that PW4 is the preferred pathway of dissociation. In addition, the peptide dissociation was coupled with the formation of metastable states that resulted from a network of salt bridges formed between the arginine amino acids and the phosphate groups of the RNA backbone as well as from the hydrogen bonding. Specifically, I identified that the R8, R14, and R15 amino acids were important for the peptide recognition by the RRE RNA. Our results also suggest the R8 amino acid to be the most critical amino acid out of the three arginine amino acids due to its increased flexibility and the ability to form a primary/terminal salt bridging interaction with the U72 nucleotide during the binding/unbinding process in PW4. These observations are potentially important for the recognition mechanism between the RNA molecules and the proteins/peptides that have charged amino acids.

5.8 Supporting Information

Additional data and figures are shown in Appendix E along with the information about the helicity of the peptide in the absence of the RRE RNA. In Appendix F, I also provide example scripts that I used to set up, conduct, and analyze my simulations. I have also included in Appendix F the scripts for creating figures.

5.9 Publication

The work described in this chapter has been submitted for review:
Levintov, L., and Vashisth, H. (2021). Role of Salt-bridging Interactions in Recognition of Viral RNA by Arginine-rich Peptides. *Biophys. J.* (Under review).

CHAPTER 6
STUDY ON THE BASE FLIPPING MECHANISM IN RNA MOLECULES

6.1 Abstract

Base flipping is a key biophysical event involved in recognition of various ligands by RNA molecules. However, the mechanism of base flipping in RNA remains poorly understood, in part due to the lack of atomistic details on complex rearrangements in neighboring bases. For studies described in this chapter, I applied TPS methods to study base flipping in a dsRNA molecule that is known to interact with RNA-editing enzymes through this mechanism. I obtained an ensemble of 1000 transition trajectories to describe the base-flipping process. I used the likelihood maximization method to determine the refined RC consisting of two CVs, a distance and a dihedral angle between nucleotides that form stacking interactions with the flipping base. The free energy profile projected along the refined RC revealed three minima, two corresponding to the initial and final states and one for a metastable state. I suggest that the metastable state likely represents a wobbled conformation of nucleobases observed in NMR studies that is often characterized as the flipped state. The analyses of reactive trajectories further revealed that the base flipping is coupled to a global conformational change in a stem loop of dsRNA.

6.2 Significance

In the studies presented in this chapter, I determined the refined RC that can characterize base flipping mechanism in RNA molecules. I investigated a set of CVs which were previously used as single-variable RC models to characterize base flipping and I revealed that the refined RC is a more complex combination of two CVs. Additionally, using the refined RC, I observed that flipping of a single base resulted in local rearrangements in the neighboring bases which in return were coupled with global structural transitions of the RNA stem loop. I suggest that the refined RC could be applied to study base flipping mechanism in other RNA systems using different enhanced sampling methods.

6.3 Background

Interactions between nucleic acids and proteins play an essential role in various cellular processes including post-transcriptional modifications [324–326], repair mechanisms [327, 328], and replication [329, 330]. Some proteins bind to nucleic acids without introducing significant structural changes but in other cases binding is associated with large distortions in the structures of nucleic acids. Among other examples are enzymes that bind to nucleic acids upon opening of a specific base pair to perform a chemical reaction on the target base [103, 331]. It means that the bases involved in chemical modifications have to be accessible to enzymes, preferably in a flipped out (extrahelical) state. However, it remains unclear whether base flipping occurs spontaneously or not [332, 333]. Therefore, resolving atomistic details of a base flipping event remains a fundamental problem of interest in biophysics of nucleic acids.

Studying spontaneous base flipping is a challenging process both for experimental and computational methods due to a lower likelihood of observation of flipping in a single nucleobase in otherwise stable structures of nucleic acids. On the experimental side, NMR spectroscopy has become the leading method to study dynamics in nucleic acids due to its ability to probe fluctuations at the level of individual nucleobases [46–48].Specifically, NMR has been applied to study the base flipping mechanism through the imino proton exchange assay [108] where the exchange of the imino proton with the catalysts in the solution is assumed to occur only when the base flips out [109]. It has been applied to study base flipping [73, 100, 101], along with other experimental techniques including X-ray crystallography [103], fluorescence-based assays [104, 105], melting point studies [106], and combined approaches [107,108]. In DNA, NMR studies have shown that the lifetime of the extrahelical state of a base can be on the order of μs, and that of the intrahelical state in the range of ms depending on the stability of individual bases [109, 110]. Additionally, several studies revealed that the target imino proton on the base becomes accessible to the the solvent for proton when the base pair opens to a pseudo-dihedral angle of at least 30°, thereby indicating that the bases are still within the cutoff of a hydrogen bond formation [111,112]. Specifically, using a combination of NMR and MD methods, the authors compared the free-energy profile, the solvent accessibility of the imino proton for base flipping based on MD simulations, and the exchange time of the imino proton based on NMR data. Using the above comparison the authors proposed that the imino proton exchange occurs when the base opens at ± 30° which means that it is not fully flipped out [112]. Therefore, the fluctuations measured by NMR may need to be reassigned to base wobbling as opposed to flipping and the mechanistic understanding may not be directly applicable to a base flipping process [108]. Thus, despite key mechanistic information emerging from the application of NMR methods, there remains the need for additional analyses at the atomic level for obtaining further insights into this molecular mechanism.

On the computational side, due to limitations in conformational sampling by conventional MD simulations, enhanced sampling methods have been applied to probe this event [108, 111–119]. Among previous studies of base flipping, some have used external forces to induce base flipping transitions [111, 115], which likely leads to a loss of critical information on key variables that may contribute to base flipping. Enhanced sampling methods also rely on the definition of an appropriate RC which is a single variable to discriminate between a given pair of stable states and using which the key thermodynamic (e.g. free energy) properties can be computed. Although establishing an appropriate RC is challenging [120], once it is identified the multidimensional free energy surface can be reduced to a one-dimensional profile along the RC to obtain crucial mechanistic insights into the transition mechanism.

Many computational methods have been applied to study nucleobase stacking/unstacking in nucleic acids [113, 114, 117, 121–125]. Several significant studies have been conducted to study the base flipping process in DNA in association with protein binding [126, 127]. Several of these studies explored simplified systems that consisted only up to three base pairs and may be limited in describing the dynamics in a larger RNA system with many base pairs [121–123]. Additionally, several previous studies were reported over a decade ago and the force fields for nucleic acids have significantly improved in recent years [36, 128]. Moreover, the candidate variables that potentially contribute to RC have not been examined systematically. Therefore, the application of simulation methods that permit systematic testing of a suitable RC is needed to improve the understanding of the mechanism of base flipping in nucleic acids.

Such techniques include the method of TPS [196, 206, 334], which has been successfully applied to study the flipping of a terminal pyrimidine base in a short DNA chain with three base pairs [114]. While most previous studies have focused on DNA due to its structural stability, I study RNA as a model system given its conformational flexibility and emerging importance in drug discovery [213]. TPS has also been applied to explore other biophysical problems including folding [209, 246, 335], flipping of amino acids in enzymes [336], DNA synthesis [337], water dynamics [338], catalysis [207, 208, 339], nucleation [340, 341], and chemical reactions [342]. In this work, I applied TPS [196, 206, 334] simulations and the likelihood maximization methods [204] to study the base flipping mechanism in a dsRNA molecule which has a nucleobase that can flip out (Figure 1.6).

6.4 Methods

6.4.1 System Preparation and Simulation Details

The initial coordinates for dsRNA were obtained from the first frame of the NMR structure (PDB code: 2L2K) [100]. The system was solvated in a 72 Å × 72 Å × 83 Å periodic box

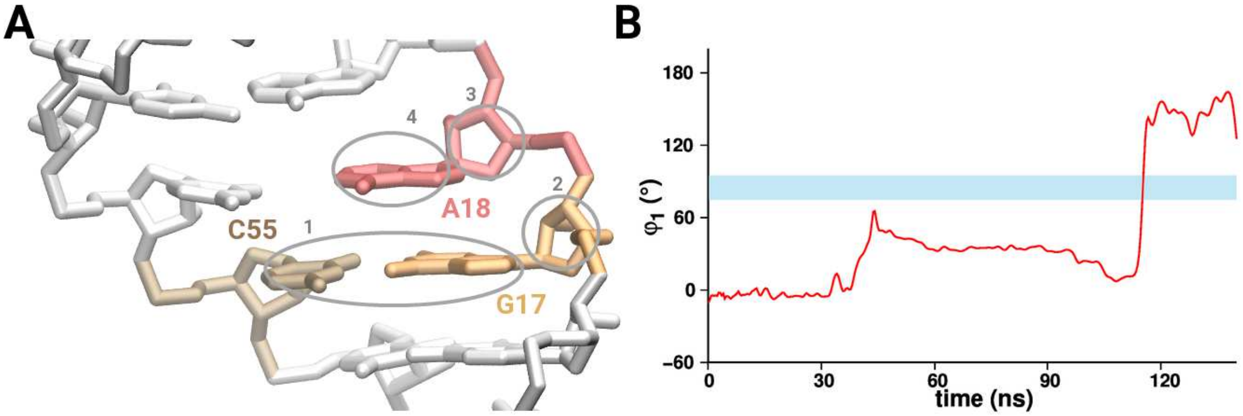

Figure 6.1: **Details on the primary order parameter.** (A) The OP is defined by the center of mass of each of the following four groups: the nitrogenous bases of C55 and G17 (labeled 1), sugar moiety attached to G17 (labeled 2), sugar moiety attached to A18 (labeled 3), and the nitrogenous base of A18 (labeled 4). Each key nucleotide is also uniquely colored and labeled. (B) Shown is a time trace of the primary OP in the seed trajectory (red). A cyan rectangle highlights the shooting region. See also Figure G.2.

of TIP3P water molecules and was comprised of 39305 atoms (Figure G.1). The system was neutralized with 21 Mg^{2+} ions. The temperature and pressure were maintained at 310 K and 1 atm using the Langevin thermostat and the Berendsen barostat, respectively. All MD simulations were carried out using the Amber [154] software combined with the recent RNA Amber force-field developed by a Rochester group (RNA.ROC) [157]. The analyses of all trajectories were carried out using the CPPTRAJ module in Amber and using the VMD software [184, 185].

6.4.2 Transition Path Sampling

TPS [195, 196, 206] is a method to generate an ensemble of transition paths that connect a pair of initial (reactant) and final (product) states that are separated by a free energy barrier. More details on the TPS method can be found in section 2.4.2.

6.4.3 Seed Trajectory and Definitions of Stable States

I conducted four conventional MD simulations, each 150 ns long, to obtain a seed trajectory. Only one out of these four simulations exhibited a spontaneous base flipping event (Figure

G.2) and therefore it was used as the seed trajectory for building an ensemble of transition paths. For these MD simulations, I first performed 1000 steps of steepest descent minimization followed by 500 steps of conjugate gradient minimization. Then, I conducted all MD simulations in the NPT ensemble using a 2 fs timestep and saved configurations every 10 ps. Prior to launching shooting simulations, I defined an OP that can unambiguously discriminate between the two stable states, the inward (*I*, state 1) and the outward (*O*, state 2) states, and determined the ranges of the OP for defining two stable states. These ranges were chosen to clearly separate stable basins, accommodate system fluctuations, and prevent sampling of non-reactive trajectories [195, 196, 209]. I selected a pseudo-dihedral angle as the OP that is defined by the centers of mass of four groups of atoms (Figure 6.1A) which has been previously identified as a potential collective variable (CV) for this system [124]. For the configuration *I*, the range of the OP was defined as -70° $< \text{OP}_1 <$ 70° and for the configuration *O* as 100° $< \text{OP}_2 <$ 180° and -120° $< \text{OP}_2 <$ -180°. From the shooting region identified in the seed trajectory, I then launched 1000 shooting trajectories, each 1 ns long. I carried out all shooting simulations in the NPT ensemble using a 2 fs timestep and saved configurations every 0.25 ps. Based on the definition of the OP, 748 of them terminated in the state *I* and 252 in the state *O* [203, 204].

6.4.4 List of Collective Variables

In addition to the primary OP, a list of other potential CVs was created and monitored in all shooting trajectories. The distributions of the values of each of the CVs (q_k) at terminal points in shooting trajectories were examined to select those CVs to be included in the construction of RC that discriminated between the I/O states. I used equation 2.15 to normalize the CVs (section 2.4.2e).

The 12 CVs (k = 1, 2,$\cdots$, 12) that were identified (Table G.1) are:

1. ϕ_1: the pseudo-dihedral angle that describes the position of A18 relative to G17. It also served as our primary OP.

2. ϕ_2: the pseudo-dihedral angle that describes the position of A18 relative to A19.

3. d_1: the distance between the centers of mass of G17 and A18.

4. d_2: the distance between the centers of mass of A18 and A19.

5. d_3: the hydrogen bond distance between the N_1 atom of A18 and the N_3 atom of C54.

6. α_1: the angle between A18 and C28 defined using the following three atoms: N_9 and C_1' of A18, and C_1' of C54.

7. α_2: the interplane angle between G17 and A18. Only heavy atoms were used to define the plane.

8. α_3: the interplane angle between A18 and A19. Only heavy atoms were used to define the plane.

9. N_W: the number of water molecules within 8 Å of A18.

10. E_1: the stacking energy between bases G17 and A18.

11. E_3: the interaction energy between bases A18 and C54.

12. E_2: the stacking energy between bases A18 and A19.

6.4.5 Refined Reaction Coordinate

The RC is defined as a linear combination of the identified and normalized CVs using equation 2.16 (section 2.4.2f). Subsequently, I applied the likelihood maximization method [203,204] to find the best set of CVs and associated a_k's following the procedure described in section 2.4.2f. By varying m in equation 2.16, different models of the RC were investigated and presented in Tables G.2-G.4.

6.4.6 Free Energy Profile along RC

The PMF/free energy profile was obtained along the RC using equation 2.20 and additional details are provided in section 2.4.2g.

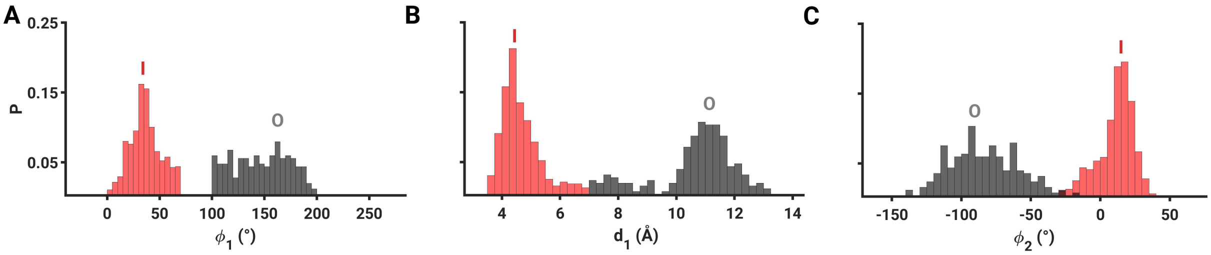

Figure 6.2: **Population distributions of CVs at terminal points.** Shown are the distributions of CVs at terminal points of transition paths for the inward (red) and outward (gray) states. (A) The pseudo-dihedral angle (ϕ_1) that describes the position of A18 relative to G17. (B) The distance (d_1) between the centers of mass of G17 and A18. (C) The pseudo-dihedral angle (ϕ_2) that describes the position of A18 relative to A19. See also Figure G.3.

6.5 Results and Discussion

6.5.1 Fluctuations of the OP (ϕ_1)

After defining a list of all potential CVs that can be used to describe the transition between the I and O configurations, I picked the pseudo-dihedral angle (ϕ_1) between G17 and A18 as our primary OP, which was used to identify the seed trajectory (Figure 6.1). A time trace of the OP in the seed trajectory is shown in Figure 6.1B and the distribution of its values at terminal points in the transition paths in Figure 6.2A. I defined the shooting region as the range between 75° and 95° and observed a transition in this region (at ~117 ns) in one of four conventional MD simulations (Figure 6.1B). Three other conventional MD simulations that were launched from the same initial structure did not exhibit base flipping (Figure G.2). The configurations from the shooting region in the seed trajectory were then used as input structures to build an ensemble of 1000 shooting trajectories.

6.5.2 Identification of Other Potential CVs

All of the predefined CVs (Table G.1) were monitored at the terminal points of each shooting trajectory for their suitability in discriminating between two states and thus for inclusion in construction of the RC. The population distributions of all tested CVs at terminal points in shooting trajectories are shown in Figures 6.2 and G.3. The CVs in Figure 6.2 (ϕ_1, d_1, and ϕ_2) were found to be the most important for constructing a model of the refined RC because these CVs exhibited distinct bimodal distributions at the terminal points where one peak was more populated at the inward state and the other at the outward state (Figure 6.2). These variables collectively describe the relative position of the nucleobase A18 with respect to G17 and A19 with which A18 forms stacking interactions (Figure 1.6). Since ϕ_1 and d_1 likely provide similar information, I anticipated that one of them may be omitted from the refined RC. Many CVs shown in Figure G.3 exhibited overlapping distributions between the two states with several CVs showing larger overlaps in their distributions (e.g. Figures G.3G,I) due to which these CVs were not included in the refined RC.

6.5.3 Refined Reaction Coordinate (r)

The refined RC was then determined using the likelihood maximization and the BIC [203,204] by testing various models of increasing complexity constructed from 12 CVs (Tables G.1-G.4). I found that the refined RC was a linear combination of 2 CVs, d_1 and ϕ_2. The final equation for the refined RC is:

$$r = -0.84 + 0.4753 d_1 - 0.3941 \phi_2 \tag{6.1}$$

The addition of a third variable to the RC improves it but according to the likelihood maximization tests, the improvement is not significant (Table G.4), meaning that 2 CVs are sufficient to formulate the refined RC. This can be seen in Figure G.4A which shows the histograms of the RC values across all shooting trajectories and the resulting free energy profiles in Figure G.4B. These data show that, for the three-variable RC models, the free energy profiles are similar to the refined RC (Figure G.4B). The evolution of the refined RC in the transition paths is shown in Figures 6.3A and G.5. The time evolution of the RC further confirms its validity by showing that the trajectories initiated from the region near $r = 0$ terminated in one of the two stable states and the RC is divided into two segments of the configuration space by terminating either at the inward state or at the outward state.

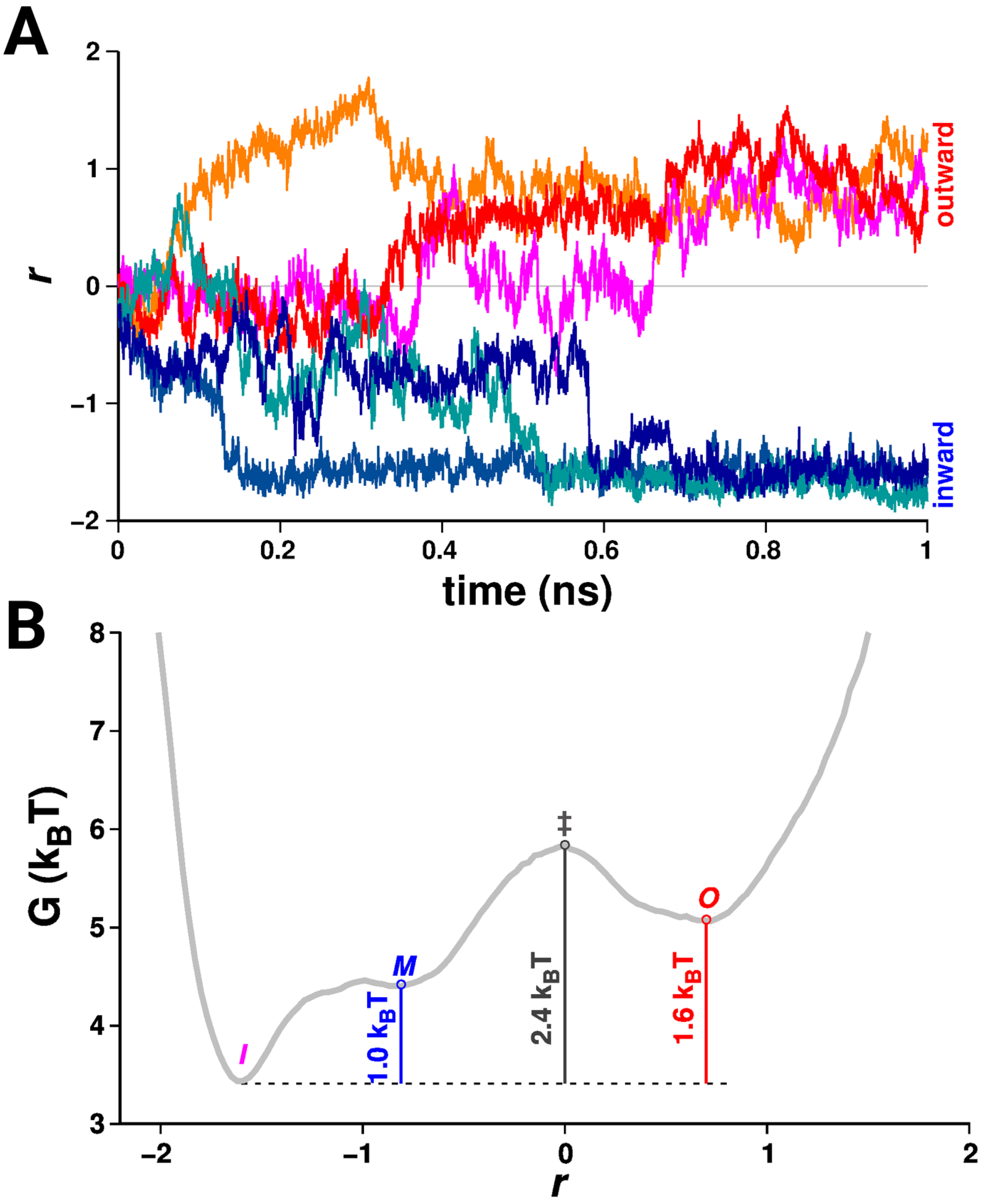

Figure 6.3: **Evolution of the refined RC and the potential of mean force (PMF) profile.** (A) The evolution of the RC along representative trajectories. See also Figure G.5. (B) PMF as a function of the RC. Three vertical lines mark the free energy difference between the inward (labeled I) and metastable (labeled M) states (blue), the activation energy (dark gray; labeled ‡), and the energy difference between the inward I and outward (labeled O) states (red).

6.5.4 Free Energy Profile

The PMF profile (Figure 6.3B) was estimated based on the population distribution of the refined RC computed across all transition paths (Figure G.6). The PMF profile exhibited three minima corresponding to the inward (I) state (-1.65 $<$ r $<$ -1.56), a metastable (M) state (-0.87 $<$ r $<$ -0.75), and the outward (O) state (0.63 $<$ r $<$ 0.81). The transition state is represented by $r = 0$. The activation free energy ($\Delta G^{\ddagger}$) and the free energy difference (ΔG) between states I and O were determined to be 1.48 kcal/mol and 1.0 kcal/mol, respectively. Based on the free energy profile, the outward state is less stable than the inward state which could be important for the deamination process performed by the ADAR2 enzyme [100]. Importantly, the metastable state represented the wobbling movement of A18 when the nucleotide is partially flipped out with ϕ_1~50°-65°. I suggest that the metastable states of these types are likely observed by NMR and mischaracterized as the flipped out (outward/extra-helical) states [100, 111]. The final RC also discriminates between the metastable (wobbled) and the flipped out (outward) states.

6.5.5 Conformational Properties of dsRNA in the Transition Path Ensemble

I launched shooting trajectories from the shooting region that is located close to the transition region and trajectories landed either in the reactant state (I and M) or in the product state (O). Based on the OP, 748 trajectories terminated in the I state and 252 trajectories terminated in the O state. The conformations of bases in dsRNA at the state I and the shooting region are shown in Figure G.7A,B. In the shooting region, the positions of A18 and U53 are perturbed compared to the initial (I) conformation by ~75-95° (ϕ_1) and ~30-40° (using the flipping angle definition similar to ϕ_1), respectively. I observed that the flipping motion of A18 resulted not only in rearrangements in neighboring bases but also in a conformational change in a stem loop of dsRNA (the loop highlighted in magenta/blue/red in Figure 6.4). Below, I discuss how the flipping of A18 affected the conformation of the stem loop, motion of nucleotides, and hydrogen bonds between various bases in three different

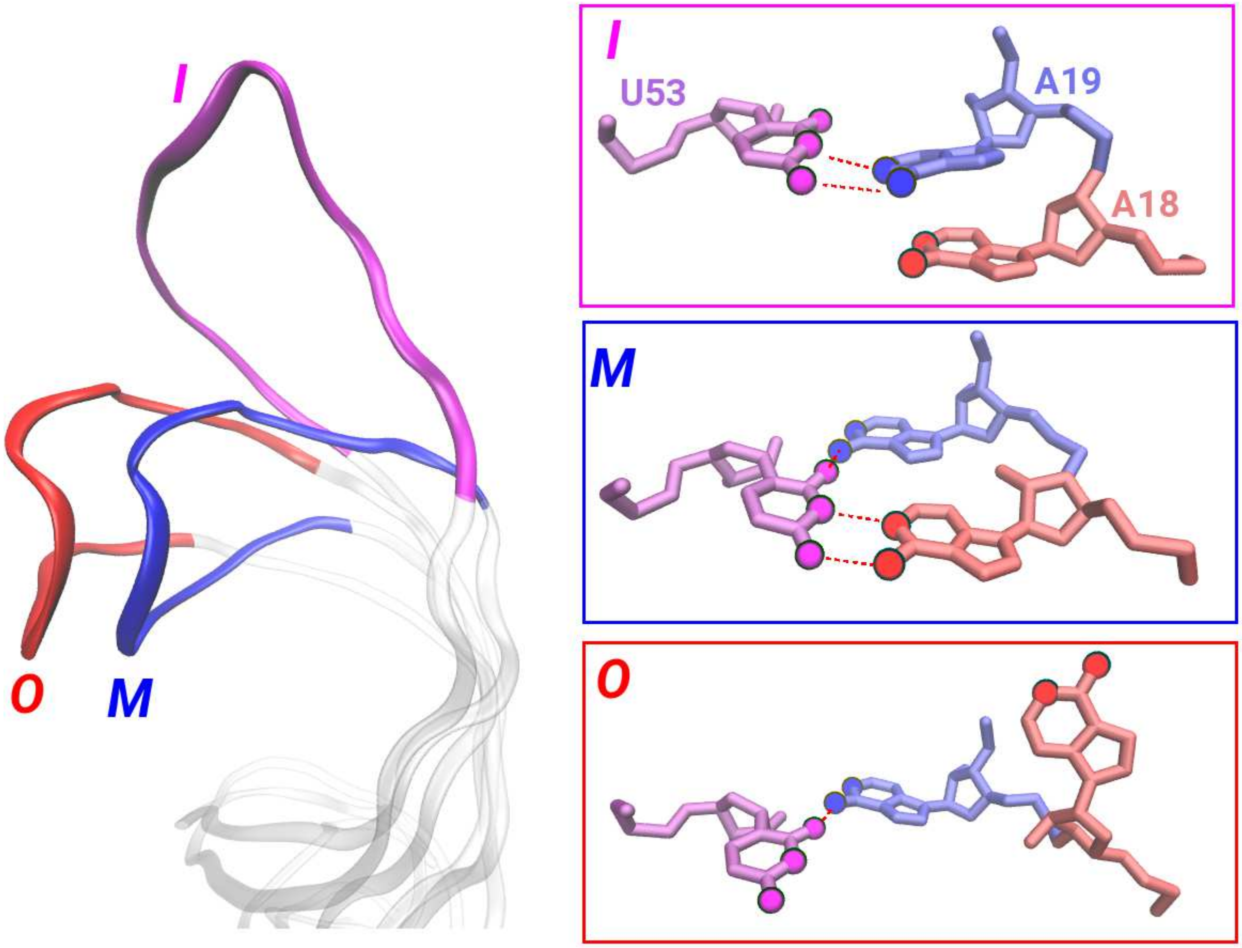

Figure 6.4: **Global and local conformational dynamics in dsRNA.** (*left*) Snapshot of global conformational changes in the RNA stem loop derived from shooting trajectories at three different states: *I* (magenta), *M* (blue), and *O* (red). (*right*) Snapshots of the flipping site in three different states. Each key nucleotide and atoms that participate in hydrogen bonding (marked by dotted red lines) are uniquely colored.

states (*I*, *M*, and *O*).

State *I***:** During inward flipping of A18 from the transition barrier region, A18 and A19 formed the base pairs with C54 and U53, respectively, and the RNA stem loop had an elongated conformation (*I*; Figure 6.4). The formation of a base pair between A19 and U53 was measured via a hydrogen bond distance of 3.5 Å between the O4 atom of U53 and the N6 atom of A19 (Figure G.8A). Concomitantly, C54 partially flipped out by ∼55-60° to provide space for A18 to flip back in (Figure G.8B and *I*; Figure 6.4). The flipping of C54 outward (Figure G.8A) was not observed in the initial configuration (Figure 1.6).

State *M***:** When A18 was in the *M* state (i.e. wobbled conformation), the RNA stem loop was in a bent conformation relative to state *I* (*M*; Figure 6.4). This conformation resulted due to the interactions between a triplet of bases: A18, A19 and U53 (*M*; Figure 6.4). A18 formed a hydrogen bond (3.5 Å long) with U53, which was partially flipped out at the transition barrier (Figure G.8C). At the same time, the initial hydrogen bond between the O4 atom of U53 and the N6 atom of A19 broke and a new hydrogen bond formed between the O2 atom of U53 and the N6 atom of A19 (Figure G.8A,D). As a result of these rearrangements, U53 formed hydrogen bonds with both A18 and A19, thus creating a triplet, which caused the RNA stem loop to bend (*M*; Figure 6.4).

State *O***:** In shooting trajectories that resulted in outward flipping of A18, the RNA stem loop was also observed to undergo a bent conformation (*O*; Figure 6.4). Similar to the *M* state, U53 partially flipped out and disrupted the initial hydrogen bond with A19 and formed another between the O2 atom of U53 and the N6 atom of A19 (Figure G.8D). A18 did not form any interactions with U53 but the flipping of A18 outward likely perturbed U53 and caused U53 to partially flip out. Thus, even minor conformational change in A18 by ∼45° (*M* state) caused local rearrangements in U53 while breaking the initial hydrogen bonds with A19 which in return resulted in a bent conformation of the RNA stem-loop. Overall, our mechanistic analyses of conformations of bases in the transition path ensemble revealed that the flipping of a single base (A18) in RNA is not only coupled with rearrangements in local bases, but also global conformational changes in common motifs (e.g. stem loops) found in nucleic acids.

6.5.6 Comparison to Previous Work

I note that Hart *et al.* [124] have previously studied this base flipping event by focusing on ϕ_1 as their hypothesized RC. However, the search for the refined RC in my work is systematic and exhaustive since I have examined a large number of CVs and their combinations using the likelihood maximization method [203,204]. The ensemble of trajectories that I have generated (totaling over 1000 ns) exceeded what was used in the previous work (14.4 ns) which further helped me in identifying a refined RC. Importantly, my results showed that, for a single variable RC model, ϕ_2 is a more important CV than ϕ_1 because ϕ_2 was ranked 2nd, while ϕ_1 was ranked 11th (Table G.2). Additionally, I estimated that a two-variable RC model is more significant for capturing the base flipping process than a single-variable model, whether it consisted of ϕ_1 or any other variable (Table G.3). In fact, even a three-variable RC model did not indicate that ϕ_1 was the most important CV out of the remaining CVs for model improvement since the model with ϕ_1 was ranked 4th (Table G.4). Moreover, I also observed that the local base flipping event in dsRNA is coupled with a global conformational transition in the stem-loop of this dsRNA. In the previous work [124], only local rearrangements of A18 and the neighboring bases were reported but in my work I showed that even a partial flipping of A18 caused the stem loop of dsRNA to bend. I also revealed that in the metastable state, U53 forms a base triplet with A18 and A19 through hydrogen bonding interactions (Figure 6.4) which has not been reported previously.

6.6 Conclusions

Using transition path sampling combined with the likelihood maximization methods, I developed a refined RC to describe the base flipping mechanism in dsRNA. The refined RC is comprised of two CVs that collectively describe the relative position of the flipping base with respect to the neighboring bases, thereby showing an improved description of the base-flipping mechanism. Outside of conformational variables, I did not observe any significant improvements in my RC models on including the solvent molecules or stacking energies between the bases. However, a further examination of these coordinates may be needed for other RNA motifs (e.g. bulges) if the flipping nucleotides are not involved in base-pairing interactions unlike the system studied in this work. My results emphasize the importance of systematic examination of CVs in constructing RC models of complex biophysical processes. I also observed that the flipping of a single base caused local rearrangements in the neighboring bases which then resulted in global structural transitions in a stem loop of the dsRNA. I suggest that the approaches described in this chapter are potentially applicable to other RNA/ligand systems, for example, conformational transitions coupled to binding of a ligand molecule in an RNA element from HIV-1, as reported in chapter 4 [226].

6.7 Supporting Information

Additional data and figures are shown in Appendix G. I have performed preliminary estimates on kinetics of base flipping which are also presented in Appendix G. In Appendix H, I also provide example scripts that I used to set up, conduct, and analyze my simulations. Specifically, all the scripts with descriptions that were used to perform TPS simulations are shown in Appendix H.

6.8 Publication

The work described in this chapter is reproduced from Ref. [343], with permission from the American Chemical Society. The citation is as follows:

Levintov, L., Paul, S., and Vashisth, H. (2021). Reaction coordinate and thermodynamics of base flipping in RNA. *J. Chem. Theory Comput.* 17:1914-1921.

CHAPTER 7
STUDY ON THE SELF-ASSEMBLY AND DYNAMICS OF PORPHYRIN/DNA SYSTEMS

7.1 Abstract

In this chapter, I discuss the results of MD simulations of porphyrin/DNA nanoassemblies. This work was carried out in collaboration with Dr. Shambhavi Tannir at University of Wyoming, with Dr. Krisztina Varga at University of New Hampshire, with Dr. Mark Townley at University of New Hampshire, with Dr. Milan Balaz at Yonsei University, with Dr. Brian Leonard at University of Wyoming, and with Dr. Jan Kubelka at University of Wyoming. I only describe the MD simulation part which includes simulations of the 40 DPD molecules and 2 DNA - 40 DPD systems. Other details can be found in our collaborative publication [344].

7.2 Significance

In the studies presented in this chapter, I identified the self-assembly mechanism and orientation of the porphyrin/DNA systems. Additionally, I measured various physical variables that characterize the final conformation of the assembly and these parameters agree with the experimentally observed behavior of the porphyrin/DNA nanoassembly. These results enhance our understanding of the self-assembly mechanisms of supramolecular nanoassemblies and of the roles of porphyrin and DNA molecules in these processes.

7.3 Background

I provide a brief introduction on the porphyrin/DNA system in section 1.5.3. In this study I conducted MD simulations of an achiral stack of porphyrin molecules in the presence and absence of the DNA strands. The primary goal was to identify the mechanism of self-assembly of a porphyrin/DNA system as well as the overall structure of the assembly and characterize it using various conformational metrics.

7.4 System Setup and Simulation Details

All MD simulations were carried out and analyzed using the NAMD/VMD software suite [152, 184]. The Amber bsc1 force-field [158] was used to simulate the DNA strands. The initial structure of an oligothymidylic acid T40 was created using the psfgen tool in VMD by using the topology information on a single thymine nucleotide. To prepare the initial system, coordinates, charges, and Amber force-field parameters of the porphyrindiaminopurine (DPD) molecule were developed. At first, the initial structure of the DPD molecule was created using CHARMM-GUI Ligand-Reader and Modeler [345, 346] followed by 100 steps of conjugate gradient energy minimization in NAMD. Then, the Antechamber program [163, 243] in Amber was used to obtain force field parameters, and the AM1-BCC charge method [162] was used to obtain atomic charges. After developing parameters, 40 DPD molecules were positioned on top of each other in an achiral stacked conformation, and two T40 DNA strands were added on the opposite sides of the stack. In the other system, a stack of 40 DPD molecules without DNA strands was positioned in an achiral pattern.

All systems were solvated using explicit water (TIP3P) molecules, and the overall charge of the DNA strands was neutralized by adding Mg^{2+} ions. Detailed information on the system size and trajectory length is provided in Table 7.1. The box volume was optimized in the NPT ensemble by first running a 1000 step conjugate gradient energy minimization that was followed by a 400 ps MD run with a 2 fs time step. The temperature in all simulations

Table 7.1: Details of simulation systems.

System	2 DNA - 40 DPD	40 DPD
Number of atoms	423,882	173,313
Simulation time (ns)	200	75
Size (Å^3)	219 × 156 × 133	226 × 98 × 95
Temperature (K)	310	310
Number of Mg^{2+}	40	0
Minimization steps	1000	1000
Force-field	Amber	Amber

was maintained at 310 K and controlled using the Langevin thermostat, and the pressure was controlled by the NoseHoover barostat in all NPT runs. All simulations were carried out using periodic boundary conditions. The simulations were further run in the NVT ensemble after brief initial equilibration in the NPT ensemble. Long-range electrostatic interactions were treated by the particle-mesh Ewald method.

7.5 Results

MD simulations were carried out to determine the probable mechanism of the assembly process. I hypothesized that since the DNA strands consist of 40 thymine nucleotides (T40), each nucleotide potentially interacts with a single molecule of DPD, and thereby 40 DNA bases potentially interact with 40 DPD molecules. Based upon this hypothesis, I assembled a system containing 2 DNA strands and 40 DPD molecules positioned between the DNA strands in an achiral columnar nanostack (Figure 7.1A).

The simulation of the 2 DNA 40 DPD system showed that the DPD molecules preferred to remain stacked and formed a predominantly helical type of structure without the application of an external physical stimulus (Figure 7.1A). After the simulation was initiated, the two T40 DNA strands diffused through the aqueous environment toward the achiral DPD stack as they interacted with it. At ∼90 ns, the counterclockwise twist of the 40 DPD stack started to appear as one T40 DNA strand interacted with the whole DPD stack, while the other interacted partially (Figure 7.1A). Detailed analysis showed that DPD molecules

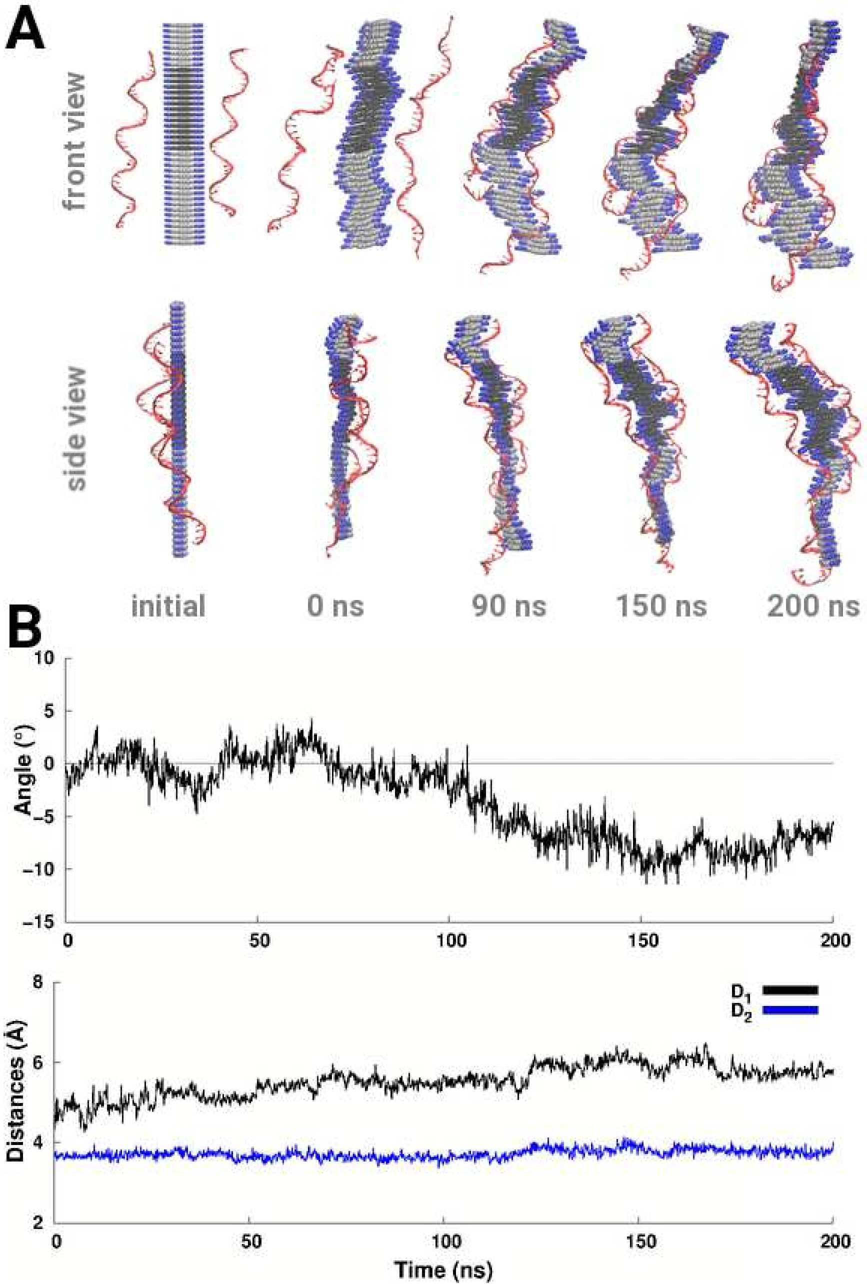

Figure 7.1: **2 DNA - 40 DPD system.** (A) Snapshots of the 2 DNA 40 DPD system are shown in two different views before the simulation was initiated, after minimization, as well as at t = 90 ns, 150 ns, and 200 ns. The DNA strands are shown as red cartoon and the DPD molecules are shown in space-filling representation. DPD molecules, that have been selected for the CD spectra calculations, are highlighted in black. (B) The orientation angle and two distances, the center-to-center distance of two adjacent DPDs (D_1) and the rise per DPD along the assemblys axis (D_2), are shown for the 2 DNA - 40 DPD system.

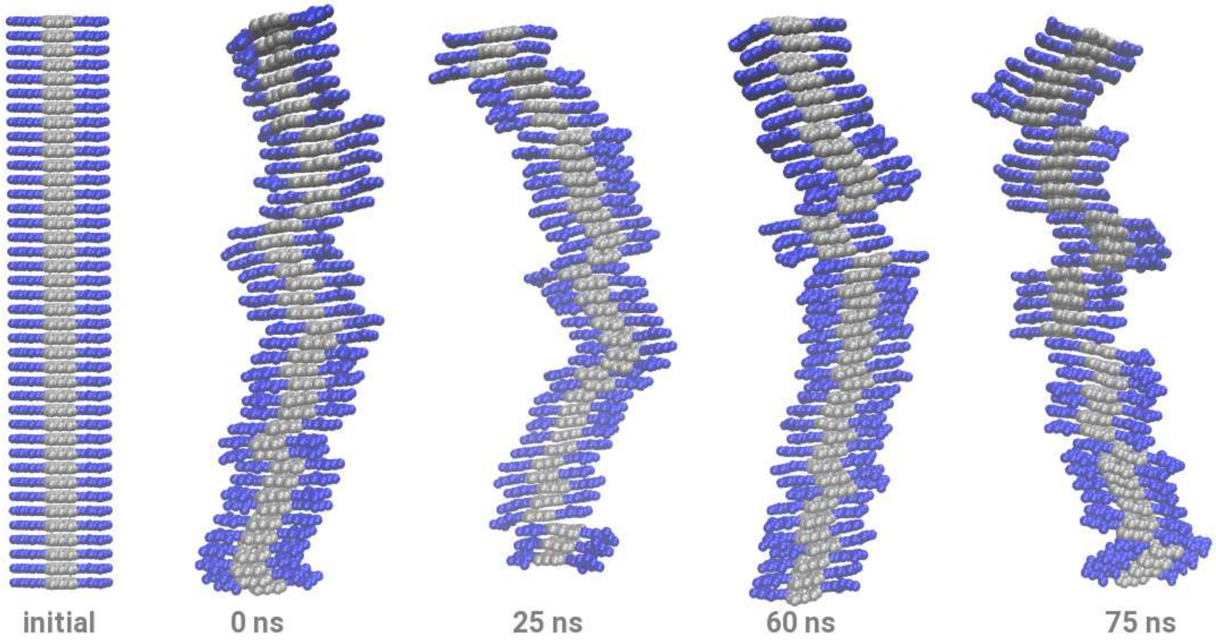

Figure 7.2: **Snapshots of the 40 DPD system.** Snapshots of the 40 DPD system are shown before the simulation was initiated, after minimization, as well as at t = 25 ns, 60 ns, and 75 ns. The DPD molecules are shown in space-filling representation.

924 of the 40 DPD stack (Figure 7.1A) interacted with both T40 DNA strands and continued to rotate counterclockwise until a relatively stable left-handed helix was formed at ∼140 ns and remained for the rest of the simulation (200 ns, Figure 7.1A). The MD simulation thus successfully reproduced the experimentally observed formation of the left-handed nanoassemblies by fast cooling in the absence of NaCl. The analysis of the system over the last 60 ns revealed the center-to-center distance of two adjacent DPD molecules of $D_1 = 5.9$ Å, the rise per DPD molecule along the assemblys axis of $D_2 = 3.6$ Å, the rotation per DPD molecule of $-8.1°$, and the length of the overall 40 DPD stack ∼145 Å (Figure 7.1B).

The simulation of the 40 DPD system without any DNA strands showed that the DPD molecules remain stacked during the entire simulation which is consistent with the ability of porphyrin-type molecules to self-assemble and stack through $\pi - \pi$ interactions (Figure 7.2). Importantly, I observed that the DPD molecules formed smaller groups of 3-5 molecules that moved together but did not dissociate from the overall assembly (Figure 7.2). Overall, the 2 DNA 40 DPD system was the only system where a stable helical shape was observed and the 40 DPD nanoassembly remained achiral.

7.6 Conclusions

In this study, I conducted explicit-solvent MD simulations of the porphyrin (DPD) molecules with and without DNA strands. These simulations revealed a left-handed orientation of the nanoassembly in the presence of two DNA strands with the center-to-center distance of two adjacent DPD molecules of 5.9 Å, the rise per DPD molecule along the assemblys axis of 3.6 Å, the rotation per DPD molecule of -8.1°, and the length of the overall 40 DPD stack of $\sim$145 Å. I did not observe the formation of a helix in the simulation of the 40 DPD molecules without any DNA strands but, importantly, the DPD molecules did not dissociate from the nanostack. Overall, DNA strands facilitated the formation of a helical shape of the porphyrin/DNA system.

7.7 Publication

The work described in this chapter is reproduced from Ref. [344] with permission from the American Chemical Society.

Tannir, S., Levintov, L., Townley, M. A., Leonard, B. M., Kubelka, J., Vashisth, H., Varga, K., and Balaz, M. (2020). "Functional nanoassemblies with mirror-image chiroptical properties templated by a single homochiral DNA strand." *Chem. Mater.*, 32:22722281.

CHAPTER 8
FUTURE WORK

In this chapter, I provide suggestions for future research work.

For work described in chapter 3, I conducted explicit-solvent MD simulations of different unliganded TAR RNA conformations which revealed the formation of transient binding pockets that can accommodate ligands of various sizes. I suggest that future researchers can use my simulations or conduct additional MD simulations of unliganded TAR RNA systems and perform compound docking in the pockets that I have identified in my work (see Figures 3.8 and A.16). As a first step, one can construct a large library of inhibitors which are known to bind TAR RNA and then expand the library by modifying the known binders and test them using virtual screening method.

In chapter 4, I studied the (un)binding process of a small molecule from the TAR RNA using non-equilibrium cv-SMD simulations. In the majority of cases, ligand binding is only investigated in terms of binding sites and bound poses while the kinetics and mechanisms of binding have not been studied to a large extent [144, 347]. A study can be conducted which combines the two approaches, SMD simulations and the TPS method, to develop a reaction coordinate which can describe the dissociation process of a small molecule from the viral RNA molecule. SMD simulations can be used to generate a seed trajectory which connects the bound (reactant) and unbound (product) states. The TPS method with the likelihood maximization method can be used to systematically explore a set of CVs in the construction of the RC.

In chapter 5, I studied the (un)binding process of the RSG-1.2 peptide from the RRE

RNA using non-equilibrium cv-SMD simulations. I think that another similar peptide that can be studied is the the Rev peptide which is a conventional RRE RNA binder. Specifically, it is known that the RSG-1.2 peptide, which was synthesized by mutating the Rev peptide, has an increased binding affinity and specificity to the RRE RNA in comparison to the Rev peptide [323]. It was previously hypothesized that increased binding affinity of the RSG-1.2 was coupled with a lower number of arginine amino acids in comparison to the Rev peptide [323]. A study can be conducted using free energy perturbation methods to identify key energetic and mechanistic differences between the RSG-1.2 peptide and the Rev peptide which facilitate an increased binding affinity of the RSG-1.2 peptide by the RRE RNA. Another suggestion is to mutate a single or several arginine amino acids in the RSG-1.2 peptide to further explore the importance of arginine amino acids.

In chapter 6, I studied the base flipping mechanism in a dsRNA molecule using path sampling methods. In my work, the flipping base was mismatched, or in other words it was not forming a canonical Watson-Crick base pair. I think that future researchers can focus on investigating systems in which a nucleotide is not involved in base-pairing interactions, for example nucleotides that constitute bulge motifs or hairpins. A good model system in that case would be the HIV-1 TAR RNA. The base flipping process can also be studied in association with protein binding to explore to what extent proteins affect the flipping of bases in RNA molecules. A reaction coordinate can be determined for the RNA/protein system and it would be of interest to compare it to the reaction coordinate that was reported in my work.

LIST OF REFERENCES

[1] Ralf Dahm. Discovering DNA: Friedrich Miescher and the early years of nucleic acid research. *Hum. Genet.*, 122(6):565–581, 2008.

[2] Oswald T Avery, Colin M MacLeod, and Maclyn McCarty. Studies on the chemical nature of the substance inducing transformation of pneumococcal types. Induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. *J. Exp. Med.*, 79(2):137158, 1944.

[3] Erwin Chargaff, Rakoma Lipshitz, and Charlotte Green. Composition of the desoxypentose nucleic acids of four genera of sea-urchin. *J. Biol. Chem.*, 195(1):155–60, 1952.

[4] Rosalind E Franklin and Raymond Gosling. Molecular configuration in sodium thymonucleate. *Nature*, 171:740–741, 1953.

[5] James D Watson and Francis H C Crick. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature*, 171:737738, 1953.

[6] Marshall W Nirenberg and J Heinrich Matthaei. The dependence of cell-free protein synthesis in e. coli upon naturally occurring or synthetic polyribonucleotides. *Proc. Natl. Acad. Sci. U. S. A.*, 47(10):15881602, 1961.

[7] Robert W Holley, Jean Apgar, George A Everett, James T Madison, Mark Marquisee, Susan H Merrill, John R Penswick, and Ada Zamir. Structure of a ribonucleic acid. *Science*, 147(3664):1462–1465, 1965.

[8] Francis Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–563, 1970.

[9] Stephen J Sharp, Jerone Schaack, Lyan Cooley, Debroh J Burke, and Dieter Sll. Structure and transcription of eukaryotic tRNA genes. *CRC Crit. Rev. Biochem.*, 19(2):107–144, 1985.

[10] Kelly Kruger, Paula J Grabowski, Arthur J Zaug, Julie Sands, Daniel E Gottschling, and Thomas R Cech. Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena. *Cell*, 31(1):147–157, 1982.

[11] Gail M Emilsson, Shingo Nakamura, Adam Roth, and Ronald R Breaker. Ribozyme speed limits. *RNA*, 9(8):907918, 2003.

[12] Ada E Yonath, Jutta Mssig, Bernd Tesche, Siegfried Lorenz, Volker A Erdmann, and Heinz G Wittmann. Crystallization of the large ribosomal subunits from *Bacillus stearothermophilus*. *Biochem. Int.*, 1(5):428–435, 1980.

[13] Ada E Yonath, Jutta Mssig, Bernd Tesche, Siegfried Lorenz, Volker A Erdmann, and Heinz G Wittmann. Several crystal forms of the *Bacillus stearothermophilus* 50 S ribosomal particles. *FEBS Lett.*, 154(1):15–20, 1983.

[14] Seth Stern, Bryn Weiser, and Harry F Noller. Model for the three-dimensional folding of 16 S ribosomal RNA. *J. Mol. Biol.*, 204(2):447–481, 1988.

[15] Joachim Frank, Jun Zhu, Pawel Penczek, Yanhong Li, Suman Srivastava, Adriana Verschoor, Michael Radermacher, Robert Grassucci, Ramani K Lata, and Rajendra K Agrawal. A model of protein synthesis based on cryo-electron microscopy of the *E. coli* ribosome. *Nature*, 376(6539):441–444, 1995.

[16] Nenad Ban, Poul Nissen, Jeffrey Hansen, Peter B Moore, and Thomas A Steitz. The complete atomic structure of the large ribosomal subunit at 2.4 resolution. *Science*, 289(5481):905–920, 2000.

[17] Frank Schluenzen, Ante Tocilj, Raz Zarivach, Joerg Harms, Marco Gluehmann, Daniela Janell, Anat Bashan, Heike Bartels, Ilana Agmon, Franois Franceschi, and Ada Yonath. Structure of functionally activated small ribosomal subunit at 3.3 angstroms resolution. *Cell*, 102(5):615–623, 2000.

[18] Brian T Wimberly, Ditlev E Brodersen, William M Clemons Jr, Robert J Morgan-Warren, Andrew P Carter, Clemens Vonrhein, Thomas Hartsch, and Venki Ramakrishnan. Structure of the 30S ribosomal subunit. *Nature*, 407(6802):327339, 2000.

[19] Andrew P Carter, William M Clemons, Ditlev E Brodersen, Robert J Morgan-Warren, Brian T Wimberly, and Venki Ramakrishnan. Functional insights from the structure of the 30S ribosomal subunit and its interactions with antibiotics. *Nature*, 407(6802):340–348, 2000.

[20] Marat M Yusupov, Gulnara Zh Yusupova, Albion Baucom, Kate Lieberman, Thomas N Earnest, Jamie H D Cate, and Harry F Noller. Crystal structure of the ribosome at 5.5 resolution. *Science*, 292(5518):883–896, 2001.

[21] James M Ogle, Andrew P Carter, and Venki Ramakrishnan. Insights into the decoding mechanism from recent ribosome structures. *Trends Biochem. Sci.*, 28(5):259–266, 2003.

[22] Andrei Korostelev, Dmitri N Ermolenko, and Harry F Noller. Structural dynamics of the ribosome. *Curr. Opin. Chem. Biol.*, 12(6):674–683, 2008.

[23] Adam Ben-Shem, Nicolas Garreau de Loubresse, Sergey Melnikov, Lasse Jenner, Gulnara Yusupova, and Marat Yusupov. The structure of the eukaryotic ribosome at 3.0 resolution. *Science*, 334(6062):1524–1529, 2011.

[24] Natalia Demeshkina, Lasse Jenner, Eric Westhof, Marat Yusupov, and Gulnara Yusupova. A new understanding of the decoding principle on the ribosome. *Nature*, 484(7393):256–259, 2012.

[25] Mitchell Guttman, Ido Amit, Manuel Garber, Courtney French, Michael F Lin, David Feldser, Maite Huarte, Or Zuk, Bryce W Carey, John P Cassady, Moran N Cabili, Rudolf Jaenisch, Tarjei S Mikkelsen, Tyler Jacks, Nir Hacohen, Bradley E Bernstein, Manolis Kellis, Aviv Regev, John L Rinn, and Eric S Lander. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*, 458(7235):223–227, 2009.

[26] Ross C Wilson and Jennifer A Doudna. Molecular mechanisms of RNA interference. *Annu. Rev. Biophys.*, 42:217–239, 2013.

[27] Sethuramasundaram Pitchiaya, Laurie A Heinicke, Thomas C Custer, and Nils G Walter. Single molecule fluorescence approaches shed light on intracellular RNAs. *Chem. Rev.*, 114(6):32243265, 2014.

[28] Thomas R Cech and Joan R Steitz. The noncoding RNA revolution-trashing old rules to forge new ones. *Cell*, 157(1):77–94, 2014.

[29] Adam M Schmitt and Howard Y Chang. Long noncoding RNAs in cancer pathways. *Cancer Cell*, 29(4):452–463, 2016.

[30] Maite Huarte. The emerging role of lncRNAs in cancer. *Nat. Med.*, 21:12531261, 2016.

[31] Alicia J Angelbello, Jonathan L Chen, and Matthew D Disney. Small molecule targeting of RNA structures in neurological disorders. *Ann. N. Y. Acad. Sci.*, 1471(1):57–71, 2019.

[32] Ageliki Laina, Aikaterini Gatsiou, Georgios Georgiopoulos, Kimon Stamatelopoulos, and Konstantinos Stellos. RNA therapeutics in cardiovascular precision medicine. *Front. Physiol.*, 9:205220, 2018.

[33] Christina E Lnse, Anna Schller, and Gnter Mayer. The promise of riboswitches as potential antibacterial drug targets. *Int. J. Med. Microbiol.*, 304(1):79–92, 2014.

[34] Eric O Freed. HIV-1 assembly, release and maturation. *Nat. Rev. Microbiol.*, 13(8):484–496, 2015.

[35] Thomas Hermann. *RNA Therapeutics. Topics in Medicinal Chemistry.* Springer, Cham, 2017.

[36] Jiri Sponer, Giovanni Bussi, Miroslav Krepl, Pavel Banas, Sandro Bottaro, Richard A Cunha, Alejandro Gil-Ley, Giovanni Pinamonti, Simn Poblete, Petr Jureka, Nils G Walter, and Michal Otyepka. RNA structural dynamics as captured by molecular simulations: a comprehensive overview. *Chem. Rev.*, 118(8):4177–4338, 2018.

[37] Neocles B Leontis and Eric Westhof. Analysis of RNA motifs. *Curr. Opin. Struct. Biol.*, 13(3):300–308, 2003.

[38] Philippe Brion and Eric Westhof. Hierarchy and dynamics of RNA folding. *Annu. Rev. Biophys. Biomol. Struct.*, 26:113–137, 1997.

[39] Ignacio Tinoco Jr and Carlos Bustamante. How RNA folds. *J. Mol. Biol.*, 293(2):271–281, 1999.

[40] Poul Nissen, Jeffrey Hansen, Nenad Ban, Peter B Moore, and Thomas A Steitz. The structural basis of ribosome activity in peptide bond synthesis. *Science*, 289(5481):920–930, 2000.

[41] Martha J Fedor and James R Williamson. The catalytic diversity of RNAs. *Nat. Rev. Mol. Cell Biol.*, 6(5):399–412, 2005.

[42] Jos A Cruz and Eric Westhof. The dynamic landscapes of RNA architecture. *Cell*, 136(4):604–609, 2009.

[43] Elizabeth A Dethoff, Jeetender Chugh, Anthony M Mustoe, and Hashim M Al-Hashimi. Functional complexity and regulation through RNA dynamics. *Nature*, 482(7385):322330, 2012.

[44] Hashim M Al-Hashimi and Nils G Walter. RNA dynamics: it is about time. *Curr. Opin. Struct. Biol.*, 18(3):321329, 2008.

[45] Anthony M Mustoe, Charles L Brooks III, and Hashim M Al-Hashimi. Hierarchy of RNA functional dynamics. *Annu. Rev. Biochem.*, 83:441466, 2014.

[46] Loic Salmon, Shan Yang, and Hashim M Al-Hashimi. Advances in the determination of nucleic acid conformational ensembles. *Annu. Rev. Phys. Chem.*, 65:293–316, 2014.

[47] Loic Salmon, George M Giambau, Evgenia N Nikolova, Katja Petzold, Akash Bhattacharya, David A Case, and Hashim M Al-Hashimi. Modulating RNA alignment using directional dynamic kinks: application in determining an atomic-resolution ensemble for a hairpin using NMR residual dipolar couplings. *J. Am. Chem. Soc.*, 137(40):1295412965, 2015.

[48] Laura R Ganser, Megan L Kelly, Daniel Herschlag, and Hashim M Al-Hashimi. The roles of structural dynamics in the cellular functions of RNAs. *Nat. Rev. Mol. Cell. Biol.*, 20(8):474–489, 2019.

[49] Aline U Juru, Neeraj N Patwardhan, and Amanda E Hargrove. Understanding the contributions of conformational changes, thermodynamics, and kinetics of RNAsmall molecule interactions. *ACS Chem. Biol.*, 14(5):824838, 2019.

[50] Thomas A Cooper, Lili Wan, and Gideon Dreyfuss. RNA and disease. *Cell*, 136(4):777–793, 2009.

[51] Manel Esteller. Non-coding RNAs in human disease. *Nat. Rev. Genet.*, 12(12):861–874, 2011.

[52] Ilyas Yildirim, Debayan Chakraborty, Matthew D Disney, David J Wales, and George C Schatz. Computational investigation of RNA CUG repeats responsible for Myotonic Dystrophy 1. *J. Chem. Theory Comput.*, 11(10):49434958, 2015.

[53] Coronaviridae Study Group of the International Committee on Taxonomy of Viruses. The species *Severe acute respiratory syndrome-related coronavirus*: classifying 2019-nCoV and naming it SARS-CoV-2. *Nat. Microbiol.*, 5:536–544, 2020.

[54] Palmiro Poltronieri, Binlian Sun, and Massimo Mallardoc. RNA viruses: RNA roles in pathogenesis, coreplication and viral load. *Curr Genomics*, 16(5):327335, 2015.

[55] John A Howe, Hao Wang, Thierry O Fischmann, Carl J Balibar, Li Xiao, Andrew M Galgoci, Juliana C Malinverni, Todd Mayhood, Artjohn Villafania, Ali Nahvi, Nicholas Murgolo, Christopher M Barbieri, Paul A Mann, Donna Carr, Ellen Xia, Paul Zuck, Dan Riley, Ronald E Painter, Scott S Walker, Brad Sherborne, Reynalda de Jesus, Weidong Pan, Michael A Plotkin, Jin Wu, Diane Rindgen, John Cummings, Charles G Garlisi, Rumin Zhang, Payal R Sheth, Charles J Gill, Haifeng Tang, and Terry Roemer. Selective small-molecule inhibition of an RNA structural element. *Nature*, 526(7575):672–677, 2015.

[56] Colleen M Connelly, Michelle H Moon, and John S Schneekloth Jr. The emerging role of RNA as a therapeutic target for small molecules. *Cell Chem. Biol.*, 23(9):1077–1090, 2016.

[57] Matthew D Disney. Targeting RNA with small molecules to capture opportunities at the intersection of chemistry, biology, and medicine. *J. Am. Chem. Soc.*, 141(17):6776–6790, 2019.

[58] Elizabeth A Dethoff, Katja Petzold, Jeetender Chugh, Anette Casiano-Negroni, and Hashim M Al-Hashimi. Visualizing transient low-populated structures of RNA. *Nature*, 491(7426):724–728, 2012.

[59] Fareed Aboul-ela, Jonathan Karn, and Gabriele Varani. The structure of the Human Immunodeficiency Virus Type-1 TAR RNA reveals principles of RNA recognition by Tat protein. *J. Mol. Biol.*, 253(2):313–332, 1995.

[60] Fareed Aboul-ela, Jonathan Karn, and Gabriele Varani. Structure of HIV-1 TAR RNA in the absence of ligands reveals a novel conformation of the trinucleotide bulge. *Nucleic Acids Res.*, 24(20):3974–3981, 1996.

[61] Janghyun Lee, Elizabeth A Dethoff, and Hashim M Al-Hashimi. Invisible RNA state dynamically couples distant motifs. *Proc. Natl. Acad. Sci. U.S.A.*, 111(26):9485–9490, 2014.

[62] Elizabeth A Dethoff, Alexandar L Hansen, Catherine Musselman, Eric D Watt, Ioan Andricioaei, and Hashim M Al-Hashimi. Characterizing complex dynamics in the transactivation response element apical loop and motional correlations with the bulge by NMR, molecular dynamics, and mutagenesis. *Biophys. J.*, 95(8):3906–3915, 2008.

[63] Rasmus Fonseca, Dimitar V Pachov, Julie Bernauer, and Henry van den Bedem. Characterizing RNA ensembles from NMR data with kinematic models. *Nucleic Acids Res.*, 42(15):9562–9572, 2014.

[64] Hashim M Al-Hashimi, Stephen W Pitt, Ananya Majumdar, Weijun Xu, and Dinshaw J Patel. Mg^{2+}-induced variations in the conformation and dynamics of HIV-1 TAR RNA probed using NMR residual dipolar couplings. *J. Mol. Biol.*, 329(5):867–873, 2003.

[65] Wei Huang, Gabriele Varani, and Gary P Drobny. $^{13}C/^{15}N$-^{19}F intermolecular REDOR NMR study of the interaction of TAR RNA with Tat peptides. *J. Am. Chem. Soc.*, 132:17643–17645, 2010.

[66] Qi Zhang, Andrew C Stelzer, Charles K Fisher, and Hashim M Al-Hashimi. Visualizing spatially correlated dynamics that directs RNA conformational transitions. *Nature*, 450:1263–1267, 2007.

[67] Nicole I Orlovsky, Hashim M Al-Hashimi, and Terrence G Oas. Exposing hidden high-affinity RNA conformational states. *J. Am. Chem. Soc.*, 142(2):907921, 2020.

[68] Laura R Ganser, Chia-Chieh Chu, Hal P Bogerd, Megan L Kelly, Bryan R Cullen, and Hashim M Al-Hashimi. Probing RNA conformational equilibria within the functional cellular context. *Cell Rep.*, 30:2472–2480, 2020.

[69] Anthony M Mustoe, Hashim M Al-Hashimi, and Charles L Brooks III. Coarse grained models reveal essential contributions of topological constraints to the conformational free energy of RNA bulges. *J. Phys. Chem. B*, 118(10):2615–2627, 2014.

[70] Nak-Kyoon Kim, Ayaluru Murali, and Victoria J DeRose. A distance ruler for RNA using EPR and site-directed spin labeling. *Chem. Biol.*, 11(7):939–948, 2004.

[71] Fiona A Riordan, Anamitra Bhattacharyya, Sean McAteer, and David M J Lilley. Kinking of RNA helices by bulged bases, and the structure of the human immunodeficiency virus transactivator response element. *J. Mol. Biol.*, 226(2):305–310, 1992.

[72] Loic Salmon, Gavin Bascom, Ioan Andricioaei, and Hashim M Al-Hashimi. A general method for constructing atomic-resolution RNA ensembles using NMR residual dipolar couplings: the basis for interhelical motions revealed. *J. Am. Chem. Soc.*, 135(14):5457–5466, 2013.

[73] Honglue Shi, Atul Rangadurai, Hala Abou Assi, Rohit Roy, David A Case, Daniel Herschlag, Joseph D Yesselman, and Hashim M Al-Hashimi. Rapid and accurate determination of atomistic RNA dynamic ensemble models using NMR and structure prediction. *Nat. Commun.*, 11:5531, 2020.

[74] Dawn K Merriman, Yi Xue, Shan Yang, Isaac J Kimsey, Anisha Shakya, Mary Clay, and Hashim M Al-Hashimi. Shortening the HIV1 TAR RNA bulge by a single nucleotide preserves motional modes over a broad range of time scales. *Biochemistry*, 55(32):4445–4456, 2016.

[75] Amy Davidson, Thomas C Leeper, Zafiria Athanassiou, Krystyna Patora-Komisarka, Jonathan Karn, John A Robinson, and Gabriele Varani. Simultaneous recognition of HIV-1 TAR RNA bulge and loop sequences by cyclic peptide mimics of Tat protein. *Proc. Natl. Acad. Sci. U.S.A.*, 106(29):11931–11936, 2009.

[76] Amy Davidson, Krystyna Patora-Komisarka, John A Robinson, and Gabriele Varani. Essential structural requirements for specific recognition of HIV TAR RNA by peptide mimetics of Tat protein. *Nucleic Acids Res.*, 39(1):248–256, 2011.

[77] Amy Davidson, Darren W Begley, Carmen Lau, and Gabriele Varani. A small-molecule probe induces a conformation in HIV TAR RNA capable of binding drug-like fragments. *J. Mol. Biol.*, 410(5):984–996, 2011.

[78] Aditi N Borkar, Michael F Bardaro Jr, Carlo Camilloni, Francesco A Aprile, Gabriele Varani, and Michele Vendruscolo. Structure of a low-population binding intermediate in protein-RNA recognition. *Proc. Natl. Acad. Sci. U.S.A.*, 113(26):7171–7176, 2016.

[79] Matthew D Shortridge, Paul T Wille, Alisha N Jones, Amy Davidson, Jasmina Bogdanovic, Eric Arts, Jonathan Karn, John A Robinson, and Gabriele Varani. An ultrahigh affinity ligand of HIV-1 TAR reveals the RNA structure recognized by P-TEFb. *Nucleic Acids Res.*, 47(3):1523–1531, 2019.

[80] Ben Davis, Mohammad Afshar, Gabriele Varani, Alastair I H Murchie, Jonathan Karn, Georg Lentzen, Martin Drysdale, Justin Bower, Andrew J Potter, Ian D Starkey, Terry M Swarbrick, and Fareed Aboul-ela. Rational design of inhibitors of HIV-1 TAR RNA through the stabilisation of electrostatic "hot spots". *J. Mol. Biol.*, 336(2):343–356, 2004.

[81] Cornelius Faber, Heinrich Sticht, Kristian Schweimer, and Paul Rösch. Structural rearrangements of HIV-1 Tat-responsive RNA upon binding of neomycin B. *J. Mol. Biol.*, 275(27):20660–20666, 2000.

[82] Alastair I H Murchie, Ben Davis, Catherine Isel, Mohammad Afshar, Martin J Drysdale, Justin Bower, Andrew J Potter, Ian D Starkey, Terry M Swarbrick, Shabana Mirza, Catherine D Prescott, Philippe Vaglio, Fareed Aboul-ela, and Jonathan Karn. Structure-based drug design targeting an inactive RNA conformation: exploiting the flexibility of HIV-1 TAR RNA. *J. Mol. Biol.*, 336(3):625–638, 2004.

[83] Zhihua Du, Kenneth E Lind, and Thomas L James. Structure of TAR RNA complexed with a Tat-TAR interaction nanomolar inhibitor that was identified by computational screening. *Chem. Biol.*, 9(6):707–712, 2002.

[84] Ivan A Belashov, David W Crawford, Chapin E Cavender, Peng Dai, Patrick C Beardslee, David H Mathews, Bradley L Pentelute, Brian R McNaughton, and Joseph E Wedekind. Structure of HIV TAR in complex with a Lab-Evolved RRM provides insight into duplex RNA recognition and synthesis of a constrained peptide that impairs transcription. *Nucleic Acids Res.*, 46(13):64016415, 2018.

[85] Sai S Chavali, Sachitanand M Mali, Jermaine L Jenkins, Rudi Fasan, and Joseph E Wedekind. Co-crystal structures of HIV TAR RNA bound to lab-evolved proteins show key roles for arginine relevant to the design of cyclic peptide TAR inhibitors. *J. Biol. Chem.*, 49(13):16470–16486, 2020.

[86] Vincent V Pham, Carolina Salguero, Shamsun N Khan, Jennifer L Meagher, W Clay Brown, Nicolas Humbert, Hugues de Rocquigny, Janet L Smith, and Victoria M D'Souza. HIV-1 Tat interactions with cellular 7SK and viral TAR RNAs identifies dual structural mimicry. *Nat. Commun.*, 9(1):4266, 2018.

[87] Ursula Schulze-Gahmen and James H Hurley. Structural mechanism for HIV-1 TAR loop recognition by Tat and the super elongation complex. *Proc. Natl. Acad. Sci. U. S. A.*, 115(51):12973–12978, 2018.

[88] Joseph A Ippolito and Thomas A Steitz. A 1.3-Å resolution crystal structure of the HIV-1 trans-activation response region RNA stem reveals a metal ion-dependent bulge conformation. *Proc. Natl. Acad. Sci. U.S.A.*, 95(17):9819–9824, 1998.

[89] Hashim M Al-Hashimi, Yuying Gosser, Andrey Gorin, Weidong Hu, Ananya Majumdar, and Dinshaw J Patel. Concerted motions in HIV-1 TAR RNA may allow access to bound state conformations: RNA dynamics from NMR residual dipolar couplings. *J. Mol. Biol.*, 315(2):95–102, 2002.

[90] Laura R Ganser, Janghyun Lee, Atul Rangadurai, Dawn K Merriman, Megan L Kelly, Aman D Kansal, Bharathwaj Sathyamoorthy, and Hashim M Al-Hashimi. High-performance virtual screening by targeting a high-resolution RNA dynamic ensemble. *Nat. Struct. Mol. Biol.*, 25:425–434, 2018.

[91] Laura R Ganser, Megan L Kelly, Neeraj N Patwardhan, Amanda E Hargrove, and Hashim M Al-Hashimi. Demonstration that small molecules can bind and stabilize low-abundance short-lived RNA excited conformational states. *J. Mol. Biol.*, 432(4):1297–1304, 2020.

[92] Jason Fernandes, Bhargavi Jayaraman, and Alan Frankel. The HIV-1 Rev response element. *RNA Biol.*, 9(1):611, 2012.

[93] Yuying Gosser, Thomas Hermann, Ananya Majumdar, Weidong Hu, Ronnie Frederick, Feng Jiang, Weijun Xu, and Dinshaw J Patel. Peptide-triggered conformational switch in HIV-1 RRE RNA complexes. *Nat. Struct. Mol. Biol.*, 8(2):146150, 2001.

[94] Kazuo Harada, Shelley S Martin, Ruoying Tan, and Alan Frankel. Molding a peptide into an RNA site by *in vivo* peptide evolution. *Proc. Natl. Acad. Sci. U. S. A.*, 94(22):11887–11892, 1997.

[95] Matthew D Daugherty, Bella Liu, and Alan Frankel. Structural basis for cooperative RNA binding and export complex assembly by HIV Rev. *Nat. Struct. Mol. Biol.*, 17(11):1337–1342, 2010.

[96] Michael A DiMattia, Norman R Watts, Stephen J Stahl, Christoph Rader, Paul T Wingfield, David I Stuart, Alasdair C Steven, and Jonathan M Grimes. Implications of the HIV-1 Rev dimer structure at 3.2 Å resolution for multimeric binding to the Rev response element. *Proc. Natl. Acad. Sci. U. S. A.*, 107(13):5810–5814, 2010.

[97] Matthew D Daugherty, David S Booth, Bhargavi Jayaraman, Yifan Cheng, and Alan Frankel. HIV Rev response element (RRE) directs assembly of the Rev homooligomer into discrete asymmetric complexes. *Proc. Natl. Acad. Sci. U. S. A.*, 107(28):12481–12486, 2010.

[98] Stephanie J K Pond, William K Ridgeway, Rae Robertson, Jun Wang, and David P Millar. HIV-1 Rev protein assembles on viral RNA one molecule at a time. *Proc. Natl. Acad. Sci. U. S. A.*, 106(5):1404–1408, 2009.

[99] Lauren A Michael, Jessica A Chenault, Billy R Miller III, Ann M Knolhoff, and Maria C Nagan. Water, shape recognition, salt bridges, and cationpi interactions differentiate peptide recognition of the HIV Rev-responsive element. *J. Mol. Biol.*, 392(3):774–786, 2009.

[100] Richard Stefl, Florian C Oberstrass, Jennifer L Hood, Muriel Jourdan, Michal Zimmermann, Lenka Skrisovska, Christophe Maris, Li Peng, Ctirad Hofr, Ronald B Emeson, and Frdric H-T Allain. The solution structure of the ADAR2 dsRBM-RNA complex reveals a sequence-specific readout of the minor groove. *Cell*, 143(2):225–237, 2010.

[101] Chunyang Cao, Yu Lin Jiang, James T Stivers, and Fenhong Song. Dynamic opening of DNA during the enzymatic search for a damaged base. *Nat. Struct. Mol. Biol.*, 11:1230–1236, 2004.

[102] Atul Rangadurai, Eric S Szymanski, Isaac Kimsey, Honglue Shi, and Hashim Al-Hashimi. Probing conformational transitions towards mutagenic WatsonCrick-like GT mismatches using off-resonance sugar carbon $R_{1\rho}$ relaxation dispersion. *J. Biomol. NMR*, 74:457471, 2020.

[103] Akram Alian, Tom T Lee, Sarah L Griner, Robert M Stroud, and Janet Finer-Moore. Structure of a TrmA-RNA complex: a consensus RNA fold contributes to substrate selectivity and catalysis in m^5U methyltransferases. *Proc. Natl. Acad. Sci. U.S.A.*, 105:6876–6881, 2008.

[104] Gregoire Altan-Bonnet, Albert Libchaber, and Oleg Krichevsky. Bubble dynamics in double-stranded DNA. *Phys. Rev. Lett.*, 90:138101, 2003.

[105] Xudong Chen, Yan Zhou, Peng Qu, and Xin S Zhao. Base-by-base dynamics in DNA hybridization probed by fluorescence correlation spectroscopy. *J. Am. Chem. Soc.*, 130:16947–16952, 2008.

[106] M. Ashley Spies and Richard L Schowen. The trapping of a spontaneously flipped-out base from double helical nucleic acids by host-guest complexation with β-cyclodextrin: the intrinsic base-flipping rate constant for DNA and RNA. *J. Am. Chem. Soc.*, 124:14049–14053, 2002.

[107] Francesco Colizzi, Cibran Perez-Gonzalez, Remi Fritzen, Yaakov Levy, Malcolm F White, J. Carlos Penedo, and Giovanni Bussi. Asymmetric base-pair opening drives helicase unwinding dynamics. *Proc. Natl. Acad. Sci. U.S.A.*, 116:22471–22477, 2019.

[108] Yandong Yin, Lijiang Yang, Guanqun Zheng, Chan Gu, Chengqi Yi, Chuan He, Yi Q Gao, and Xin S Zhao. Dynamics of spontaneous flipping of a mismatched base in DNA duplex. *Proc. Natl. Acad. Sci. U.S.A.*, 111(22):8043–8048, 2014.

[109] Maurice Gueron, Michel Kochoyan, and Jean-Louis Leroy. A single mode of DNA base-pair opening drives imino proton exchange. *Nature*, 328(6125):89–92, 1987.

[110] James G Moe and Irina M Russu. Kinetics and energetics of base-pair opening in 5'-d(CGCGAATTCGCG)-3' and a substituted dodecamer containing GT mismatches. *Biochemistry*, 31(36):8421–8428, 1992.

[111] U Deva Priyakumar and Alexander D MacKerell. Computational approaches for investigating base flipping in oligonucleotides. *Chem. Rev.*, 106(2):489–505, 2006.

[112] Peter Varnai, Muriel Canalia, and Jean-Louis Leroy. Opening mechanism of GT/U pairs in DNA and RNA duplexes: a combined study of imino proton exchange and molecular dynamics simulation. *J. Am. Chem. Soc.*, 126(44):14659–14667, 2004.

[113] Niu Huang, Nilesh K Banavali, and Alexander D MacKerell, Jr. Protein-facilitated base flipping in DNA by cytosine-5-methyltransferase. *Proc. Natl. Acad. Sci. U.S.A.*, 100(1):68–73, 2003.

[114] Michael F Hagan, Aaron R Dinner, David Chandler, and Arup K Chakraborty. Atomistic understanding of kinetic pathways for single base-pair binding and unbinding in DNA. *Proc. Natl. Acad. Sci. U.S.A.*, 100(24):13922–13927, 2003.

[115] Benjamin Bouvier and Helmut Grubmüller. A molecular dynamics study of slow base flipping in DNA using conformational flooding. *Biophys. J.*, 93(3):770–786, 2007.

[116] Lin-Tai Da and Jin Yu. Base-flipping dynamics from an intrahelical to an extrahelical state exerted by thymine DNA glycosylase during DNA repair process. *Nucleic Acids Res.*, 46(11):5410–5425, 2018.

[117] Addie Kingsland and Lutz Maibaum. DNA base pair mismatches induce structural changes and alter the free-energy landscape of base flip. *J. Phys. Chem. B*, 122(51):12251–12259, 2018.

[118] Kun Song, Arthur J Campbell, Christina Bergonzo, Carlos de los Santos, Arthur P Grollman, and Carlos Simmerling. An improved reaction coordinate for nucleic acid base flipping studies. *J. Chem. Theory Comput.*, 5(11):31053113, 2009.

[119] Haoquan Li, Anton V Endutkin, Christina Bergonzo, Lin Fu, Arthur Grollman, Dmitry O Zharkov, and Carlos Simmerling. DNA deformation-coupled recognition of 8-oxoguanine: conformational kinetic gating in human DNA glycosylase. *J. Am. Chem. Soc.*, 139(7):26822692, 2017.

[120] Baron Peters. Reaction coordinates and mechanistic hypothesis tests. *Annu. Rev. Phys. Chem.*, 67:669–690, 2016.

[121] Jan Norberg and Lennart Nilsson. Conformational free energy landscape of ApApA from molecular dynamics simulations. *J. Phys. Chem.*, 100(7):2550–2554, 1996.

[122] Elena Cubero, Edward C Sherer, F Javier Luque, Modesto Orozco, and Charles A Laughton. Observation of spontaneous base pair breathing events in the molecular dynamics simulation of a difluorotoluene-containing DNA oligonucleotide. *J. Am. Chem. Soc.*, 121(37):8653–8654, 1999.

[123] Reid F Brown, Casey T Andrews, and Adrian H Elcock. Stacking free energies of all DNA and RNA nucleoside pairs and dinucleoside-monophosphates computed using recently revised AMBER parameters and compared with experiment. *J. Chem. Theory Comput.*, 11(5):2315–2328, 2015.

[124] Katarina Hart, Boel Nystrm, Marie hman, and Lennart Nilsson. Molecular dynamics simulations and free energy calculations of base flipping in dsRNA. *RNA*, 11(4):609–618, 2005.

[125] Florian Hase and Martin Zacharias. Free energy analysis and mechanism of base pair stacking in nicked DNA. *Nucleic Acids Res.*, 44(15):7100–7108, 2016.

[126] Nikita A Kuznetsov, Christina Bergonzo, Arthur J Campbell, Haoquan Li, Grigory V Mechetin, Carlos de los Santos, Arthur P Grollman, Olga S Fedorova, Dmitry O Zharkov, and Carlos Simmerling. Active destabilization of base pairs by a DNA glycosylase wedge initiates damage recognition. *Nucleic Acids Res.*, 43(1):272281, 2015.

[127] James Byrnes, Kevin Hauser, Leah Norona, Edison Mejia, Carlos Simmerling, and Miguel Garcia-Diaz. Base flipping by MTERF1 can accommodate multiple conformations and occurs in a stepwise fashion. *J. Mol. Biol.*, 428(12):2542–2556, 2016.

[128] Petra Kuhrova, Vojtech Mlynsky, Marie Zgarbova, Miroslav Krepl, Giovanni Bussi, Robert B Best, Michal Otyepka, Jiri Sponer, and Pavel Banas. Improving the performance of the Amber RNA force field by tuning the hydrogen-bonding interactions. *J. Chem. Theory Comput.*, 15(5):3288–3305, 2019.

[129] Mitsunobu Nakamura, Tsukasa Okaue, Tadao Takada, and Kazushige Yamana. DNA-templated assembly of naphthalenediimide arrays. *Chem. Eur. J.*, 18(1):196–201, 2012.

[130] Eugen Stulz. Nanoarchitectonics with porphyrin functionalized DNA. *Acc. Chem. Res.*, 50(4):823–831, 2017.

[131] Martin Karplus and J Andrew McCammon. Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol.*, 9(9):646–652, 2002.

[132] J Andrew McCammon, Bruce R Gelin, and Martin Karplus. Dynamics of folded proteins. *Nature*, 267(5612):585–590, 1977.

[133] Axel T Brunger, Charles L Brooks III, and Martin Karplus. Active site dynamics of ribonuclease. *Proc. Natl. Acad. Sci. U. S. A.*, 82(24):84588462, 1985.

[134] Jana Khandogin and Charles L Brooks III. Linking folding with aggregation in Alzheimer's beta-amyloid peptides. *Proc. Natl. Acad. Sci. U. S. A.*, 104(43):16880–16885, 2007.

[135] Yinglong Miao, Ferran Feixas, Changsun Eun, and J Andrew McCammon. Accelerated molecular dynamics simulations of protein folding. *J. Comput. Chem.*, 36(20):1536–1549, 2015.

[136] Robert B Best, Wenwei Zheng, and Jeetain Mittal. Balanced proteinwater interactions improve properties of disordered proteins and non-specific protein association. *J. Chem. Theory Comput.*, 10(11):51135124, 2014.

[137] Starr C Harvey, M Prabhakaran, and J Andrew McCammon. Phenylalanine transfer RNA: molecular dynamics simulation. *Science*, 223(4641):1189–1191, 1984.

[138] Eric J Sorin, Mark A Engelhardt, Daniel Herschlag, and Vijay S Pande. RNA simulations: probing hairpin unfolding and the dynamics of a GNRA tetraloop. *J. Mol. Biol.*, 317(4):493–506, 2002.

[139] Michael V Schrodt, Casey T Andrews, and Adrian H Elcock. Large-scale analysis of 48 DNA and 48 RNA tetranucleotides studied by 1 μs explicit-solvent molecular dynamics simulations. *J. Chem. Theory Comput.*, 11(12):59065917, 2015.

[140] David A Case and Martin Karplus. Dynamics of ligand binding to heme proteins. *J. Mol. Biol.*, 132(3):343–368, 1979.

[141] Anthony J Clark, Pratyush Tiwary, Ken Borrelli, Shulu Feng, Edward B Miller, Robert Abel, Richard A Friesner, and B J Berne. Prediction of protein-ligand binding poses via a combination of induced fit docking and metadynamics simulations. *J. Chem. Theory Comput.*, 12(6):2990–2998, 2016.

[142] Morten Jensen, Vishwanath Jogini, David W Borhani, Abba E Leffler, Ron O Dror, and David E Shaw. Mechanism of voltage gating in potassium channels. *Science*, 336(6078):229–233, 2012.

[143] Marco De Vivo, Matteo Masetti, Giovanni Bottegoni, and Andrea Cavalli. Role of molecular dynamics and related methods in drug discovery. *J. Med. Chem.*, 59(9):40354061, 2016.

[144] Alex Dickson, Pratyush Tiwary, and Harish Vashisth. Kinetics of ligand binding through advanced computational approaches: a review. *Curr. Top. Med. Chem.*, 17(23):2626–2641, 2017.

[145] Daan Frenkel and Berend Smit. *Understanding Molecular Simulation.* Academic Press, Inc. 6277 Sea Harbor Drive Orlando, FL, United States, 2001.

[146] M Scott Shell. *Thermodynamics and Statistical Mechanics: an Integrated Approach.* Cambridge University Press, 2015.

[147] Richard J Loncharich, Bernard R Brooks, and Richard W Pastor. Langevin dynamics of peptides: the frictional dependence of isomerization rates of N-acetylalanyl-N′-methylamide. *Biopolymers*, 32(5):523–535, 1992.

[148] Young M Rhee and Vijay S Pande. Solvent viscosity dependence of the protein folding dynamics. *J. Phys. Chem. B*, 112(19):62216227, 2008.

[149] Jess A Izaguirre, Daniel P Catarello, Justin M Wozniak, and Robert D Skeel. Langevin stabilization of molecular dynamics. *J. Chem. Phys.*, 114:2090, 2000.

[150] Axel Brnger, Charles L Brooks III, and Martin Karplus. Stochastic boundary conditions for molecular dynamics simulations of ST2 water. *Chem. Phys. Lett*, 105(5):495–500, 1984.

[151] Tamar Schlick, Eric Barth, and Margaret Mandziuk. Biomolecular dynamics at long timesteps: bridging the timescale gap between simulation and experimentation. *Annu. Rev. Biophys. Biomol. Struct.*, 26:181222, 1997.

[152] James C Phillips, Rosemary Braun, Wei Wang, James Gumbart, Emad Tajkhorshid, Elizabeth Villa, Christophe Chipot, Robert D Skeel, Laxmikant Kal, and Klaus Schulten. Scalable molecular dynamics with NAMD. *J. Comput. Chem.*, 26(16):17811802, 2005.

[153] Jiri Sponer, Pavel Banas, Petr Jurecka, Marie Zgarbova, Petra Kuhrova, Marek Havrila, Miroslav Krepl, Petr Stadlbauer, and Michal Otyepka. Molecular dynamics simulations of nucleic acids. From tetranucleotides to the ribosome. *J. Phys. Chem. Lett.*, 5(10):1771–1782, 2014.

[154] David A Case, Ido Y Ben-Shalom, Scott R Brozell, David S Cerutti, Thomas E Cheatham III, Vinicius W D Cruzeiro, Tom A Darden, Robert E Duke, Delaram Ghoreishi, Mike K Gilson, Holger Gohlke, Andreas W Goetz, D'Artagnan Greene, Robert Harris, Nadine Homeyer, Saeed Izadi, Andriy Kovalenko, Tom Kurtzman, Taisung S Lee, Scott LeGrand, Pengfei Li, Charles Lin, Jian Liu, Tyler Luchko, Ray Luo, Daniel J Mermelstein, Kenneth M Merz, Yinglong Miao, Grald Monard, Crystal Nguyen, Hai Nguyen, Igor Omelyan, Alexey Onufriev, Feng Pan, Ruxi Qi, Daniel R Roe, Adrian Roitberg, Celeste Sagui, Stephan Schott-Verdugo, Jana Shen, Carlos L Simmerling, Jamie Smith, Romelia Salomon-Ferrer, Jason Swails, Ross C Walker, Junmei Wang, Haixin Wei, Romain M Wolf, Xiongwu Wu, Li Xiao, Darrin M York, and Peter A Kollman. *AMBER 2018, University of California, San Francisco.* 2018.

[155] Alberto Perez, Ivan Marchn, Daniel Svozil, Jiri Sponer, Thomas E Cheatham, III, Charles A Laughton, and Modesto Orozco. Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. *Biophys. J.*, 92(11):3817–3829, 2007.

[156] Marie Zgarbova, Michal Otyepka, Jiri Sponer, Arnost Mladek, Pavel Banas, Thomas E Cheatham, III, and Petr Jurecka. Refinement of the Cornell et al. nucleic acids force field based on reference quantum chemical calculations of glycosidic torsion profiles. *J. Chem. Theory Comput.*, 7(9):2886–2902, 2011.

[157] Asaminew H Aytenfisu, Aleksandar Spasic, Alan Grossfield, Harry A Stern, and David H Mathews. Revised RNA dihedral parameters for the Amber force field improve RNA molecular dynamics. *J. Chem. Theory Comput.*, 13:900915, 2017.

[158] Ivan Ivani, Pablo D. Dans, Agnes Noy, Alberto Perez, Ignacio Faustino, Adam Hospital, Jurgen Walther, Pau Andrio, Ramon Goni, Alexandra Balaceanu, Guillem Portella, Federica Battistini, Josep L Gelp, Carlos Gonzlez, Michele Vendruscolo, Charles A Laughton, Sarah A Harris, David A Case, and Modesto Orozco. Parmbsc1: a refined force-field for DNA simulations. *Nat. Methods.*, 13(1):55–58, 2016.

[159] William L Jorgensen, Jayaraman Chandrasekhar, Jeffry D Madura, Roger W Impey, and Michael L Klein. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, 79(2):926, 1983.

[160] Pengfei Li, Lin F Song, and Kenneth M Merz Jr. Systematic parameterization of monovalent ions employing the nonbonded model. *J. Chem. Theory Comput.*, 11:1645–1657, 2015.

[161] James Maier, Carmenza Martinez, Koushik Kasavajhala, Lauren Wickstrom, Kevin E Hauser, and Carlos Simmerling. ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. *J. Chem. Theory Comput.*, 11(8):3696–3713, 2015.

[162] Araz Jakalian, David B Jack, and Christopher I Bayly. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *J. Comput. Chem.*, 23(16):1623–1641, 2002.

[163] Junmei Wang, Romain M Wolf, James W Caldwell, Peter A Kollman, and David A Case. Development and testing of a general Amber force field. *J. Comp. Chem.*, 25(9):1157–1173, 2004.

[164] Wendy D Cornell, Piotr Cieplak, Christopher I Bayly, Ian R Gould, Kenneth M Merz, David M Ferguson, David C Spellmeyer, Thomas Fox, James W Caldwell, and Peter A Kollman. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.*, 117:51795197, 1995.

[165] Scott J Weiner, Peter A Kollman, David A Case, U Chandra Singh, Caterina Ghio, Guliano Alagona, Salvatore Profeta, and Paul Weiner. A new force field for molecular

mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.*, 106:765784, 1984.

[166] Thomas E Cheatham III, Piotr Cieplak, and Peter A Kollman. A modified version of the Cornell *et al.* force field with improved sugar pucker phases and helical repeat. *J. Biomol. Struct. Dyn.*, 16:845–862, 1999.

[167] Junmei Wang, Piotr Cieplak, and Peter A Kollman. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *J. Comput. Chem.*, 21:1049–1074, 2000.

[168] Rodrigo Galindo-Murillo, James C Robertson, Marie Zgarbova, Jiri Sponer, Michal Otyepka, Petr Jurecka, and Thomas E Cheatham III. Assessing the current state of amber force field modifications for DNA. *J. Chem. Theory Comput.*, 12:41144127, 2016.

[169] Pavel Banas, Daniel Hollas, Marie Zgarbova, Petr Jurecka, Modesto Orozco, Thomas E Cheatham III, Jiri Sponer, and Michal Otyepka. Performance of molecular mechanics force fields for RNA simulations: stability of UUCG and GNRA hairpins. *J. Chem. Theory Comput.*, 6:38363849, 2010.

[170] Miroslav Krepl, Marie Zgarbova, Petr Stadlbauer, Michal Otyepka, Pavel Banas, Jaroslav Koca, Thomas E Cheatham III, Petr Jurecka, and Jiri Sponer. Reference simulations of noncanonical nucleic acids with different χ variants of the AMBER force field: quadruplex DNA, quadruplex RNA, and Z-DNA. *J. Chem. Theory Comput.*, 8:25062520, 2012.

[171] Marie Zgarbova, F Javier Luque, Jiri Sponer, Thomas E Cheatham III, Michal Otyepka, and Petr Jurecka. Toward improved description of DNA backbone: revisiting epsilon and zeta torsion force field parameters. *J. Chem. Theory Comput.*, 9:23392354, 2013.

[172] Marie Zgarbova, Jiri Sponer, Michal Otyepka, Thomas E Cheatham III, Rodrigo Galindo-Murillo, and Petr Jurecka. Refinement of the sugarphosphate backbone torsion beta for AMBER force fields improves the description of Z- and B-DNA. *J. Chem. Theory Comput.*, 11:57235736, 2015.

[173] Petra Kuhrova, Vojtech Mlynsky, Marie Zgarbova, Miroslav Krepl, Giovanni Bussi, Robert B Best, Michal Otyepka, Jiri Sponer, and Pavel Banas. Improving the performance of the Amber RNA force field by tuning the hydrogen-bonding interactions. *J. Chem. Theory Comput.*, 15:32883305, 2019.

[174] Vojtech Mlynsky, Petra Kuhrova, Tomas Kuhr, Michal Otyepka, Giovanni Bussi, Pavel Banas, and Jiri Sponer. Fine-tuning of the AMBER RNA force field with a new term adjusting interactions of terminal nucleotides. *J. Chem. Theory Comput.*, 16:39363946, 2020.

[175] Dazhi Tan, Stefano Piana, Robert M Dirks, and David E Shaw. RNA force field with accuracy comparable to state-of-the-art protein force fields. *Proc. Natl. Acad. Sci. U. S. A.*, 115:E1346–E1355, 2018.

[176] Andrea Cesari, Sandro Bottaro, Kresten Lindorff-Larsen, Pavel Banas, Jiri Sponer, and Giovanni Bussi. Fitting corrections to an RNA force field using experimental data. *J. Chem. Theory Comput.*, 15:34253431, 2019.

[177] Michael J Robertson, Yue Qian, Matthew C Robinson, Julian Tirado-Rives, and William L Jorgensen. Development and testing of the OPLS-AA/M force field for RNA. *J. Chem. Theory Comput.*, 15:27342742, 2019.

[178] Wilfred F van Gunsteren, Philippe H Hnenberger, Alan E Mark, Paul E Smith, and Ilario G Tironi. Computer simulation of protein motion. *Comput. Phys. Commun.*, 91:305–319, 1995.

[179] Tamar Schlick. *Molecular modeling and simulation: an interdisciplinary guide.* Springer Science and Business Media, 2 edition, 2010.

[180] Shuichi Nose. A molecular dynamics method for simulations in the canonical ensemble. *Mol. Phys.*, 52(2):255–268, 1984.

[181] William Q Hoover, Anthony J C Ladd, and Bill Moran. High-strain-rate plastic flow studied via nonequilibrium molecular dynamics. *Phys. Rev. Lett*, 48(26):1818–1820, 1982.

[182] William Q Hoover. Canonical dynamics: equilibrium phase-space distributions. *Phys. Rev. A*, 31(3):1695–1697, 1985.

[183] Romelia Salomon-Ferrer, David A Case, and Ross C Walker. An overview of the Amber biomolecular simulation package. *WIREs Comput. Mol. Sci.*, 3:198–210, 2013.

[184] William Humphrey, Andrew Dalke, and Klaus Schulten. VMD: visual molecular dynamics. *J. Molec. Graphics*, 14(1):33–38, 1996.

[185] Daniel R Roe and Thomas E Cheatham III. Ptraj and cpptraj: software for processing and analysis of molecular dynamics trajectory data. *J. Chem. Theory Comput.*, 9(7):30843095, 2013.

[186] Peter Schmidtke, Axel Bidon-Chanal, F Javier Luque, and Xavier Barril. MDpocket: open-source cavity detection and characterization on molecular dynamics trajectories. *Bioinformatics*, 27(23):3276–3285, 2011.

[187] Yi I Yang, Qiang Shao, Jun Zhang, Lijiang Yang, and Yi Q Gao. Enhanced sampling in molecular dynamics. *J. Chem. Phys.*, 151:070902, 2019.

[188] Sergei Izrailev, Sergey Stepaniants, Barry Isralewitz, Dorina Kosztin, Hui Lu, Ferenc Molnar, Willy Wriggers, and Klaus Schulten. Computational molecular dynamics: challenges, methods. *Lect. Notes Comput. Sci. Eng.*, 4:3965, 1999.

[189] Phuc-Chau Do, Eric H Lee, and Ly Le. Steered molecular dynamics simulation in rational drug design. *J. Chem. Inf. Model.*, 58(8):14731482, 2018.

[190] Barry Isralewitz, Mu Gao, and Klaus Schulten. Steered molecular dynamics and mechanical functions of proteins. *Curr. Opin. Struct. Biol.*, 11(2):224–230, 2001.

[191] Christopher Jarzynski. Nonequilibrium equality for free energy differences. *Phys. Rev. Lett.*, 78(14):2690–2694, 1997.

[192] Morten Jensen, Sanghyun Park, Emad Tajkhorshid, and Klaus Schulten. Energetics of glycerol conduction through aquaglyceroporin GlpF. *Proc. Natl. Acad. Sci. USA*, 99(10):6731–6736, 2002.

[193] Sanghyun Park, Fatemeh Khalili-Araghi, Emad Tajkhorshid, and Klaus Schulten. Free energy calculation from steered molecular dynamics simulations using Jarzynskis equality. *J. Chem. Phys.*, 119(6):3559–3566, 2003.

[194] Sanghyun Park and Klaus Schulten. Calculating potentials of mean force from steered molecular dynamics simulations. *J. Chem. Phys.*, 120(13):5946–5961, 2004.

[195] Sanjib Paul, Nisanth N Nair, and Harish Vashisth. Phase space and collective variable based simulation methods for studies of rare events. *Mol. Simulat.*, 45:1273–1284, 2019.

[196] Peter G Bolhuis, David Chandler, Christoph Dellago, and Phillip L Geissler. Transition path sampling: throwing ropes over rough mountain passes, in the dark. *Annu. Rev. Phys. Chem.*, 53:291–318, 2002.

[197] Giovanni Bussi, Francesco L Gervasio, Alessandro Laio, and Michele Parrinello. Free-energy landscape for β hairpin folding from combined parallel tempering and metadynamics. *J. Am. Chem. Soc.*, 128(41):1343513441, 2006.

[198] Eric Darve and Andrew Pohorille. Calculating free energies using average force. *J. Chem. Phys.*, 115(20):9169, 2001.

[199] Wanli You, Zhiye Tang, and Chia-en A Chang. Potential mean force from umbrella sampling simulations: what can we learn and what is missed? *J. Chem. Theory Comput.*, 15(4):24332443, 2019.

[200] Shankar Kumar, Djamal Bouzida, Robert H Swendsen, Peter A Kollman, and John M Rosenberg. The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *J. Comput. Chem.*, 13(8):1011–1021, 1992.

[201] Douglas J Tobias and Charles L Brooks III. Calculation of free energy surfaces using the methods of thermodynamic perturbation theory. *Chem. Phys. Lett.*, 142(6):472–476, 1987.

[202] Xianjun Kong and Charles L Brooks III. λdynamics: a new approach to free energy calculations . *J. Chem. Phys.*, 105(6):2414, 1996.

[203] Baron Peters, Gregg T Beckham, and Bernhardt L Trout. Extensions to the likelihood maximization approach for finding reaction coordinates. *J. Chem. Phys.*, 127:034109, 2007.

[204] Baron Peters and Bernhardt L Trout. Obtaining reaction coordinates by likelihood maximization. *J. Chem. Phys.*, 125:054108, 2006.

[205] Christoph Dellago, Peter G Bolhuis, Felix S Csajka, and David Chandler. Transition path sampling and the calculation of rate constants. *J. Chem. Phys.*, 108:1964, 1997.

[206] Christoph Dellago, Peter G Bolhuis, Flix S Csajka, and David Chandler. Efficient transition path sampling: application to Lennard-Jones cluster rearrangements. *J. Chem. Phys.*, 108:9236–9245, 1998.

[207] Sanjib Paul, Tanmoy K Paul, and Srabani Taraphder. Orthogonal order parameters to model the reaction coordinate of an enzyme catalyzed reaction. *J. Mol. Graph. Model.*, 90:18–32, 2019.

[208] Sanjib Paul, Tanmoy K Paul, and Srabani Taraphder. Reaction coordinate, free energy, and rate of intramolecular proton transfer in human carbonic anhydrase II. *J. Phys. Chem. B*, 122:2851–2866, 2018.

[209] Peter G Bolhuis. Transition-path sampling of β-hairpin folding. *Proc. Natl. Acad. Sci. U.S.A.*, 100(21):12129–12134, 2003.

[210] Jarek Juraszek, Jocelyne Vreede, and Peter G Bolhuis. Transition path sampling of protein conformational changes. *Chem. Phys.*, 396(2):30–44, 2012.

[211] Joseph M Watts, Kristen K Dang, Robert J Gorelick, Christopher W Leonard, Julian W Bess Jr, Ronald Swanstrom, Christina L Burch, and Kevin M Weeks. Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature*, 460(7256):711–716, 2009.

[212] Jonathan L Chen, Damian M VanEtten, Matthew A Fountain, Ilyas Yildirim, and Matthew D Disney. Structure and dynamics of RNA repeat expansions that cause Huntington's disease and Myotonic Dystrophy Type 1. *Biochemistry*, 56(27):3463–3474, 2017.

[213] Katherine D Warner, Christine E Hajdin, and Kevin M Weeks. Principles for targeting RNA with drug-like small molecules. *Nat. Rev. Drug Discovery*, 17:547–558, 2018.

[214] Nicole M Bouvier and Peter Palese. The biology of Influenza viruses. *Vaccine*, 26:D49D53, 2008.

[215] Nicolas Leulliot and Gabriele Varani. Current topics in RNA-protein recognition: control of specificity and biological function through induced fit and conformational capture. *Biochemistry*, 40(27):7947–7956, 2001.

[216] James R Williamson. Induced fit in RNA-protein recognition. *Nat. Struct. Biol.*, 7:834–837, 2000.

[217] Matthew D Disney, Ilyas Yildirim, and Jessica L Childs-Disney. Methods to enable the design of bioactive small molecules targeting RNA. *Org. Biomol. Chem.*, 12:1029–1039, 2014.

[218] Tamar Schlick and Anna M Pyle. Opportunities and challenges in RNA structural modeling and design. *Biophys. J.*, 113(2):225–234, 2017.

[219] Hashim M Al-Hashimi. NMR studies of nucleic acid dynamics. *J. Magn. Reson.*, 237:191–204, 2013.

[220] Aaron T Frank, Andrew C Stelzer, Hashim M Al-Hashimi, and Ioan Andricioaei. Constructing RNA dynamical ensembles by combining MD and motionally decoupled NMR RDCs: new insights into RNA dynamics and adaptive ligand recognition. *Nucleic Acids Res.*, 37:3670–3679, 2009.

[221] Bo Zhao and Qi Zhang. Characterizing excited conformational states of RNA by NMR spectroscopy. *Curr. Opin. Struct. Biol.*, 30:134–146, 2015.

[222] Neeraj N Patwardhan, Laura R Ganser, Gary J Kapral, Christopher S Eubanks, Janghyun Lee, Bharathwaj Sathyamoorthy, Hashim M Al-Hashimi, and Amanda E Hargrove. Amiloride as a new RNA-binding scaffold with activity against HIV-1 TAR. *Med. Chem. Commun.*, 8:1022–1036, 2017.

[223] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In E. Simoudis, J. Han, and U. Fayyad, editors, *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pages 226–231. AAAI Press: Menlo Park, CA, 1996, 1996.

[224] Nilesh K Banavali and Alexander D MacKerell, Jr. Free energy and structural pathways of base flipping in a DNA GCGC containing sequence. *J. Mol. Biol.*, 319(1):141–160, 2002.

[225] Michael F Bardaro Jr, Zahra Shajani, Krystyna Patora-Komisarska, John A Robinson, and Gabriele Varani. How binding of small molecule and peptide ligands to HIV-1 TAR alters the RNA motional landscape. *Nucleic Acids Res.*, 37(5):1529–1540, 2009.

[226] Lev Levintov and Harish Vashisth. Ligand recognition in viral RNA necessitates rare conformational transitions. *J. Phys. Chem. Lett.*, 11:5426–5432, 2020.

[227] Lev Levintov and Harish Vashisth. Role of conformational heterogeneity in ligand recognition by viral RNA molecules. *Phys. Chem. Chem. Phys.*, 2021.

[228] Phillip A Sharp. The centrality of RNA. *Cell*, 136(4):577–580, 2009.

[229] Ageliki Laina, Aikaterini Gatsiou, Georgios Georgiopoulos, Kimon Stamatelopoulos, and Konstantinos Stellos. RNA therapeutics in cardiovascular precision medicine. *Front. Physiol.*, 123(2):205–220, 2018.

[230] Jeffrey S Kieft, Kaihong Zhou, Ronald Jubin, and Jennifer A Doudna. Mechanism of ribosome recruitment by hepatitis C IRES RNA. *RNA*, 7(2):194–206, 2001.

[231] Ewan P Plant and Jonathan D Dinman. The role of programmed-1 ribosomal frameshifting in coronavirus propagation. *Front. Biosci.*, 13:4873–4881, 2008.

[232] Neeraj N Patwardhan, Zhengguo Cai, Aline U Juru, and Amanda E Hargrove. Driving factors in amiloride recognition of HIV RNA targets. *Org. Biomol. Chem.*, 17:93139320, 2019.

[233] Qi Zhang, Xiaoyan Sun, Eric D Watt, and Hashim M Al-Hashimi. Resolving the motional modes that code for rna adaptation. *Science*, 311(5761):653–656, 2006.

[234] Ramona Ettig, Nick Kepper, Rene Stehr, Gero Wedemann, and Karsten Rippe. Dissecting DNA-histone interactions in the nucleosome by molecular dynamics simulations of DNA unwrapping. *Biophys. J.*, 101(8):1999–2008, 2011.

[235] Asmita Gupta and Manju Bansal. The role of sequence in altering the unfolding pathway of an RNA pseudoknot: a steered molecular dynamics study. *Phys. Chem. Chem. Phys.*, 18(41):28767–28780, 2016.

[236] Harish Vashisth and Cameron F Abrams. Ligand escape pathways and (un)binding free energy calculations for the hexameric insulin-phenol complex. *Biophys. J.*, 95(9):4193–4204, 2008.

[237] Anna M Capelli and Gabriele Costantino. Unbinding pathways of VEGFR2 inhibitors revealed by steered molecular dynamics. *J. Chem. Inf. Model.*, 54(11):3124–3136, 2014.

[238] Francesco Di Palma, Francesco Colizzi, and Giovanni Bussi. Ligand-induced stabilization of the aptamer terminal helix in the add adenine riboswitch. *RNA*, 19:1517–1524, 2013.

[239] Trang N Do, Paolo Carloni, Gabriele Varani, and Giovanni Bussi. RNA/peptide binding driven by electrostatics-insight from bidirectional pulling simulations. *J. Chem. Theory Comput.*, 9(3):1720–1730, 2013.

[240] Francesco Zonta, Damiano Buratto, Chiara Cassini, Mario Bortolozzi, and Fabio Mammano. Molecular dynamics simulations highlight structural and functional alterations in deafness-related M34T mutation of connexin 26. *Front. Physiol.*, 5:85, 2014.

[241] Yasutaka Nishihara and Akito Kitao. Gate-controlled proton diffusion and protonation-induced ratchet motion in the stator of the bacterial flagellar motor. *Proc. Natl. Acad. Sci. U.S.A.*, 112(25):7737–7742, 2015.

[242] Sylvie Bannwarth and Anne Gatignol. Hiv-1 tar rna: the target of molecular interactions between the virus and its host. *Curr. HIV Res.*, 3(1):61–71, 2005.

[243] Junmei Wang, Wei Wang, Peter A Kollman, and David A Case. Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graph. Model.*, 25(2):247–260, 2006.

[244] Junmei Wang, Romain M Wolf, James W Caldwell, Peter A Kollman, and David A Case. Development and testing of a general amber force field. *J. Comput. Chem.*, 25(9):1157–1174, 2004.

[245] Karim Snoussi and Jean L Leroy. Imino proton exchange and base-pair kinetics in RNA duplexes. *Biochemistry*, 40(30):8898–8904, 2001.

[246] Frederico C Freitas, Angelica N Lima, Vincius G Contessoto, Paul C Whitford, and Ronaldo J Oliveira. Drift-diffusion (DrDiff) framework determines kinetics and thermodynamics of two-state folding trajectory and tunes diffusion models. *J. Chem. Phys.*, 151:114106, 2019.

[247] Mariana Levi and Paul C Whitford. Dissecting the energetics of subunit rotation in the ribosome. *J. Phys. Chem. B*, 123(13):2812–2823, 2019.

[248] Yu Chen and Gabriele Varani. Protein families and RNA recognition. *FEBS J.*, 272(9):2088–2097, 2005.

[249] Donny D Licatalosi and Robert B Darnell. RNA processing and its regulation: global insights into biological networks. *Nat. Rev. Genet.*, 11:7587, 2010.

[250] Eric Holmqvist and Jörg Vogel. RNA-binding proteins in bacteria. *Nat. Rev. Microbiol.*, 16:601–615, 2018.

[251] Mitchell Guttman and John L Rinn. Modular regulatory principles of large non-coding RNAs. *Nature*, 482:339346, 2012.

[252] John L Rinn and Howard Y Chang. Genome regulation by long noncoding RNAs. *Annu. Rev. Biochem.*, 81:145–166, 2012.

[253] Rosario FranciscoVelilla, Embarc-Buh Azman, and Encarnacion MartinezSalas. Impact of RNAprotein interaction modes on translation control: the versatile multidomain protein Gemin5. *Bioessays*, 41:1800241, 2019.

[254] Paul Bieniasz and Alice Telesnitsky. Multiple, switchable protein:RNA interactions regulate human immunodeficiency virus type 1 assembly. *Annu. Rev. Virol.*, 5:165–183, 2018.

[255] Roland Ivanyi-Nagy, Igor Kanevsky, Caroline Gabus, Jean-Pierre Lavergne, Damien Ficheux, Franois Penin, Philippe Foss, and Jean-Luc Darlix. Analysis of hepatitis C virus RNA dimerization and core-RNA interactions. *Nucleic Acids Res.*, 34:2618–2633, 2006.

[256] Nora Schmidt, Caleb A Lareau, Hasmik Keshishian, Sabina Ganskih, Cornelius Schneider, Thomas Hennig, Randy Melanson, Simone Werner, Yuanjie Wei, Matthias Zimmer, Jens Ade, Luisa Kirschner, Sebastian Zielinski, Lars Dölken, Eric S Lander, Neva Caliskan, Utz Fischer, Jörg Vogel, Steven A Carr, Jochen Bodem, and Mathias Munschauer. The SARS-CoV-2 RNA-protein interactome in infected human cells. *Nat. Microbiol.*, 2020.

[257] Joseph D Puglisi, Lily Chen, Scott Blanchard, and Alan D Frankel. Solution structure of a bovine immunodeficiency virus Tat-TAR peptide-RNA complex. *Science*, 270(5239):1200–1203, 1995.

[258] Roberto Valverde, Laura Edwards, and Lynne Regan. Structure and function of KH domains. *FEBS J.*, 275:2712–2726, 2008.

[259] Dinshaw J Patel. Adaptive recognition in RNA complexes with peptides and protein modules. *Curr. Opin. Struct. Biol.*, 9:74–87, 1999.

[260] Yumin Dai, Jessica E Wynn, Ashley N Peralta, Chringma Sherpa, Bhargavi Jayaraman, Hao Li, Astha Verma, Alan D Frankel, Stuart F Le Grice, and Webster L Santos. Discovery of a branched peptide that recognizes the Rev response element (RRE) RNA and blocks HIV-1 replication. *J. Med. Chem.*, 61:96119620, 2018.

[261] Matthew J Walker and Gabriele Varani. Structure-based design of RNA-binding peptides. *Methods Enzymol.*, 623:339372, 2019.

[262] Alan D Frankel. Fitting peptides into the RNA world. *Curr. Opin. Struct. Biol.*, 10:332–340, 2000.

[263] Moran Frenkel-Pinter, Jay W Haynes, Ahmad M Mohyeldin, Martin C, Alyssa B Sargon, Anton S Petrov, Ramanarayanan Krishnamurthy, Nicholas V Hud, Loren D Williams, and Luke J Leman. Mutually stabilizing interactions between proto-peptides and RNA. *Nat. Commun.*, 11:3137, 2020.

[264] Meredith Corley, Margaret C Burns, and Gene W Yeo. How RNA-binding proteins interact with RNA: molecules and mechanisms. *Mol. Cell*, 78:9–29, 2020.

[265] Chandreyee Das and Alan D Frankel. Sequence and structure space of RNA-binding peptides. *Biopolymer*, 70:80–85, 2003.

[266] Tetsuya Iwazaki, Xianglan Li, and Kazuo Harada. Evolvability of the mode of peptide binding by an RNA. *RNA*, 11:13641373, 2005.

[267] Peter Hemmerich, Stefan Bosbach, Anna von Mikecz, and Ulrich Krawinkel. Human ribosomal protein L7 binds RNA with an alpha-helical arginine-rich and lysine-rich domain. *Eur. J. Biochem.*, 245:549–556, 1997.

[268] Katherine S Godin and Gabriele Varani. How arginine-rich domains coordinate mRNA maturation events. *RNA Biol.*, 4:69–75, 2007.

[269] Ray Truant and Bryan R Cullen. The arginine-rich domains present in human immunodeficiency virus type 1 Tat and Rev function as direct Importin β-dependent nuclear localization signals. *Mol. Cell. Biol.*, 19:12101217, 1999.

[270] Fabio Casu, Brendan M Duggan, and Mirko Hennig. The arginine-rich RNA-binding motif of HIV-1 Rev is intrinsically disordered and folds upon RRE binding. *Biophys. J.*, 105:1004–1017, 2013.

[271] Martin J Scanlon, David P Fairlie, D J Craik, D R Englebretsen, and M L West. NMR solution structure of the RNA-binding peptide from human immunodeficiency virus (type 1) Rev. *Biochemistry*, 34:8242–8249, 1995.

[272] John L Battiste, Hongyuan Mao, N Sambasiva Rao, Ruoying Tan, D R Muhandiram, Lewis E Kay, Alan D Frankel, and James Williamson. α helix-RNA major groove recognition in an HIV-1 rev peptide-RRE RNA complex. *Science*, 273:1547–1551, 1996.

[273] Pascale Legault, Joyce Li, Jeremy Mogridge, Lewis E Kay, and Jack Greenblatt. NMR structure of the bacteriophage λ N peptide/boxB RNA complex: recognition of a GNRA fold by an arginine-rich motif. *Cell*, 93:289–299, 1998.

[274] Matthew D Daugherty, Ivan DOrso, and Alan D Frankel. A solution to limited genomic capacity: using adaptable binding surfaces to assemble the functional HIV Rev oligomer on RNA. *Mol. Cell*, 31:824834, 2008.

[275] Ruoying Tan, Lily Chen, Joseph A Buettner, Derek Hudson, and Alan D Frankel. RNA recognition by an isolated alpha helix. *Cell*, 73:1031–1040, 1993.

[276] Travis S Bayer, Lauren N Booth, Scott M Knudsen, and Andrew D Ellington. Arginine-rich motifs present multiple interfaces for specific binding by RNA. *RNA*, 11:18481857, 2005.

[277] Roderico Acevedo, Declan Evans, Katheryn A Penrod, and Scott A Showalter. Binding by trbp-dsrbd2 does not induce bending of double-stranded rna. *Biophys. J.*, 110:2610–2617, 2016.

[278] Miroslav Krepl, Markus Blatter, Antoine Clery, Fred F Damberger, Frederic H T Allain, and Jiri Sponer. Structural study of the Fox-1 RRM protein hydration reveals a role for key water molecules in RRM-RNA recognition. *Nucleic Acids Res.*, 45:80468063, 2017.

[279] Malgorzata Figiel, Miroslav Krepl, Sangwoo Park, Jaroslaw Poznanski, Krzysztof Skowronek, Agnieszka Golab, Taekjip Ha, Jiri Sponer, and Marcin Nowotny. Mechanism of polypurine tract primer generation by HIV-1 reverse transcriptase. *J. Biol. Chem.*, 293:191–202, 2018.

[280] David E Draper. Themes in RNA-protein recognition. *J. Mol. Biol.*, 293:255–270, 1995.

[281] Katalin Nadassy, Shoshana J Wodak, and Joël Janin. Structural features of protein-nucleic acid recognition sites. *Biochemistry*, 38:19992017, 1999.

[282] Junichi Iwakiri, Hiroki Tateishi, Anirban Chakraborty, Prakash Patil, and Naoya Kenmochi. Dissecting the proteinRNA interface: the role of protein surface shapes and RNA secondary structures in proteinRNA recognition. *Nucleic Acids Res.*, 40:32993306, 2011.

[283] Dennis M Krüger, Saskia Neubacher, and Tom N Grossmann. ProteinRNA interactions: structural characteristics and hotspot amino acids. *RNA*, 24:14571465, 2018.

[284] Alexander D Mackerell Jr and Lennart Nilsson. Molecular dynamics simulations of nucleic acid-protein complexes. *Curr. Opin. Struct. Biol.*, 18:194–199, 2008.

[285] Carolina M Reyes and Peter A Kollman. Molecular dynamics studies of U1A-RNA complexes. *RNA*, 5:235244, 1999.

[286] Ikuo Kurisaki, Masayoshi Takayanagi, and Masataka Nagaoka. Combined mechanism of conformational selection and induced fit in U1ARNA molecular recognition. *Biochemistry*, 53:36463657, 2014.

[287] Jian-xin Guo and William H Gmeiner. Molecular dynamics simulation of the human U2B' protein complex with U2 snRNA hairpin IV in aqueous solution. *Biophys. J.*, 81:630642, 2001.

[288] Nathan Schmid, Bojan Zagrovic, and Wilfred F van Gunsteren. Mechanism and thermodynamics of binding of the polypyrimidine tract binding protein to RNA. *Biochemistry*, 46:65006512, 2007.

[289] Miroslav Krepl, Antoine Clery, Markus Blatter, Frederic H-T Allain, and Jiri Sponer. Synergy between NMR measurements and MD simulations of protein/RNA complexes: application to the RRMs, the most common RNA recognition motifs. *Nucleic Acids Res.*, 44:6452–6470, 2016.

[290] Nana D dit Konte, Miroslav Krepl, Fred F Damberger, Nina Ripin, Olivier Duss, Jiri Sponer, and Frederic H-T Allain. Aromatic side-chain conformational switch on the surface of the RNA recognition motif enables RNA discrimination. *Nat. Commun.*, 8:e654, 2017.

[291] Carolina M Reyes and Peter A Kollman. Structure and thermodynamics of RNA-protein binding: using molecular dynamics and free energy analyses to calculate the free energies of binding and conformational change. *J. Mol. Biol.*, 297:1145–1158, 2000.

[292] Dukagjin M Blakaj, Kevin J McConnell, David L Beveridge, and Anne M Baranger. Molecular dynamics and thermodynamics of protein-RNA interactions: mutation of a conserved aromatic residue modifies stacking interactions and structural adaptation in the U1A-stem loop 2 RNA complex. *J. Am. Chem. Soc.*, 123:25482551, 2001.

[293] Bethany L Kormos, Yulia Benitex, Anne M Beveridge, and David L Baranger. Affinity and specificity of protein U1A-RNA complex formation based on an additive component free energy model. *J. Mol. Biol.*, 371:1405–1419, 2007.

[294] Tiziana Castrignano, Giovanni Chillemi, Gabriele Varani, and Alessandro Desideri. Molecular dynamics simulation of the RNA complex of a double-stranded RNA-binding domain reveals dynamic features of the intermolecular interface and its hydration. *Nat. Commun.*, 8:e654, 2017.

[295] Zhen Xia, Zhihong Zhu, Jun Zhu, and Ruhong Zhou. Recognition mechanism of siRNA by viral p19 suppressor of RNA silencing: a molecular dynamics study. *Biophys. J.*, 96:17611769, 2009.

[296] Junru Yang, Jianing Song, John Z H Zhang, and Changge Ji. Effect of mismatch on binding of ADAR2/GluR-2 pre-mRNA complex. *J. Mol. Model.*, 21:222, 2015.

[297] Xinlei Wang, Lela Vukovic, Hye R Koh, Klaus Schulten, and Sua Myong. Dynamic profiling of double-stranded RNA binding proteins. *Nucleic Acids Res.*, 43:75667576, 2015.

[298] Salvador I Drusin, Irina P Suarez, Diego F Gauto, Rodolfo M Rasia, and Diego M Moreno. dsRNA-protein interactions studied by molecular dynamics techniques. Unravelling dsRNA recognition by DCL1. *Arch. Biochem. Biophys.*, 596:118–125, 2016.

[299] Qiao Xue, Qing-Chuan Zheng, Ji-Long Zhang, Ying-Lu Cui, and Hong-Xing Zhang. Exploring the mechanism how Marburg virus VP35 recognizes and binds dsRNA by molecular dynamics simulations and free energy calculations. *Biopolymers*, 101:849–860, 2014.

[300] S Harikrishna and P I Pradeepkumar. Probing the binding interactions between chemically modified siRNAs and human Argonaute 2 using microsecond molecular dynamics simulations. *J. Chem. Inf. Model.*, 57:883896, 2017.

[301] Kamila Reblova, Nada Spackova, Jaroslav Koca, Neocles B Leontis, and Jiri Sponer. Long-residency hydration, cation binding, and dynamics of loop E/helix IV rRNA-L25 protein complex. *Biophys. J.*, 87:33973412, 2002.

[302] Thomas Crety and Therese Malliavin. The conformational landscape of the ribosomal protein S15 and its influence on the protein interaction with 16S RNA. *Biophys. J.*, 92:2647–2465, 2007.

[303] Ke Chen, John Eargle, Krishnarjun Sarkar, Martin Gruebele, and Zaida Luthey-Schulten. Functional role of ribosomal signatures. *Biophys. J.*, 99:39303940, 2010.

[304] Miroslav Krepl, Kamila Reblova, Jaroslav Koca, and Jiri Sponer. Bioinformatics and molecular dynamics simulation study of L1 stalk non-canonical rRNA elements: kink-turns, loops, and tetraloops. *J. Phys. Chem. B*, 117:55405555, 2013.

[305] Satoshi Yamasaki, Shugo Nakamura, Tohru Terada, and Kentaro Shimizu. Mechanism of the difference in the binding affinity of E. coli tRNAGln to glutaminyl-tRNA synthetase caused by noninterface nucleotides in variable loop. *Biophys. J.*, 92:192–200, 2007.

[306] Amit Ghosh and Saraswathi Vishveshwara. A study of communication pathways in methionyl- tRNA synthetase by molecular dynamics simulations and structure network analysis. *Proc. Natl. Acad. Sci. U. S. A.*, 104:15711–15716, 2007.

[307] Anurag Sethi, John Eargle, Alexis A Black, and Zaida Luthey-Schulten. Dynamical networks in tRNA:protein complexes. *Proc. Natl. Acad. Sci. U. S. A.*, 106:6620–6625, 2009.

[308] Moitrayee Bhattacharyya, Amit Ghosh, Priti Hansia, and Saraswathi Vishveshwara. Allostery and conformational free energy changes in human tryptophanyl-tRNA synthetase from essential dynamics and structure networks. *Proteins*, 78:506–517, 2010.

[309] Amit Ghosh, Reiko Sakaguchi, Cuiping Liu, Saraswathi Vishveshwara, and Ya-Ming Hou. Allosteric communication in cysteinyl tRNA synthetase: a network of direct and indirect readout. *J. Biol. Chem.*, 286:3772137731, 2011.

[310] Eric A C Bushnell, WenJuan Huang, Jorge Llano, and James W Gauld. Molecular dynamics investigation into substrate binding and identity of the catalytic base in the mechanism of Threonyl-tRNA synthetase. *J. Phys. Chem. B*, 116:5205–5212, 2012.

[311] Rongzhong Li, Lindsay M Macnamara, Jessica D Leuchter, Rebecca W Alexander, and Samuel S Cho. MD simulations of tRNA and aminoacyl-tRNA synthetases: dynamics, folding, binding, and allostery. *Int. J. Mol. Sci.*, 16:1587215902, 2015.

[312] Carolina Estarellas, Michal Otyepka, Jaroslav Koca, Pavel Banas, Miroslav Krepl, and Jiri Sponer. Molecular dynamic simulations of protein/RNA complexes: CRISPR/Csy4 endoribonuclease. *Biochim. Biophys. Acta.*, 1850:1072–1090, 2015.

[313] Malgorzata Figiel, Miroslav Krepl, Jaroslaw Poznanski, Agnieszka Golab, Jiri Sponer, and Marcin Nowotny. Coordination between the polymerase and RNase H activity of HIV-1 reverse transcriptase. *Nucleic Acids Res.*, 45:3341–3352, 2017.

[314] Giulia Palermo, Yinglong Miao, Ross C Walker, Martin Jinek, and J Andrew McCammon. Striking plasticity of CRISPR-Cas9 and key role of non-target DNA, as revealed by molecular simulations. *ACS Cent. Sci.*, 2:756763, 2016.

[315] Giulia Palermo, Yinglong Miao, Ross C Walker, Martin Jinek, and J Andrew McCammon. CRISPR-Cas9 conformational activation as elucidated from enhanced molecular simulations. *Proc. Natl. Acad. Sci. U. S. A.*, 114:7260–7265, 2017.

[316] Miroslav Krepl, Marek Havrila, Petr Stadlbauer, Pavel Banas, Michal Otyepka, J Pasulka, Richard Stefl, and Jiri Sponer. Can we execute stable microsecond-scale atomistic simulations of proteinRNA complexes? *J. Chem. Theory Comput.*, 11:12201243, 2015.

[317] Riccardo Nifos, Carolina M Reyes, and Peter A Kollman. Molecular dynamics studies of the HIV-1 TAR and its complex with argininamide. *Nucleic Acids Res.*, 28:49444955, 2000.

[318] Carolina M Reyes, Riccardo Nifos, Alan D Frankel, and Peter A Kollman. Molecular dynamics and binding specificity analysis of the bovine immunodeficiency virus BIV Tat-TAR complex. *Biophys. J.*, 80:28332842, 2001.

[319] Yuguang Mu and Gerhard Stock. Conformational dynamics of RNA-peptide binding: a molecular dynamics simulation study. *Biophys. J.*, 90:391–399, 2006.

[320] Mattia Mori, Ursula Dietrich, Fabrizio Manetti, and Maurizio Botta. Molecular dynamics and DFT study on HIV-1 nucleocapsid protein-7 in complex with viral genome. *J. Chem. Inf. Model.*, 50:638650, 2010.

[321] Trang N Do, Emiliano Ippoliti, Paolo Carloni, Gabriele Varani, and Michele Parrinello. Counterion redistribution upon binding of a Tat-protein mimic to HIV-1 TAR RNA. *. Chem. Theory Comput.*, 8:688694, 2012.

[322] Chun H Li, Zhi C Zuo, Ji G Su, Xian J Xu, and Cun X Wang. The interactions and recognition of cyclic peptide mimetics of Tat with HIV-1 TAR RNA: a molecular dynamics simulation study. *J. Biomol. Struct. Dyn.*, 31:276287, 2013.

[323] Kazuo Harada, Shelley S Martin, Ruoying Tan, and Alan D Frankel. Molding a peptide into an RNA site by *in vivo* peptide evolution. *Proc. Natl. Acad. Sci. U. S. A.*, 94:11887–11892, 1997.

[324] Sigrid Nachtergaele and Chuan He. The emerging biology of RNA post-transcriptional modifications. *RNA Biol.*, 14:156–163, 2017.

[325] Brenda L Bass. RNA editing by adenosine deaminases that act on RNA. *Annu. Rev. Biochem.*, 71:817–846, 2002.

[326] Kazuko Nishikura. Editor meets silencer: crosstalk between RNA editing and RNA interference. *Nat. Rev. Mol. Cell Biol.*, 7:919–931, 2006.

[327] Richard J Roberts and Xiaodong Cheng. Base flipping. *Annu. Rev. Biochem.*, 67:181–198, 1998.

[328] Cai-Guang Yang, Chengqi Yi, Erica M Duguid, Christopher T Sullivan, Xing Jian, Phoebe A Rice, and Chuan He. Crystal structures of DNA/RNA repair enzymes AlkB and ABH2 bound to dsDNA. *Nature*, 452:961–965, 2008.

[329] Maria Spies and Brian O Smith. Protein-nucleic acids interactions: new ways of connecting structure, dynamics and function. *Biophys. Rev.*, 9:289–291, 2017.

[330] Jason L J Lin, Chyuan-Chuan Wu, Wei-Zen Yang, and Hanna S Yuan. Crystal structure of endonuclease G in complex with DNA reveals how it nonspecifically degrades DNA as a homodimer. *Nucleic Acids Res.*, 44:10480–10490, 2016.

[331] Samuel Hong and Xiaodong Cheng. DNA base flipping: a general mechanism for writing, reading, and erasing DNA modifications. *Adv. Exp. Med. Biol.*, 945:321341, 2016.

[332] Paul C Blainey, Antoine M van Oijen, Anirban Banerjee, Gregory L Verdine, and X Sunney Xie. A base-excision DNA-repair protein finds intrahelical lesion bases by fast sliding in contact with DNA. *Proc. Natl. Acad. Sci. U.S.A.*, 103:5752–5757, 2006.

[333] Yuzong Z Chen, V Mohan, and Richard H Griffey. Spontaneous base flipping in DNA and its possible role in methyltransferase binding. *Phys. Rev. E Stat. Phys. Plasmas Fluids Relat. Interdiscip. Topics*, 62:1133–1137, 2000.

[334] Christoph Dellago, Peter G Bolhuis, and Phillip L Geissler. Transition path sampling. *Adv. Chem. Phys.*, 123:1–78, 2002.

[335] Rose Du, Vijay S Pande, Alexander Y Grosberg, Toyoichi Tanaka, and Eugene S Shakhnovich. On the transition coordinate for protein folding. *J. Chem. Phys.*, 108:334–350, 1998.

[336] Sanjib Paul and Srabani Taraphder. Determination of the reaction coordinate for a key conformational fluctuation in human carbonic anhydrase II. *J. Phys. Chem. B*, 119:11403–11415, 2015.

[337] Ravi Radhakrishnan and Tamar Schlick. Orchestration of cooperative events in DNA synthesis and repair mechanism unraveled by transition path sampling of DNA polymerase β's closing. *Proc. Natl. Acad. Sci. U.S.A.*, 101:59705975, 2004.

[338] Li Xi, Manas Shah, and Bernhardt L Trout. Hopping of water in a glassy polymer studied via transition path sampling and likelihood maximization. *J. Phys. Chem. B*, 117:3634–3647, 2013.

[339] Sara L Quaytman and Steven D Schwartz. Reaction coordinate of an enzymatic reaction revealed by transition path sampling. *Proc. Natl. Acad. Sci. U.S.A.*, 104:12253–12258, 2007.

[340] Brandon C Knott, Valeria Molinero, Michael F Doherty, and Baron Peters. Homogeneous nucleation of methane hydrates: unrealistic under realistic conditions. *J. Am. Chem. Soc.*, 134:19544–19547, 2012.

[341] Laura Lupi, Arpa Hudait, Baron Peters, Michael Grünwald, Ryan G Mullen, Andrew H Nguyen, and Valeria Molinero. Role of stacking disorder in ice nucleation. *Nature*, 551:218–222, 2017.

[342] Christian Leitold, Christopher J Mundy, Marcel D Baer, Gregory K Schenter, and Baron Peters. Solvent reaction coordinate for an S_N2 reaction. *J. Chem. Phys.*, 153:024103, 2020.

[343] Lev Levintov, Sanjib Paul, and Harish Vashisth. Reaction coordinate and thermodynamics of base flipping in RNA. *J. Chem. Theory Comput.*, 17:1914–1921, 2021.

[344] Shambhavi Tannir, Lev Levintov, Mark A Townley, Brian M Leonard, Jan Kubelka, Harish Vashisth, Krisztina Varga, and Milan Balaz. Functional nanoassemblies with mirror-image chiroptical properties templated by a single homochiral DNA strand. *Chem. Mater.*, 32(6):22722281, 2020.

[345] Sunhwan Jo, Taehoon Kim, Vidyashankara G Iyer, and Wonpil Im. CHARMM-GUI: a web-based graphical user interface for CHARMM. *J. Comput. Chem.*, 29(11):1859–1865, 2008.

[346] Seonghoon Kim, Jumin Lee, Sunhwan Jo, Charles L Brooks III, Hui S Lee, and Wonpil Im. CHARMM-GUI ligand reader and modeler for CHARMM force field generation of small molecules. *J. Comput. Chem.*, 38(21):1879–1886, 2017.

[347] Alex Dickson and Samuel D Lotz. Ligand release pathways obtained with WExplore: residence times and mechanisms. *J. Phys. Chem. B*, 120(24):53775385, 2016.

[348] Eugene Wigner. The transition state method. *J. Chem. Soc. Faraday Trans.*, 34:29–41, 1938.

[349] Henry Eyring. The activated complex in chemical reactions. *J. Chem. Phys.*, 3:107–115, 1935.

[350] Baron Peters. Recent advances in transition path sampling: accurate reaction coordinates, likelihood maximisation and diffusive barrier-crossing dynamics. *Mol. Simul.*, 36:1265–1281, 2010.

APPENDIX A
SUPPORTING INFORMATION FOR CHAPTER 3

A.1 Principal Component Analysis (PCA)

I conducted a principal component analysis (PCA) to reveal the dominant modes of motion in each unliganded and liganded simulation and to compare PC projections between different simulations. PCA can be used as a metric of similarity between the dominant modes of motion sampled across various unliganded and liganded simulations. Figure A.17 shows the overlap of histograms of the first PC projections for the unliganded and liganded simulations. In the unliganded simulations, I observe a significant overlap between systems with PDB codes 1ARJ, 1QD3, 1UUD, 1UUI, 2KX5, 2L8H, and 5J0M signifying that these systems exhibited similar motions in their RNA structures (Figure A.17A). I also observed overlap of the first PC in the systems with PDB codes 1ANR, 1LVJ, 2KDQ, 5J2W, and 6D2U (Figure A.17A). One system exhibited a completely different shape of the first PC which is coupled with increased mobility of all three bulge bases (1UTS; Figures 3.7, A.14, and A.17A). In the liganded simulations, I observed a significant overlap among a larger fraction of systems in comparison to unliganded simulations, specifically, systems with PDB codes 1ARJ, 1QD3, 1UTS, 1UUI, 2L8H, 5J0M, 5J1O, 5J2W, and 6D2U (Figure A.17B). Two systems (PDB codes 2KDQ and 2KX5) formed a second group of systems which exhibited an overlap in the projections of the first PC. Overall, this data shows that there are similarities in motion among various unliganded and liganded simulations but liganded systems have more overlaps across the projections of the first PC in comparison to the unliganded simulations. This means that most of the liganded RNA systems exhibit the same type of motions while interacting with the ligands.

Table A.1: **Details on ligands studied.** For each ligand, shown is the PDB code, ligand name, chemical formula, number of atoms, molecular weight, and the dissociation constant (K_D). See also Figure A.1.

PDB ID	Name	Chemical Formula	Number of atoms	Molecular weight	K_D (μM)
1ARJ	Arginine amide	$C_6H_{15}N_4O_2$	27	175	1000
1LVJ	Acetylpromazine	$C_{19}H_{22}N_2OS$	45	326	0.1
1QD3	Neomycin B	$C_{23}H_{46}N_6O_{13}$	88	614	5.9
1UTS	RBT550	$C_{24}H_{33}N_5O$	63	407	0.039
1UUD	RBT203	$C_{16}H_{31}N_7O_2$	56	353	1.54
1UUI	RBT158	$C_{16}H_{29}N_5O_2$	52	323	>50
2L8H	MV2003	$C_{17}H_{25}N_5O_2$	49	331	NA
2KDQ 5J0M 5J1O 5J2W	L-22	$C_{76}H_{145}N_{33}O_{15}$	269	1759	0.03
2KX5	KP-Z-41	$C_{94}H_{179}N_{41}O_{19}S_2$	335	2249	0.001
6D2U	JB181	$C_{72}H_{140}N_{31}O_{15}$	258	1678	<0.00018

Table A.2: **Details on all simulation systems.** For each TAR structure (see PDB codes), listed are RNA sequences in the original structure (column labeled *RNA Before*) and in the modeled structure (column labeled *RNA After*). The last column lists the total number of atoms in solvated and ionized unliganded as well as liganded (marked in bold) systems except for the PDB code 1ANR (first row) which is the experimental *apo* structure. See also section 3.4.1.

	PDB ID	RNA Before	RNA After	System Size
apo	1ANR	930 GGCAGAUCUGAGCCUGGGAGCUCUCUGCC	931 GGCAGAUCUGAGCCUGGGAGCUCUCUGCC	22959
liganded/unliganded	5J2W	929 GGCAGAUCUGAGCCUGGGAGCUCUCUGCC	931 GGCAGAUCUGAGCCUGGGAGCUCUCUGCC	20586 **20542**
	1UUI	930 GGCAGAUCUGAGCCUGGGAGCUCUCUGCC	931 GGCAGAUCUGAGCCUGGGAGCUCUCUGCC	23076 **23078**
	5J1O	929 GGCAGAUCUGAGCCUGGGAGCUCUCUGCC	931 GGCAGAUCUGAGCCUGGGAGCUCUCUGCC	21159 **21127**
	6D2U	930 GGCAGAUCUGAGCCUGGGAGCUCUCUGCC	931 GGCAGAUCUGAGCCUGGGAGCUCUCUGCC	21742 **21761**
	1UUD	929 GGCAGAUCUGAGCCUGGGAGCUCUCUGCC	931 GGCAGAUCUGAGCCUGGGAGCUCUCUGCC	24633 **24645**
	1UTS	930 GGCAGAUCUGAGCCUGGGAGCUCUCUGCC	931 GGCAGAUCUGAGCCUGGGAGCUCUCUGCC	24165 **24159**
	2KX5	930 GGCAGAUCUGAGCCUGGGAGCUCUCUGCC	931 GGCAGAUCUGAGCCUGGGAGCUCUCUGCC	23475 **25647**
	1QD3	927 G**C**CAGAU**U**UGAGCCUGGGAGCUCUCUG**G**C	931 GGCAGAUCUGAGCCUGGGAGCUCUCUGCC	19443 **19483**
	5J0M	929 GGCAGAUCUGAGCCUGGGAGCUCUCUGCC	931 GGCAGAUCUGAGCCUGGGAGCUCUCUGCC	22446 **22444**
	2KDQ	930 GGCAGAUCUGAGCCUGGGAGCUCUCUGCC	931 GGCAGAUCUGAGCCUGGGAGCUCUCUGCC	24471 **24529**
	1ARJ	930 GGCAGAUCUGAGCCUGGGAGCUCUCUGCC	931 GGCAGAUCUGAGCCUGGGAGCUCUCUGCC	22845 **22865**
	2L8H	930 GGCAGAUCUGAGCCUGGGAGCUCUCUGCC	931 GGCAGAUCUGAGCCUGGGAGCUCUCUGCC	20691 **20714**
	1LVJ	999 GGC**C**AGAUCUGAGCCUGGGAGCUCUCU**G**GCC	931 GGCAGAUCUGAGCCUGGGAGCUCUCUGCC	22023 **22053**

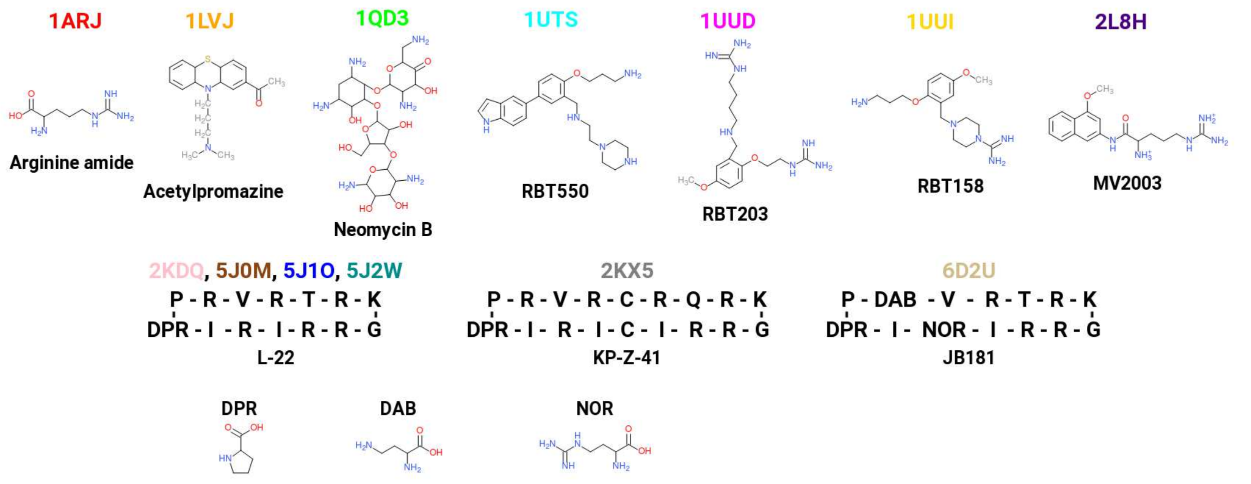
1ARJ
1LVJ
1QD3
1UTS
1UUD
1UUI
2L8H
Arginine amide
Acetylpromazine
Neomycin B
RBT550
RBT203
RBT158
MV2003
2KDQ, 5J0M, 5J1O, 5J2W
P - R - V - R - T - R - K
DPR - I - R - I - R - R - G
L-22
2KX5
P - R - V - R - C - R - Q - R - K
DPR - I - R - I - C - I - R - R - G
KP-Z-41
6D2U
P - DAB - V - R - T - R - K
DPR - I - NOR - I - R - R - G
JB181
DPR
DAB
NOR

Figure A.1: **Chemical structures of ligands studied:** (*top row*) small molecules; (*bottom row*) peptides.

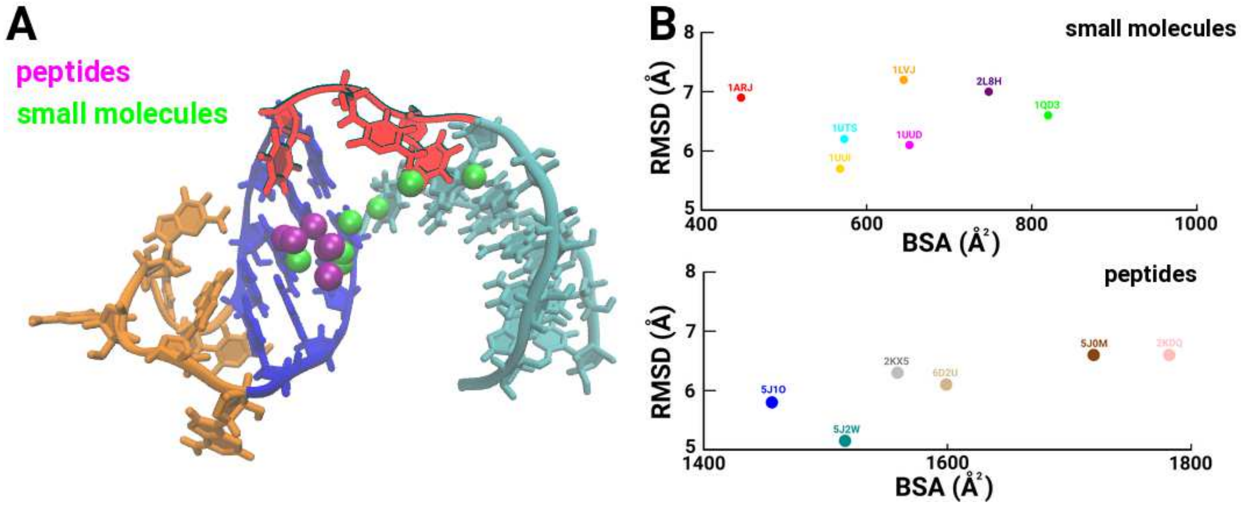

Figure A.2: **Conformational metrics of ligands.** (A) The centers of mass of ligands are represented by spheres (colored and labeled) and overlaid on the *apo* structure of TAR. (B) The all-atom root-mean-squared-deviation (RMSD) of each liganded TAR structure relative to the apo structure (PDB code 1ANR) *vs.* buried surface area (BSA) of each ligand are shown for small-molecules (*top panel*) and peptides (*bottom panel*).

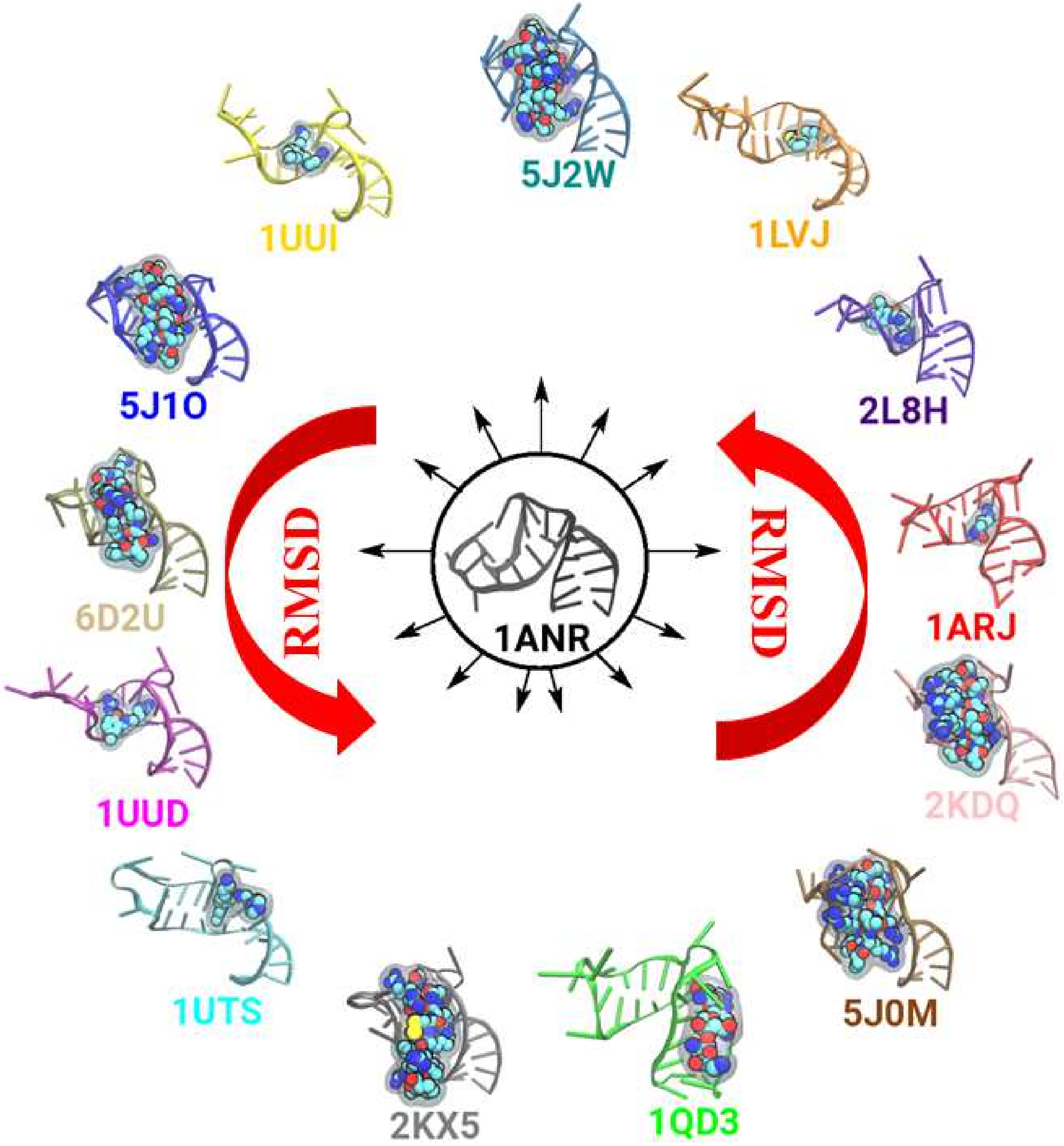

Figure A.3: **Snapshots of the initial systems in liganded simulations:** RNA, cartoon representation; ligands, space-filling. Snapshot of the *apo* TAR structure (PDB code 1ANR) is located at the center (black cartoon). See also Figure 3.1.

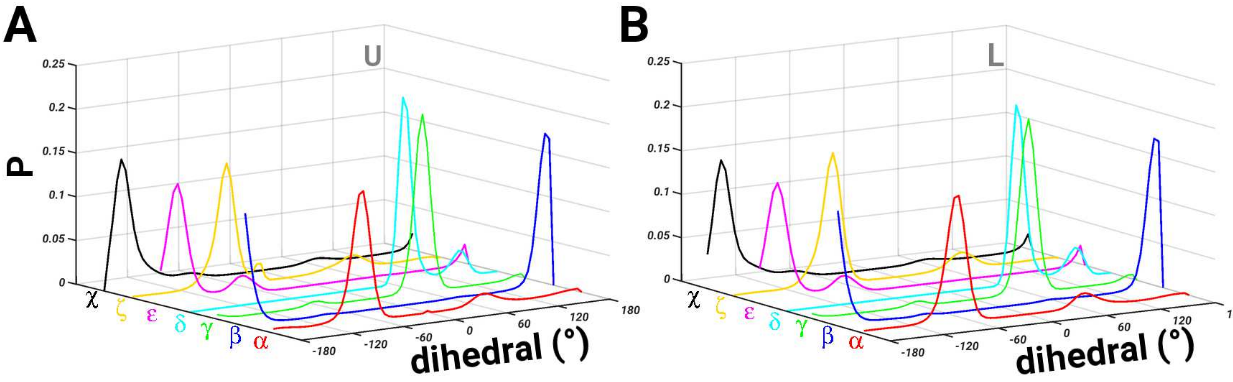

Figure A.4: **Torsional flexibility.** The normalized distributions of each RNA backbone dihedral angle (α, β, γ, δ, ϵ, ζ) and the glycosidic dihedral angle (χ) for unliganded (labeled U in panel A) and liganded (labeled L in panel B) simulations.

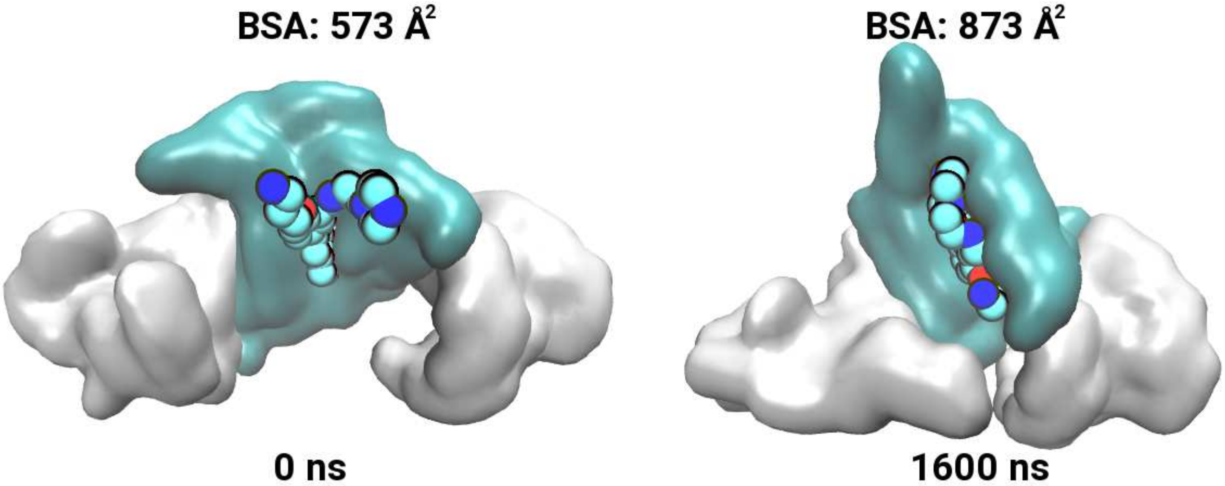

Figure A.5: **Snapshots highlighting an increase in BSA.** Shown are the snapshots of the TAR RNA conformations (surface map) and the ligand (RBT550; space-filling) from a liganded simulation (PDB code 1UTS) highlighting an increase in BSA in comparison to the initial BSA due to conformational rearrangements of the ligand in the binding pocket. A cyan surface indicates the nucleotides of the binding pocket in close contact with the ligand and a white surface represents the rest of the RNA structure.

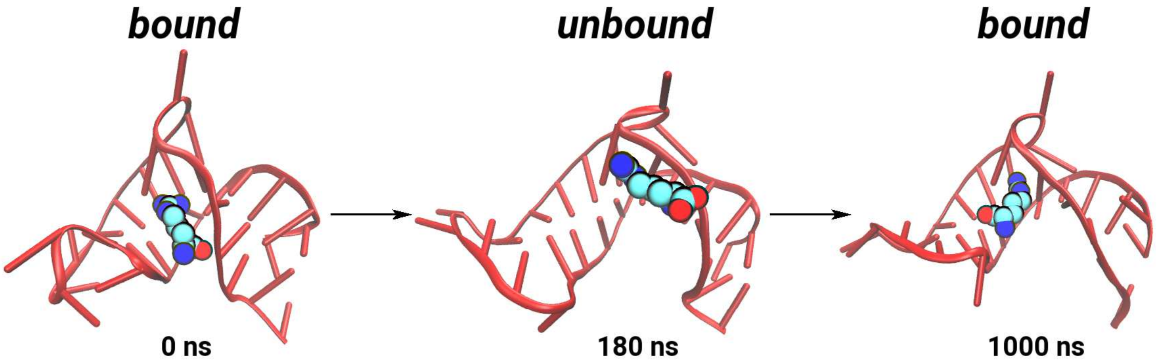

Figure A.6: **Snapshots highlighting partial ligand dissociation.** Shown are the snapshots of the TAR RNA conformations (red cartoon) and the ligand (arginine amide; space-filling) from a liganded simulation (PDB code 1ARJ) highlighting the partial dissociation of the ligand at t = 180 ns and then rebinding again in the original binding pocket (t = 1000 ns).

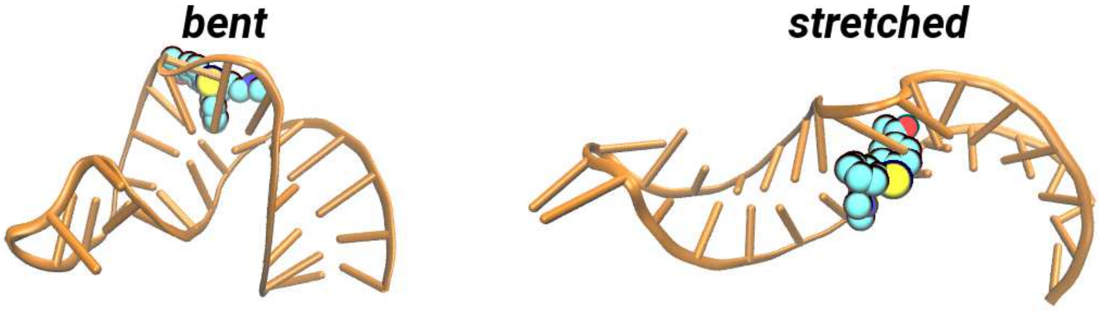

Figure A.7: **Conformational change in TAR RNA in a liganded simulation.** Shown are the *bent* and *stretched* conformations of TAR (orange cartoon) with the ligand (space-filling) from a liganded simulation (PDB code 1LVJ).

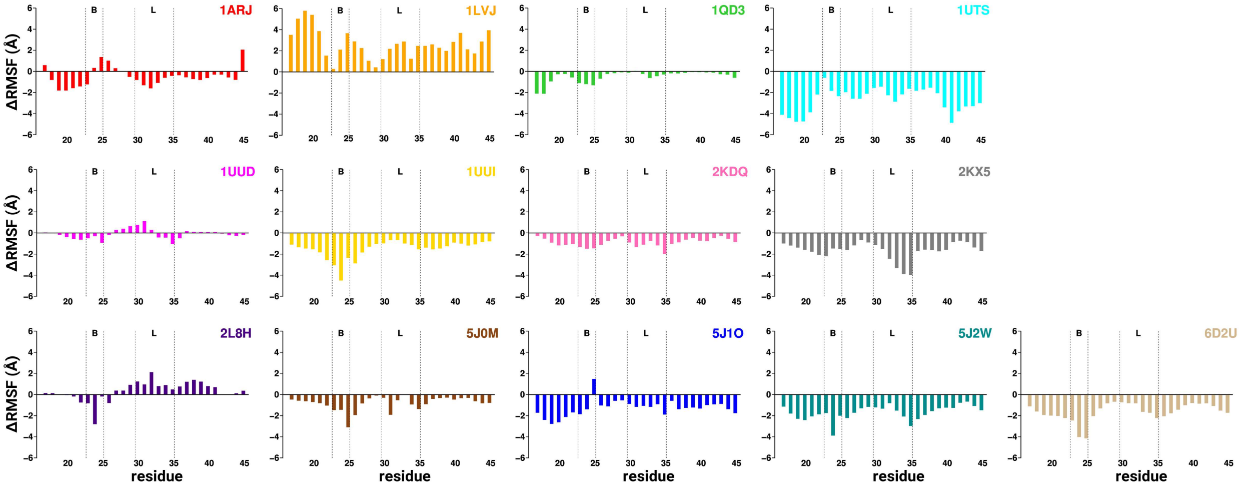

Figure A.8: The ΔRMSF per residue data are presented highlighting the differences between the unliganded and liganded simulations. Each system is uniquely colored. The bulge (B) and the loop (L) motifs are marked with the dashed lines.

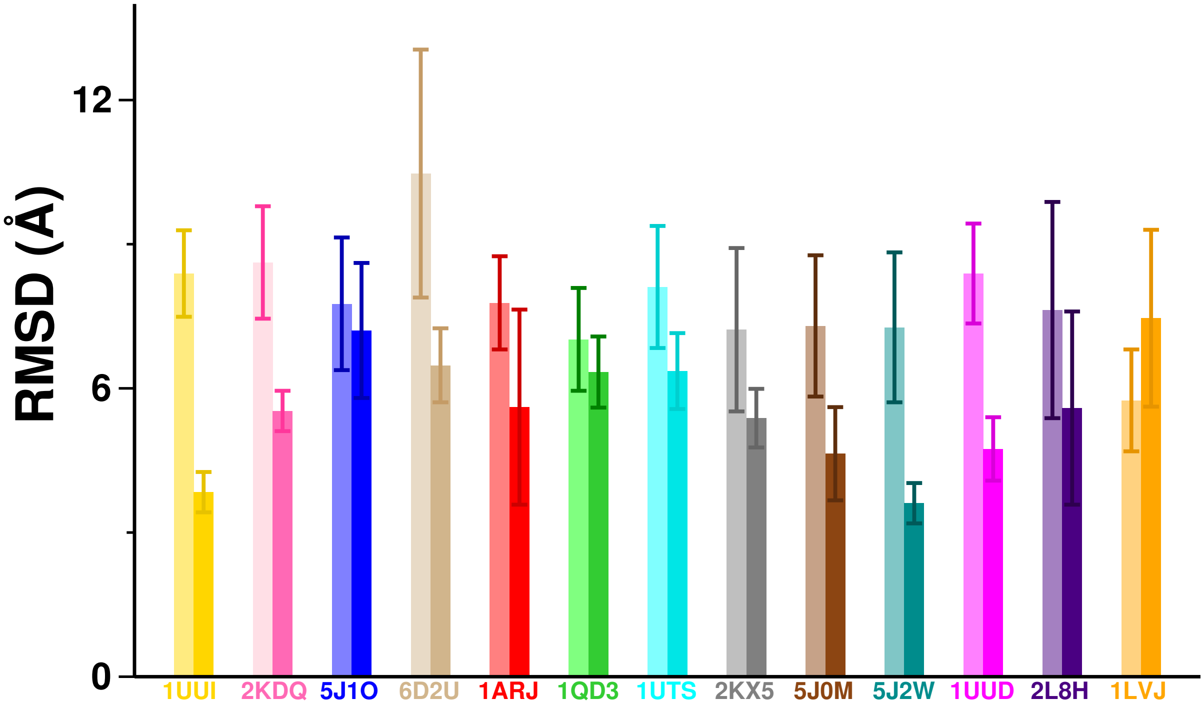

Figure A.9: **Flexibility of the bulge motif in unliganded and liganded states.** RMSD data with error bars, similar to Figure 3.3A, are shown for the bulge motif nucleotides (*lighter shades*, unliganded; *darker shades*, liganded).

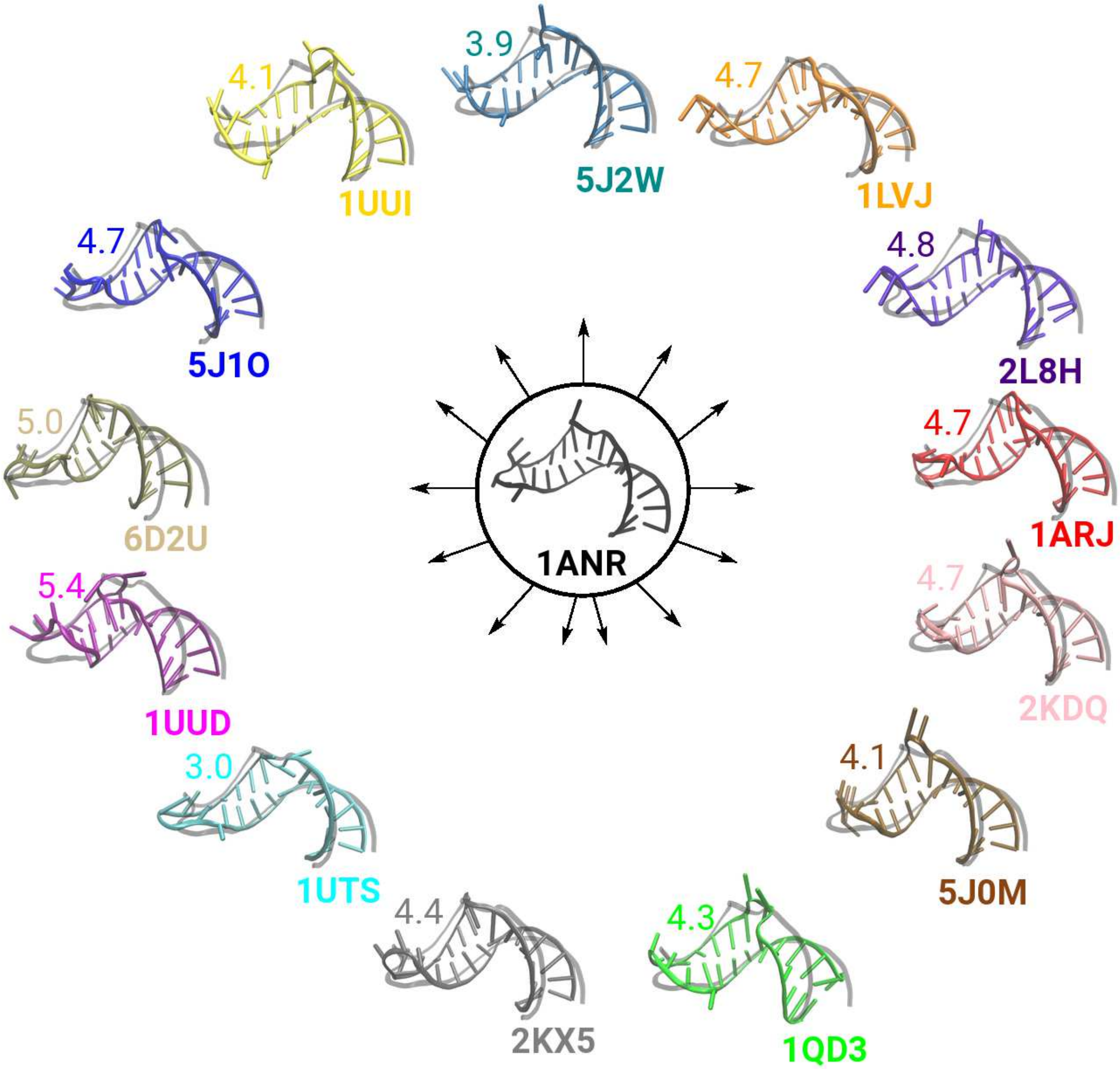

Figure A.10: **Comparison of average structures of TAR RNA.** The average structures of TAR RNA from unliganded simulations are overlaid on the average structures from the simulation of the *apo* TAR RNA structure (PDB code 1ANR). The RMSD values between the average structures are also labeled in color along with the PDB codes.

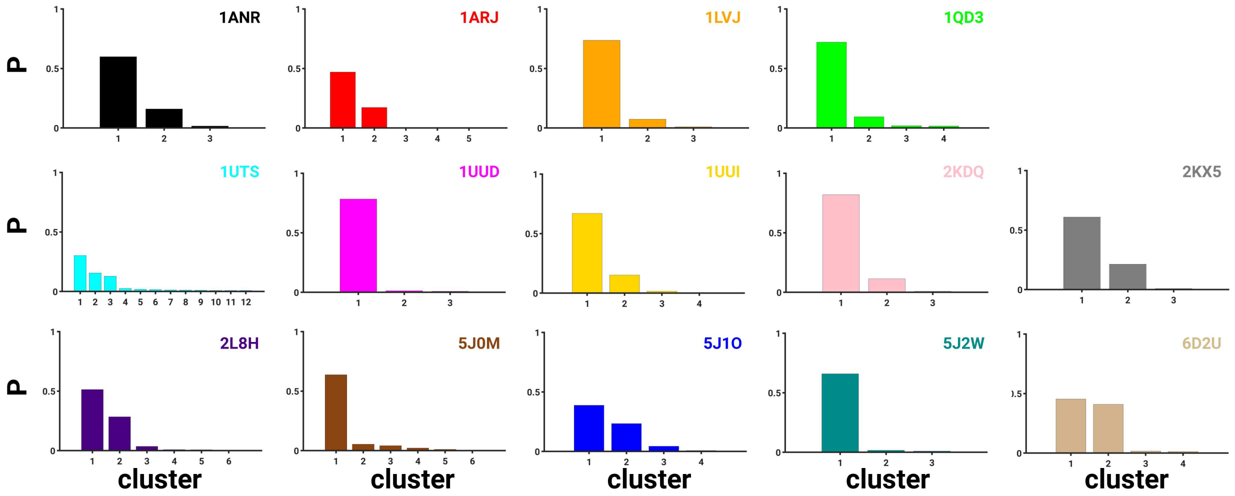

Figure A.11: **Cluster analysis of unliganded simulations.** The distributions of all clusters computed from conformations sampled via MD simulations are shown for the unliganded state of each system. Histograms are shown in the same color as the labeled PDB code. See also Figure 3.3B.

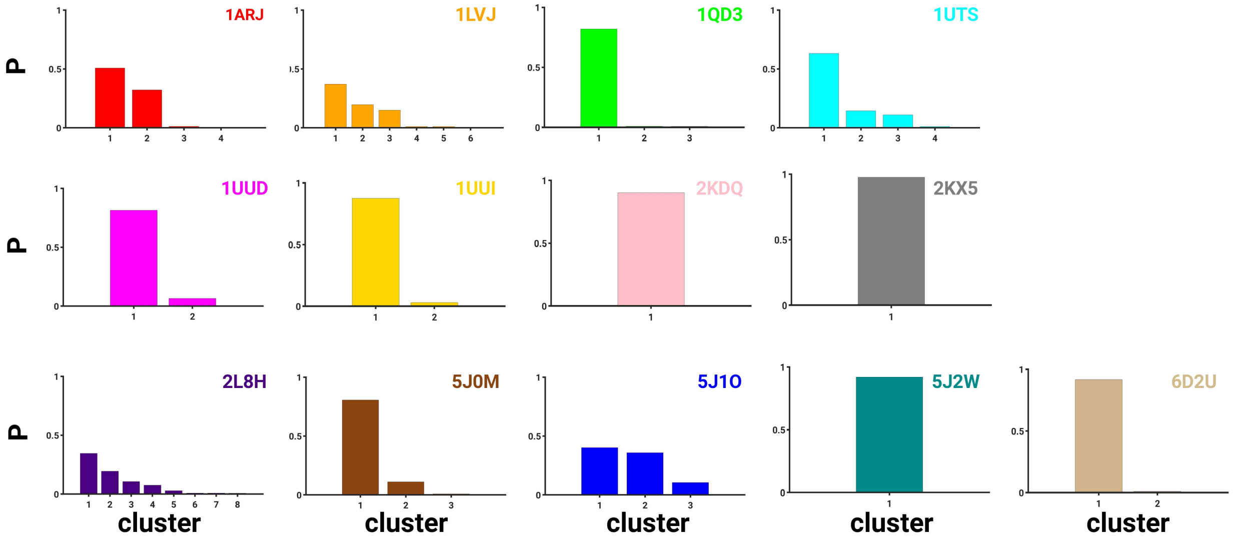

Figure A.12: **Cluster analysis of liganded simulations.** Data similar to Figure A.11 are shown for the liganded state of each system. See also Figure 3.3B.

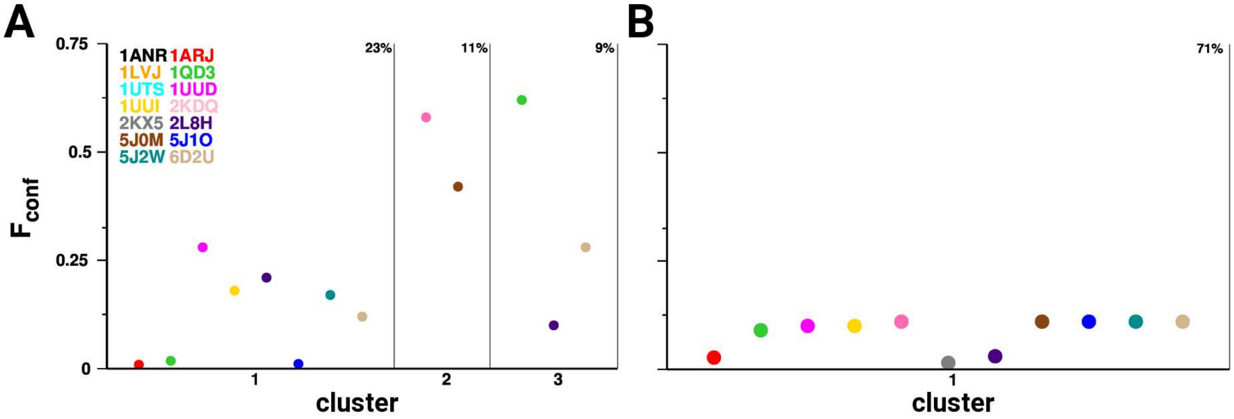

Figure A.13: **Combined cluster analysis.** The fraction of conformations (F_{conf}) from each system that populate each cluster from a set of (A) unliganded and (B) liganded simulations. Each system is uniquely colored. The numbers at the top of each cluster signify the percentage of the total number of frames that constitute that specific cluster.

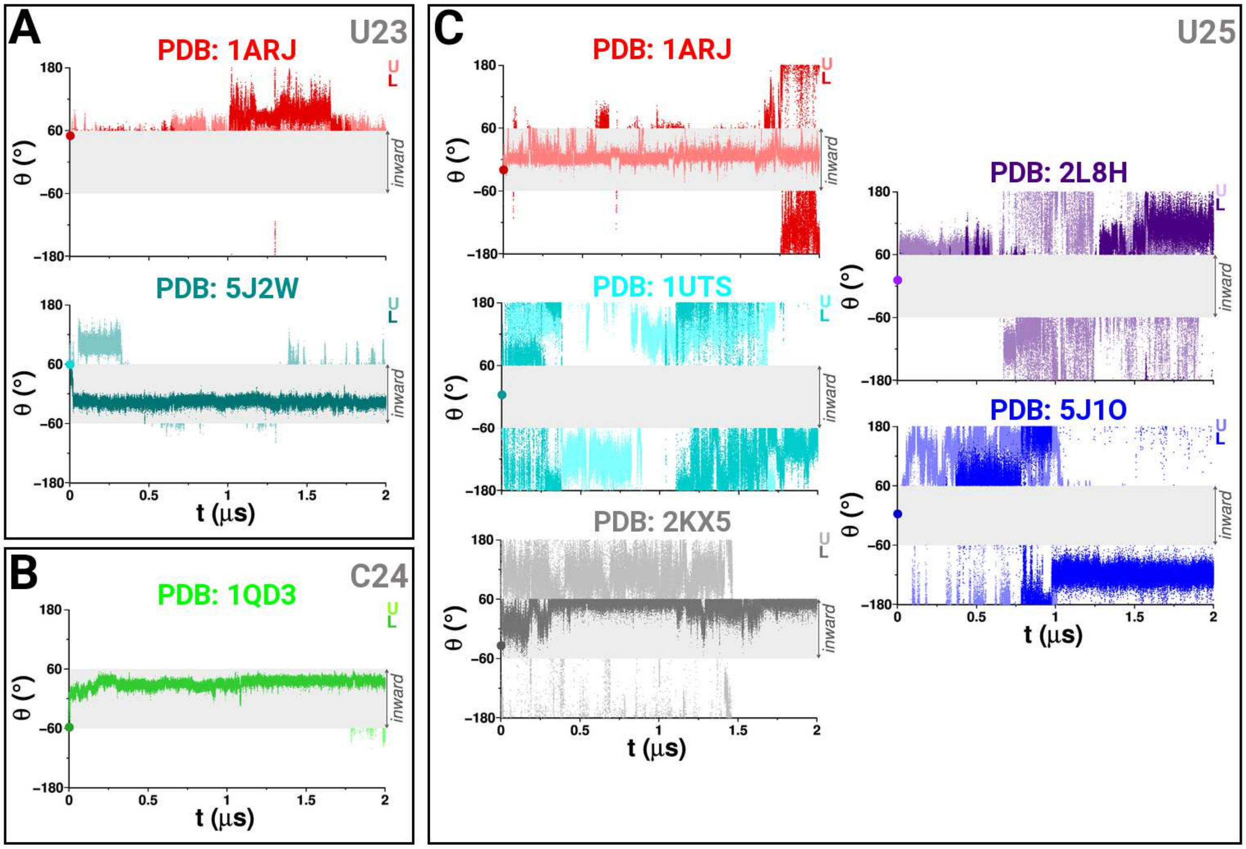

Figure A.14: **Conformational transitions in bulge nucleotides (U23, C24, and U25) in unliganded (U) and liganded (L) simulations.** Data similar to Figure 3.7 are shown for additional systems.

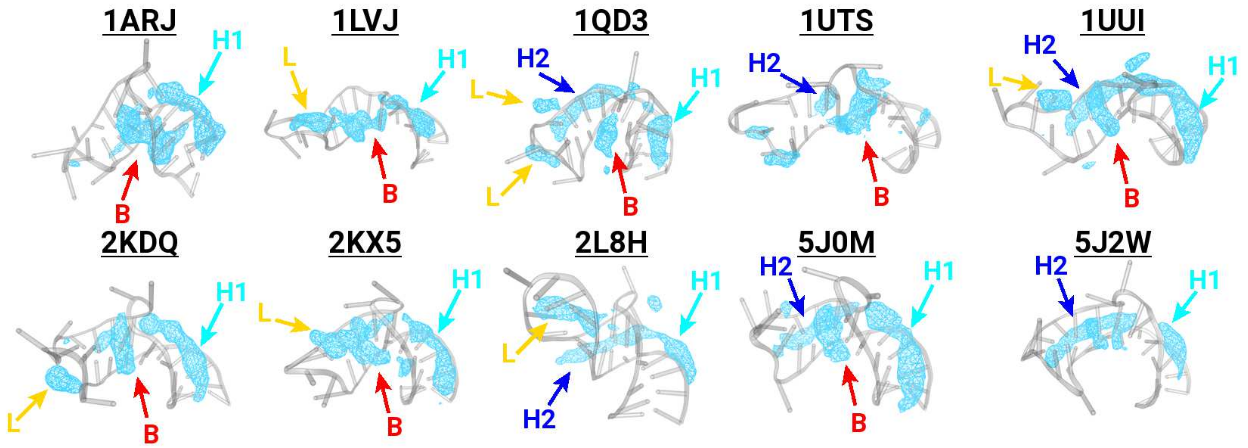

Figure A.15: **Predicted binding pockets in unliganded TAR structures.** Data similar to Figure 3.8A are shown for additional systems.

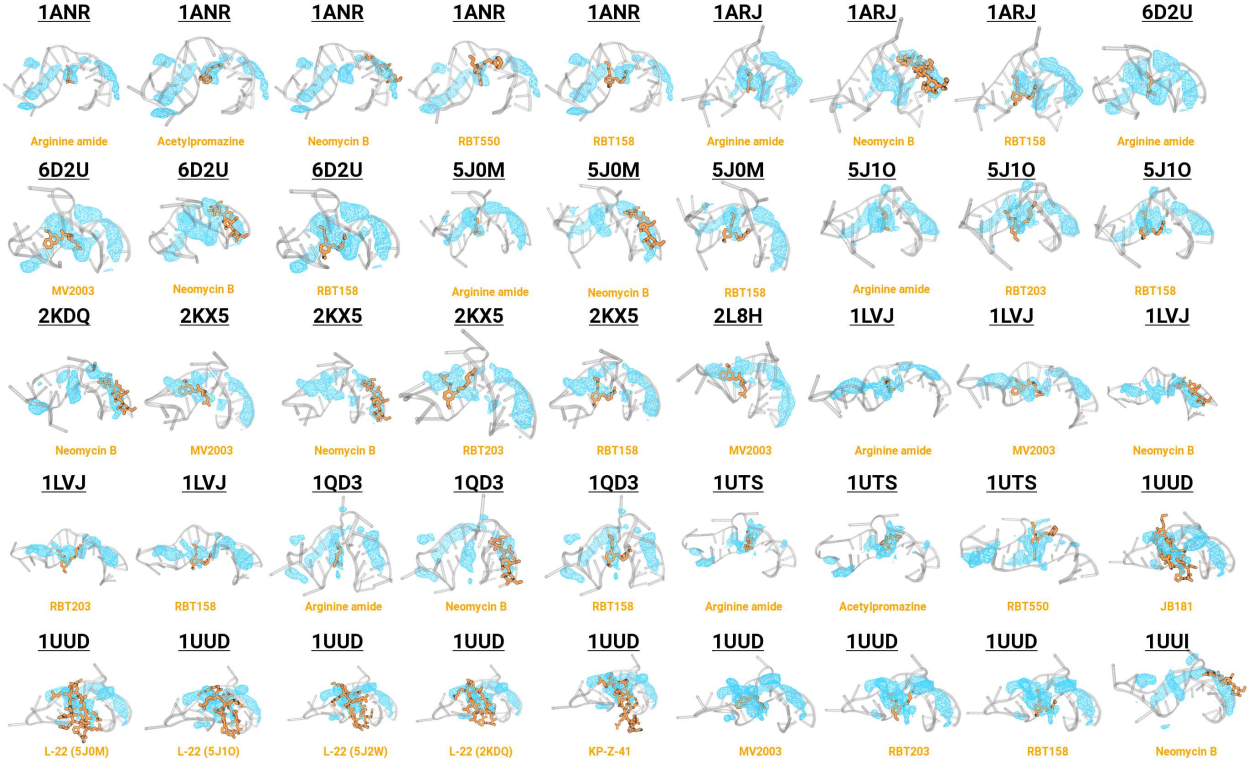

Figure A.16: **Overlays of ligands in predicted binding pockets.** Shown are the snapshots of predicted binding pockets (cyan surfaces) with an overlay of each ligand (orange sticks) on various TAR structures (transparent gray cartoons). See also Figure 3.8.

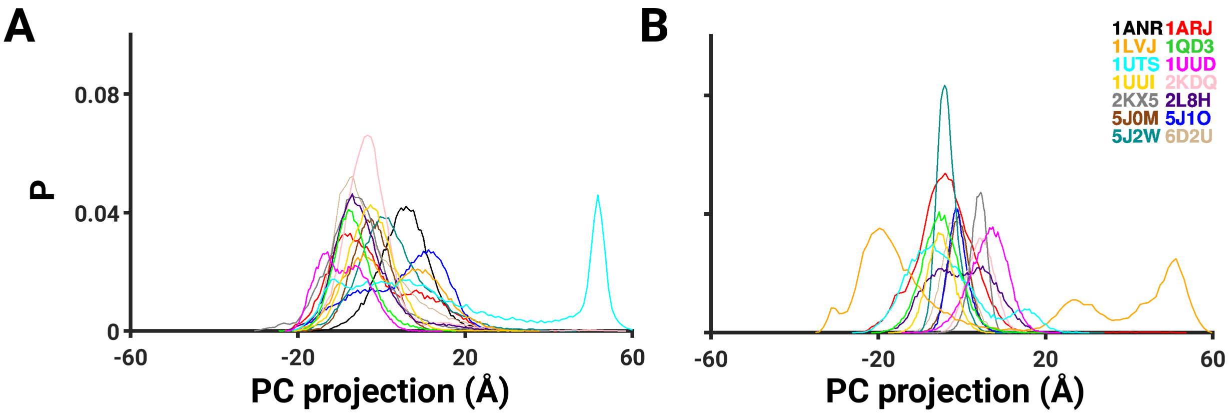

Figure A.17: **Principal component analysis.** Shown are the normalized histograms of the first principal component projection from each of the (A) unliganded and (B) liganded simulations. Each system is uniquely colored.

APPENDIX B
SCRIPTS FOR CHAPTER 3

B.1 Overview

In this appendix, I provide the scripts that I have used to set up and analyze simulations in chapter 3 using various software packages.

B.2 Amber Simulation

To set up an Amber simulation, I modified the residue names in the initial PDB file that was downloaded from the Protein Data Bank using a simple SED script. If this step is not performed, the TLEAP program will not be able to read the PDB file and set up the simulation domain.

```
sed ’s/G /RG/g’ 1anr_frame0.pdb > modified.pdb
sed -i ’s/C /RC/g’ modified.pdb
sed -i ’s/U /RU/g’ modified.pdb
sed -i ’s/A /RA/g’ modified.pdb
```

After that, I used the TLEAP program in Ambertools to solvate the system, to ionize the system, and to generate the input files for an MD simulation.

```
addPath /home/harishv/bin/amber16/dat/leap/parm/
source leaprc.RNA.ROC
source leaprc.water.tip3p
loadamberparams frcmod.ionsjc_tip3p
loadamberparams frcmod.ions234lm_126_tip3p
loadamberprep drug.prepi      !load force field file for the ligand
loadamberparams drug.frcmod  !load force field file for the ligand
mol = loadPdb "modified.pdb"
check mol
solvateBox mol TIP3PBOX 15
```

```
addIons2 mol Na+ 0
savepdb mol apo_final.pdb
saveAmberParm mol apo_final.prmtop apo.inpcrd
quit
```

The minimization of a system should be performed prior to conducting an MD simulation. Below, I provide the Amber minimization script.

```
Minimize
 $cntrl
  imin=1,
  ntx=1,
  irest=0,
  maxcyc=2000,
  ncyc=1000,
  ntpr=100,
  ntwx=0,
  cut=8.0,
/
```

It is critical to note that the simulations involving a minimization using Amber can be conducted only on **CPUs**. The Amber configuration file for conducting a simulation in the NPT ensemble is provided below.

```
Production
 &cntrl
 imin=0,
 iwrap=1,
 ntx=1,
 irest=0,
 nstlim=1000000000,
 dt=0.002,
 ntf=2,
 ntc=2,
 temp0=300,
 ntpr=10000,
 ntwr=10000,
 ntwx=10000,
 ntwv=10000,
 ntxo=1,
 cut=9.,
 ntb=2,
 ntp=1,
 ntt=3,
 gamma_ln=2,
 /
```

I conducted all simulations in this chapter on a local supercomputer at UNH (Premise). To copy my input files I used the following script:

```
scp -pr /home/levintov/project_dynamics/common/6d2u/setup
levintov@premise.sr.unh.edu:/mnt/home/chem-eng/levintov/tar/
```

I also provide jobscripts to submit Amber simulations on a local supercomputer at UNH (Premise). A script for launching a minimization job on a CPU node:

```
#!/bin/csh

#SBATCH --partition=harish -N 1 --time=00:15:00 --job-name=1arj_min
#SBATCH --output=slurm_%x-%j.log  --error=slurm_%x-%j.err

module load mpi
module load Amber

set AMBER="pmemd"

set parmfile="arj_final.prmtop"
set initial="arj.inpcrd"

echo -n "Starting Script at: "
date
echo ""

mpirun $AMBER -O -i min.in -o min.out -p $parmfile -c $initial -r min.rst
-inf min.info

echo "ALL DONE"
```

A script for launching an MD simulation on a GPU node:

```
#!/bin/csh

#SBATCH --partition=harish --gres=gpu:1 --job-name=1arj

#SBATCH --output=slurm_%x-%j.log  --error=slurm_%x-%j.err

module load mpi
module load Amber

set AMBER="pmemd.cuda.MPI"

set parmfile="arj_final.prmtop"
##set input=""
set initial="min.rst"

echo -n "Starting Script at: "
date
echo ""
```

```

mpirun $AMBER -O -i run.in -o run.out -p $parmfile -c $initial -r run.rst
-x run.mdcrd  -v run.mdvel -inf run.info

echo "ALL DONE"
```

B.3 Analysis Scripts

In this study, I used various conformational metrics to analyze the system. In this section, I provide all the scripts that I have generated. VMD [184] scripts can be executed from the terminal window: **vmd -dispdev text -e input.tcl**. CPPTRAJ [185] scripts can also be executed from the terminal window: **cpptraj -i input.in**. MATLAB scripts are executed in the MATLAB GUI.

B.3.1 RMSD Script

```
###VMD SCRIPT
mol new ../arj_raw.prmtop
mol addfile ../no_wat.nc waitfor all

set nf [molinfo top get numframes]
set ref0 [atomselect top "nucleic" frame 0]

set sel [atomselect top "nucleic"]
set sel1 [atomselect top "nucleic"]

##Apo
set out0 [open rmsd_arj.dat "w"]

for {set i 0} { $i < $nf } { incr i } {

$sel frame $i
$sel1 frame $i

$sel move [measure fit $sel1 $ref0]
set rmsd [measure rmsd $sel1 $ref0]
puts $out0 "$i $rmsd"
}

exit
```

B.3.2 RMSF Script

```
###VMD SCRIPT
mol new ../arj_raw.prmtop
mol addfile ../no_wat.nc waitfor all

set outfile [open "arj_rmsf_bound.dat" w]
set sel [atomselect top all]
set sel0 [$sel num]
set sel [atomselect top "resid 1 to $sel0 and name P"]
set ref0 [atomselect top all frame 0]
set sel1 [atomselect top all]

set stepsize 1

set nframes [molinfo top get numframes]
set nframes2 [expr $nframes - 1]

for {set i 0} {$i < [$sel num]} {incr i} {
    $sel1 move [measure fit $sel1 $ref0]
    set rmsf [measure rmsf $sel first 1 last $nframes2 step $stepsize]
    puts $outfile "[expr {$i+1}] \t [lindex $rmsf $i]"
}

close $outfile
```

B.3.3 BSA Script

```
###VMD SCRIPT
mol new ../arj_raw.prmtop
mol addfile ../no_wat.nc waitfor all

set nf [molinfo top get numframes]

set ref0 [atomselect top "nucleic" frame 0]

set rna [atomselect top "nucleic"]
set drug [atomselect top "not nucleic"]
set together [atomselect top all]

set out0 [open bsa_arj.dat "w"]

for {set i 0} { $i < $nf } { incr i } {
$rna frame $i
$drug frame $i
$together frame $i

$together move [measure fit $rna $ref0]

set sasa_rna [measure sasa 1.4 $rna -restrict $rna]
set sasa_drug [measure sasa 1.4 $drug -restrict $drug]
set sasa_together [measure sasa 1.4 $together -restrict $together]
set bsa [expr $sasa_rna+$sasa_drug-$sasa_together]
puts $out0 "$i $bsa"
}

exit
```

B.3.4 Script to Compute an Average Structure

```
###CPPTRAJ SCRIPT
parm ../arj_rna.prmtop
trajin ../no_wat.nc
rms first mass @P
average arj_aver_bound.pdb pdb
run
exit
```

B.3.5 Script to Conduct a Cluster Analysis

```
###CPPTRAJ SCRIPT
parm ../arj_rna.prmtop
trajin ../aligned.nc
cluster C0 \
        dbscan minpoints 25 epsilon 1.5 sievetoframe \
        rms @P \
        sieve 10 \
        out cnumvtime.dat \
        summary summary.dat \
        info info.dat \
        cpopvtime cpopvtime.agr normframe \
        repout rep repfmt pdb \
        singlerepout singlerep.nc singlerepfmt netcdf \
        avgout Avg avgfmt restart
```

B.3.6 Script to Compute a Dihedral Angle of a Flipping Base

```
###CPPTRAJ SCRIPT
parm ../arj_rna.prmtop
trajin ../rna.nc
dihedral r8_r6 :6@N1,C2,N3,C4,C5,C6 :7@C1’,O4’,C4’,C3’,C2’
:8@C2’,C3,C4’,O4,C1’ :8@N1,C2,N3,C4,C5,C6 out arj_r8_r6.dat mass
run
exit
```

B.3.7 Script to Compute ΔRMSF

```
###MATLAB SCRIPT
A_free = importdata(’data/arj_rmsf_free.dat’);
A_bound = importdata(’data/arj_rmsf_bound.dat’);

RMSF_free = A_free(:,2);
RMSF_bound = A_bound(:,2);

del_RMSF = RMSF_bound - RMSF_free;

fileID = fopen(’arj_delRMSF.dat’,’w’);
fprintf(fileID,’%f\n’,del_RMSF);
```

B.4 Scripts to Generate Figures

I primarily used GNUPLOT and MATLAB to generate data plots in my work. In this section, I provide several example scripts to generate the plots presented in chapter 3.

B.4.1 RMSD Plot

```
###GNUPLOT SCRIPT
#!/usr/bin/gnuplot
set encoding iso_8859_1
set term post eps enh color size 20,12 "HelveticaBold" 50 solid
set output "rmsd_main_v2.eps"
unset key
unset tics
unset border

set xrange [-2.8:27]
set yrange [-0.5:10]

set arrow 1 from 0,0 to 27,0 nohead lw 8
set arrow 2 from -0.001,0 to 0,10 nohead lw 8
set arrow 3 from 0,5 to -0.4,5 nohead lw 6
set arrow 4 from 0,9.98 to -0.4,9.98 nohead lw 6
set arrow 5 from 0,0 to -0.4,0 nohead lw 6
set arrow 100 from 0,2.5 to -0.2,2.5 nohead lw 6
set arrow 101 from 0,7.5 to -0.2,7.5 nohead lw 6

set label 1 "0" at -1.1,0 font "HelveticaBold,85"
set label 2 "5" at -1.1,5 font "HelveticaBold,85"
set label 3 "10" at -1.6,10 font "HelveticaBold,85"
set label 4 "RMSD ({\305})" at -2.5,3.25 font "HelveticaBold,120"
rotate by 90
```

```

set label 5 "1UUI" at 0.7,-0.3 font "HelveticaBold,60" tc rgb "#ffd600"
set label 6 "2KDQ" at 2.7,-0.3 font "HelveticaBold,60" tc rgb "#FF69B4"
set label 7 "5J1O" at 4.7,-0.3 font "HelveticaBold,60" tc rgb "#0000ff"
set label 8 "6D2U" at 6.7,-0.3 font "HelveticaBold,60" tc rgb "#d2b48c"
set label 9 "1ARJ" at 8.7,-0.3 font "HelveticaBold,60" tc rgb "red"
set label 10 "1QD3" at 10.7,-0.3 font "HelveticaBold,60" tc rgb "#32CD32"
set label 11 "1UTS" at 12.7,-0.3 font "HelveticaBold,60" tc rgb "#00FFFF"
set label 12 "2KX5" at 14.7,-0.3 font "HelveticaBold,60" tc rgb "#7f7f7f"
set label 13 "5J0M" at 16.7,-0.3 font "HelveticaBold,60" tc rgb "#8b4513"
set label 14 "5J2W" at 18.7,-0.3 font "HelveticaBold,60" tc rgb "#008b8b"
set label 15 "1UUD" at 20.65,-0.3 font "HelveticaBold,60" tc rgb "#ff00ff"
set label 16 "2L8H" at 22.7,-0.3 font "HelveticaBold,60" tc rgb "#4b0082"
set label 17 "1LVJ" at 24.7,-0.3 font "HelveticaBold,60" tc rgb "orange"

set label 100 "{/ZapfDingbats \153}" at 26.1,7 font "HelveticaBold,50"
tc rgb "orange"
set label 101 "1" at 4.3,8.8 font "HelveticaBold,80"
set label 102 "2" at 14.3,8.8 font "HelveticaBold,80"
set label 103 "3" at 23.2,8.8 font "HelveticaBold,80"

#set label 11 "PDB" at 6,-200 font "HelveticaBold,120"

#STD
#1UUI
set arrow 12 from 1.75,3.3 to 1.75,4.06 nohead lw 9 lc rgb "#e6c100"
set arrow 13 from 1.55,3.3 to 1.95,3.3 nohead lw 9 lc rgb "#e6c100"
set arrow 14 from 1.55,4.06 to 1.95,4.06 nohead lw 9 lc rgb "#e6c100"

set arrow 15 from 1.25,4.32 to 1.25,5.6 nohead lw 9 lc rgb "#e6c100"
set arrow 16 from 1.05,4.32 to 1.45,4.32 nohead lw 9 lc rgb "#e6c100"
set arrow 17 from 1.05,5.6 to 1.45,5.6 nohead lw 9 lc rgb "#e6c100"

#2KDQ
set arrow 18 from 3.75,3.85 to 3.75,4.39 nohead lw 9 lc rgb "#ff369b"
set arrow 19 from 3.55,3.85 to 3.95,3.85 nohead lw 9 lc rgb "#ff369b"
set arrow 20 from 3.55,4.39 to 3.95,4.39 nohead lw 9 lc rgb "#ff369b"

set arrow 21 from 3.25,4.9 to 3.25,6.28 nohead lw 9 lc rgb "#ff369b"
set arrow 22 from 3.05,4.9 to 3.45,4.9 nohead lw 9 lc rgb "#ff369b"
set arrow 23 from 3.05,6.28 to 3.45,6.28 nohead lw 9 lc rgb "#ff369b"

#5J1O
set arrow 24 from 5.75,3.22 to 5.75,4.18 nohead lw 9 lc rgb "#0000b3"
set arrow 25 from 5.55,3.22 to 5.95,3.22 nohead lw 9 lc rgb "#0000b3"
set arrow 26 from 5.55,4.18 to 5.95,4.18 nohead lw 9 lc rgb "#0000b3"

set arrow 27 from 5.25,4.41 to 5.25,6.23 nohead lw 9 lc rgb "#0000b3"
set arrow 28 from 5.05,4.41 to 5.45,4.41 nohead lw 9 lc rgb "#0000b3"
set arrow 29 from 5.05,6.23 to 5.45,6.23 nohead lw 9 lc rgb "#0000b3"

#6D2U
set arrow 30 from 7.75,2.96 to 7.75,3.5 nohead lw 9 lc rgb "#c49c67"
set arrow 31 from 7.55,2.96 to 7.95,2.96 nohead lw 9 lc rgb "#c49c67"
set arrow 32 from 7.55,3.5 to 7.95,3.5 nohead lw 9 lc rgb "#c49c67"

set arrow 33 from 7.25,4 to 7.25,6.16 nohead lw 9 lc rgb "#c49c67"
set arrow 34 from 7.05,4 to 7.45,4 nohead lw 9 lc rgb "#c49c67"
set arrow 35 from 7.05,6.16 to 7.45,6.16 nohead lw 9 lc rgb "#c49c67"

#1ARJ
set arrow 36 from 9.75,4.87 to 9.75,6.33 nohead lw 9 lc rgb "#cc0000"
set arrow 37 from 9.55,4.87 to 9.95,4.87 nohead lw 9 lc rgb "#cc0000"
set arrow 38 from 9.55,6.33 to 9.95,6.33 nohead lw 9 lc rgb "#cc0000"

set arrow 39 from 9.25,5.61 to 9.25,7.65 nohead lw 9 lc rgb "#cc0000"
set arrow 40 from 9.05,5.61 to 9.45,5.61 nohead lw 9 lc rgb "#cc0000"
set arrow 41 from 9.05,7.65 to 9.45,7.65 nohead lw 9 lc rgb "#cc0000"

#1QD3
set arrow 42 from 11.75,4.04 to 11.75,4.86 nohead lw 9 lc rgb "#008000"
set arrow 43 from 11.55,4.04 to 11.95,4.04 nohead lw 9 lc rgb "#008000"
set arrow 44 from 11.55,4.86 to 11.95,4.86 nohead lw 9 lc rgb "#008000"

set arrow 45 from 11.25,4.65 to 11.25,6.01 nohead lw 9 lc rgb "#008000"
set arrow 46 from 11.05,4.65 to 11.45,4.65 nohead lw 9 lc rgb "#008000"
set arrow 47 from 11.05,6.01 to 11.45,6.01 nohead lw 9 lc rgb "#008000"

#1UTS
set arrow 48 from 13.75,4.85 to 13.75,6.39 nohead lw 9 lc rgb "#00cece"
set arrow 49 from 13.55,4.85 to 13.95,4.85 nohead lw 9 lc rgb "#00cece"
set arrow 50 from 13.55,6.39 to 13.95,6.39 nohead lw 9 lc rgb "#00cece"

set arrow 51 from 13.25,5.28 to 13.25,8.24 nohead lw 9 lc rgb "#00cece"
set arrow 52 from 13.05,5.28 to 13.45,5.28 nohead lw 9 lc rgb "#00cece"
set arrow 53 from 13.05,8.24 to 13.45,8.24 nohead lw 9 lc rgb "#00cece"

#2KX5
set arrow 54 from 15.75,3.79 to 15.75,4.49 nohead lw 9 lc rgb "#666666"
set arrow 55 from 15.55,3.79 to 15.95,3.79 nohead lw 9 lc rgb "#666666"
set arrow 56 from 15.55,4.49 to 15.95,4.49 nohead lw 9 lc rgb "#666666"

set arrow 57 from 15.25,3.87 to 15.25,5.65 nohead lw 9 lc rgb "#666666"
set arrow 58 from 15.05,3.87 to 15.45,3.87 nohead lw 9 lc rgb "#666666"
set arrow 59 from 15.05,5.65 to 15.45,5.65 nohead lw 9 lc rgb "#666666"

#5J0M
set arrow 60 from 17.75,4.1 to 17.75,4.8 nohead lw 9 lc rgb "#5e2f0d"
set arrow 61 from 17.55,4.1 to 17.95,4.1 nohead lw 9 lc rgb "#5e2f0d"
set arrow 62 from 17.55,4.8 to 17.95,4.8 nohead lw 9 lc rgb "#5e2f0d"

set arrow 63 from 17.25,4.14 to 17.25,5.52 nohead lw 9 lc rgb "#5e2f0d"
set arrow 64 from 17.05,4.14 to 17.45,4.14 nohead lw 9 lc rgb "#5e2f0d"
set arrow 65 from 17.05,5.52 to 17.45,5.52 nohead lw 9 lc rgb "#5e2f0d"

#5J2W
set arrow 66 from 19.75,3.07 to 19.75,3.55 nohead lw 9 lc rgb "#005858"
set arrow 67 from 19.55,3.07 to 19.95,3.07 nohead lw 9 lc rgb "#005858"
set arrow 68 from 19.55,3.55 to 19.95,3.55 nohead lw 9 lc rgb "#005858"

set arrow 69 from 19.25,3.44 to 19.25,5.34 nohead lw 9 lc rgb "#005858"
set arrow 70 from 19.05,3.44 to 19.45,3.44 nohead lw 9 lc rgb "#005858"
set arrow 71 from 19.05,5.34 to 19.45,5.34 nohead lw 9 lc rgb "#005858"

#1UUD
set arrow 72 from 21.75,3.99 to 21.75,5.11 nohead lw 9 lc rgb "#d800d8"
set arrow 73 from 21.55,3.99 to 21.95,3.99 nohead lw 9 lc rgb "#d800d8"
set arrow 74 from 21.55,5.11 to 21.95,5.11 nohead lw 9 lc rgb "#d800d8"

set arrow 75 from 21.25,4.3 to 21.25,5.34 nohead lw 9 lc rgb "#d800d8"
set arrow 76 from 21.05,4.3 to 21.45,4.3 nohead lw 9 lc rgb "#d800d8"
set arrow 77 from 21.05,5.34 to 21.45,5.34 nohead lw 9 lc rgb "#d800d8"

#2L8H
set arrow 78 from 23.75,3.9 to 23.75,6.12 nohead lw 9 lc rgb "#2e004f"
set arrow 79 from 23.55,3.9 to 23.95,3.9 nohead lw 9 lc rgb "#2e004f"
set arrow 80 from 23.55,6.12 to 23.95,6.12 nohead lw 9 lc rgb "#2e004f"

set arrow 81 from 23.25,4.04 to 23.25,6.2 nohead lw 9 lc rgb "#2e004f"
set arrow 82 from 23.05,4.04 to 23.45,4.04 nohead lw 9 lc rgb "#2e004f"
set arrow 83 from 23.05,6.2 to 23.45,6.2 nohead lw 9 lc rgb "#2e004f"

#1LVJ
set arrow 84 from 25.75,5.73 to 25.75,8.17 nohead lw 9 lc rgb "#e69500"
set arrow 85 from 25.55,5.73 to 25.95,5.73 nohead lw 9 lc rgb "#e69500"
set arrow 86 from 25.55,8.17 to 25.95,8.17 nohead lw 9 lc rgb "#e69500"

set arrow 87 from 25.25,4.4 to 25.25,5.72 nohead lw 9 lc rgb "#e69500"
set arrow 88 from 25.05,4.4 to 25.45,4.4 nohead lw 9 lc rgb "#e69500"
set arrow 89 from 25.05,5.72 to 25.45,5.72 nohead lw 9 lc rgb "#e69500"

#1UUI
set object 1 rect from 1,0 to 1.5,4.96 fc rgb "#ffd600" fs solid 0.5
noborder
set object 2 rect from 1.5,0 to 2,3.68 fc rgb "#ffd600" fs solid 1
noborder

#2KDQ
set object 3 rect from 3,0 to 3.5,5.59 fc rgb "#ffc0cb" fs solid 0.5
noborder
set object 4 rect from 3.5,0 to 4,4.12 fc rgb "#FF69B4" fs solid 1
noborder

#5J1O
set object 5 rect from 5,0 to 5.5,5.32 fc rgb "#0000ff" fs solid 0.5
noborder
set object 6 rect from 5.5,0 to 6,3.7 fc rgb "#0000ff" fs solid 1 noborder

#6D2U
set object 7 rect from 7,0 to 7.5,5.08 fc rgb "#d2b48c" fs solid 0.5
noborder
set object 8 rect from 7.5,0 to 8,3.23 fc rgb "#d2b48c" fs solid 1
```

```
noborder

#1ARJ
set object 9 rect from 9,0 to 9.5,6.63 fc rgb "red" fs solid 0.5
noborder
set object 10 rect from 9.5,0 to 10,5.6 fc rgb "red" fs solid 1
noborder

#1QD3
set object 11 rect from 11,0 to 11.5,5.33 fc rgb "green" fs solid 0.5
noborder
set object 12 rect from 11.5,0 to 12,4.45 fc rgb "#32CD32" fs solid 1
noborder

#1UTS
set object 13 rect from 13,0 to 13.5,6.76 fc rgb "#00FFFF" fs solid 0.5
noborder
set object 14 rect from 13.5,0 to 14,5.62 fc rgb "#00e5e5" fs solid 1
noborder

#2KX5
set object 15 rect from 15,0 to 15.5,4.76 fc rgb "#7f7f7f" fs solid 0.5
noborder
set object 16 rect from 15.5,0 to 16,4.14 fc rgb "#7f7f7f" fs solid 1
noborder

#5J0M
set object 17 rect from 17,0 to 17.5,4.83 fc rgb "#8b4513" fs solid 0.5
noborder
set object 18 rect from 17.5,0 to 18,4.45 fc rgb "#8b4513" fs solid 1
noborder

#5J2W
set object 19 rect from 19,0 to 19.5,4.39 fc rgb "#008b8b" fs solid 0.5
noborder
set object 20 rect from 19.5,0 to 20,3.31 fc rgb "#008b8b" fs solid 1
noborder

#1UUD
set object 21 rect from 21,0 to 21.5,4.82 fc rgb "#ff00ff" fs solid 0.5
noborder
set object 22 rect from 21.5,0 to 22,4.55 fc rgb "#ff00ff" fs solid 1
noborder

#2L8H
set object 23 rect from 23,0 to 23.5,5.12 fc rgb "#4b0082" fs solid 0.5
noborder
set object 24 rect from 23.5,0 to 24,5.01 fc rgb "#4b0082" fs solid 1
noborder

#1LVJ
set object 25 rect from 25,0 to 25.5,5.06 fc rgb "orange" fs solid 0.5
noborder
set object 26 rect from 25.5,0 to 26,6.95 fc rgb "orange" fs solid 1
```

```
noborder

#Lines to outline groups
set arrow 104 from 0.73,8.5 to 8.27,8.5 nohead lw 6
set arrow 105 from 0.75,8.5 to 0.75,8.3 nohead lw 6
set arrow 106 from 8.25,8.5 to 8.25,8.3 nohead lw 6

set arrow 107 from 8.73,8.5 to 20.27,8.5 nohead lw 6
set arrow 108 from 8.75,8.5 to 8.75,8.3 nohead lw 6
set arrow 109 from 20.25,8.5 to 20.25,8.3 nohead lw 6

set arrow 110 from 20.73,8.5 to 26.27,8.5 nohead lw 6
set arrow 111 from 20.75,8.5 to 20.75,8.3 nohead lw 6
set arrow 112 from 26.25,8.5 to 26.25,8.3 nohead lw 6

p "test.dat" u 1:2 w l lt rgb "gray" lw 8 notitle
```

B.4.2 Average Bar Plot

The following MATLAB script generates a bar plot from chapter 3 (see Figure 3.4A).

```
###MATLAB SCRIPT
A1 = importdata(’matrix_initial.dat’);
figure
width = 0.4;
h = bar3(A1,width);
set(gca,’box’,’off’,’TickDir’,’out’,’fontweight’,’bold’,’FontAngle’,
’italic’,’fontsize’,24,’linewidth’,3,’FontName’,’Bookman’,
’FontSmoothing’,’on’,’xtick’,[],’ytick’,[]);
ax = gca;
ax.ZMinorTick       = ’off’;
ylim([0.5,14.5]);
cm = get(gcf,’colormap’);  % Use the current colormap.
cnt = 0;
for jj = 1:length(h)
    xd = get(h(jj),’xdata’);
    yd = get(h(jj),’ydata’);
    zd = get(h(jj),’zdata’);
    delete(h(jj))
    idx = [0;find(all(isnan(xd),2))];
    if jj == 1
        S = zeros(length(h)*(length(idx)-1),1);
        dv = floor(size(cm,1)/length(S));
    end
    for ii = 1:length(idx)-1
        cnt = cnt + 1;
        S(cnt) = surface(xd(idx(ii)+1:idx(ii+1)-1,:),...
                         yd(idx(ii)+1:idx(ii+1)-1,:),...
                         zd(idx(ii)+1:idx(ii+1)-1,:),...
                         ’facecolor’,cm((cnt-1)*dv+1,:));
    end
end
```

```
zlim([0,10]);
% 1anr
set(S(1),'facecolor','white');
set(S(2),'facecolor','black');
set(S(3),'facecolor','black');
set(S(4),'facecolor','black');
set(S(5),'facecolor','black');
set(S(6),'facecolor','black');
set(S(7),'facecolor','black');
set(S(8),'facecolor','black');
set(S(9),'facecolor','black');
set(S(10),'facecolor','black');
set(S(11),'facecolor','black');
set(S(12),'facecolor','black');
set(S(13),'facecolor','black');
set(S(14),'facecolor','black');
% 5j2w
set(S(15),'facecolor','white');
set(S(16),'facecolor','white');
set(S(17),'facecolor','[0 0.545 0.545]');
set(S(18),'facecolor','[0 0.545 0.545]');
set(S(19),'facecolor','[0 0.545 0.545]');
set(S(20),'facecolor','[0 0.545 0.545]');
set(S(21),'facecolor','[0 0.545 0.545]');
set(S(22),'facecolor','[0 0.545 0.545]');
set(S(23),'facecolor','[0 0.545 0.545]');
set(S(24),'facecolor','[0 0.545 0.545]');
set(S(25),'facecolor','[0 0.545 0.545]');
set(S(26),'facecolor','[0 0.545 0.545]');
set(S(27),'facecolor','[0 0.545 0.545]');
set(S(28),'facecolor','[0 0.545 0.545]');
% 1uui
set(S(29),'facecolor','white');
set(S(30),'facecolor','white');
set(S(31),'facecolor','white');
set(S(32),'facecolor','yellow');
set(S(33),'facecolor','yellow');
set(S(34),'facecolor','yellow');
set(S(35),'facecolor','yellow');
set(S(36),'facecolor','yellow');
set(S(37),'facecolor','yellow');
set(S(38),'facecolor','yellow');
set(S(39),'facecolor','yellow');
set(S(40),'facecolor','yellow');
set(S(41),'facecolor','yellow');
set(S(42),'facecolor','yellow');
% 5j1o
set(S(43),'facecolor','white');
set(S(44),'facecolor','white');
set(S(45),'facecolor','white');
set(S(46),'facecolor','white');
set(S(47),'facecolor','[0 0 1]');
set(S(48),'facecolor','[0 0 1]');
set(S(49),'facecolor','[0 0 1]');
set(S(50),'facecolor','[0 0 1]');
set(S(51),'facecolor','[0 0 1]');
set(S(52),'facecolor','[0 0 1]');
set(S(53),'facecolor','[0 0 1]');
set(S(54),'facecolor','[0 0 1]');
set(S(55),'facecolor','[0 0 1]');
set(S(56),'facecolor','[0 0 1]');
% 6d2u
set(S(57),'facecolor','white');
set(S(58),'facecolor','white');
set(S(59),'facecolor','white');
set(S(60),'facecolor','white');
set(S(61),'facecolor','white');
set(S(62),'facecolor','[0.823 0.706 0.549]');
set(S(63),'facecolor','[0.823 0.706 0.549]');
set(S(64),'facecolor','[0.823 0.706 0.549]');
set(S(65),'facecolor','[0.823 0.706 0.549]');
set(S(66),'facecolor','[0.823 0.706 0.549]');
set(S(67),'facecolor','[0.823 0.706 0.549]');
set(S(68),'facecolor','[0.823 0.706 0.549]');
set(S(69),'facecolor','[0.823 0.706 0.549]');
set(S(70),'facecolor','[0.823 0.706 0.549]');
% 1uud
set(S(71),'facecolor','white');
set(S(72),'facecolor','white');
set(S(73),'facecolor','white');
set(S(74),'facecolor','white');
set(S(75),'facecolor','white');
set(S(76),'facecolor','white');
set(S(77),'facecolor','[1 0 1]');
set(S(78),'facecolor','[1 0 1]');
set(S(79),'facecolor','[1 0 1]');
set(S(80),'facecolor','[1 0 1]');
set(S(81),'facecolor','[1 0 1]');
set(S(82),'facecolor','[1 0 1]');
set(S(83),'facecolor','[1 0 1]');
set(S(84),'facecolor','[1 0 1]');
% 1uts
set(S(85),'facecolor','white');
set(S(86),'facecolor','white');
set(S(87),'facecolor','white');
set(S(88),'facecolor','white');
set(S(89),'facecolor','white');
set(S(90),'facecolor','white');
set(S(91),'facecolor','white');
set(S(92),'facecolor','[0 1 1]');
set(S(93),'facecolor','[0 1 1]');
set(S(94),'facecolor','[0 1 1]');
set(S(95),'facecolor','[0 1 1]');
set(S(96),'facecolor','[0 1 1]');
set(S(97),'facecolor','[0 1 1]');
set(S(98),'facecolor','[0 1 1]');
% 2kx5
set(S(99),'facecolor','white');
set(S(100),'facecolor','white');
set(S(101),'facecolor','white');
set(S(102),'facecolor','white');
set(S(103),'facecolor','white');
set(S(104),'facecolor','white');
set(S(105),'facecolor','white');
set(S(106),'facecolor','white');
set(S(107),'facecolor','[0.498 0.498 0.498]');
set(S(108),'facecolor','[0.498 0.498 0.498]');
set(S(109),'facecolor','[0.498 0.498 0.498]');
set(S(110),'facecolor','[0.498 0.498 0.498]');
set(S(111),'facecolor','[0.498 0.498 0.498]');
set(S(112),'facecolor','[0.498 0.498 0.498]');
% 1qd3
set(S(113),'facecolor','white');
set(S(114),'facecolor','white');
set(S(115),'facecolor','white');
set(S(116),'facecolor','white');
set(S(117),'facecolor','white');
set(S(118),'facecolor','white');
set(S(119),'facecolor','white');
set(S(120),'facecolor','white');
set(S(121),'facecolor','white');
set(S(122),'facecolor','green');
set(S(123),'facecolor','green');
set(S(124),'facecolor','green');
set(S(125),'facecolor','green');
set(S(126),'facecolor','green');
% 5j0m
set(S(127),'facecolor','white');
set(S(128),'facecolor','white');
set(S(129),'facecolor','white');
set(S(130),'facecolor','white');
set(S(131),'facecolor','white');
set(S(132),'facecolor','white');
set(S(133),'facecolor','white');
set(S(134),'facecolor','white');
set(S(135),'facecolor','white');
set(S(136),'facecolor','white');
set(S(137),'facecolor','[0.545 0.271 0.074]');
set(S(138),'facecolor','[0.545 0.271 0.074]');
set(S(139),'facecolor','[0.545 0.271 0.074]');
set(S(140),'facecolor','[0.545 0.271 0.074]');
% 2kdq
set(S(141),'facecolor','white');
set(S(142),'facecolor','white');
set(S(143),'facecolor','white');
set(S(144),'facecolor','white');
set(S(145),'facecolor','white');
set(S(146),'facecolor','white');
set(S(147),'facecolor','white');
set(S(148),'facecolor','white');
set(S(149),'facecolor','white');
set(S(150),'facecolor','white');
set(S(151),'facecolor','white');
set(S(152),'facecolor','[1 0.753 0.796]');
set(S(153),'facecolor','[1 0.753 0.796]');
set(S(154),'facecolor','[1 0.753 0.796]');
% 1arj
set(S(155),'facecolor','white');
set(S(156),'facecolor','white');
set(S(157),'facecolor','white');
set(S(158),'facecolor','white');
set(S(159),'facecolor','white');
set(S(160),'facecolor','white');
set(S(161),'facecolor','white');
set(S(162),'facecolor','white');
set(S(163),'facecolor','white');
set(S(164),'facecolor','white');
set(S(165),'facecolor','white');
set(S(166),'facecolor','white');
set(S(167),'facecolor','red');
set(S(168),'facecolor','red');
% 2l8h
set(S(169),'facecolor','white');
set(S(170),'facecolor','white');
set(S(171),'facecolor','white');
set(S(172),'facecolor','white');
set(S(173),'facecolor','white');
set(S(174),'facecolor','white');
set(S(175),'facecolor','white');
set(S(176),'facecolor','white');
set(S(177),'facecolor','white');
set(S(178),'facecolor','white');
set(S(179),'facecolor','white');
set(S(180),'facecolor','white');
set(S(181),'facecolor','white');
set(S(182),'facecolor','[0.294 0 0.51]]');
% 1lvj
set(S(183),'facecolor','white');
set(S(184),'facecolor','white');
set(S(185),'facecolor','white');
set(S(186),'facecolor','white');
set(S(187),'facecolor','white');
set(S(188),'facecolor','white');
set(S(189),'facecolor','white');
set(S(190),'facecolor','white');
set(S(191),'facecolor','white');
set(S(192),'facecolor','white');
set(S(193),'facecolor','white');
set(S(194),'facecolor','white');
set(S(195),'facecolor','white');
set(S(196),'facecolor','white');
```

B.4.3 Dihedral Angle Plot of a Bulge Nucleotide

The following GNUPLOT script generates an individual dihedral angle plot (see Figure 3.7A).

```
###GNUPLOT SCRIPT
#!/usr/bin/gnuplot
set encoding iso_8859_1
set term post eps enh color size 20,12 "HelveticaBold" 50 solid
set output "1arj_u23.eps"
unset key
unset tics
unset border

#set style fill  transparent solid 0.3 noborder
set style circle radius 0.0025
set xrange [-0.3:2.05]
set yrange [-230:190]
set arrow 1 from 0,-180 to 2,-180 nohead lw 8
set arrow 2 from 0,-180 to 0,180 nohead lw 8
set arrow 3 from 0,0 to -0.015,0 nohead lw 6
set arrow 4 from 0,60 to -0.03,60 nohead lw 6
set arrow 5 from 0,120 to -0.015,120 nohead lw 6
set arrow 6 from 0,179.5 to -0.03,179.5 nohead lw 6
set arrow 7 from 0,-60 to -0.03,-60 nohead lw 6
set arrow 8 from 0,-120 to -0.015,-120 nohead lw 6
set arrow 9 from 0,-179.5 to -0.03,-179.5 nohead lw 6

set arrow 11 from 1.999,-180 to 1.999,-188 nohead lw 6
set arrow 13 from 0.999,-180 to 0.999,-188 nohead lw 6
set arrow 14 from 0.5,-180 to 0.5,-188 nohead lw 6
set arrow 15 from 1.5,-180 to 1.5,-188 nohead lw 6
set arrow 19 from 1.75,-180 to 1.75,-184 nohead lw 6
set arrow 18 from -0.001,-180 to -0.001,-188 nohead lw 6
set arrow 20 from 0.25,-180 to 0.25,-184 nohead lw 6
set arrow 21 from 1.25,-180 to 1.25,-184 nohead lw 6
set arrow 22 from 0.75,-180 to 0.75,-184 nohead lw 6

#set arrow 16 from 0,60 to 2,60  nohead lw 12 dt 2
#set arrow 17 from 0,-60 to 2,-60  nohead lw 12 dt 2

#set label 23 "1ARJ" at 0.01,188 font "HelveticaBold,70" tc rgb "red"
set label 24 "U" at 2.01,182 font "HelveticaBold,120" tc rgb "#ff8080"
set label 25 "L" at 2.01,155 font "HelveticaBold,120" tc rgb "#e60000"

set label 2 "60" at -0.14,60 font "HelveticaBold,100"
set label 4 "180" at -0.19,180 font "HelveticaBold,100"
set label 5 "-60" at -0.19,-60 font "HelveticaBold,100"
set label 7 "-180" at -0.23,-180 font "HelveticaBold,100"
#set label 8 "2" at 1.98,-200 font "HelveticaBold,100"
#set label 9 "1" at 0.98,-200 font "HelveticaBold,100"
```

```
#set label 10 "0" at -0.02,-200 font "HelveticaBold,100"
#set label 13 "1.5" at 1.45,-200 font "HelveticaBold,100"
#set label 14 "0.5" at 0.45,-200 font "HelveticaBold,100"
set label 15 "inward" at 2.075,-50 front font "Helvetica-Italic,120"
tc rgb "#595959" rotate by 90

set object 1 rect from 0,-60 to 2,60 fc rgb "#a5a5a5" fs solid 0.2
noborder
set arrow 23 from 2.01,-58 to 2.01,58 heads size 0.02,30 lw 6
lt rgb "#595959"

set label 1000 "{/ZapfDingbats \154}" at -0.027,50 front
font "HelveticaBold,90" tc rgb "#cd0000"
#set object 2 circle at screen 1,-90 size screen 30 fc rgb "#cd0000"
##c49c67

set label 11 "{/Symbol \161} ({\260})" at -0.3,-45
font "HelveticaBold,180" rotate by 90

p  "arj_u23_unl.dat" u ($1*0.00002):2  w circles lt rgb "red"
fs solid 0.5 noborder notitle, \
"arj_u23_lig.dat" u ($1*0.00002):2 w circles lt rgb "#e60000"
fs solid 1 noborder notitle, \
```

B.4.4 Dihedral Angle Plot of a Bulge Nucleotide

The following MATLAB script generates a backbone torsion angle plot (see Figure 3.2A).

```
###MATLAB SCRIPT
theta_1 = linspace(-1.92,-0.96);
rho_1 = theta_1-theta_1+1;
alpha_1 = polarplot(theta_1,rho_1,’LineWidth’,10,’Color’,’red’);
hold on
theta_2 = linspace(0.96,1.48);
rho_2 = theta_2 - theta_2 + 1;
alpha_2 = polarplot(theta_2,rho_2,’LineWidth’,10,’Color’,’red’);
set(gca,’box’,’off’,’TickDir’,’out’,’fontweight’,’bold’,’fontsize’
,40,’linewidth’,3);
ax=gca;
ax.RMinorTick=’off’;
ax.RTickLabelMode=’manual’;
ax.RTickLabel=’{}’;
ax.RTickMode=’manual’;
ax.RTick=[1 2 3 4 5 6 7];
ax.ThetaTickLabel={’0’; ’’; ’’; ’90’; ’’; ’’; ’180’;
’’; ’’; ’-90’;’’;’’;’’};


theta_3 = linspace(-3.14,-2.62);
rho_3 = theta_3-theta_3+2;
beta_1 = polarplot(theta_3,rho_3,’LineWidth’,10,’Color’,’blue’);
```

```
theta_4 = linspace(2.53,3.14);
rho_4 = theta_4-theta_4+2;
beta_2 = polarplot(theta_4,rho_4,'LineWidth',10,'Color','blue');

theta_5 = linspace(0.7,1.4);
rho_5 = theta_5-theta_5+3;
gamma = polarplot(theta_5,rho_5,'LineWidth',10,'Color','green');

theta_6 = linspace(0.96,1.83);
rho_6 = theta_6-theta_6+4;
delta_1 = polarplot(theta_6,rho_6,'LineWidth',10,'Color','cyan');

theta_7 = linspace(2.09,2.62);
rho_7 = theta_7-theta_7+4;
delta_2 = polarplot(theta_7,rho_7,'LineWidth',10,'Color','cyan');

theta_8 = linspace(-3.14,-2.09);
rho_8 = theta_8-theta_8+5;
epsilon_1 = polarplot(theta_8,rho_8,'LineWidth',10,'Color','magenta');

theta_9 = linspace(-1.83,-1.22);
rho_9 = theta_9-theta_9+5;
epsilon_2 = polarplot(theta_9,rho_9,'LineWidth',10,'Color','magenta');

theta_10 = linspace(2.97,3.14);
rho_10 = theta_10-theta_10+5;
epsilon_3 = polarplot(theta_10,rho_10,'LineWidth',10,'Color','magenta');

theta_11 = linspace(-1.83,-0.7);
rho_11 = theta_11-theta_11+6;
zeta_1 = polarplot(theta_11,rho_11,'LineWidth',10,'Color','[1 0.84 0]');

theta_12 = linspace(0.96,1.57);
rho_12 = theta_12-theta_12+6;
zeta_2 = polarplot(theta_12,rho_12,'LineWidth',10,'Color','[1 0.84 0]');

theta_13 = linspace(-3.14,-1.4);
rho_13 = theta_13-theta_13+7;
chi_1 = polarplot(theta_13,rho_13,'LineWidth',10,'Color','black');

theta_14 = linspace(3.05,3.14);
rho_14 = theta_14-theta_14+7;
chi_2 = polarplot(theta_14,rho_14,'LineWidth',10,'Color','black');

%%%% Experimental lines
%%%% Alpha
theta_15 = linspace(-0.39,-2);
rho_15 = theta_15-theta_15+1;
alpha_15 = polarplot(theta_15,rho_15,'LineWidth',15,'Color',
'[0.498 0.498 0.498]');
alpha_15.Color(4) = 0.4;
theta_16 = linspace(2.53,2.79);
rho_16 = theta_16-theta_16+1;
alpha_16 = polarplot(theta_16,rho_16,'LineWidth',15,'Color',
’[0.498 0.498 0.498]’);
alpha_16.Color(4) = 0.4;
theta_17 = linspace(1,1.05);
rho_17 = theta_17-theta_17+1;
alpha_17 = polarplot(theta_17,rho_17,’LineWidth’,15,’Color’,
’[0.498 0.498 0.498]’);
alpha_17.Color(4) = 0.4;
theta_18 = linspace(1.35,1.4);
rho_18 = theta_18-theta_18+1;
alpha_18 = polarplot(theta_18,rho_18,’LineWidth’,15,’Color’,
’[0.498 0.498 0.498]’);
alpha_18.Color(4) = 0.4;

%%%% Beta
theta_19 = linspace(2.53,3.14);
rho_19 = theta_19-theta_19+2;
beta_19 = polarplot(theta_19,rho_19,’LineWidth’,15,’Color’,
’[0.498 0.498 0.498]’);
beta_19.Color(4) = 0.4;
theta_20 = linspace(-2.79,-3.14);
rho_20 = theta_20-theta_20+2;
beta_20 = polarplot(theta_20,rho_20,’LineWidth’,15,’Color’,
’[0.498 0.498 0.498]’);
beta_20.Color(4) = 0.4;
theta_21 = linspace(1.44,1.53);
rho_21 = theta_21-theta_21+2;
beta_21 = polarplot(theta_21,rho_21,’LineWidth’,15,’Color’,
’[0.498 0.498 0.498]’);
beta_21.Color(4) = 0.4;
theta_22 = linspace(1.74,1.88);
rho_22 = theta_22-theta_22+2;
beta_22 = polarplot(theta_22,rho_22,’LineWidth’,15,’Color’,
’[0.498 0.498 0.498]’);
beta_22.Color(4) = 0.4;
theta_23 = linspace(-2.7,-2.75);
rho_23 = theta_23-theta_23+2;
beta_23 = polarplot(theta_23,rho_23,’LineWidth’,15,’Color’,
’[0.498 0.498 0.498]’);
beta_23.Color(4) = 0.4;
theta_24 = linspace(-1.96,-2);
rho_24 = theta_24-theta_24+2;
beta_24 = polarplot(theta_24,rho_24,’LineWidth’,15,’Color’,
’[0.498 0.498 0.498]’);
beta_24.Color(4) = 0.4;

%%%% Gamma
theta_25 = linspace(0.39,1.57);
rho_25 = theta_25-theta_25+3;
gamma_25 = polarplot(theta_25,rho_25,’LineWidth’,15,’Color’,
’[0.498 0.498 0.498]’);
gamma_25.Color(4) = 0.4;
theta_26 = linspace(2.97,3.14);
rho_26 = theta_26-theta_26+3;
gamma_26 = polarplot(theta_26,rho_26,’LineWidth’,15,’Color’,
'[0.498 0.498 0.498]');
gamma_26.Color(4) = 0.4;
theta_27 = linspace(-2.79,-3.14);
rho_27 = theta_27-theta_27+3;
gamma_27 = polarplot(theta_27,rho_27,'LineWidth',15,'Color',
'[0.498 0.498 0.498]');
gamma_27.Color(4) = 0.4;
theta_28 = linspace(0,-0.04);
rho_28 = theta_28-theta_28+3;
gamma_28 = polarplot(theta_28,rho_28,'LineWidth',15,'Color',
'[0.498 0.498 0.498]');
gamma_28.Color(4) = 0.4;
theta_29 = linspace(0.31,0.35);
rho_29 = theta_29-theta_29+3;
gamma_29 = polarplot(theta_29,rho_29,'LineWidth',15,'Color',
'[0.498 0.498 0.498]');
gamma_29.Color(4) = 0.4;
theta_30 = linspace(1.66,1.7);
rho_30 = theta_30-theta_30+3;
gamma_30 = polarplot(theta_30,rho_30,'LineWidth',15,'Color',
'[0.498 0.498 0.498]');
gamma_30.Color(4) = 0.4;
theta_31 = linspace(1.96,2);
rho_31 = theta_31-theta_31+3;
gamma_31 = polarplot(theta_31,rho_31,'LineWidth',15,'Color',
'[0.498 0.498 0.498]');
gamma_31.Color(4) = 0.4;
theta_32 = linspace(2.27,2.31);
rho_32 = theta_32-theta_32+3;
gamma_32 = polarplot(theta_32,rho_32,'LineWidth',15,'Color',
'[0.498 0.498 0.498]');
gamma_32.Color(4) = 0.4;
theta_33 = linspace(2.79,2.83);
rho_33 = theta_33-theta_33+3;
gamma_33 = polarplot(theta_33,rho_33,'LineWidth',15,'Color',
'[0.498 0.498 0.498]');
gamma_33.Color(4) = 0.4;

%%%% Delta
theta_34 = linspace(1.13,1.74);
rho_34 = theta_34-theta_34+4;
delta_34 = polarplot(theta_34,rho_34,'LineWidth',15,'Color',
'[0.498 0.498 0.498]');
delta_34.Color(4) = 0.4;
theta_35 = linspace(2.53,2.71);
rho_35 = theta_35-theta_35+4;
delta_35 = polarplot(theta_35,rho_35,'LineWidth',15,'Color',
'[0.498 0.498 0.498]');
delta_35.Color(4) = 0.4;

%%%% Epsilon
theta_36 = linspace(-3.14,-3.05);
rho_36 = theta_36-theta_36+5;
epsilon_36 = polarplot(theta_36,rho_36,'LineWidth',15,'Color',
'[0.498 0.498 0.498]');
epsilon_36.Color(4) = 0.4;
theta_37 = linspace(-2.88,-2.27);
rho_37 = theta_37-theta_37+5;
epsilon_37 = polarplot(theta_37,rho_37,'LineWidth',15,'Color',
'[0.498 0.498 0.498]');
epsilon_37.Color(4) = 0.4;
theta_38 = linspace(-1.92,-2.05);
rho_38 = theta_38-theta_38+5;
epsilon_38 = polarplot(theta_38,rho_38,'LineWidth',15,'Color',
'[0.498 0.498 0.498]');
epsilon_38.Color(4) = 0.4;
theta_39 = linspace(-1.57,-1.61);
rho_39 = theta_39-theta_39+5;
epsilon_39 = polarplot(theta_39,rho_39,'LineWidth',15,'Color',
'[0.498 0.498 0.498]');
epsilon_39.Color(4) = 0.4;

%%%% Zeta
theta_40 = linspace(-0.87,-1.74);
rho_40 = theta_40-theta_40+6;
zeta_40 = polarplot(theta_40,rho_40,'LineWidth',15,'Color',
'[0.498 0.498 0.498]');
zeta_40.Color(4) = 0.4;
theta_41 = linspace(2.88,3.14);
rho_41 = theta_41-theta_41+6;
zeta_41 = polarplot(theta_41,rho_41,'LineWidth',15,'Color',
'[0.498 0.498 0.498]');
zeta_41.Color(4) = 0.4;
theta_42 = linspace(-1.96,-1.88);
rho_42 = theta_42-theta_42+6;
zeta_42 = polarplot(theta_42,rho_42,'LineWidth',15,'Color',
'[0.498 0.498 0.498]');
zeta_42.Color(4) = 0.4;
theta_43 = linspace(-3.14,-3.05);
rho_43 = theta_43-theta_43+6;
zeta_43 = polarplot(theta_43,rho_43,'LineWidth',15,'Color',
'[0.498 0.498 0.498]');
zeta_43.Color(4) = 0.4;
theta_44 = linspace(1.44,1.48);
rho_44 = theta_44-theta_44+6;
zeta_44 = polarplot(theta_44,rho_44,'LineWidth',15,'Color',
'[0.498 0.498 0.498]');
zeta_44.Color(4) = 0.4;
theta_45 = linspace(0.83,0.87);
rho_45 = theta_45-theta_45+6;
zeta_45 = polarplot(theta_45,rho_45,'LineWidth',15,'Color',
'[0.498 0.498 0.498]');
zeta_45.Color(4) = 0.4;
theta_46 = linspace(0.17,0.22);
rho_46 = theta_46-theta_46+6;
zeta_46 = polarplot(theta_46,rho_46,'LineWidth',15,'Color',
'[0.498 0.498 0.498]');
zeta_46.Color(4) = 0.4;
```

```
theta_47 = linspace(-0.78,-0.74);
rho_47 = theta_47-theta_47+6;
zeta_47 = polarplot(theta_47,rho_47,'LineWidth',15,'Color',
'[0.498 0.498 0.498]');
zeta_47.Color(4) = 0.4;
theta_48 = linspace(-0.52,-0.57);
rho_48 = theta_48-theta_48+6;
zeta_48 = polarplot(theta_48,rho_48,'LineWidth',15,'Color',
'[0.498 0.498 0.498]');
zeta_48.Color(4) = 0.4;

%%%% Chi
theta_49 = linspace(-2.44,-3.14);
rho_49 = theta_49-theta_49+7;
chi_49 = polarplot(theta_49,rho_49,'LineWidth',20,'Color',
'[0.498 0.498 0.498]');
chi_49.Color(4) = 0.4;
theta_50 = linspace(3.1,3.14);
rho_50 = theta_50-theta_50+7;
chi_50 = polarplot(theta_50,rho_50,'LineWidth',20,'Color',
'[0.498 0.498 0.498]');
chi_50.Color(4) = 0.4;
theta_51 = linspace(2.97,3.05);
rho_51 = theta_51-theta_51+7;
chi_51 = polarplot(theta_51,rho_51,'LineWidth',20,'Color',
'[0.498 0.498 0.498]');
chi_51.Color(4) = 0.4;
theta_52 = linspace(0.74,0.78);
rho_52 = theta_52-theta_52+7;
chi_52 = polarplot(theta_52,rho_52,'LineWidth',20,'Color',
'[0.498 0.498 0.498]');
chi_52.Color(4) = 0.4;
theta_53 = linspace(0.31,0.35);
rho_53 = theta_53-theta_53+7;
chi_53 = polarplot(theta_53,rho_53,'LineWidth',20,'Color',
'[0.498 0.498 0.498]');
chi_53.Color(4) = 0.4;
theta_54 = linspace(1.13,1.18);
rho_54 = theta_54-theta_54+7;
chi_54 = polarplot(theta_54,rho_54,'LineWidth',20,'Color',
'[0.498 0.498 0.498]');
chi_54.Color(4) = 0.4;
hold off
```

APPENDIX C

SUPPORTING INFORMATION FOR CHAPTER 4

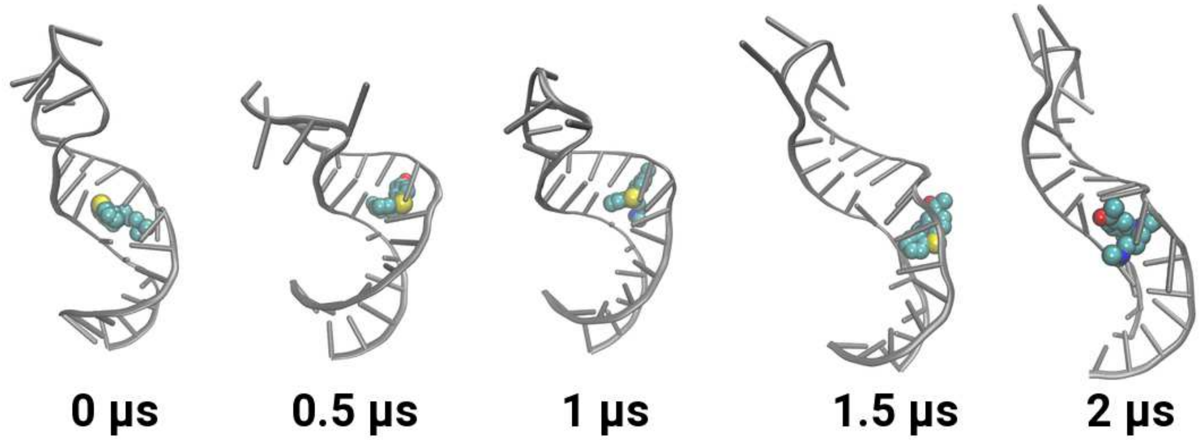

Figure C.1: **Long time-scale classical MD simulation:** Snapshots of RNA (gray cartoon) and ligand (space-filling) are shown at various timepoints from a 2 μs long classical MD simulation, where ligand remains stably bound.

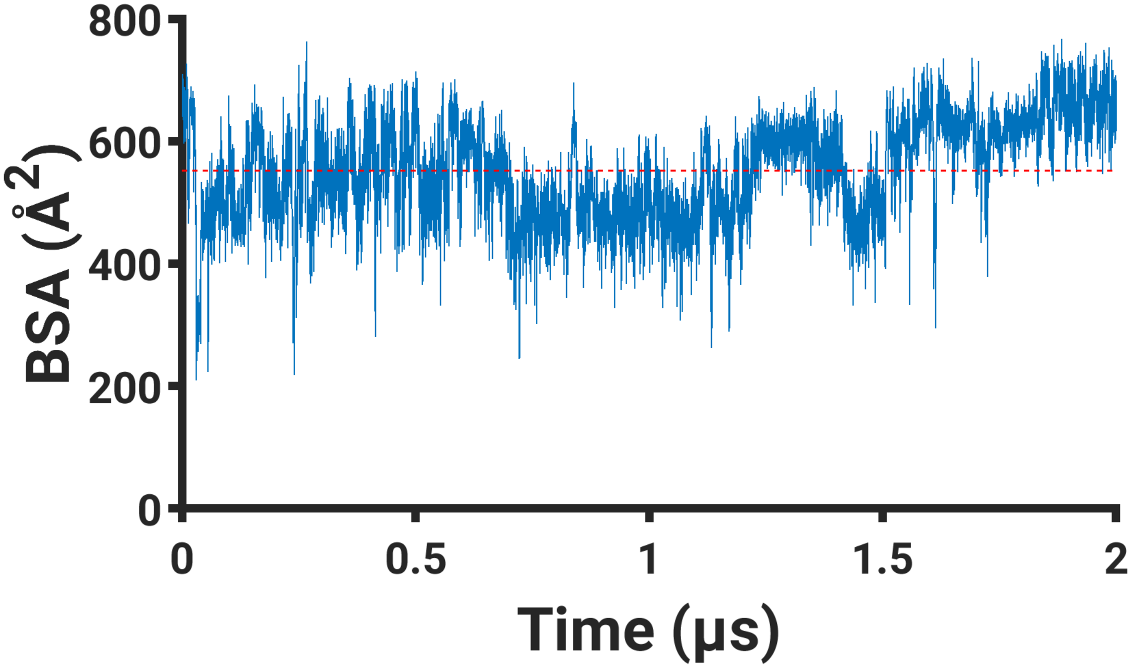

Figure C.2: The buried surface area (BSA) trace *vs.* time is shown for the ligand from a 2 μs long classical MD simulation (cf. Figure C.1). The dotted red line corresponds to the average BSA.

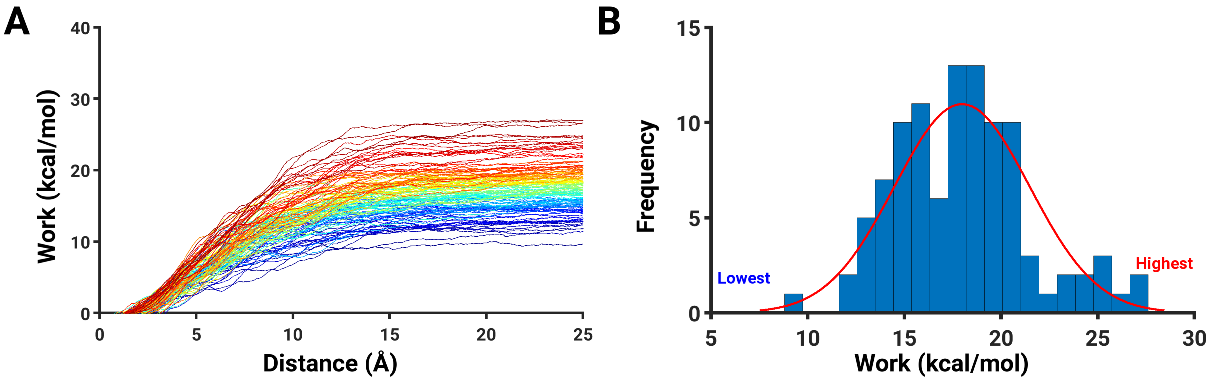

Figure C.3: **Ligand dissociation work from SMD simulations:** (A) work values are plotted along the reaction coordinate (RC) from 102 independent cv-SMD simulations. Blue to red color palette indicates lower to higher values of work required for ligand dissociation. (B) A histogram of all work values (at 25 Å) is shown with a best-fit distribution line (red trace).

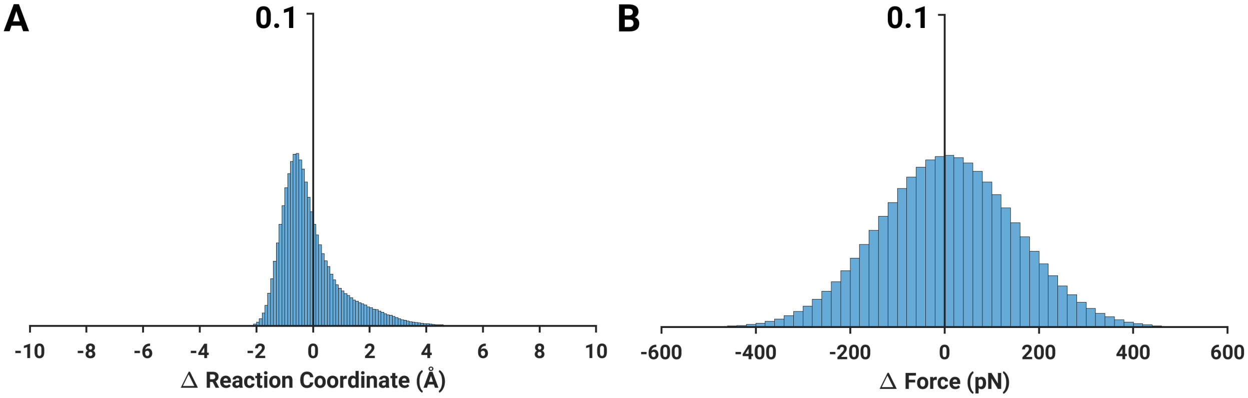

Figure C.4: **Reaction coordinate and force-convergence data:** Shown are (A) distributions of ΔRC values computed from the deviations of pathways from cv-SMD simulations with respect to the actual reaction coordinate. (B) distributions of ΔForce values after the force on average converges to 0 (at ∼17.5 Å along the reaction coordinate; cf. Figure 4.2B).

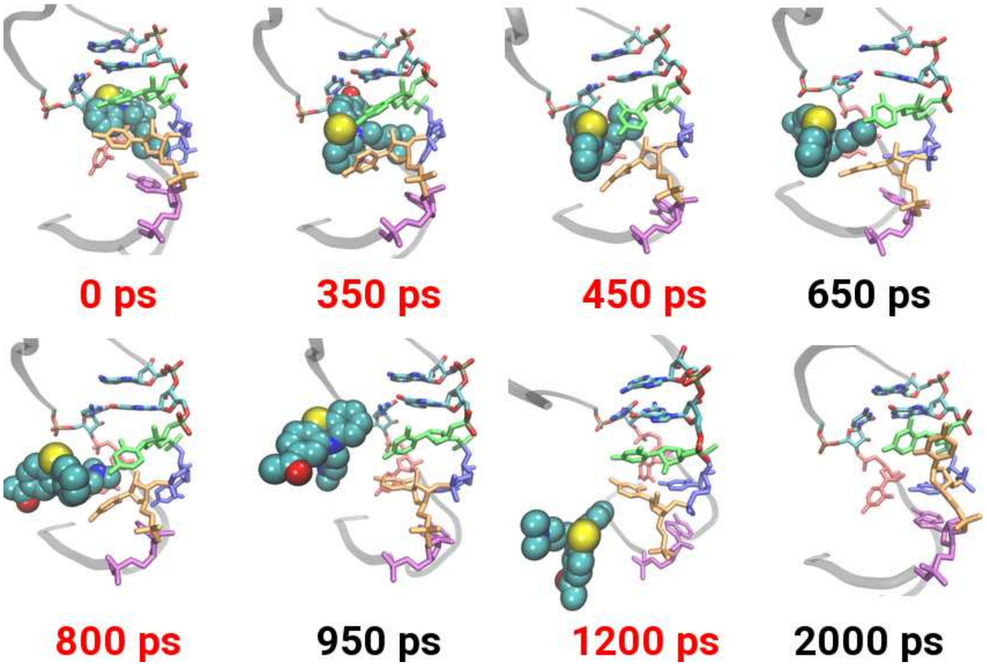

Figure C.5: **Ligand dissociation in the lowest work simulation:** Snapshots of the ligand dissociation process at various timepoints are shown. The time values highlighted in red are those snapshots that are also shown in Figure C.5. Key nucleotides are highlighted: A22 (purple), U23 (orange), C24 (blue), U25 (green) and U40 (red). Ligand is shown in a space-filling representation.

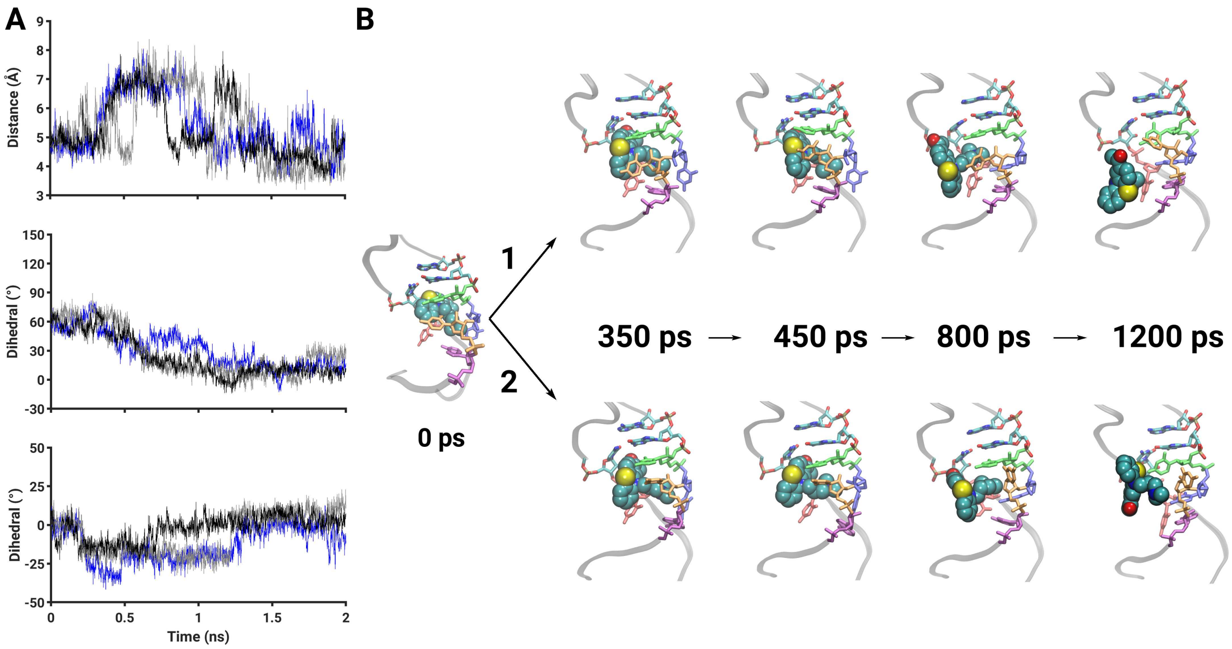

Figure C.6: **Conformational metrics and ligand dissociation mechanism from two additional simulations with lower work values:** (A) Same metrics as in Figure 4.4 are presented. The traces correspond to the lowest work simulation (blue trace with $W =$ 9.56 kcal/mol) and two additional simulations with the next lower values of work (black and gray traces with $W = 12.59$ and 12.76 kcal/mol, respectively). (B) Snapshots of ligand dissociation from two additional lower work simulations are shown. Color and labeling scheme is same as in Figures 4.1 and 4.3.

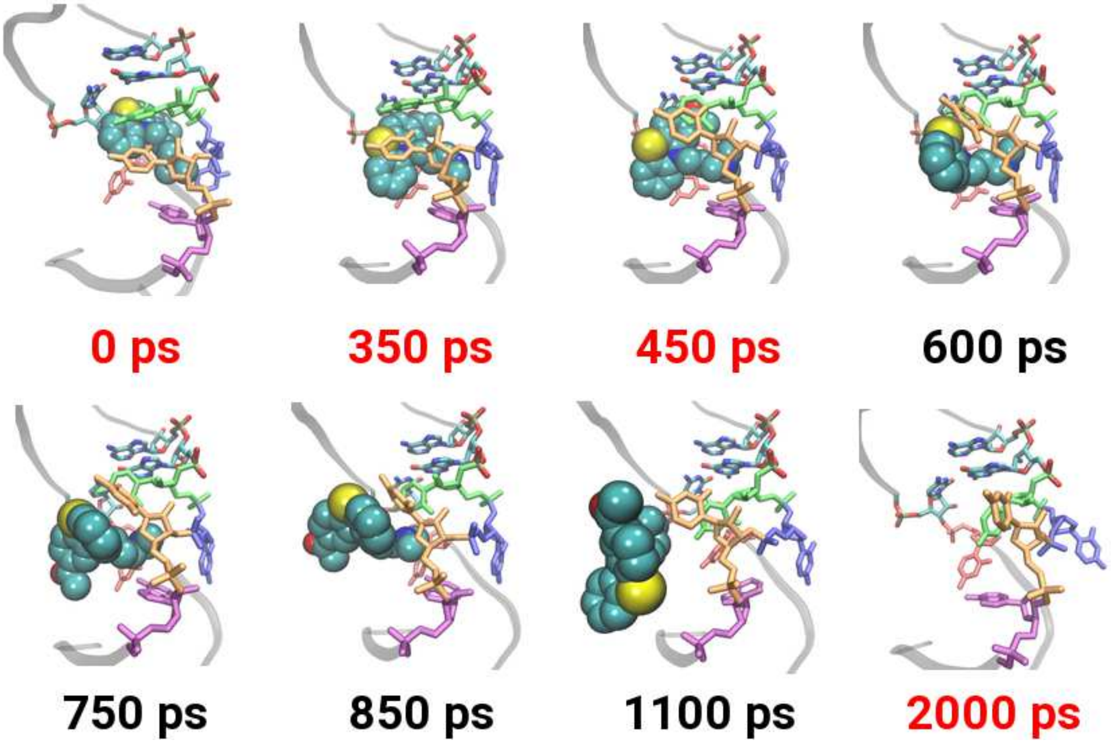

Figure C.7: **Ligand dissociation in the highest work simulation:** Snapshots of the ligand dissociation process at various timepoints are shown. cf. Figure 4.3 and Figure C.5 for other details.

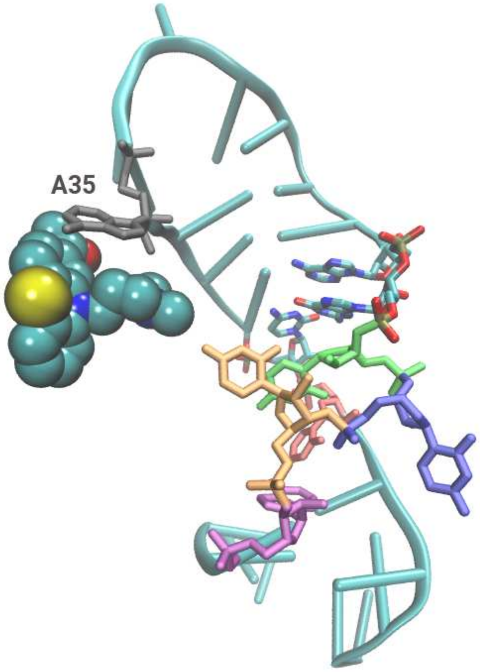

Figure C.8: A side-view snapshot of the ligand interacting with the A35 nucleotide (gray sticks) is shown. Other nucleotides shown are the same as in Figure 4.1C and Figure 4.3.

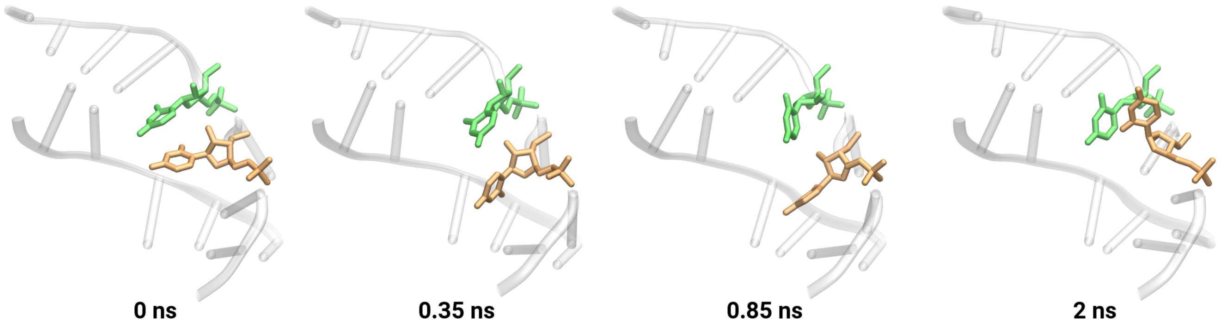

Figure C.9: Shown are representative snapshots of TAR-RNA from the lowest work simulation highlighting nucleotides U23 (orange) and U25 (green).

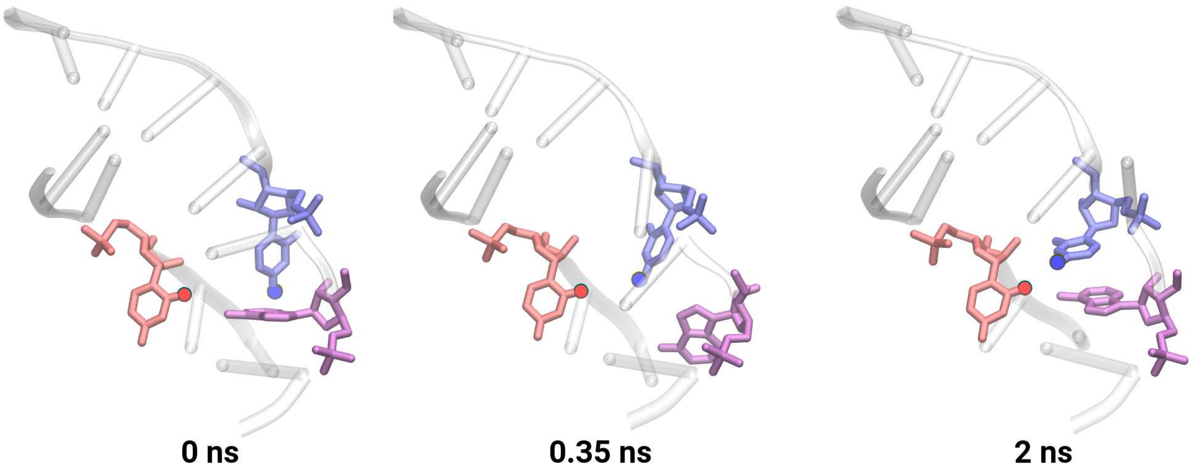

Figure C.10: Representative snapshots from the lowest work simulation highlighting A22 (purple), C24 (blue) and U40 (red). The dotted line indicates a hydrogen bond between C24 and U40.

APPENDIX D
SCRIPTS FOR CHAPTER 4

D.1 Overview

In this appendix, I provide the scripts that I have used to set up and analyze simulations in chapter 4 using various software packages.

D.2 cv-SMD Simulation using NAMD

I begin setting up a cv-SMD simulation by placing the system at the origin:

```
set sel [atomselect top all]
set ligand [atomselect top "resname PMZ"]
set gec [measure center ligand]
$sel moveby [vecscale -1.0 $gec]
```

An example script to rotate the system and to align the pulling direction along a z-axis:

```
set sel [atomselect top all]
set com [measure center $sel weight mass]
set matrix [transaxis y 180]
$sel moveby [vecscale -1.0 $com]
$sel move $matrix
$sel moveby $com
$sel writepdb rna_ligand.pdb  % Save the final orientation to a pdb file
```

An example script to generate PSF and PDB files for the peptide. It can be executed from the terminal window: **vmd -dispdev text -e psfgen.pgn**.

```
package require psfgen
topology top_all36_na.rtf
segment K {pdb rna.pdb}
coordpdb rna.pdb K
guesscoord
```

```
writepdb rna_fin.pdb
writepsf rna_fin.psf
exit
```

A script to solvate a system in VMD with an extended z-dimension:

```
package require solvate
solvate rna_drug.psf rna_drug.pdb -o rna_drug_wat -x 12.5 +x 12.5
-y 12.5 +y 12.5 -z 12.5 +z 50
```

A script to ionize a system in VMD with sodium ions:

```
package require autoionize
autoionize -psf rna_drug_wat.psf -pdb rna_drug_wat.pdb -neutralize
-cation SOD
```

Below is the script to determine the vector of the pulling direction and to label the atoms that will be pulled in the cv-SMD simulation.

```
set allatoms [atomselect top all]
$allatoms set beta 0
set fixedatoms [atomselect top "nucleic and name P"]
$fixedatoms set beta 1
$allatoms set occupancy 0
set smdatoms [atomselect top "resname PMZ"]
$smdatoms set occupancy 1
$allatoms writepdb lvj_smd.ref


set smdpos [lindex [$smdatoms get {x y z}] 0]
set z_dir_pos {6.7093586921691895 0.4800395369529724 53}
vecnorm [vecsub $smdpos $z_dir_pos]
```

After that, I used similar scripts from section B.2 to generate the Amber parameter/topology files. The script below shows an example of an SMD configuration file to conduct an SMD simulation.

```
#####################################################
coordinates   lvj_smd.pdb

set temperature   310
set outputname    lvj_smd_run
firsttimestep     0

#Continuing a run
#set inputname   lvj_smd_eq.restart ; ### Edit with previous restart file
```

```
#binCoordinates $inputname.coor;
#binVelocities  $inputname.vel;
#extendedSystem $inputname.xsc;

#Simulation Parameters
#Input
amber         on              ;##### Enables Amber format force field
parmfile      lvj_smd.prmtop  ;##### Amber parameter/topology file
temperature   $temperature

#Force field parameters
exclude         scaled1-4
1-4scaling      1.0
cutoff          10
switching       on
switchdist      8
pairlistdist    12

#Integrator Parameters
timestep              2.0   #2 fs/step
rigidBonds            all
nonBondedFreq         1
fullElectFrequency    2
stepspercycle         10


# Periodic Boundary Conditions
if {1} {
cellBasisVector1    65.0    0.0   0.0
cellBasisVector2     0.0   79.0  0.0
cellBasisVector3     0.0    0.0   133.0
cellOrigin           0.43169 -1.24704 29.93577
}
wrapAll on

#Electrostatic Force Evaluation (PME)
PME             yes
PMEGridSizeX    72
PMEGridSizeY    80
PMEGridSizeZ    135

# Constant Temperature Control
langevin           off;          ### The temperature should be off in SMD

# Constant Pressure Control
if {1} {
useGroupPressure      yes
useFlexibleCell       no
useConstantArea       no

langevinPiston        on
langevinPistonTarget  1.01325 ;#  in bar -> 1 atm
langevinPistonPeriod  100.0
langevinPistonDecay   50.0
langevinPistonTemp    $temperature
}
##############SMD Settings##############
if {1} {                     ;### Atoms are held fixed during a simulation
fixedAtoms       on          ;### if this option is on;
fixedAtomsFile   lvj_smd.ref ;### File containing fixed atom parameters;
fixedAtomsCol    B           ;### Column of the file containing fixed
}                            ;### atom parameters;

SMD on                       ;### Enables SMD simulation
SMDFile lvj_smd.ref          ;### SMD constraint reference file
SMDk 7                       ;### Force constant value
SMDVel 0.000025              ;### Velocity of the SMD pulling
SMDDir -0.0491 0.04992 0.997;### Direction of the SMD movement
SMDOutputFreq 10

########################################

#Output
outputName          $outputname
restartfreq         2500        ;# 5000steps = every 10ps
dcdfreq             1000
xstFreq             1000
outputEnergies      1000
outputPressure      1000

#Execution
reinitvels          $temperature
run     2500000
```

I conducted all cv-SMD simulations on the XSEDE supercomputer in San Diego (Comet). Below, I provide a jobscript file.

```
#!/bin/sh

#SBATCH --job-name="lvj_6"
#SBATCH --output="namd.%j.%N.out"
#SBATCH --partition=compute
#SBATCH --nodes=2
#SBATCH --ntasks-per-node=24
#SBATCH --export=ALL
#SBATCH -t 48:00:00
#SBATCH -A dxu114

module load namd/2.13

#Run NAMD job using mpirun_rsh
PROCS=$(( $SLURM_NNODES * $SLURM_NTASKS_PER_NODE))
NODEFILE=‘generate_pbs_nodefile‘
exe=‘which namd2‘

#mpirun_rsh
```

```
mpirun_rsh -export-all -hostfile $NODEFILE -np 24 $exe npt_cv.conf
```

D.3 Analysis Scripts

In this section, I provide all the scripts that I have generated to analyze SMD simulations.

D.3.1 Script to Compute Force

```
###VMD SCRIPT
### Open the log file for reading and the output .dat file for writing
set imax 1
set work_J 0
for {set i 0} { $i < $imax } {incr i} {
set file [open namd.log  r]
set output [open work.dat w]
### Gather input from user.
set nx {-0.0490834900098211}
set ny {0.0499167955671872}
set nz {0.9975465525622148}
set initx {7.27452}
set inity {-0.748136}
set initz {-5.76629}
### Loop over all lines of the log file
set file [open namd.log r]

while { [gets $file line] != -1 } {

### Determine if a line contains SMD output. If so, write the
### timestep followed by f(dot)n to the output file
  if {[lindex $line 0] == "SMD"} {
### Input
set time {0.02}
set vel [expr 0.0125*pow(10,-10)]
set dist [expr sqrt( ($initx - [lindex $line 2])*($initx -
[lindex $line 2]) + ($inity - [lindex $line 3])*($inity -
[lindex $line 3]) + ($initz - [lindex $line 4])*($initz -
[lindex $line 4]))]
set f [expr $nx*[lindex $line 5] + $ny*[lindex $line 6] +
$nz*[lindex $line 7]]
      puts $output "[lindex $line 1]  $dist $f"
    }
  }
}
### Close the log file and the output .dat file
close $file
close $output

exit
```

D.3.2 Script to Compute Work and PMF

```
###MATLAB SCRIPT

clear
ac=.015;
pa=’loess’;
tt=1000000;

cd(’/run/media/levintov/easystore/smd/1lvj/smd_1lvj_1’);
fid = fopen(’ft.dat’,’r’);
A = textscan(fid, ’%f %f’); %,’headerlines’,4
fclose(fid);
data=cell2mat(A(2));
step=cell2mat(A(1));

for i=2:102
cd([’/run/media/levintov/easystore/smd/1lvj/smd_1lvj_’,num2str(i)]);
fid = fopen(’ft.dat’,’r’);
A = textscan(fid, ’%f %f’); %,’headerlines’,4
fclose(fid);
data=[data cell2mat(A(2))];
end

v=0.0000125/1e-15; % Angestrom per fs

for i=1:102
intdata(:,i)=cumtrapz(step.*1e-15,data(:,i))*v;%pN*Angestrom
end

w=intdata*1e-12*1e-10/4184; %kcal
T=310;
K=3.29982916e-27; % kilocalorie / kelvin
pmf=-K*T*log(mean(exp(-w/T/K),2));

NA=6.022e23; %avogardo’s
w=w*NA; %kcal/mol
pmf=pmf*NA; %kcal/mol
dist=step*0.000025; % step*Angestrom/step=>Angestrom
[ss, I]=sort(w(end,:));
sd = std(w,0,2);
out = [dist pmf sd];
h1=plot(dist,w(:,I))


a=gca;
a.Box=’off’;
for i=1:102
a.Children(i).LineWidth=2;
end
a.Children(1).LineWidth=4;
a.LineWidth=6;
```

```
a.TickDir='out';
a.FontSize=50;
a.FontWeight='Bold';
a.XTick=0:5:25;
a.YTick=0:10:60;
axis([0 25 0 40])
cmp=jet(102);
for line = 1:102
set(h1(line),'Color',cmp(line,:));
end
% xlabel(['Distance (',char(197),')'])
% ylabel(['W (kcal/mol)'])
% l=legend([h1.mainLine,h1.edge(1)],{'mean','std'});
% l.Box='off';
% l.Location='northeast';
% l.FontSize=14;
figure
h2=plot(dist,pmf,'k');
a=gca;
a.Box='off';
a.Children.LineWidth=4;
a.LineWidth=6;
a.TickDir='out';
a.FontSize=40;
a.FontWeight='Bold';
a.XTick=0:5:25;
a.YTick=-5:5:25;
axis([0 25 -5 25])
xlabel(['Distance (',char(197),')'])
ylabel(['PMF (kcal/mol)'])
```

D.3.3 Script to Compute Physical Variables

```
###VMD Script
mol new lvj_smd.prmtop

mol addfile 1lvj_low_work.dcd waitfor all
set out1 [open cv.dat "w"]

set nf [molinfo top get numframes]
set u7_n1 [atomselect top "index 211"]
set u7_c6 [atomselect top "index 212"]
set u7_c4 [atomselect top "index 216"]

set u9_n1 [atomselect top "index 272"]
set u9_c6 [atomselect top "index 273"]
set u9_c4 [atomselect top "index 277"]

set u7 [atomselect top "resid 7 and (name N1 or name C2 or name N3 or
name C4 or name C5 or name C6)"]
set u9 [atomselect top "resid 9 and (name N1 or name C2 or name N3 or
name C4 or name C5 or name C6)"]
```

```
global M_PI
for {set i 0} {$i < $nf} {incr i} {

$u7_n1 frame $i
$u7_c4 frame $i
$u7_c6 frame $i
$u9_n1 frame $i
$u9_c4 frame $i
$u9_c6 frame $i
$u7 frame $i
$u9 frame $i

set g1 [measure center $u7_n1 weight mass]
set g2 [measure center $u7_c6 weight mass]
set g3 [measure center $u7_c4 weight mass]

set g4 [measure center $u9_n1 weight mass]
set g5 [measure center $u9_c6 weight mass]
set g6 [measure center $u9_c4 weight mass]

set gu7 [measure center $u7 weight mass]
set gu9 [measure center $u9 weight mass]

set dA [vecsub $g1 $g2]
set dB [vecsub $g1 $g3]
set dC [vecsub $g4 $g5]
set dD [vecsub $g4 $g6]

set u7_normal [veccross $dA $dB]
set u9_normal [veccross $dC $dD]
set dot6 [vecdot $u7_normal $u9_normal]
set cosine6 [expr $dot6/([veclength $u7_normal]*[veclength $u9_normal])]

set 1 [measure dihed {302 297 236 241} frame $i]
set 2 [measure bond {247 766} frame $i]
set 3 [measure dihed {211 206 173 178} frame $i]
set 4 [measure dihed {178 173 139 144} frame $i]
set 5 [measure dihed {269 270 272 281} frame $i]
set 6 [expr acos($cosine6)*(180/$M_PI)]
set 7 [measure dihed {208 209 211 220} frame $i]
set 8 [vecdist $gu7 $gu9]
puts $out1 "$i $1 $2 $3 $4 $5 $6 $7 $8"
}
exit
```

D.4 Scripts to Generate Figures

In this section, I provide several example scripts to generate the plots presented in chapter 4.

D.4.1 Script to Plot Force Profile

```
###Gnuplot Script
#!/usr/bin/gnuplot
#40 dpd system
#Histogram of sign
set encoding iso_8859_1
set term post eps enh color size 18,11 "HelveticaBold" 70 solid

set output "average_f_1lvj.eps"

unset key
set xtics out nomirror scale 3.5
#set xzeroaxis
set xrange [0:25]
set yrange [-800:800]
set ytics out -800,200,800 nomirror scale 3.5 offset 0.5
set border 3 lw 12
Shadecolor = "#80E0A080"
#Shadecolor2 = "#B0C4DE"

set arrow 1 from 0,0 to 25,0 nohead front dt 2 lw 6
#set yrange [2:8]
#set xlabel "Distance (\305)" font "HelveticaBold,60"
#set ylabel "Force (pN)" font "HelveticaBold,60"

#p "./center_full_da/comb_d1.dat" u ($1*0.15):2 w l lw 2 lc rgb
"black" title "D_{1}", \
#  "./center_full_da/comb_d2.dat" u ($1*0.15):2 w l lw 2 lc rgb
"#0000FF" title "D_{2}",

p "f_average.dat" u 1:($2+$3):($2-$3) with filledcurve fc rgb
Shadecolor notitle, \
"f_average.dat" u 1:2 w l lw 2 lc rgb "black" notitle
```

D.4.2 Script to Plot Distance Profile

```
%%%MATLAB Script
cd(’/run/media/levintov/easystore/smd/1lvj/smd_1lvj_1’);
fid = fopen(’dist_pass.dat’,’r’);
A = textscan(fid, ’%f %f’); %,’headerlines’,4
fclose(fid);
data=cell2mat(A(2));
step=cell2mat(A(1));
time = step/500000;
for i=2:102
cd([’/run/media/levintov/easystore/smd/1lvj/smd_1lvj_’,num2str(i)]);
fid = fopen(’dist_pass.dat’,’r’);
A = textscan(fid, ’%f %f’); %,’headerlines’,4
fclose(fid);
```

```
data=[data cell2mat(A(2))];
end

[ss, I]=sort(data(end,:));
h1=plot(data,time);

a=gca;
a.Box='off';
for i=1:102
a.Children(i).LineWidth=2;
end
%a.Children(1).LineWidth=4;
a.LineWidth=6;
a.TickDir='out';
a.FontSize=50;
a.FontWeight='Bold';
xticks([]);
% a.YTick=0:0.5:2.5;
xlim([0 25]);
% ylim([0 2]);
% axis([0 25 0 2])
cmp=jet(102);
for line = 1:102
set(h1(line),'Color',cmp(line,:));
end
hold on
h2 = plot([0 25] , [0 2]);
h2.Color='black';
h2.LineWidth=4;
xlim([0 25]);
ylim([0 2]);
```

D.4.3 Script to Plot a Physical Variable

The following MATLAB script generates physical variables traces shown in Figure 4.4.

```
%%%MATLAB Script
CVs_low = importdata('smd_1lvj_28/cv.dat');
step_low1 = CVs_low(1:309,1);
cv2_low1 = CVs_low(1:309,3);
step_low2 = CVs_low(309:310,1);
cv2_low2 = CVs_low(309:310,3);
step_low3 = CVs_low(310:2000,1);
cv2_low3 = CVs_low(310:2000,3);

CVs_high = importdata('smd_1lvj_35/cv.dat');
step_high = CVs_high(:,1);
cv2_high = CVs_high(:,3);

time_low1 = step_low1/1000;
time_low2 = step_low2/1000;
time_low3 = step_low3/1000;
time_high = step_high/1000;
```

```

p1_1 = plot(time_low1,cv2_low1);
p1_1.Color = ’0.53 0.66 0.97’;
p1_1.LineWidth = 2;
hold on
p1_11 = plot(time_low2,cv2_low2);
p1_11.Color = ’blue’;
p1_11.LineWidth = 2;
p1_12 = plot(time_low3,cv2_low3);
p1_12.Color = ’blue’;
p1_12.LineWidth = 2;
p1_2 = plot(time_high,cv2_high);
p1_2.Color = ’0.94 0.72 0.73’;
p1_2.LineWidth = 2;
set(gca,’box’,’off’,’TickDir’,’out’,’fontweight’,’bold’,’fontsize’,50,
’linewidth’,8);
% legend({’Low Work’,’High Work’},’Location’,’northeastoutside’);
% legend(’boxoff’);
xlabel(’Time (ns)’);
hold off
```

APPENDIX E
SUPPORTING INFORMATION FOR CHAPTER 5

E.1 The RSG-1.2 peptide

The structure of the RSG-1.2 peptide was studied both in the presence and absence of the RNA using NMR and CD spectroscopy [93, 323]. In the presence of the RNA, the peptide adopts a partially α-helical configuration as highlighted by the NMR spectra [93]. In the absence of the RNA, the CD spectra showed that the RSG-1.2 peptide formed a predominantly disordered conformation (only 12% helicity) [93, 323]. I conducted a 100 ns MD simulation of the peptide in a simulation domain of TIP3P water molecules with Cl^-. As shown in Figure E.11A, the α-helix of the peptide was partially distorted in the course of the MD simulation. I also computed ϕ and ψ dihedral angles from the entire MD simulation of the amino acids that constitute the α-helix in the initial structure (Figure E.11B). I determined that the dihedral angles are mostly scattered in the fourth quadrant, which does not correspond to any particular state, and around $\psi = -50$ and $\phi = -60$, which corresponds to the α-helical conformation (red circle; Figure E.11B). Based on the MD data, I suggest that the peptide exists in the predominantly disordered conformation, which confirms the experimental observations, and through the interactions with the RRE RNA the peptide adopts the α-helical conformation.

Table E.1: **Details on all simulation systems.** The simulation details on all four pathways (PWs) are presented, including the dimensions of the simulation domain of each system (column labeled *system dimensions*), number of atoms (column labeled *system size*), pulling distance (column labeled *distance*), simulation time of a single run, and the number of runs. The system size is slightly distinct in each pathway due to the initial reorientation of the RNA-peptide complex and resolvation.

PW	system dimensions	system size (atoms)	distance (Å)	time/run (ns)	# runs
1	68 Å × 84 Å × 126 Å	63919	80	13	75
2	65 Å × 79 Å × 133 Å	61171	80	13	75
3	64 Å × 84 Å × 129 Å	61618	80	13	75
4	70 Å × 83 Å × 130 Å	66511	80	13	75

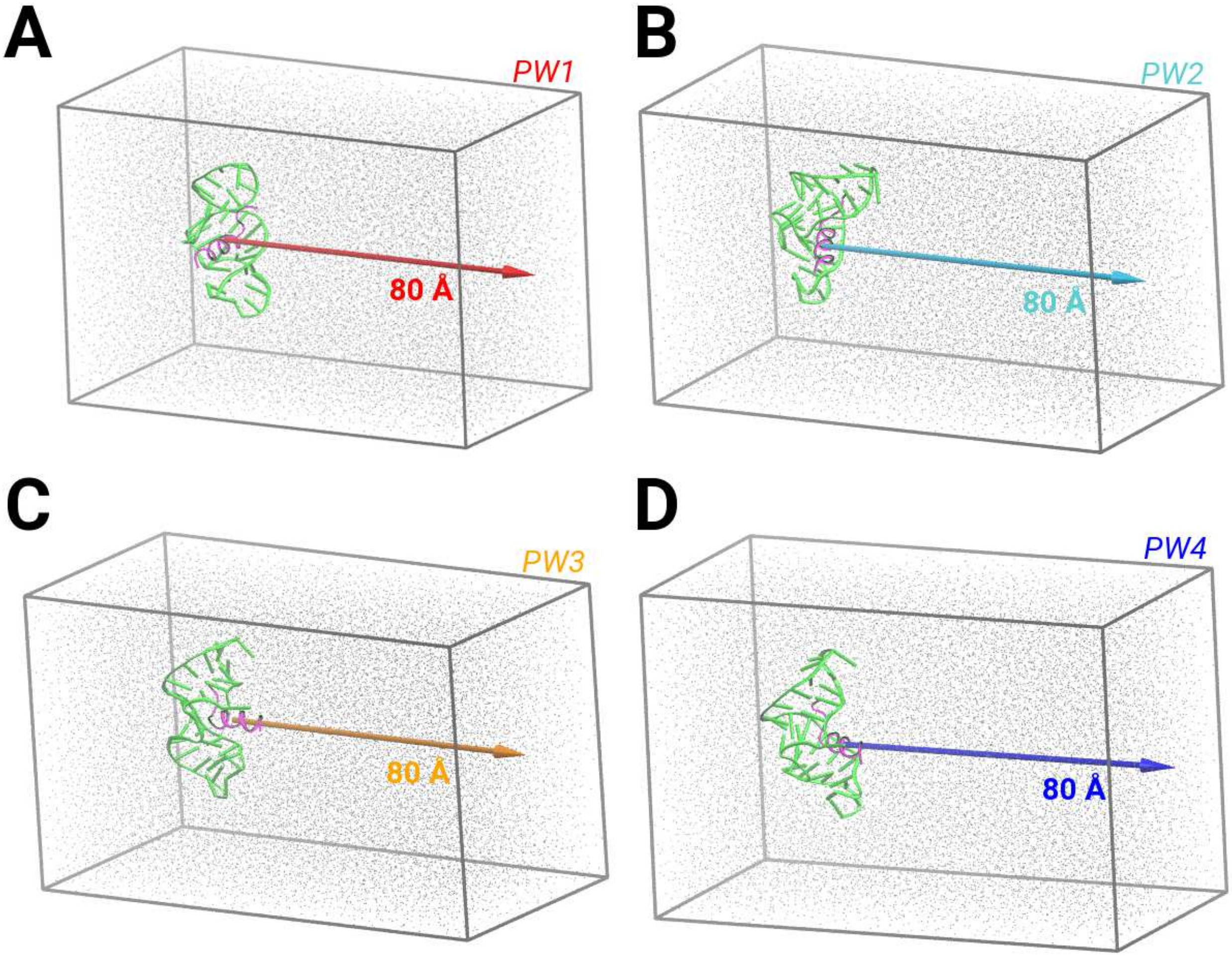

Figure E.1: **System setup:** Shown are the side-views of the simulation domains along (A) pathway 1 (PW1), (B) pathway 2 (PW2), (C) pathway 3 (PW3), and (D) pathway 4 (PW4). In each snapshot, RNA is represented as a green cartoon; peptide as a purple cartoon; water molecules as gray points; and the bounding box in gray. The arrow in each panel indicates the reaction coordinate for each pathway that was used to conduct non-equilibrium cv-SMD simulations.

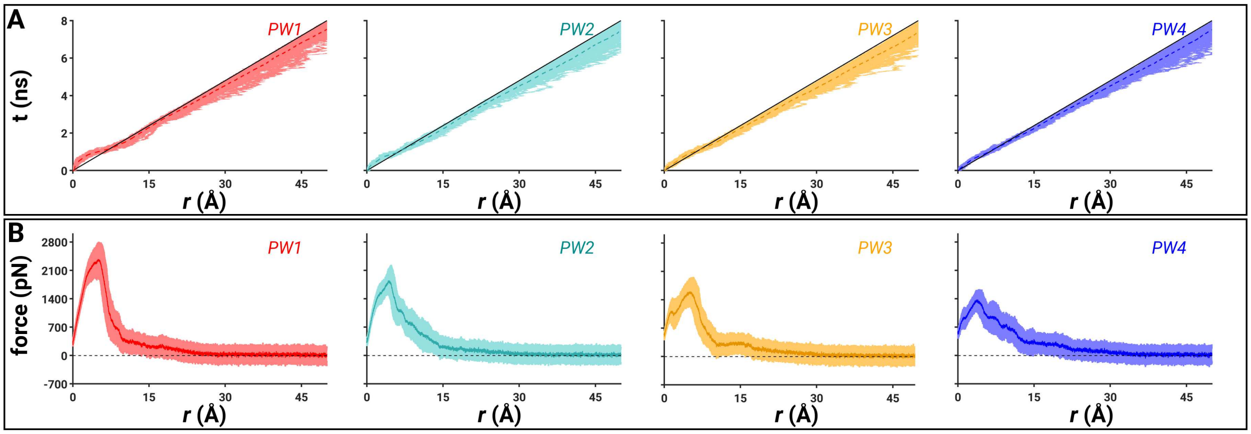

Figure E.2: **The reaction coordinates (*r*) and the unbinding force profiles.** (A) The center-of-mass (COM) trajectories of the peptide are shown in unique colors for PW1 (red), PW2 (cyan), PW3 (orange), and PW4(blue). The black solid lines represent the actual ***r***; the dark dotted lines represent the average trace across 75 trajectories for the corresponding PW; and the lighter shaded lines represent all SMD trajectories for the corresponding PW. (B) The unbinding force profiles are shown in unique colors for PW1 (red), PW2 (cyan), PW3 (orange), and PW4 (blue) with the average force traces (darker solid lines) and standard deviations (lighter shades).

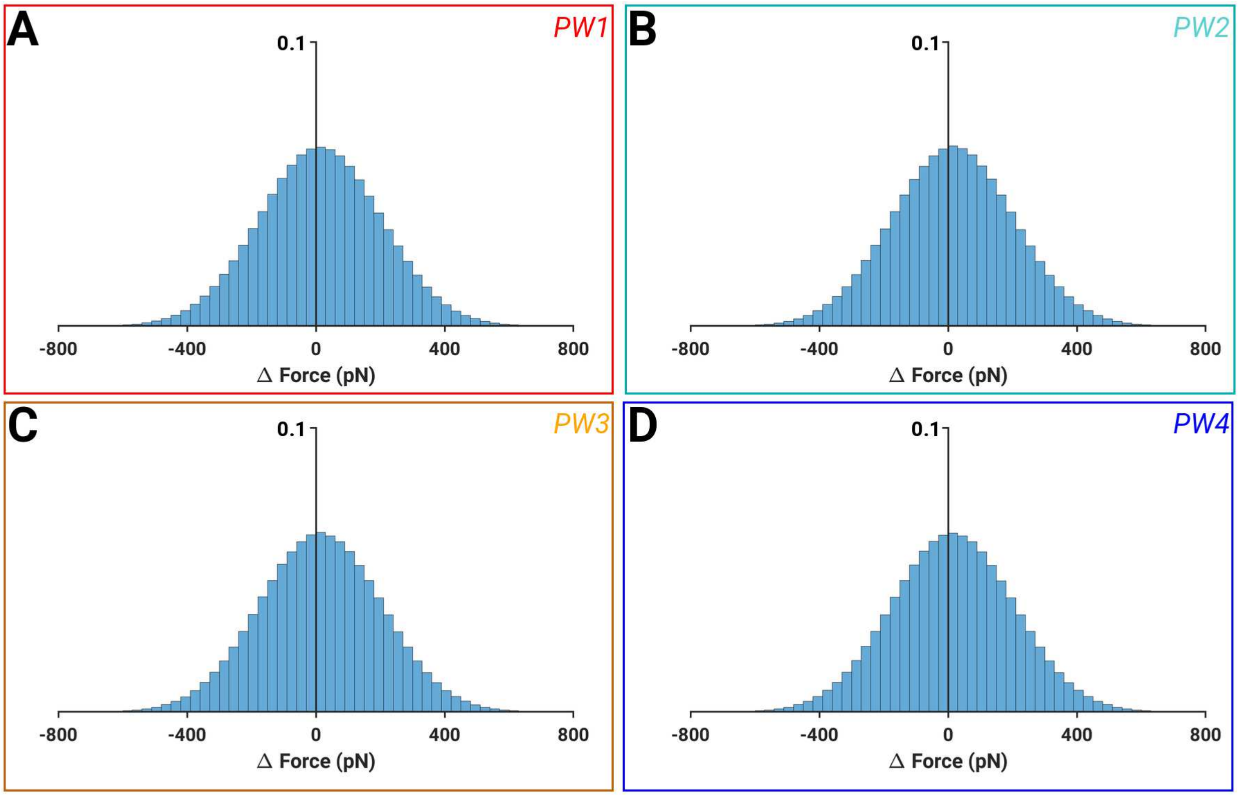

Figure E.3: **Force convergence data:** Shown are the distributions of ΔForce values, defined as a difference between zero and the actual force value after the average force profile converged to zero for (A) PW1, (B) PW2, (C) PW3, and (D) PW4. The ΔForce values were measured after a distance of 40 Å along the reaction coordinate. See also Figure E.2B.

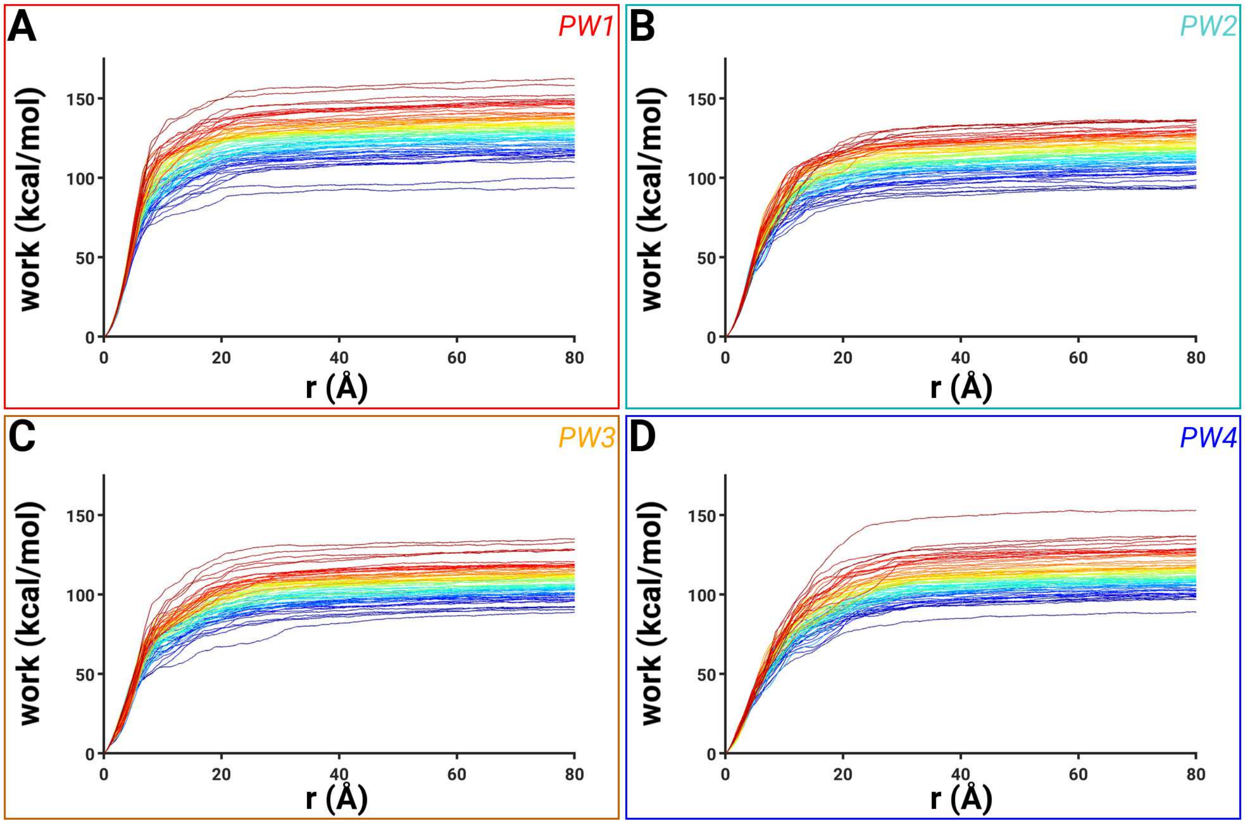

Figure E.4: **Non-equilibrium work profiles:** The non-equilibrium work values obtained from 75 independent cv-SMD simulations for (A) PW1, (B) PW2, (C) PW3, and (D) PW4. The lower work values are indicated in blue traces and the higher work values in red traces.

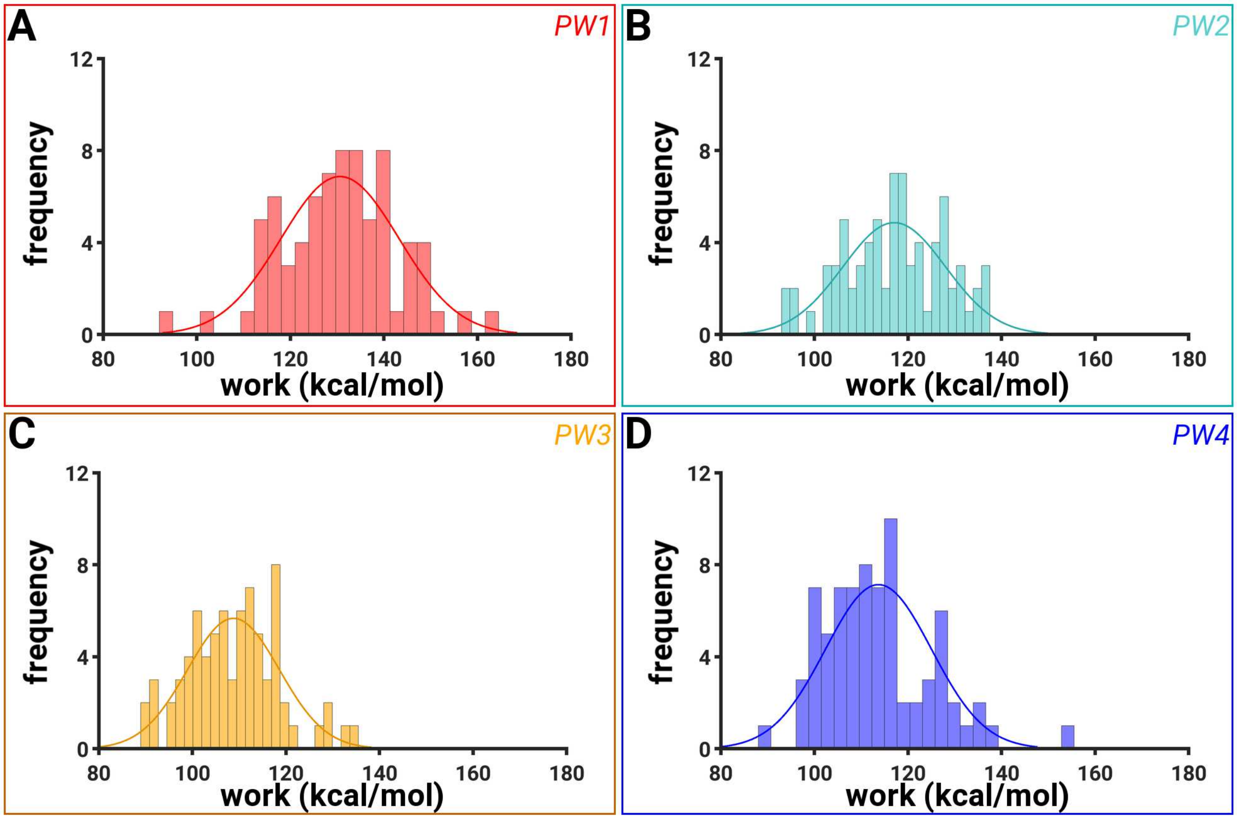

Figure E.5: **Distributions of the final work values:** The histograms of all final work values during cv-SMD simulations are shown with a best-fit distribution line for (A) PW1, (B) PW2, (C) PW3, and (D) PW4.

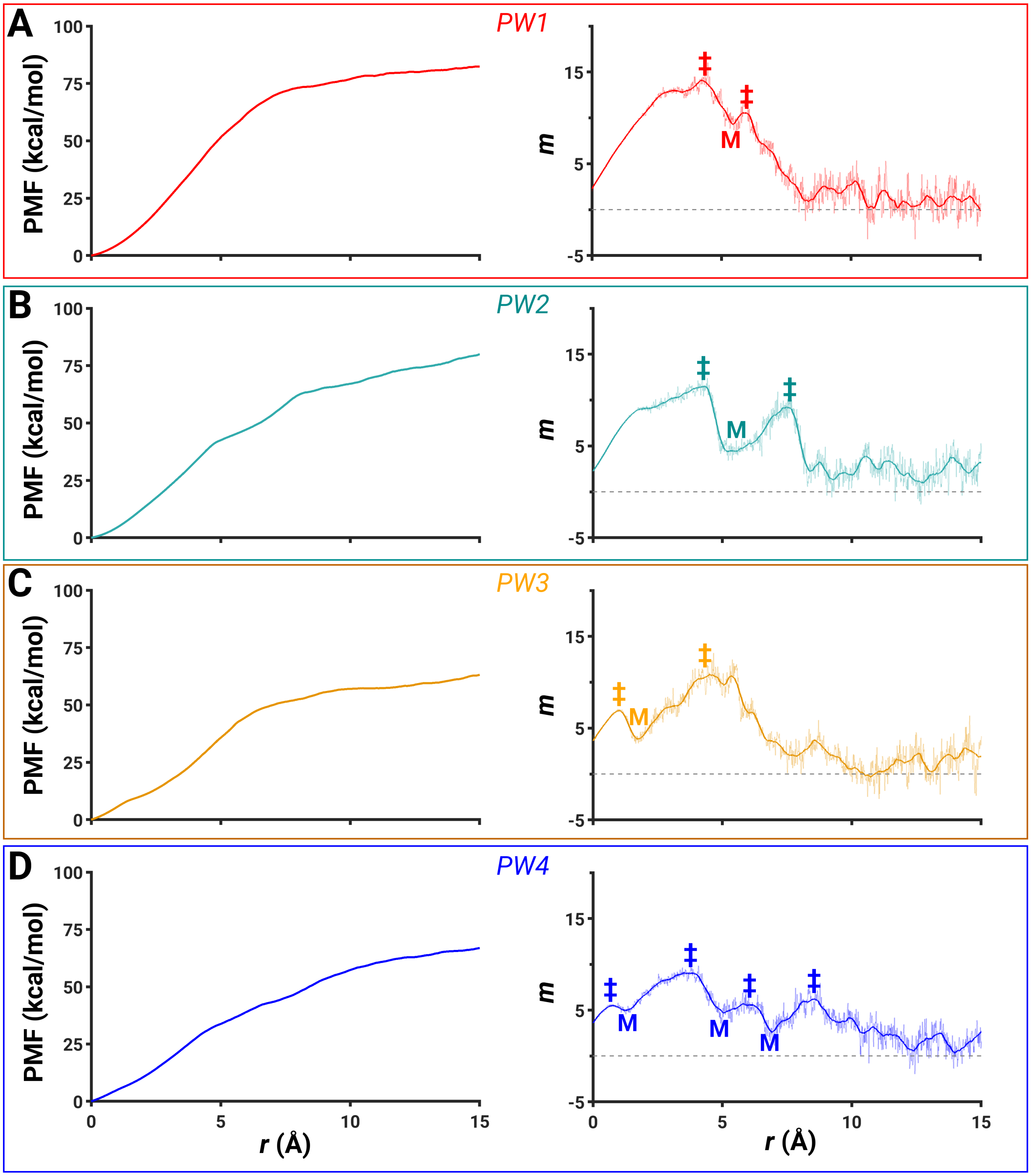

Figure E.6: **The free energy and corresponding first-order derivative profiles:** A zoomed view of each free-energy profile (*left*) and the corresponding first-order derivative profile ($\boldsymbol{m}$; *right*) computed every 100 points for r values between 0 Å and 15 Å are shown for (A) PW1, (B) PW2, (C) PW3, and (D) PW4. The fluctuations of the first-order derivatives are shown in light shaded colors. The free-energy barriers (indicated by ‡) and metastable states (indicated by M) are also shown and labeled.

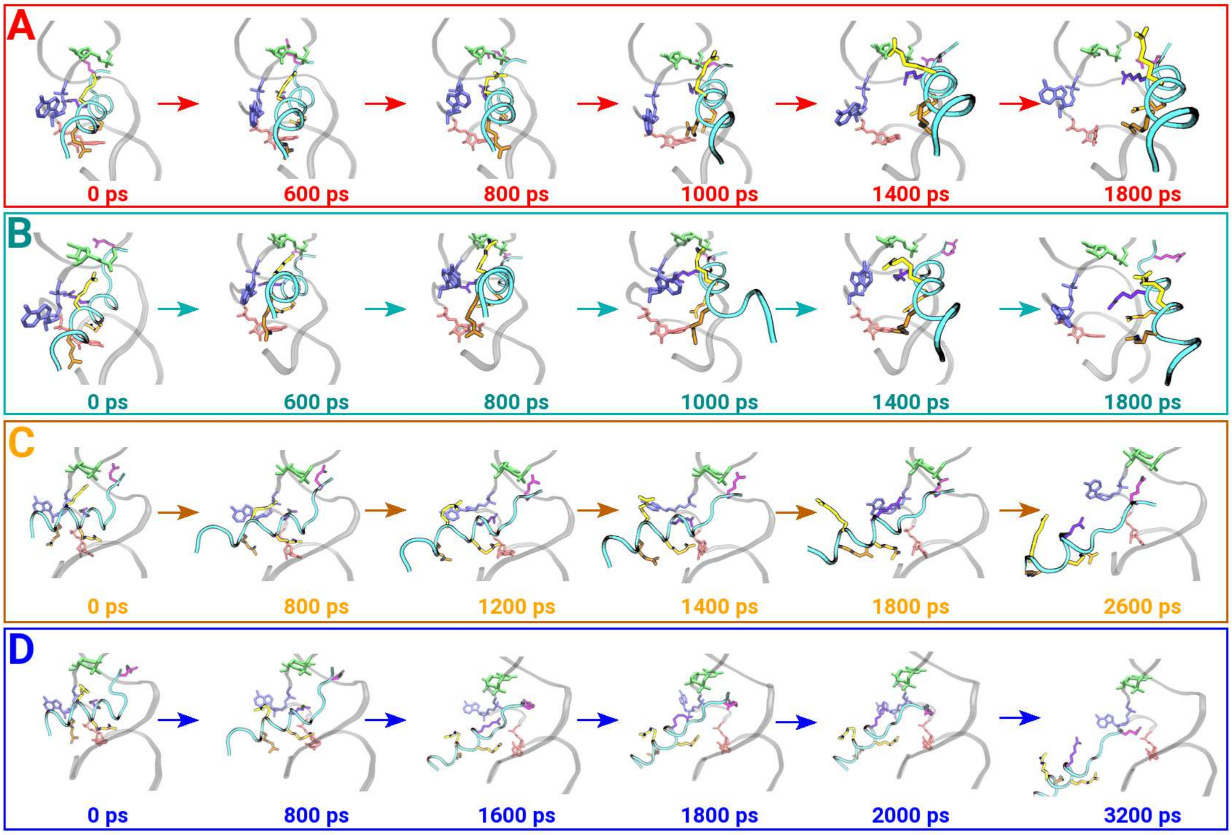

Figure E.7: **Dissociation pathways:** Shown are the snapshots of the peptide (cyan tube with key amino acids highlighted) dissociating from the RRE RNA (gray cartoon) from the least work cv-SMD simulation for (A) PW1 (red), (B) PW2 (cyan), (C) PW3 (orange), and (D) PW4 (blue).

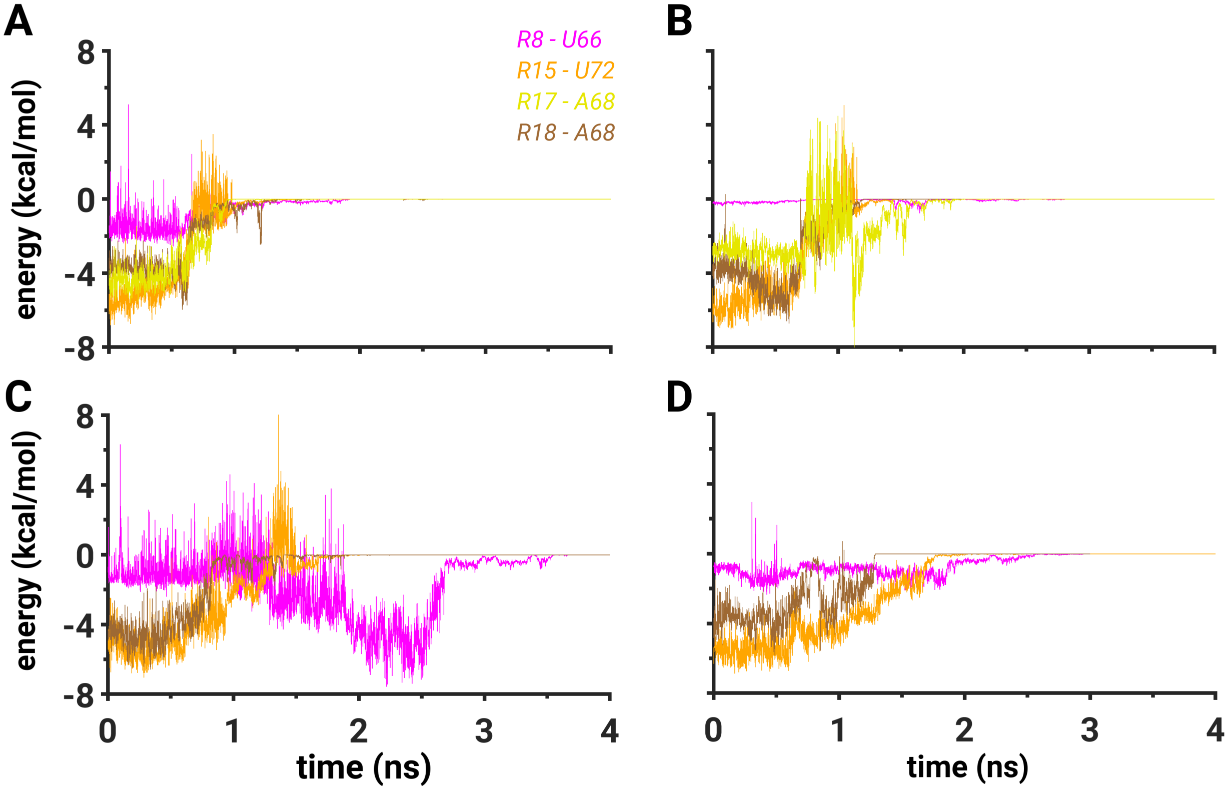

Figure E.8: **Van der Waals interaction energies:** Shown are the time traces of the van der Waals interaction energy computed between the following amino acid - nucleotide pairs: the Arg8 (R8) amino acid and the U66 nucleotide (purple); the Arg15 (R15) amino acid and the U72 nucleotide (orange); the Arg17 (R17) amino acid and the A68 nucleotide (yellow); the Arg18 (R18) amino acid and the A68 nucleotide (brown). The energies were computed from the simulation with the least work in (A) PW1, (B) PW2, (C) PW3, and (D) PW4. Data after 4 ns are truncated due to the convergence to zero of each van der Waals energy trace.

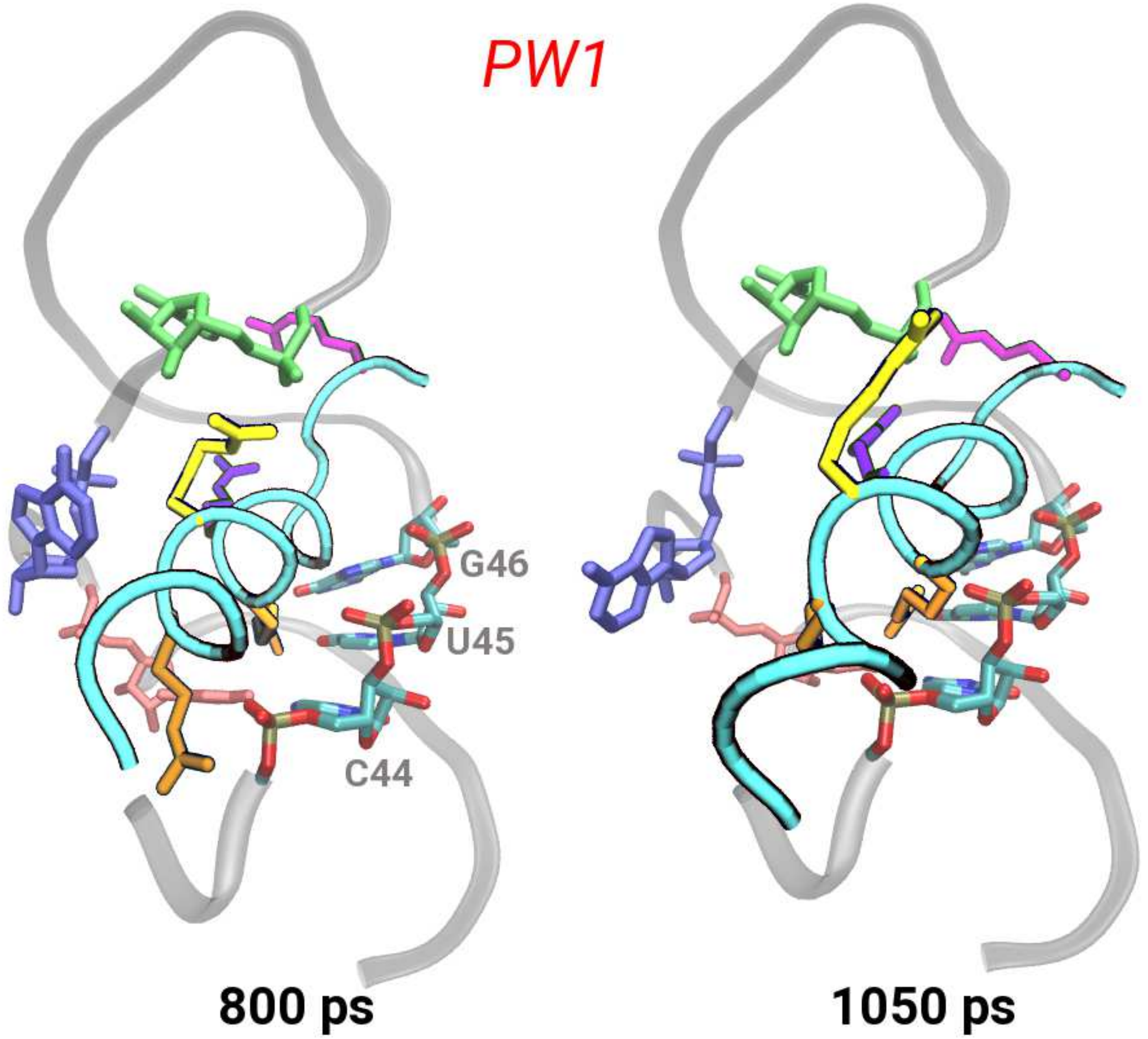

Figure E.9: **Snapshots from PW1:** Shown are the snapshots of the peptide (cyan tube with key amino acids highlighted) dissociating from the RRE RNA (gray cartoon) from the lowest work simulation of PW1. Three nucleotides (U44, U45, and G46) which interact with the peptide through the atoms in the backbone are each shown in a stick representation. The color scheme is the same as in Figure 5.1B.

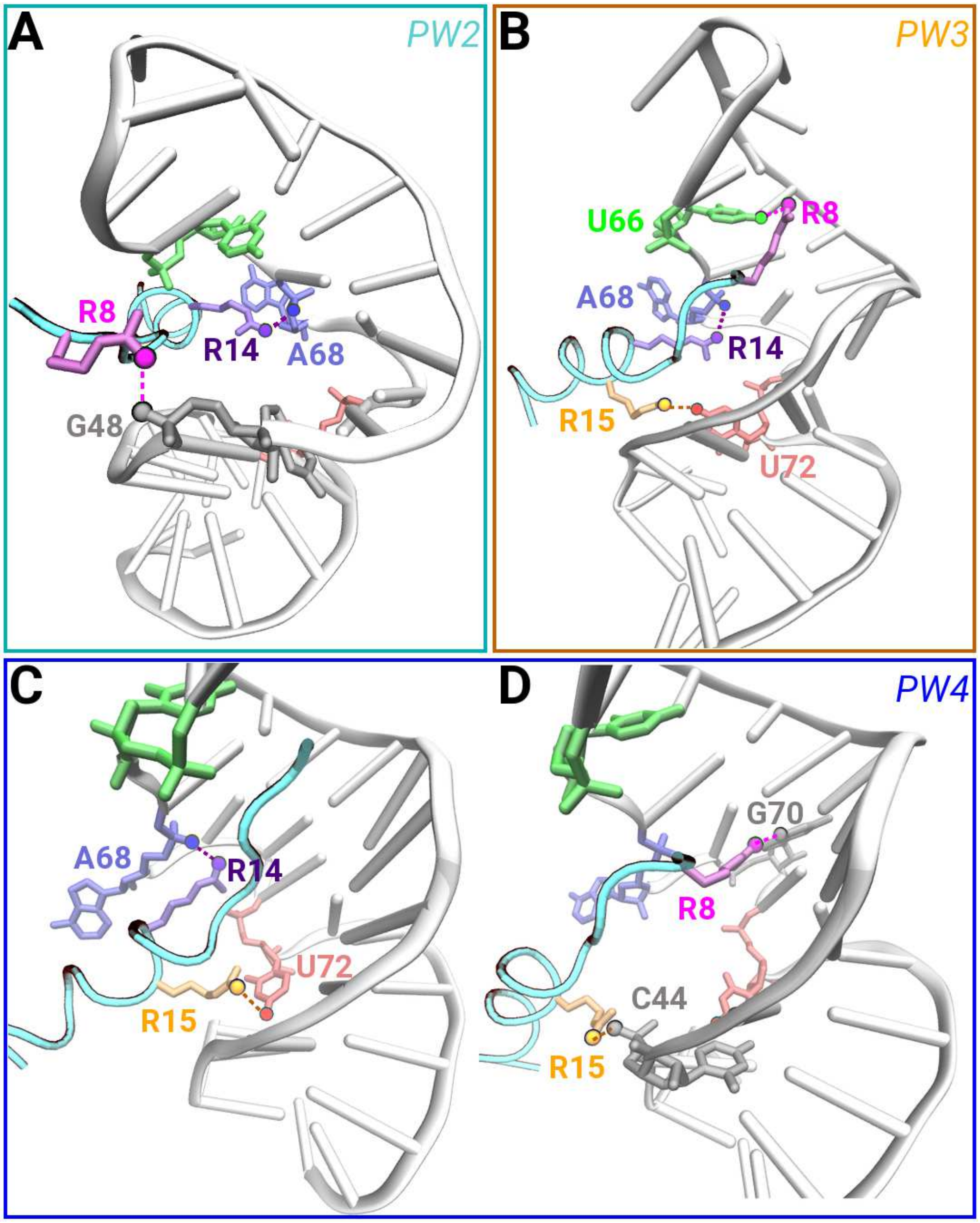

Figure E.10: **Networks of salt bridging and hydrogen bonding interactions:** Shown are the snapshots of the peptide (cyan tube with key amino acids highlighted) dissociating from the RRE RNA (white cartoon) and forming a network of salt bridging and hydrogen bonding interactions from the least work cv-SMD simulations in (A) PW2, (B) PW3, and (C-D) PW4. Each amino acid, each nucleotide or an atom that participate in hydrogen bonding or salt bridging interactions (marked by a dashed line), are uniquely colored. The color scheme is the same as in Figure 5.1B.

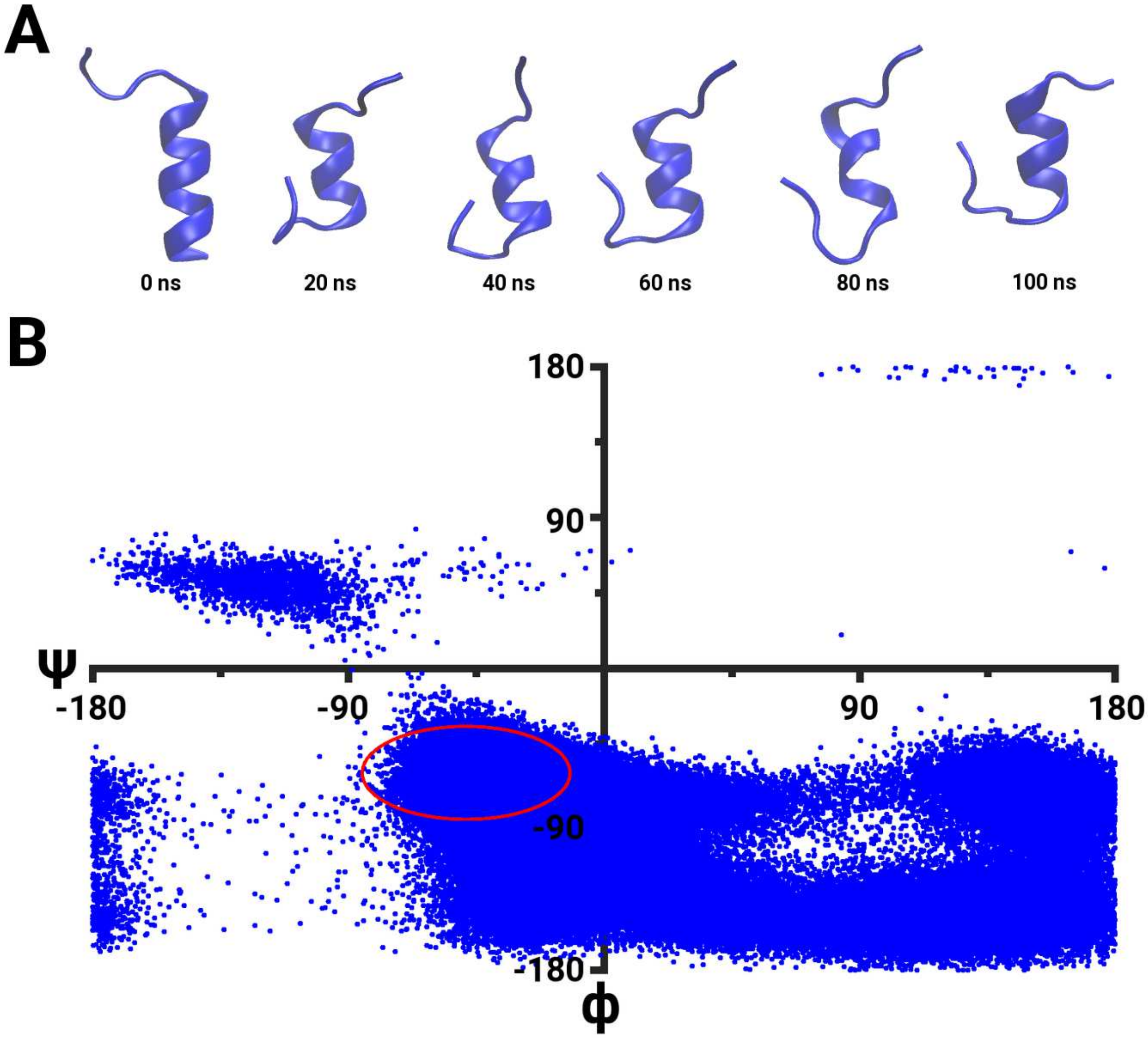

Figure E.11: **The RSG-1.2 peptide:** (A) Shown are the snapshots of the peptide (blue cartoon) from a 100 ns MD simulation. (B) The Ramachandran plot computed for ϕ and ψ dihedral angles of amino acids that constitute an α-helix in the initial structure in the course of a 100 ns MD simulation. Red circle highlights an approximate region of an α-helical conformation.

APPENDIX F

SCRIPTS FOR CHAPTER 5

F.1 Overview

In this appendix, I provide the scripts that I have used to set up and analyze simulations in chapter 5 using various software packages. The set up procedure in this study was similar to the set up procedure presented in Appendix F, thus I only provide the scripts that were new in this study.

F.2 Additional Scripts to Set Up a cv-SMD Simulation using NAMD

A script to apply restraints to the protein secondary structure based on hydrogen bonding interactions in the helix using SSRestraints plugin in VMD is shown below:

```
ssrestraints -psf rna_peptide_fin.psf -pdb rna_peptide_fin.pdb
-o g70_hbond.ref -sel helix -hbonds
```

I also provide a different cv-SMD script since I used several new simulation settings in this study in comparison to study presented in chapter 4.

```
coordinates   g70_smd.pdb

set temperature         310
set outputname    g70_smd_dir1

#Simulation Parameters
#Input
amber   on
parmfile    g70_smd.prmtop

#Continuing a run
set inputname   g70_smd_eq.restart         ;
```

```
binCoordinates $inputname.coor ;
binVelocities  $inputname.vel  ;
extendedSystem $inputname.xsc  ;
firsttimestep 0                      ;
numsteps        6500000        ;

#Force field parameters
exclude                 scaled1-4
1-4scaling              1.0
cutoff      10
switching   on
switchdist    8
pairlistdist    12

#Integrator Parameters
timestep    2.0   #2 fs/step
rigidBonds    all
nonBondedFreq   1
fullElectFrequency  2
stepspercycle         10

#Constant Temperature Control
langevin    on
langevinDamping   1
langevinTemp    $temperature
langevinHydrogen  off

# Periodic Boundary Conditions
cellBasisVector1    65.0    0.0   0.0
cellBasisVector2     0.0   79.0  0.0
cellBasisVector3     0.0    0.0   133.0
cellOrigin          0.43169 -1.247 29.9358
wrapAll       on

#Electrostatic Force Evaluation (PME)
PME               yes
PMEGridSizeX    72
PMEGridSizeY    80
PMEGridSizeZ    135

# Constant Pressure Control
useGroupPressure      yes
useFlexibleCell       no
useConstantArea       no

langevinPiston        on
langevinPistonTarget  1.01325 ;#  in bar -> 1 atm
langevinPistonPeriod  100.0
langevinPistonDecay   50.0
langevinPistonTemp    $temperature

#SMD
if {1} {
fixedAtoms          on
fixedAtomsFile g70_smd.ref
fixedAtomsCol       B
}
SMD on
SMDFile g70_smd.ref
SMDk 12
SMDVel 0.0000125
SMDDir -0.0697 -0.03259 0.997
SMDOutputFreq 10

#Output
outputName          $outputname
restartfreq         5000     ;# 5000steps = every 10ps
dcdfreq             5000
xstFreq             5000
outputEnergies      5000
outputPressure      5000

#Extra Bonds                 ### module to introduce secondary
extraBonds on                ### structure restraints
extraBondsFile g70_hbond.ref

#Colvar Module               ### module to introduce
colvars on                   ### orientational restraints
colvarsConfig colvar.in

#Execution
reinitvels          $temperature
```

The CV file **colvar.in** is shown below:

```
colvarsTrajFrequency  100
colvarsRestartFrequency 500
colvarsTrajAppend       on

colvar {
  name rotation
  orientation {
                atoms {atomNumbersRange 1-255 }
                refpositionsFile {reference.pdb}
    refPositionsCol {B}
    refPositionsColValue {2}
    closestToQuaternion  {1.0, 0.0, 0.0, 0.0}
}
}
harmonic {
         name harm
         colvars {rotation}
         centers {( 1 , 0 , 0 , 0 ) }
         forceConstant 10000
}
```

F.3 Analysis Scripts

In this section, I provide additional scripts that I have generated to analyze SMD simulations.

F.3.1 Interaction Energy

I used the NAMDEnergy plugin in VMD to calculate interaction energies.

```
mol new ../g70_smd.prmtop
mol addfile ../low_dir1.dcd first 0 last 4000 waitfor all

package require namdenergy

set sel1 [atomselect top "resid 2"]
set sel2 [atomselect top "resid 35"]

namdenergy -sel $sel1 $sel2 -nonb -tempname test -ofile low_dir1_2_35.dat
-extsys ../g70_smd_dir1.xsc -pme -cutoff 10 -switch 8 -exe
/home/levintov/NAMD/NAMD_2.12_Linux-x86_64/namd2
exit
```

F.3.2 First Order Derivative Calculation

A script to compute and plot the first order derivative of the free-energy profile is shown below:

```
%% Direction 1
A1 = importdata(’pmf_exp_aver_dir1.dat’);
dist1 = A1(:,1);
dist1_tan = A1(1:100:650000,1);
pmf1 = A1(:,2);

%%%%%First Order Derivative%%%%%
dy1=diff(pmf1)./diff(dist1);
k=6001; % point number 6001
tang1=(dist1-dist1(k))*dy1(k)+pmf1(k);
m1 = dy1(k);
dy1_tan = dy1(1:100:650000,1);
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

dy1_tan_smooth = smoothdata(dy1_tan,’sgolay’,80);
TAN1_smooth = plot(dist1_tan,dy1_tan_smooth);
TAN1_smooth.LineWidth = 5;
TAN1_smooth.Color = ’red’;
hold on
TAN1 = plot(dist1_tan,dy1_tan);
```

```
TAN1.LineWidth = 3;
TAN1.Color = ’1 0 0 0.2’;

h2 = plot([0 15] , [0 0]);
h2.Color=’[0.5 0.5 0.5]’;
h2.LineStyle = ’--’;
h2.LineWidth = 3;

set(gca,’box’,’off’,’TickDir’,’out’,’fontweight’,’bold’,’fontsize’,50,
’linewidth’,8,’Xticklabel’,[]);
a=gca;
a.YAxis.MinorTick = ’on’; % Must turn on minor ticks if they are off
a.YAxis.MinorTickValues = [-5:5:20];
xticks(0:5:15);
yticks(-5:10:20);
xlim([0 15]);
ylim([-5 20]);
```

APPENDIX G
SUPPORTING INFORMATION FOR CHAPTER 6

G.1 Preliminary Kinetics Calculations

In the study presented in chapter 6, I also performed preliminary estimates on kinetics of the base flipping in the dsRNA. The PMF profile (Figure 6.3B) was used to calculate the rate constant using the transition state theory (TST) [348, 349]

$$k_{TST} = \frac{\omega_R}{2\pi} e^{-\frac{\Delta G^{\ddagger}}{k_B T}}, \tag{G.1}$$

where ω_R represents the harmonic frequency around the free energy minimum of state I and $\Delta G^{\ddagger}$ is the activation energy estimated from the free-energy profile.

To compute the harmonic frequency (ω_R), I used the following equation:

$$\omega_R = \sqrt{\frac{k}{m}} \tag{G.2}$$

where k is related to the coefficients of the quadratic equation that is fitted to the energy minimum of the reactant state (state I) or the metastable state (state M) and m is the effective mass. To estimate k, I obtained the coefficients of the quadratic equation ($y = ax^2 + bx + c$) that described the energy minimum of the corresponding state I or M (Figure G.9A). I then took a second order derivative of the fitted quadratic equation and estimated $k = 2a$. The mass was computed using the following equation:

$$m = \frac{3k_B T}{v^2} \tag{G.3}$$

where v is the velocity that was computed from all shooting trajectories by extracting points which had RC values that were within the range of the free energy minimum of the states *I* or *M*. The velocity was thus computed by taking a difference in the subsequent RC values over time difference ($v = \frac{\Delta RC}{\Delta t}$) between those two steps that described the free energy minimum of states *I* or *M*.

The rate constant can also be estimated from the inverse of mean first passage time (MFPT) [350]:

$$k_{I \to O} = D(r^*)[\int_{\cap} e^{\frac{G(r)}{k_B T}} dr \int_{\cup} e^{-\frac{G(r)}{k_B T}} dr]^{-1} \tag{G.4}$$

where $D(r^*)$ is the diffusivity at the barrier with r^* being the RC value at the barrier and $G(r)$ is the free-energy. The diffusivity across the free-energy barrier is calculated by the mean squared displacement using the following equation:

$$\langle \omega^2 [r(t) - r^*]^2 \rangle = e^{2\omega^2 D(r^*) t} - 1 \tag{G.5}$$

where $-\omega^2$ is the curvature of the free energy barrier at r^*. I fitted the free energy barrier with an equation of the form, $y = ax^2 + bx + c$, to estimate the curvature at $x = r^*$ (Figure G.9B).

Based on the free-energy profile (Figure 6.3B), the base flipping event can be perceived as a single or as a two step process. To estimate the reaction rate for a one step reaction, from the inward state to the outward state ($I \to O$), the harmonic frequency ω_R was computed at the minimum of the free energy basin of the inward state and was calculated to be $9.28 \times 10^{10}\ s^{-1}$. The function fitted to the energy minimum of state I to obtain ω_R is shown on Figure G.9A. Using ω_R and the transition state theory [348, 349], I estimated the rate constant to be $1.34 \times 10^{9}\ s^{-1}$. Using MFPT, the rate constant was estimated to be $1.09 \times 10^{8}\ s^{-1}$. The function fitted to the energy barrier between states *I* and *O* to compute the curvature is shown on Figure G.9B.

Since the base flipping occurred in two steps, I estimated the reaction rates for each transition: between the inward and metastable states ($I \rightarrow M$) and between the metastable and outward states ($M \rightarrow O$). Using ω_R that was obtained from functions fitted to the energy minima of states I and M (Figure G.9A) and the transition state theory, I estimated that $k_{I\rightarrow M} = 5.29 \times 10^9\ s^{-1}$ and $k_{M\rightarrow O} = 1.39 \times 10^9\ s^{-1}$. Using MFPT the rate constant was estimated to be $k_{I\rightarrow M} = 9.65 \times 10^8\ s^{-1}$ and $k_{M\rightarrow O} = 3.27 \times 10^8\ s^{-1}$ using the curvature values obtained from functions fitted to the energy barriers (Figure G.9B). Based on TST and MFPT methods, I concluded that the second step ($M \rightarrow O$) is the rate determining step since it is the slowest step. Moreover, the value of the rate constant for the second step is close to the single step rate constant in each method. To my knowledge, the rate constant of base flipping in dsRNA molecules has not been reported yet. However, based on the timescales suggested by Al-Hashimi *et al.*, [45] base flipping likely occurs on *ns*-*ms* timescale range which is consistent with our prediction of the rate constant.

Table G.1: List of relevant CVs and the upper (U) and lower (L) boundaries of CVs in each of the stable states (*I*, inward; and *O*, outward)

#	CV	I_L	I_U	O_L	O_U
1	ϕ_1	-70°	70°	100°/-120°	180°/-180°
2	ϕ_2	0°	50°	-150°	-35°
3	d_1	3.5 Å	7 Å	7.5 Å	13 Å
4	d_2	3.5 Å	7.5 Å	8 Å	14 Å
5	d_3	5.5 Å	9 Å	10 Å	16 Å
6	α_1	25°	95°	100°	180°
7	α_2	0°	60°	90°	180°
8	α_3	0°	60°	90°	180°
9	N_W	60	92	97	128
10	E_1	-8	-0.5	0	1.5
11	E_2	-6.5	-0.5	0	1
12	E_3	-7.5	-0.5	0	1

Table G.2: Likelihood scores for single-variable reaction coordinate models

a_0	q_1	a_1	$\ln(L)$	rank
-0.63	d_1	0.6989	-542.839	1
-0.63	ϕ_2	-0.6328	-549.832	2
-0.30	d_3	0.5419	-577.899	3
-0.30	d_2	0.5008	-589.915	4
0.19	E_2	0.3037	-649.798	5
0.17	E_3	0.3265	-652.817	6
-0.24	α_1	0.2954	-653.589	7
-0.22	N_W	0.1979	-674.77	8
-0.01	E_1	0.1825	-676.978	9
-0.33	α_3	0.1489	-682.314	10
-0.12	ϕ_1	0.1051	-687.709	11
-0.01	α_2	0.0043	-693.138	12

Table G.3: Five two-variable reaction coordinate models with the highest likelihood scores

a_0	q_1	a_1	q_2	a_2	$\ln(L)$	rank
-0.84	d_1	0.4753	ϕ_2	-0.3941	-510.534	1
-0.62	d_1	0.5552	d_2	0.2373	-528.551	2
-0.60	d_1	0.5347	d_3	0.2310	-532.742	3
-0.62	ϕ_2	-0.6382	N_W	0.1977	-535.738	4
-0.64	d_1	0.6513	E_2	0.1324	-537.245	5

Table G.4: Five three-variable reaction coordinate models with the highest likelihood scores

a_0	q_1	a_1	q_2	a_2	q_3	a_3	$\ln(L)$	rank
-0.84	d_1	0.4476	ϕ_2	-0.4057	E_1	0.0464	-509.876	1
-0.84	d_1	0.4401	ϕ_2	-0.4136	N_W	0.0503	-509.885	2
-0.84	d_1	0.4754	ϕ_2	-0.4261	E_2	-0.0492	-510.009	3
-0.84	d_1	0.4766	ϕ_2	-0.4012	ϕ_1	-0.0332	-510.199	4
-0.84	d_1	0.4644	ϕ_2	-0.3672	d_2	0.0419	-510.265	5

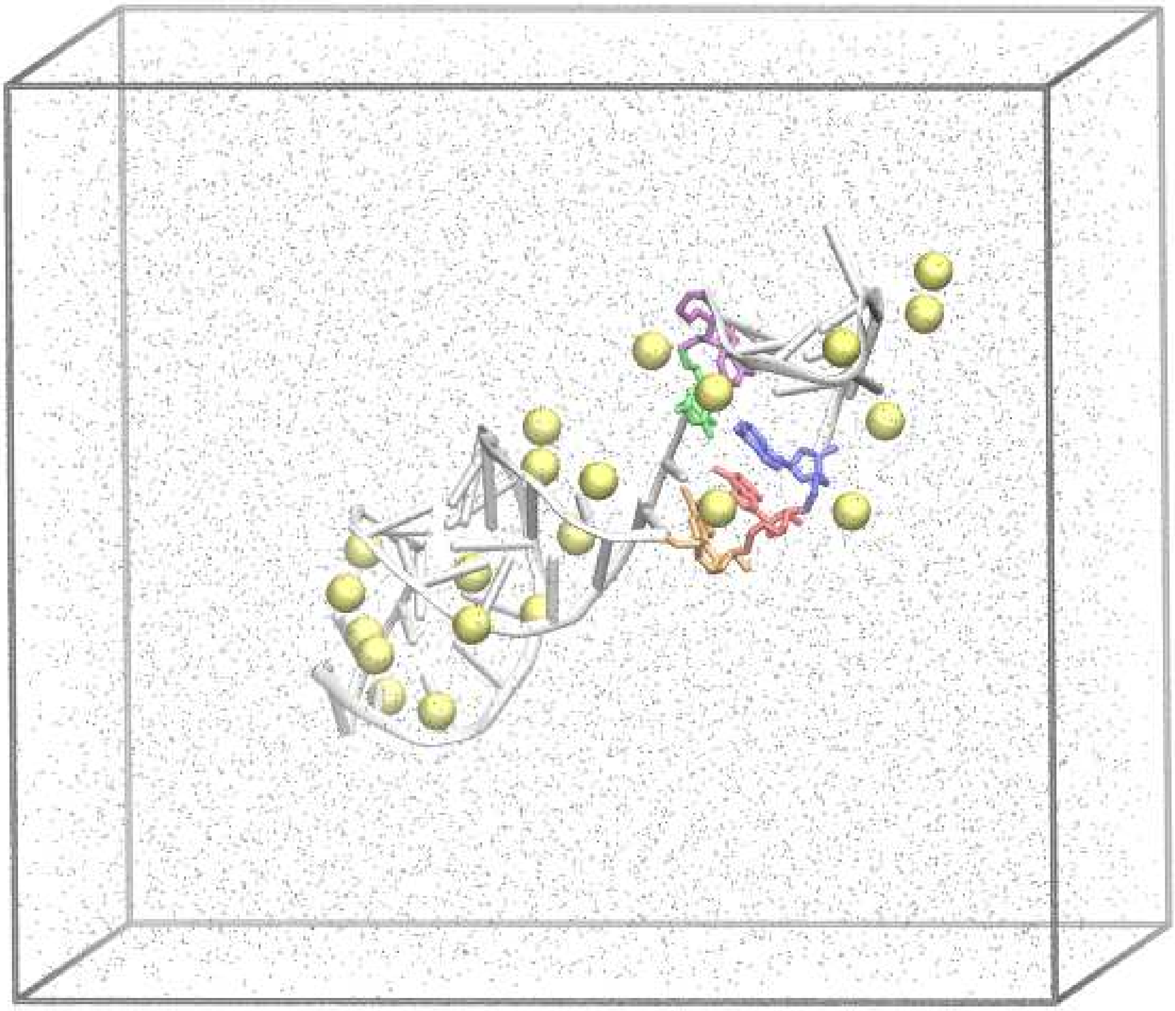

Figure G.1: **System setup:** Shown is a side-view of the simulation domain: RNA, white cartoon; water molecules, gray points; sodium ions, yellow; and the bounding box, gray. Each key nucleotide is highlighted in a unique color. See also Figure 1.6.

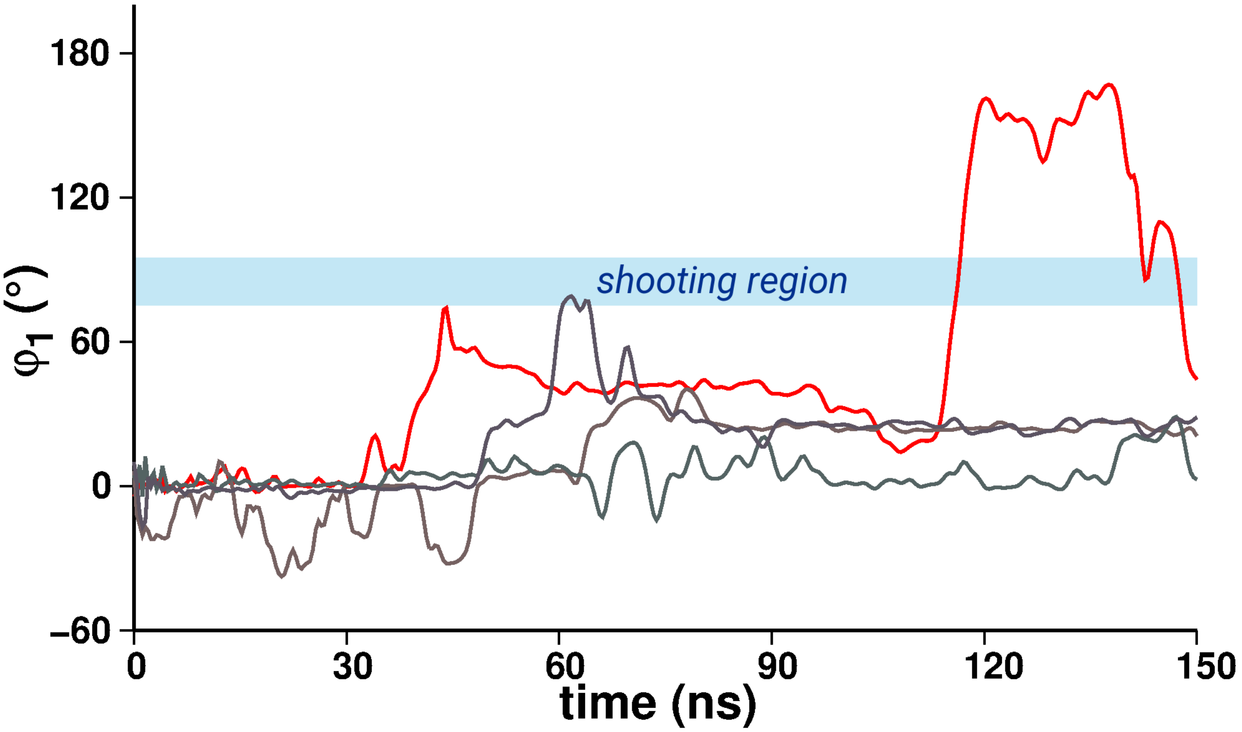

Figure G.2: Time traces of the primary OP (ϕ_1) in the seed trajectory (red) and 3 other conventional MD simulations. A cyan rectangle highlights the shooting region. See also Figure 6.1.

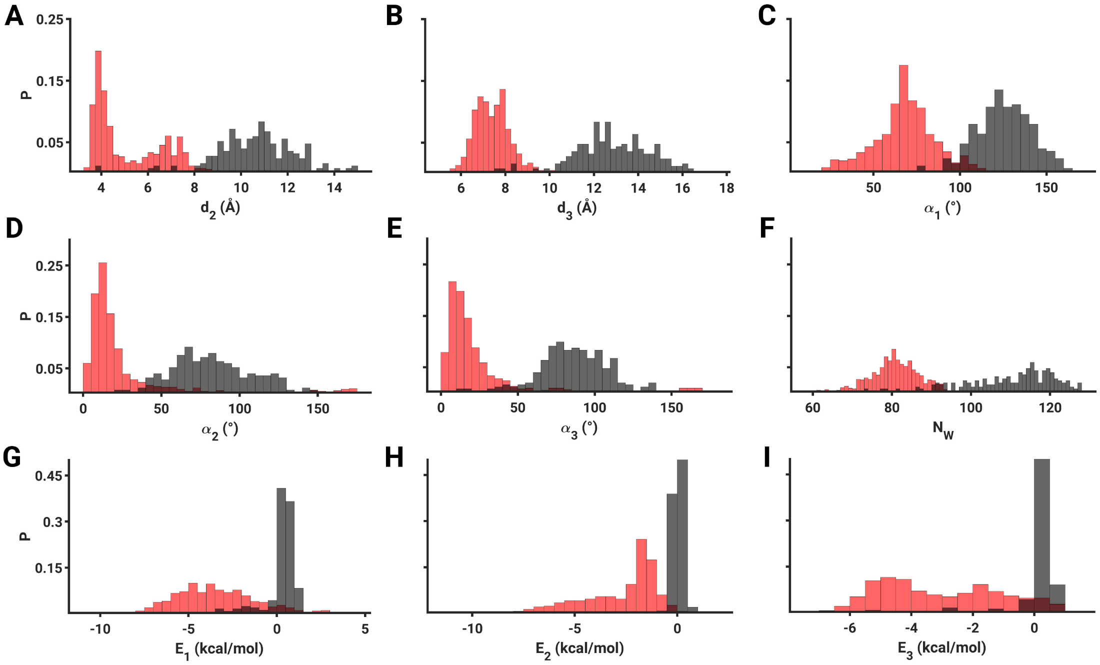

Figure G.3: The population distributions of CVs at terminal points are shown. (A) The distance between the centers of mass of bases A18 and A19. (B) The hydrogen bond distance between bases A18 and C54. (C) The angle between A18 and C28. (D) The interplane angle between G17 and A18. (E) The interplane angle between A18 and A19. (F) The number of water molecules. (G) The stacking energy between bases G17 and A18. (H) The stacking energy between bases A18 and A19. (I) The interaction energy between bases A18 and C54. See also Figure 6.2 .

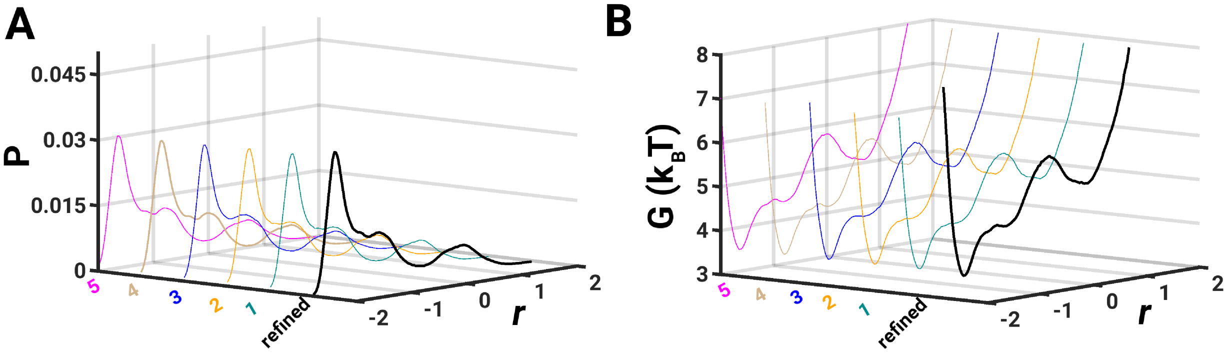

Figure G.4: Key metrics of the refined RC (black line) along with top 5 three-variable RC models are shown: (A) Distributions of the RC and (B) Free energy profiles corresponding to each individual RC model. Labels 1 through 5 indicate model rankings from Table G.4.

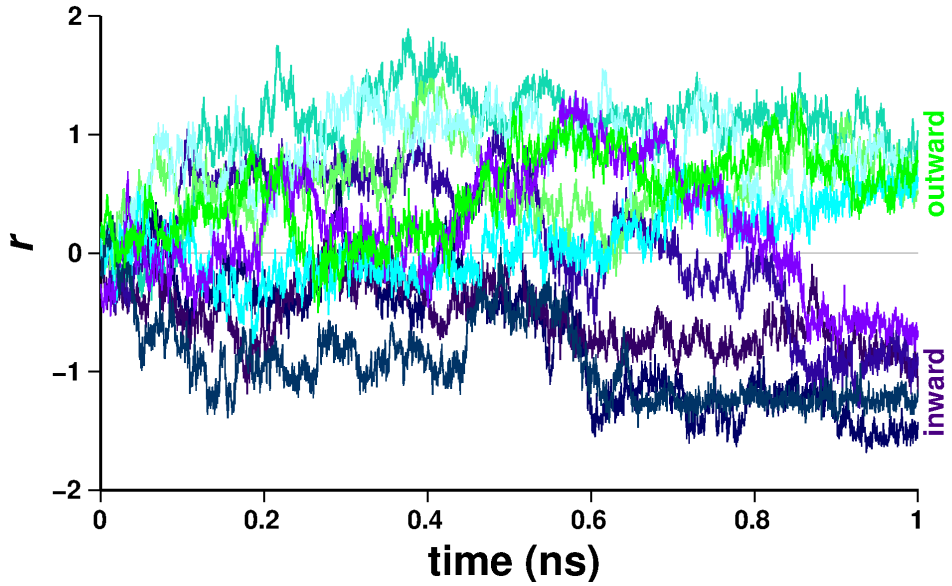

Figure G.5: Evolution of the RC along additional representative trajectories. See also Figure 6.3A.

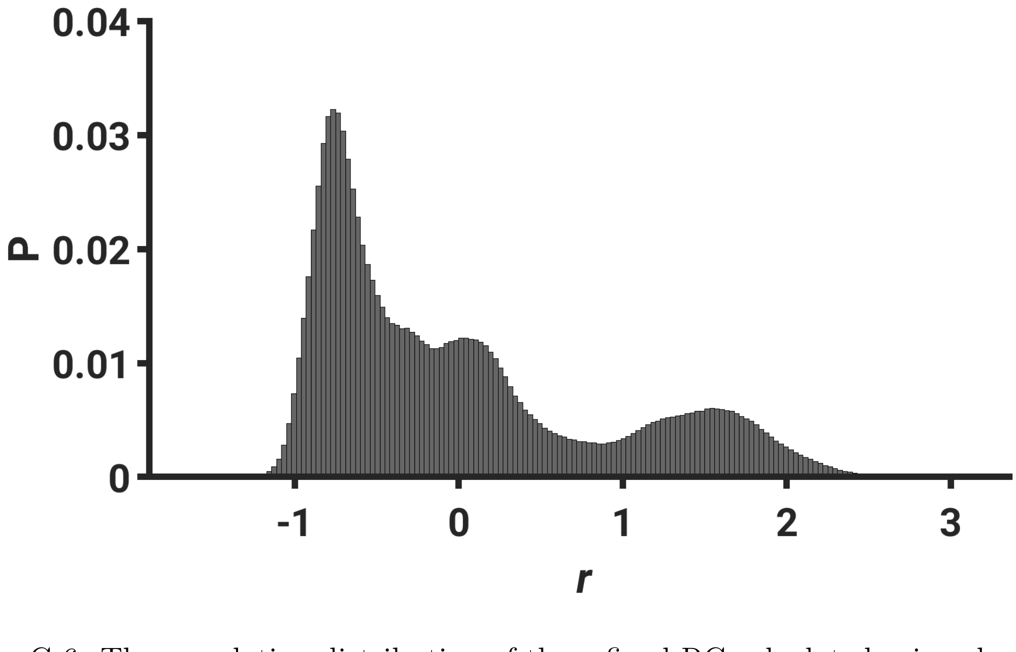

Figure G.6: The population distribution of the refined RC calculated using shooting trajectories.

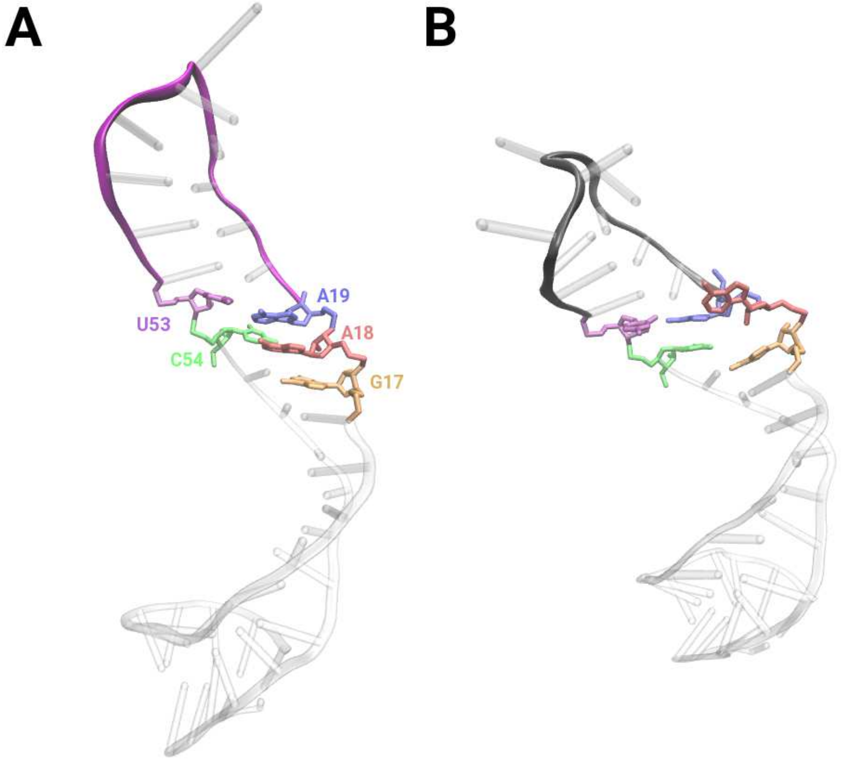

Figure G.7: Snapshots of the dsRNA are shown at (A) the initial state, and (B) the transition barrier. The stem loop is highlighted in magenta (corresponding to state *I*) and gray (corresponding to transition region; ‡). Color scheme for nucleotides is same as in Figure 1.6.

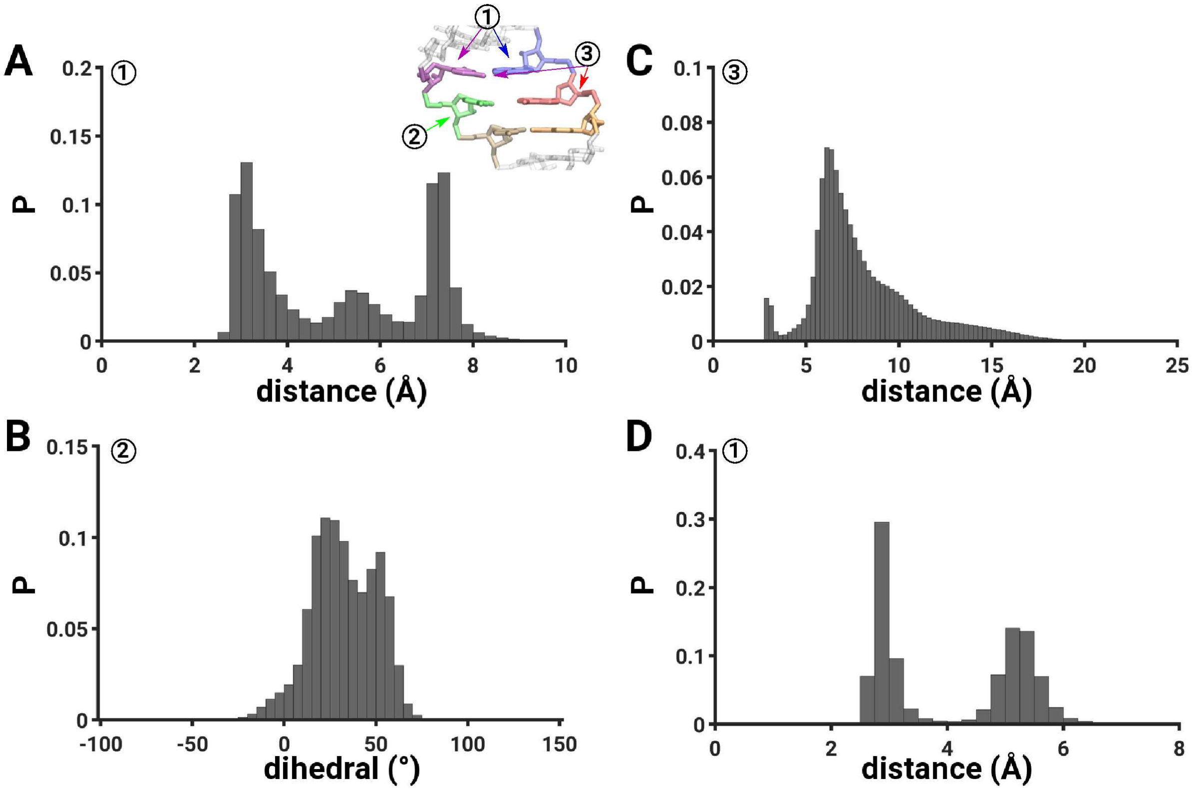

Figure G.8: The population distributions of several physical variables from the transition path ensemble. The numbers in each panel correspond to metrics computed for specific nucleotides (see *inset* in panel A). The conformational metrics shown are: (A) the hydrogen bond distance between the N6 atom of A18 and the O4 atom of U53; (B) the dihedral angle that describes the flipping of C28; (C) the hydrogen bond distance between A18 and U53; and (D) the hydrogen bond distance between the N6 atom of A18 and the O2 atom of U53.

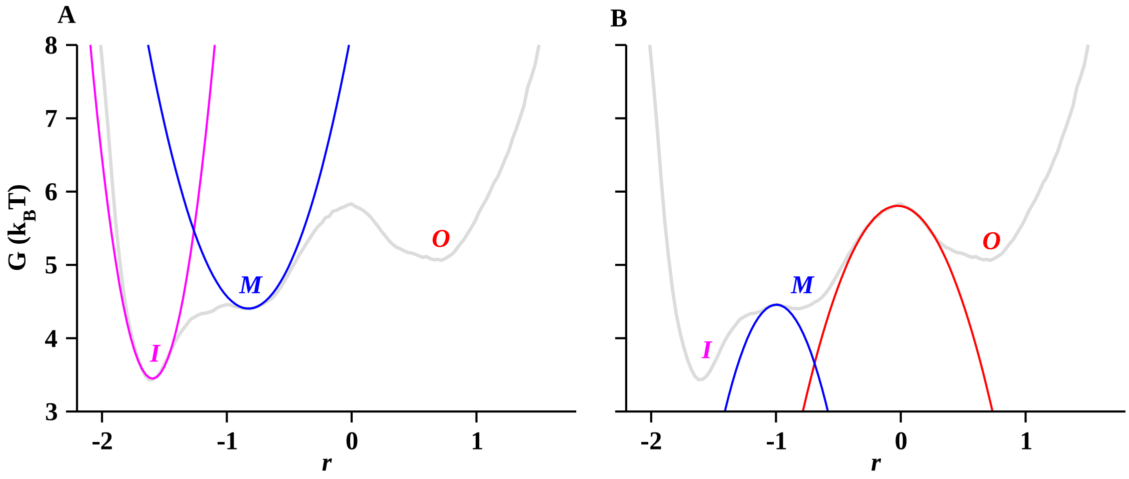

Figure G.9: Shown are the functions fitted to the free energy minima (panel A) or barrier regions (panel B) for rate calculations. See Figure 6.3B for definitions of states *I*, *M*, and *O*. Fitted curves are in the same color scheme as the label for each state. To estimate the rate constants using the transition state theory (equation G.1), the harmonic frequencies (ω_R) were calculated based on fitted curves shown in panel A, and to estimate the diffusivities using equation G.5, the curvatures of the barrier regions ($-\omega^2$) were calculated based on fitted curves shown in panel B.

APPENDIX H
SCRIPTS FOR CHAPTER 6

H.1 Overview

In this appendix, I provide the scripts that I have used to set up and conduct TPS simulations in chapter 6 using various software packages.

H.2 Set Up of TPS Simulations

A script file containing all commands to generate input files for shooting trajectories is shown below. If all the pathways are specified and all the files are located in the corresponding directories which are defined in the first block, then the script can be executed by typing **.\aimless.sh** in the terminal window. It is also possible to run each individual portion of the script separately (described below). This script was originally created by Dr. Sanjib Paul.

```
#!/bin/csh -fx

#########################################################
set echo

#########################################################
# Set environment variables
#########################################################


#BLOCK 1: location of scripts and input files to generate input files
for a given system
#setenv WORKDIR /home/levintov/ds_rna/r4/shooting_dih/job_run

# location of amber executable file
#setenv PROGRAM /home/software/AMBER12/amber12/bin/pmemd.cuda.MPI
```

```

# location of system-specific input files
#setenv INPUTDIR /home/levintov/ds_rna/r4/shooting_dih/input

# location of seed directory
#setenv SEEDDIR /home/levintov/ds_rna/r4/shooting_dih/seed

#location of Source directory
#setenv SOURCEDIR /home/levintov/ds_rna/r4/shooting_dih/Source

#location of output directory
setenv OUTDIR /home/levintov/ds_rna/r4/shooting_dih/run


# ++++++++++++++++++++++++++++++++++++++++++++
# BLOCK 2: Copy data files & execute scripts
# ++++++++++++++++++++++++++++++++++++++++++++


cp -f $INPUTDIR/*      $WORKDIR/.
cp -f $SEEDDIR/*       $WORKDIR/.
cp -f $SOURCEDIR/*     $WORKDIR/.

gfortran -O read_prmtop.f -o readprm
gfortran -O det_barrier.f -o op_barr
gfortran -O getsp.f -o getsp
gfortran -O shoot3.f -o shoot
gfortran -O det_lastframe.f -o detlast
f95 -O all_op_barrier.f -o allop


# setting up general variables

set parm_in = "dsrna_final.prmtop"

set atmid_in = "atom_id.in"

set natom = 39305

set nres = 12709

set nsnap = 20000

set ind_out = "index.out"

# extracting indices of atoms defining order parameters

./readprm $parm_in $atmid_in $nres $natom $ind_out


#looping over different shooting points

./op_barr ../shooting/ds_run.crd $ind_out barrier.in mass_dih.in
$natom $nsnap op_barrier.out
```

```

set nc = ‘wc op_barrier.out | awk ’{print $1}’‘
echo $nc

set n0  = 1

set n0max = $nc

while ($n0 <= $n0max )

    mkdir -p $OUTDIR/sp_$n0

setenv OUTDIR2 /home/levintov/ds_rna/r4/shooting_dih/run/sp_$n0


./getsp ../shooting/ds_run.crd ../shooting/ds_run.vel op_barrier.out $nc
$n0 $natom $nsnap stock_rst.in barrier.rst

#setting up loop variables for shooting trials from a given shooting point

    set ns = 1
    set nmax = 1

    while ($ns <= $nmax)

      mkdir -p $OUTDIR2/shoot_$ns

       set nr = ‘wc random.in | awk ’{print $1}’‘
        echo $nr

./shoot barrier.rst $natom mass.in momentum.in random.in $nr $ns
velocity.out shoot.out shoot.rst

/usr/mpi/gcc/openmpi-1.6.4/bin/mpirun -np 4 $PROGRAM -O -i NVT.in -o
ds_forward.out -p $parm_in -c shoot.rst -r hca_forward.rst -x
ds_forward.mdcrd -v ds_forward.mdvel

./detlast ds_forward.rst $ind_out state.in mass_dih.in $natom terstate.out

./allop shoot.rst $ind_out $natom allop_shoot.out

        bzip2 hca_forward.mdcrd
        bzip2 hca_forward.mdvel

         mv hca_forward.out       $OUTDIR2/shoot_$ns
         mv shoot.out             $OUTDIR2/shoot_$ns
         mv shoot.rst             $OUTDIR2/shoot_$ns
         cp /home/levintov/ds_rna/r4/common_dsrna/* $OUTDIR2/shoot_$n0
         mv hca_forward.rst       $OUTDIR2/shoot_$ns
         mv terstate.out          $OUTDIR2/shoot_$ns
         mv allop_shoot.out       $OUTDIR2/shoot_$ns
         mv hca_forward.mdcrd     $OUTDIR2/shoot_$ns
         mv hca_forward.mdvel     $OUTDIR2/shoot_$ns

      @ ns++
      end

@ n0 = $n0 + 7
end
#
unset echo

######################################################
```

The first step is to generate a file with masses of all particles in the system. To do that, I execute the file **getmass.f** in the terminal window by typing **gfortran getmass.f** and then **.\a.out**. This script was originally created by Dr. Sanjib Paul.

```
c     This program reads prmtop file and extract the value of mass of
c     each atom and the calculates the square root of those mass values.
c     -------------------------------------------------------------------
      implicit real*4 (a-h,o-z)
c
      parameter (n1 = 50000, natom = 39305)
      dimension amass(n1)
      character*80 pline
c
      open(11, file = "dsrna_final.prmtop", status = "old")
c
 10   read(11,3)pline
 3    format(a)
      if(pline(7:10).eq.’MASS’)then
          read(11,*)
          read(11,1)(amass(i), i = 1,natom)
 1    format(5e16.8)
      else
         go to 10
      end if
c
      close(11)
      open(13, file = "mass.out", status = "unknown")
c
      do 20 i =1, natom
          smass = sqrt(amass(i))
          write(13,*)amass(i), smass
 20   continue
c
      stop
      end
```

Next step is to generate an index file (**index.out**) which contains indices of atoms which are used as order parameters. The **index.out** file can be generated either using the script below (**getmass.f**) or by manually writing the file. This script was originally created by Dr. Sanjib Paul.

```
c     read_prmtop
c
c     This program reads prmtop file and then finds out the indices of
c     atoms & write the indices in index_out file. atom information
c     (residue numbers & atom names) is given in atomid_in file.
c-----------------------------------------------------------------------
      parameter (n1 = 15000, n2 = 50000, n3 = 100, n4=50)
c
      implicit real*4 (a-h,o-z)
      dimension ires(n1), inpres(n3,n4), iopt(n3)
      character*80 pline
      character*8 atmnam(n2),inpatm(n3,n4),nam
      character*256 prmtop_in,atomid_in,numres,numatm,index_out
c
      call getarg(1, prmtop_in)
      call getarg(2, atomid_in)
      call getarg(3, numres)
      call getarg(4, numatm)
      call getarg(5, index_out)
c
c     Reading input variables
c----------------------------------
      read(numres,*)nres                  ! # of residues
      read(numatm,*)natom                 ! # of atom
c----------------------------------
c     reading input on order parameters
c
      open(11, file = atomid_in, status= "old")
c
      read(11,*)nop                       ! # of order parameters
      do i=1,nop
         read(11,*)it                     ! type of OP (=2 dist)
!         write(*,*)i,it
          do j=1,it
            read(11,*)inpres(i,j), inpatm(i,j) ! residue no. &  atom name
!            write(*,*)inpres(i,j), inpatm(i,j) ! residue no. &  atom name
          end do
          iopt(i)=it
      end do
      close(11)
c--------------------------------------------------
c     reading input topology file
c
      open(12, file = prmtop_in, status= "unknown")
c
 100  read(12,1)pline
 1    format(a)
c
      if(pline(7:15).eq.'ATOM_NAME')then
```

```
           read(12,*)
           read(12,2) (atmnam(j), j = 1, natom)
 2         format(20a4)
      else
           go to 100
      end if
c
 110  read(12,1)pline
      if(pline(7:21).eq.'RESIDUE_POINTER')then
            read(12,*)
            read(12,3)(ires(j), j = 1, nres)
 3          format(10i8)
      else
           go to 110
      end if
c
      close(12)
c------------------------------------------------
c     retrieving indices of atoms defining order parameters
c
      open(13, file = index_out, status= "unknown")
c
      write(13,*)nop
      do 20 i = 1, nop
            it=iopt(i)
            write(13,*)it
            do 25 k=1,it
               ik=inpres(i,k)
               nam=inpatm(i,k)
!               write(*,*)ik,nam
               do 30 j = ires(ik), ires(ik+1)-1
                     if (nam.eq.atmnam(j)) write(13,*)j
 30            continue
 25         continue
 20   continue
      close(13)
c
      stop
      end
```

The **index.out** file format is shown below:

```
1    ###Two atoms constitute an OP here, type atom indices as 1,2, and so
2    ###on if necessary.
441  ###Then type the actual atom indices from the coordinate file
475
```

Next, create a **barrier.in** file which contains information on the range of values of the barrier region based on the defined order parameter. The format of the **barrier.in** file is shown below:

```
2      ###Number of OPs
1 7 8  ###List variables (1,2,...) and ranges (e.g. 7 8)
2 5 6
```

Next, create a **state.in** file which contains a slightly wider range of values compared to what was set in the **barrier.in** file. The format is shown below:

```
6.5 8.5
4.5 6.5
```

Then, run the following lines from the **aimless.csh** file:

```
gfortran -O op_barrier.f -o op_barr

# setting up general variables

set parm_in = "dsrna_final.prmtop"

set atmid_in = "atom_id.in"

set natom = 39305

set nres = 12709

set nsnap = 20000

set ind_out = "index.out"
./op_barr ds.crd $ind_out $natom $nsnap barrier.in state.in opbarrier.out
```

By executing this command, the script will output an **opbarrier.out** file which contains a list of frames from the seed trajectory (labeled as **ds.crd** in the above script) which correspond to a barrier region defined in the **barrier.in** file. The next step is to create the **stockrst.in** file which contains information on the dimensions of the water box. The format of the file is shown below:

```
default_name
39305  0.2000000E+06 #Number of atoms and water box dimensions below
  70.2068726  70.5446739  80.4372217  90.0000000  90.0000000  90.0000000
```

Next, the restart files for shooting trajectories are created by executing the following lines in the **aimless.csh** file:

```
#location of output directory
```

```
setenv OUTDIR /home/levintov/ds_rna/r4/shooting/run

gfortran -O getsp.f -o getsp
gfortran -O shoot3.f -o shoot

# setting up general variables

set parm_in = "dsrna_final.prmtop"

set atmid_in = "atom_id.in"

set natom = 39305

set nres = 12709

set nsnap = 20000

set ind_out = "index.out"

set nc = ‘wc op_barrier.out | awk ’{print $1}’‘
echo $nc

set n0  = 1

set n0max = $nc

while ($n0 <= $n0max )


mkdir -p $OUTDIR/sp_$n0

setenv OUTDIR2 /home/levintov/dsrna/r4/shooting/run/sp_$n0

./getsp ds.crd ds.vel op_barrier.out $nc $n0 $natom $nsnap
stock_rst.in barrier.rst

#setting up loop variables for shooting trials from a given shooting point

    set ns = 1
    set nmax = 1

    while ($ns <= $nmax)

       mkdir -p $OUTDIR/shoot_$n0

       set nr = ‘wc random.in | awk ’{print $1}’‘
        echo $nr

./shoot barrier.rst $natom mass.in momentum.in random.in $nr $ns
velocity.out shoot.out shoot.rst

mv shoot.rst    $OUTDIR/shoot_$n0

@ ns++
end

@ n0 = $n0 + 50 ###Define which frames to extract from seed trajectory
end
```

The script uses the **getsp.f** file which extracts the coordinate, velocity and restart information from the seed trajectory coordinate and velocity files and the **shoot3.f** file which generates new velocities for particles and generates a new restart file for a shooting simulation. The **getsp.f** file is shown below. This script was originally created by Dr. Sanjib Paul.

```
c      program getsp   (Version 2)
c
c      this version works with TPS-AMBER for aimless shooting
c      used to carry out forward propagation only
c
c      for use in calculation of reaction coordinate by likelihood
c      maximization
c*************************************************************
c      this program
c      (1) reads in a trajectory,
c      (2) generate restart files at barrier regions.
c*************************************************************
       implicit real*4 (a-h, o-z)
c
       parameter (n1=50000, n2= 1000)
c
       character*255 mdcrd_in, op_in
       character*255 rststock_in, mdvel_in
       character*255 sp_out, ibar
       character*255 numset, numatom, numbar
c
       character*80 lstock1,lstock2,lstock3
c
       dimension x0(n1),y0(n1),z0(n1)
       dimension vx0(n1),vy0(n1),vz0(n1)
       dimension vx(n1),vy(n1),vz(n1)
       dimension ibarrier(n2)
c**********************************************************************
c      reading inputs
c------------------------
c      command line inputs
c
       call getarg (1, mdcrd_in)      ! name of AMBER Coord file, input
       call getarg (2, mdvel_in)      ! name of AMBER velocity file
       call getarg (3, op_in)         ! time slices in which seed trj. in
       !barrier region.
       call getarg (4, numbar)        ! total # of sanps in which seed trj.
       !in barrier region.
```

```
      call getarg (5, ibar)            ! which no. snaps going to extract.
      call getarg (6, numatom)         ! # of atom in system.
      call getarg (7, numset)         ! # of snaps.
      call getarg (8, rststock_in)
      call getarg (9, sp_out)          ! output file at Amber .restart file
c
      read(numset,*)nset
      read(numatom,*)natom
      read(numbar,*)nbar
      read(ibar,*)ib
c
      write(*,*)'nset, natom = ',nset,natom
      write(*,*)'nbar = ',nbar
c---------------------------
c     reading information on order parameter
c     calculated along input trajectory
c
      open(unit=31,file=op_in,status='old')
c
      do i=1,nbar
        read(31,*)ibarrier(i) ! trj time slice # residing at barrier
      end do
      close(31)
      ipb = ibarrier(ib)
c----------------------------------------------------------------------------
c     reading stock information for restart file
c
      open(unit=56, file=rststock_in,status='old')
c
      read(56,9)lstock1
      read(56,9)lstock2
      read(56,9)lstock3
      close(56)
 9    format(a)
c----------------------------
c
c     READING of INPUT FILES ENDS HERE
c
c*******************************************************************
c-------------------------------------------------------------------
c     retrieving and recording at the shooting point
c     coordinate and velocity of each atom from input mdcrd file
c
      do i=1,natom
         x0(i)=0.0     ! initializing array for coordinates & velocities
         y0(i)=0.0     ! at the chosen shooting point
         z0(i)=0.0
         vx0(i)=0.0
         vy0(i)=0.0
         vz0(i)=0.0
      end do
c
      write(*,*)'retrieving mass weighted coordinates and velocities'
c
```

```
      call getcoord(mdcrd_in,nset,ipb,natom,x0,y0,z0)
      call getvel(mdvel_in,nset,ipb,natom,vx0,vy0,vz0)
c
      write(*,91)x0(1),y0(1),z0(1)
      write(*,91)vx0(1),vy0(1),vz0(1)
      write(*,91)x0(natom),y0(natom),z0(natom)
      write(*,91)vx0(natom),vy0(natom),vz0(natom)
 91   format(3f12.3)
c
c----------------------------------------------------------------------
      open(unit=68,file=sp_out,status='unknown')
c
      write(68,9)lstock1
      write(68,9)lstock2
      write(68,77)(x0(i),y0(i),z0(i),i=1,natom)
      write(68,77)(vx0(i),vy0(i),vz0(i),i=1,natom)
      write(68,9)lstock3
c
      close(68)
 77   format(6f12.7)
c--------------------------------------------------------------------
      stop
      end
c*************************************************************************
c     Subroutines start here
c*************************************************************************
      subroutine getcoord(mdcrd_in,nset,iset,natom,x0,y0,z0)
c
      implicit real*4 (a-h, o-z)
c
      character*255 mdcrd_in
      character*80 mline
c
      dimension x0(natom),y0(natom),z0(natom)
      dimension x(natom),y(natom),z(natom)
c
      open(unit=12,file=mdcrd_in,status='old')
c
      read(12,88)mline
 88   format(a)
c
      do i=1,nset
         read(12,99)(x(j),y(j),z(j),j=1,natom)
 99      format(10f8.3)
         read(12,*)
         if(i.eq.iset) then
            do j=1,natom
              x0(j)=x(j)
              y0(j)=y(j)
              z0(j)=z(j)
            end do

         endif
c
```

```
      end do
      close(12)
c
      return
      end
c*************************************************************************
      subroutine getvel(mdvel_in,nset,iset,natom,vx0,vy0,vz0)
c
      implicit real*4 (a-h, o-z)
c
      character*255 mdvel_in
      character*80 mline
c
      dimension vx0(natom),vy0(natom),vz0(natom)
      dimension vx(natom),vy(natom),vz(natom)
c
      open(unit=22,file=mdvel_in,status='old')
c
      read(22,98)mline
 98   format(a)
c
      do i=1,nset
         read(22,99)(vx(j),vy(j),vz(j),j=1,natom)
 99      format(10f8.3)
         if(i.eq.iset) then
            do j=1,natom
              vx0(j)=vx(j)
              vy0(j)=vy(j)
              vz0(j)=vz(j)
            end do
         endif
cc
      end do
      close(22)
c
      return
      end
c*************************************************************************
```

The **shoot3.f** file is shown below. This script was originally created by Dr. Sanjib Paul.

```
c     program shoot_amber   (Version 2)
c
c     this version works with TPS-AMBER for aimless shooting
c     used to carry out forward propagation only
c
c     for use in calculation of reaction coordinate by likelihood
c     maximization
c*************************************************************
c     this program
c     (1) reads coordinates,velocites and box information from file
c         generated by getsp.f.
c     (2) generate new velocity.
```

```
c      (3) prepares the input for subsequent forward AMBER run
c
c      Important note:
c
c      After a random momentum displacement, this program sets the
c      c.o.m at rest and then introduces correction for changes
c      in total angular momentum. Finally a temperature rescaling is
c      carried out and ensured that c.o.m. is at rest
c*************************************************************
       implicit real*4 (a-h, o-z)
c
       parameter (n1=50000, n2= 100)
c
       character*255 mass_in, momentum_in,random_in
       character*255 shoot_rst
       character*255 velocity_out, shoot_out, sp_out
       character*255 numatom, numran, num_call
c
       character*80 lstock1,lstock2,lstock3
c
       dimension x0(n1),y0(n1),z0(n1)
       dimension vx0(n1),vy0(n1),vz0(n1)
       dimension vx(n1),vy(n1),vz(n1)
       dimension vxn(n1),vyn(n1),vzn(n1)
       dimension ibarrier(n2)
       dimension amass(n1), smass(n1)
c*********************************************************************
c      reading inputs
c-----------------------
c      command line inputs
c
       call getarg (1, sp_out)          ! configuration at barrier region
       call getarg (2, numatom)         ! # of atoms in the system
       call getarg (3, mass_in)         ! mass of each atom
       call getarg (4, momentum_in)     ! input for momentum displacement
       call getarg (5, random_in)       ! input for idum
       call getarg (6, numran)          ! # of entries in random_in
       call getarg (7, num_call)        ! shooting step #
       call getarg (8, velocity_out)    ! output
       call getarg (9, shoot_out)       ! record of time slice of shooting
       call getarg (10, shoot_rst)      ! file for restarting shooting step
c
       read(numatom,*)natom
       read(numran,*)nrtot
       read(num_call,*)nshoot
c
       write(*,*)' natom = ', natom
       write(*,*)' nrtot = ', nrtot
       write(*,*)' nshoot = ',nshoot
c------------------------------------------------------------------------
c        reading input for momentum displacement
c
         open(unit=13,file=momentum_in,status='old')
c
```

```
        read(13,*)idum            ! seed for random number generator
        read(13,*)iscale          ! For iscale = 1, updated velocities
                                  ! are scaled to maintain constant temp
        read(13,*)sigvel          ! unit conversion factor for velocity
        read(13,*)temp            ! temperature
        read(13,*)nc              ! # of constraints
        read(13,*)pdbvelfact      ! pdb vel factor
        read(13,*)constfact       ! a constant factor
        !added JACS, 2005, 127, 13822
        close(13)
c------------------------------------------------------------------------
c       reading input for mass of atoms
c
        open(unit=76,file=mass_in,status='old')
c
        totmass=0.0
        do i=1,natom
   read(76,*)amass(i),smass(i)  ! mass & square root of mass of each atom
           totmass=totmass+amass(i)
        end do
        totmass2 = sqrt(totmass)
        close(76)
c
        write(*,*)'total mass of the system =',totmass
        write(*,*)'square root of total mass of the system =',totmass2
c--------------------------
c     reading input of random number
c
      open(unit=54,file=random_in,status='old')
c
      do i=1,nrtot
         read(54,*)xrr
         if(i.eq.nshoot) xran=xrr   !! caution: nrtot > = nshoot
      end do
      close(54)
c---------------------------
c
c     READING of INPUT FILES ENDS HERE
c
c******************************************************************
c     retrieving and recording at the shooting point
c     coordinate and velocity of each atom from input mdcrd file
c
      do i=1,natom
         x0(i)=0.0      ! initializing array for coordinates & velocities
         y0(i)=0.0      ! at the chosen shooting point
         z0(i)=0.0
         vx0(i)=0.0
         vy0(i)=0.0
         vz0(i)=0.0
         vxn(i)=0.0
         vyn(i)=0.0
         vzn(i)=0.0
      end do
```

```
c
      write(*,*)'retrieving mass weighted coordinates and velocities'
c
      open(12, file = sp_out, status = "old")
      read(12,9)lstock1
      read(12,9)lstock2
      read(12,77)(x0(j),y0(j),z0(j),j=1,natom)
      read(12,77)(vx0(j),vy0(j),vz0(j), j=1,natom)
      read(12,9)lstock3
 9    format(a)
c
      write(*,91)x0(1),y0(1),z0(1)
      write(*,91)vx0(1),vy0(1),vz0(1)
      write(*,91)x0(natom),y0(natom),z0(natom)
      write(*,91)vx0(natom),vy0(natom),vz0(natom)
 91   format(3f12.3)
c
c------------------------------------------------------------------
c     modification of velocity using the momentum displacement
c
      write(*,*)'applying momentum displacement'
c
      call momdisp(pdbvelfact,temp,natom,sigvel,constfact,
     #    smass,amass,vx0,vy0,vz0,x0,y0,z0,nc,vxn,vyn,vzn,idum,totmass,
     $    totmass2)
c
      write(*,*)'new and old momenta'
      write(*,*)'particle 1'
      write(*,91)vxn(1),vyn(1),vzn(1)
      write(*,91)vx0(1),vy0(1),vz0(1)
      write(*,*)'particle',natom
      write(*,91)vxn(natom),vyn(natom),vzn(natom)
      write(*,91)vx0(natom),vy0(natom),vz0(natom)
c
      call scale_velocity(natom,vxn,vyn,vzn,amass,totmass,temp,iscale)
c
      write(*,*)'after temperature rescaling'
      write(*,*)'new and old momenta'
      write(*,*)'particle 1'
      write(*,91)vxn(1),vyn(1),vzn(1)
      write(*,91)vx(1),vy(1),vz(1)
      write(*,*)'particle ',natom
      write(*,91)vxn(natom),vyn(natom),vzn(natom)
      write(*,91)vx(natom),vy(natom),vz(natom)
c**************************************************************************
c     writing the output
c---------------------------------------------------------------------
c
c     recording modified velocities
c
      open(unit=67,file=velocity_out,status='unknown')
c
      do i=1,natom
         write(67,91)vxn(i),vyn(i),vzn(i)
      end do
      close(67)
c----------------------------------------------------------------------
c     preparation of restart files for further dynamical propagation
c----------------------------------------------------------------------
      open(unit=68,file=shoot_rst,status='unknown')
c
      write(68,9)lstock1
      write(68,9)lstock2
      write(68,77)(x0(i),y0(i),z0(i),i=1,natom)
      write(68,77)(vxn(i),vyn(i),vzn(i),i=1,natom)
      write(68,9)lstock3
c
      close(68)
 77   format(6f12.7)
c--------------------------------------------------------------------
      stop
      end
c*************************************************************************
c     Subroutines start here
c*************************************************************************
      subroutine momdisp(pdbvelfact,temp,nat,sigvel,constfact,
     #smass1,amass1,vx,vy,vz,x0,y0,z0,nc,vxnew,vynew,vznew,idum,totmass,
     $totmass2)
c
        implicit real*4 (a-h, o-z)
c
       dimension x(nat),y(nat),z(nat)
       dimension x0(nat),y0(nat),z0(nat)
       dimension vx(nat),vy(nat),vz(nat)
       dimension vx1(nat),vy1(nat),vz1(nat)
       dimension vx2(nat),vy2(nat),vz2(nat)
       dimension vx0(nat),vy0(nat),vz0(nat)
       dimension delw0(3)
       dimension fact1(nat),xc(nat),yc(nat),zc(nat)
       dimension vxnew(nat),vynew(nat),vznew(nat), smass1(nat)
       dimension amass1(nat)
       dimension a(3,3),ym(3,3)
       dimension u(3),v(3),cr(3)
       dimension u1(3),v1(3),cr1(3)
       dimension qa(3),pa(3),qcp(3),totl(3)
       dimension omega(3,1),omp(3),delvl(3)
c----------------------------------------------------------------------------
c      checking input information
c
       npart=nat
       xnp=float(npart)
       write(*,*)'within momdisp ',npart,xnp,totmass, totmass2
c---------------------------------------------------------------------------
c   Step 1: scaling coordinates and velocities with individual atom masses

       xcom = 0.0
       ycom = 0.0
       zcom = 0.0
      do i=1,npart
            sm=smass1(i)         ! square root of m_{i}
            am = amass1(i)
            vx1(i)=vx(i)*sm
            vy1(i)=vy(i)*sm      ! mass weighted velocities
            vz1(i)=vz(i)*sm
c
            xcom = xcom + x0(i)*am
            ycom = ycom + y0(i)*am
            zcom = zcom + z0(i)*am

c           write(42,455)x(i),y(i),z(i),vx1(i),vy1(i),vz1(i)
      end do

      xcom = xcom / totmass
      ycom = ycom / totmass
      zcom = zcom / totmass

      do i = 1, npart
        sm=smass1(i)
        xc(i) = (x0(i)-xcom)*sm
        yc(i) = (y0(i)-ycom)*sm
        zc(i) = (z0(i)-zcom)*sm
      enddo
c-------------------------------------------------------------------
c Step 2: generating the random velocity from
c Maxwell-Boltzmann distribution
c
      np3=3*npart                ! 3 * # of particles
c
      do j=1,npart
        sm = smass1(j)
        fact1(j)=sqrt(temp)*sigvel*constfact    ! sigvel: to convert
        !velocity to (amu)^{1/2} Ang ps-1
      end do                       ! given temp is in K and mass is in amu
c
      vxsum=0.0
      vysum=0.0
      vzsum=0.0
c
      do i=1,npart
c
         do j=1,3
            xi1=ran3(idum)
            xi2=ran3(idum)
c
            x1=log(xi1)
            x2=-2.0*x1
            xpart1=sqrt(x2)
c
            pi=4.0*atan(1.0)
            x3=2.0*pi*xi2
            xpart2=cos(x3)

```

```
c
             delw0(j)=xpart1*xpart2*fact1(i)
c
          end do

          vx2(i)=vx1(i)+delw0(1)          ! new momenta after adding
          vy2(i)=vy1(i)+delw0(2)          ! momentum displacement from
          vz2(i)=vz1(i)+delw0(3)          ! gaussian distribution
c
c         write(43,455)vx2(i),vy2(i),vz2(i),(delw0(j),j=1,3)
c
          sm=smass1(i)
          am = amass1(i)
          vxsum=vxsum+(sm*vx2(i))
          vysum=vysum+(sm*vy2(i))
          vzsum=vzsum+(sm*vz2(i))
       end do
c
       vxs=vxsum/totmass2                     ! com momentum
       vys=vysum/totmass2
       vzs=vzsum/totmass2
c
       write(*,*)'velocities of c.o.m after random displacement'
       write(*,*)vxs,vys,vzs
c-------------------------------------------------------------------
c Step 3: Construction of new momenta setting velocity of
c the c.o.m. to zero
c
       do i=1,npart
          sm = smass1(i)
          vx0(i)=(vx2(i)-vxs)
          vy0(i)=(vy2(i)-vys)
          vz0(i)=(vz2(i)-vzs)
c
c         write(45,455)x(i),y(i),z(i),vx0(i),vy0(i),vz0(i)
c         write(45,455)vx2(i),vy2(i),vz2(i),vx0(i),vy0(i),vz0(i)
c455      format(6f15.5)
       end do             ! end of correction for total linear momentum
c-------------------------------------------------------------------------
c       Step 4: Calculation of the correction from angular momentum
c
c       calculation of the moment of inertia tensor
c
        n=3                   ! initialization
        np=3
        do i=1,n
           do j=1,n
              a(i,j)=0.0
           end do
        end do
c
        sum2yz=0.0       ! initializing summations needed
        sum2xz=0.0       ! to evaluate elements of moment
        sum2xy=0.0       ! of inertia tensor as a 3X3 matrix
```

```fortran
        sumxy=0.0
        sumxz=0.0
        sumyz=0.0
c
        do i=1,npart
           xi=xc(i)
           yi=yc(i)
           zi=zc(i)
           sum2yz=sum2yz+(yi*yi)+(zi*zi)
           sum2xz=sum2xz+(xi*xi)+(zi*zi)
           sum2xy=sum2xy+(xi*xi)+(yi*yi)
           sumxy=sumxy+(xi*yi)
           sumxz=sumxz+(xi*zi)
           sumyz=sumyz+(yi*zi)
        end do
c
        a(1,1)=sum2yz
        a(2,2)=sum2xz
        a(3,3)=sum2xy
        a(1,2)=-sumxy
        a(2,1)=-sumxy
        a(1,3)=-sumxz
        a(3,1)=-sumxz
        a(2,3)=-sumyz
        a(3,2)=-sumyz
c
        write(*,*)'elements of MOI matrix'
        do i=1,3
           write(*,*)(a(i,j),j=1,3)
        end do
c
c       calculating the inverse of moment of inertia tensor
c
        call ainverse(a,n,np,ym)
c
        write(*,*)'elements of inverted MOI matrix'
        do i=1,3
           write(*,*)(ym(i,j),j=1,3)
        end do
c
c       calculation of the total angular momentum
c
        do i=1,3
           totl(i)=0.0
        end do
c
        do i=1,npart
           qa(1)=xc(i)
           qa(2)=yc(i)
           qa(3)=zc(i)
           pa(1)=vx0(i)
           pa(2)=vy0(i)
           pa(3)=vz0(i)
c
```

```
c          write(46,455)(qa(m),m=1,3),(pa(k),k=1,3)
c
           call cross(qa,pa,qcp)
           do j=1,3
              totl(j)=totl(j)+qcp(j)
           end do
        end do
c
        write(*,*)'x,y,z components of total angular momentum'
        write(*,*)(totl(j),j=1,3)
c
c       calculation of the angular velocity
c
        call matmult(3,3,ym,3,1,totl,omega)
c
        write(*,*)'components of angular velocity'
        write(*,*)(omega(i,1),i=1,3)
c
c       modified velocity after constraining angular momentum
c       and removal of mass weighting
c
        omp(1)=omega(1,1)
        omp(2)=omega(2,1)
        omp(3)=omega(3,1)
c
        do i=1,npart
           qa(1)=xc(i)
           qa(2)=yc(i)
           qa(3)=zc(i)
           call cross(omp,qa,delvl)
c
           sm=smass1(i)
           vxnew(i)=(vx0(i)-delvl(1))/sm
           vynew(i)=(vy0(i)-delvl(2))/sm
           vznew(i)=(vz0(i)-delvl(3))/sm
        end do
c
        write(*,*)'completed momdisp'
c
        return
        end
c*************************************************************
        subroutine cross(u,v,z)
c
c       calculation of the components of a vector z resulting from
c       a cross product of the 3 dimensional vectors u and v
c
        implicit real*4 (a-h, o-z)
        dimension u(3),v(3),z(3)
c
        z(1)=u(2)*v(3)-u(3)*v(2)
c
        z(2)=u(3)*v(1)-u(1)*v(3)
c
```

```
        z(3)=u(1)*v(2)-u(2)*v(1)
c
        return
        end
c*************************************************************
       subroutine ainverse(a,n,np,y)
c
       implicit real*4 (a-h, o-z)
c
       dimension indx(3)
       dimension a(3,3),y(3,3)
c
c      INTEGER np,indx(np)
c      REAL a(np,np),y(np,np)
c
        do i=1,n
             do j=1,n
                y(i,j)=0.0
             enddo
             y(i,i)=1.0
        enddo
c
        call ludcmp(a,n,np,indx,d)
        do j=1,n
              call lubksb(a,n,np,indx,y(1,j))
        enddo
c
        return
        end
c------------------------------------------------------
      subroutine ludcmp(a,n,np,indx,d)
c
      implicit real*4 (a-h, o-z)
c
      PARAMETER (NMAX=500,TINY=1.0e-20)
      dimension indx(3),a(3,3)
      dimension vv(NMAX)
c
c     INTEGER n,np,indx(n),NMAX
c     REAL d,a(np,np),TINY
c     PARAMETER (NMAX=500,TINY=1.0e-20)
c     INTEGER i,imax,j,k
c     REAL aamax,dum,sum,vv(NMAX)
c
      d=1.
      do 12 i=1,n
        aamax=0.
        do 11 j=1,n
          if (abs(a(i,j)).gt.aamax) aamax=abs(a(i,j))
11      continue
        if (aamax.eq.0.) then
           write(*,*) 'singular matrix in ludcmp'
           stop
        else
```

```
           vv(i)=1./aamax
        endif
12    continue
      do 19 j=1,n
        do 14 i=1,j-1
          sum=a(i,j)
          do 13 k=1,i-1
            sum=sum-a(i,k)*a(k,j)
13        continue
          a(i,j)=sum
14      continue
        aamax=0.
        do 16 i=j,n
          sum=a(i,j)
          do 15 k=1,j-1
            sum=sum-a(i,k)*a(k,j)
15        continue
          a(i,j)=sum
          dum=vv(i)*abs(sum)
          if (dum.ge.aamax) then
            imax=i
            aamax=dum
          endif
16      continue
        if (j.ne.imax)then
          do 17 k=1,n
            dum=a(imax,k)
            a(imax,k)=a(j,k)
            a(j,k)=dum
17        continue
          d=-d
          vv(imax)=vv(j)
        endif
        indx(j)=imax
        if(a(j,j).eq.0.)a(j,j)=TINY
        if(j.ne.n)then
          dum=1./a(j,j)
          do 18 i=j+1,n
            a(i,j)=a(i,j)*dum
18        continue
        endif
19    continue
      return
      END
c------------------------------------
      subroutine lubksb(a,n,np,indx,b)
c
      implicit real*4 (a-h, o-z)
c
      dimension indx(3), a(3,3),b(3)
c
c     INTEGER n,np,indx(n)
c     REAL a(np,np),b(n)
c     INTEGER i,ii,j,ll
```

```
c      REAL sum
c
      ii=0
      do 12 i=1,n
        ll=indx(i)
        sum=b(ll)
        b(ll)=b(i)
        if (ii.ne.0)then
          do 11 j=ii,i-1
            sum=sum-a(i,j)*b(j)
11        continue
        else if (sum.ne.0.) then
          ii=i
        endif
        b(i)=sum
12    continue
      do 14 i=n,1,-1
        sum=b(i)
        do 13 j=i+1,n
          sum=sum-a(i,j)*b(j)
13      continue
        b(i)=sum/a(i,i)
14    continue
      return
      END
c******************************************************************
c subroutine to carry out matrix multiplication
c
        subroutine matmult(m1,n1,a,m2,n2,b,e)
c
        implicit real*4(a-h, o-z)
c
        dimension a(3,3),b(3,1),e(3,1)
c
        if(n1.ne.m2) then
          write(*,*)'multiplication not possible'
          stop
        endif
c
c       write(*,*)'inside matmult'
        do i=1,m1
           do j=1,n2
              sumij=0.0
              do k=1,n1
                 sumij=sumij+(a(i,k)*b(k,j))
              end do
              e(i,j)=sumij
c             write(*,*)sumij
           end do
        end do
c
        return
        end
```

```

c********************************************************
c  subroutine for scaling of velocities to maintain overall kinetic energy
c
       subroutine scale_velocity(npart,vx,vy,vz,amass1,totmass,
     +  temp,iscale)
c
        implicit real*4 (a-h, o-z)
c
        dimension vx(npart),vy(npart),vz(npart)
        dimension amass1(npart)
c
        xn=1.0/(3.0*float(npart))
c
        if(iscale.eq.1) then
          sumvx2=0.0
          sumvy2=0.0
          sumvz2=0.0
          sumvx=0.0
          sumvy=0.0
          sumvz=0.0
c
          do i=1,npart
             am = amass1(i)
             vxi=vx(i)
             vyi=vy(i)
             vzi=vz(i)
c
             sumvx=sumvx+(vxi*am)
             sumvy=sumvy+(vyi*am)
             sumvz=sumvz+(vzi*am)
c
             sumvx2=sumvx2+(am*(vxi**2))
             sumvy2=sumvy2+(am*(vyi**2))
             sumvz2=sumvz2+(am*(vzi**2))
          end do
c
          sumvx = sumvx/totmass
          sumvy = sumvy/totmass
          sumvz = sumvz/totmass
          sumv2=(sumvx2+sumvy2+sumvz2)*xn
c
          fs=sqrt((0.001988*temp)/sumv2)
          write(*,*)’temperature scaling factor =’,fs
c
          do i=1,npart
             vx(i)=fs*(vx(i)-sumvx)
             vy(i)=fs*(vy(i)-sumvy)
             vz(i)=fs*(vz(i)-sumvz)
          end do

        endif
c
        return
        end
c*********************************************************************
      FUNCTION ran3(idum)
      INTEGER idum
      INTEGER MBIG,MSEED,MZ
C     REAL MBIG,MSEED,MZ
      REAL ran3,FAC
      PARAMETER (MBIG=1000000000,MSEED=161803398,MZ=0,FAC=1./MBIG)
C     PARAMETER (MBIG=4000000.,MSEED=1618033.,MZ=0.,FAC=1./MBIG)
      INTEGER i,iff,ii,inext,inextp,k
      INTEGER mj,mk,ma(55)
C     REAL mj,mk,ma(55)
      SAVE iff,inext,inextp,ma
      DATA iff /0/
      if(idum.lt.0.or.iff.eq.0)then
        iff=1
        mj=MSEED-iabs(idum)
        mj=mod(mj,MBIG)
        ma(55)=mj
        mk=1
        do 11 i=1,54
          ii=mod(21*i,55)
          ma(ii)=mk
          mk=mj-mk
          if(mk.lt.MZ)mk=mk+MBIG
          mj=ma(ii)
11      continue
        do 13 k=1,4
          do 12 i=1,55
            ma(i)=ma(i)-ma(1+mod(i+30,55))
            if(ma(i).lt.MZ)ma(i)=ma(i)+MBIG
12        continue
13      continue
        inext=0
        inextp=31
        idum=1
      endif
      inext=inext+1
      if(inext.eq.56)inext=1
      inextp=inextp+1
      if(inextp.eq.56)inextp=1
      mj=ma(inext)-ma(inextp)
      if(mj.lt.MZ)mj=mj+MBIG
      ma(inext)=mj
      ran3=mj*FAC
      return
      END
```

Restart files for shooting trajectories are generated using the above scripts which can then be used for conducting a short MD simulation. In the study presented in chapter 6, each shooting trajectory was 1 ns long. After the shooting trajectories were completed, I checked the last frame of the simulation and determined a value of the OP to understand if the simulation terminated in the reactant or product state. If the simulation terminates in one of the defined states, then the shooting trajectory is used as a new seed trajectory and the whole procedure is repeated from the step when the **opbarrier.out** file is generated.

H.2.1 Likelihood Maximization Method

I used a MATLAB script to perform the likelihood maximization method. Firstly, I executed the following script (**getCVs.csh**) to collect values of a set of CVs from each shooting trajectory. This script was originally created by Dr. Sanjib Paul.

```
setenv PATHDIR /run/media/levintov/easystore2/tps_dih/paths
gfortran -O All_Op_Barrier.f -o allop
set natom = 39305

set n0  = 1

set n0max = 1000

while ($n0 <= $n0max )
./allop $PATHDIR/path$n0/shoot.rst cv.in
$PATHDIR/path$n0/terstate.out $natom CVin.out CVout.out

@ n0++
end
```

This script also requires the **AllOpBarrier.f** file which is shown below. This script was originally created by Dr. Sanjib Paul.

```
c     This program reads rst file generated from shoot.f and then
c     calculates OP values at shooting point.
c---------------------------------------------------------------------

      parameter (n1 = 40000, n2 = 100)
      implicit real*4 (a-h,o-z)
c
      dimension x(n1), y(n1), z(n1), indx(n2,n2), iopt(n2)
      dimension xind(n2), yind(n2), zind(n2), rop(n2)
      dimension vx(n1), vy(n1), vz(n1)
c
      character*256 rst_in,numatom,op_in,op1_out,op2_out,terstate_out

      call getarg(1, rst_in)
      call getarg(2, op_in)
      call getarg(3, terstate_out)
      call getarg(4, numatom)
      call getarg(5, op1_out)
      call getarg(6, op2_out)
c
      open(18, file = op1_out, status = "unknown",access = "append")
      open(19, file = op2_out, status = "unknown",access = "append")
      open(91, file = terstate_out, status = "old")
      read(91,*)iter
c     -----------------------------------
c     Reading input variables from scripts.
c     ------------------------------------
      read(numatom,*)natom     ! # of atoms
c----------------------------------------------
c     reading input on barier and individual states from file
c
      open(11, file = op_in, status = "old")
c
      read(11,*)nop                    ! # of OP
      do 10 i = 1, nop
         read(11,*)it                    ! type of OP
         do 15 j = 1, it
            read(11,*)indx(i,j)
 15      continue
         iopt(i) = it              ! type of i-th OP
 10   continue
      close(11)
c
!       do 110 i = 1, nop
!         do 120 j = 1, iopt(i)
!             write(*,*)indx(i,j)
!120      continue
!110      continue
c
c   ---------------------------------------------------------------
c     reading of rst file & calculation of OP.
c  ----------------------------------------------------------------
c
      open(12, file = rst_in, status = "unknown")
c
      read(12,*)                       ! to skip the heading
      read(12,*)
c
        read(12,99)(x(j), y(j), z(j), j = 1, natom) ! coordinate reading
99      format(6f12.7)
c
        read(12,99)(vx(j), vy(j), vz(j), j = 1, natom)
        read(12,99)a, b, c, aa, ba, ca

c
        do 30 j = 1,nop
              ip = iopt(j)
              do 50 k = 1,ip
                    ik=indx(j,k)
                    xind(k) = x(ik)
                    yind(k) = y(ik)
                    zind(k) = z(ik)
 50           continue
              call opbar(ip,xind,yind,zind,a,b,c,vop)
              rop(j) = vop
 30     end do
         if(iter.eq.1)then
         write(18,1)(rop(j),j=1,nop)
         end if
         if(iter.eq.-1)then
         write(19,1)(rop(j),j=1,nop)
         end if
1        format(13f10.3)
      stop
      end
c**********************************************************
c     Subroutines starts here.
c-----------------------------
      subroutine opbar(ip,xind,yind,zind,a,b,c,vop)
c
      implicit real*4 (a-h, o-z)
c
      dimension xind(ip),yind(ip),zind(ip)
c
      if(ip.eq.2) call getdis (xind, yind, zind, a,b,c, vop)
      if(ip.eq.3) call getang (xind, yind, zind, a,b,c, vop)
      if(ip.eq.4) call getdih (xind, yind, zind, a,b,c, vop)
c

      return
      end
c---------------------------------------------------------
      subroutine getdih (xi, yi, zi, a,b,c,dih)
      real xi(4), yi(4), zi(4), v21(3), v23(3), v34(3)

         v21(1) = xi(1) - xi(2)
         v21(2) = yi(1) - yi(2)
         v21(3) = zi(1) - zi(2)

         v23(1) = xi(3) - xi(2)
         v23(2) = yi(3) - yi(2)
         v23(3) = zi(3) - zi(2)

         v34(1) = xi(4) - xi(3)
         v34(2) = yi(4) - yi(3)
         v34(3) = zi(4) - zi(3)

         call orthonorm (v23, v21)
         call orthonorm (v23, v34)

         cdih = dot(v21, v34)

         call xcross (v23, v21, pn)
         sdih = dot(pn, v34)

         dih = atan2(sdih, cdih)

         pi = 4.0*atan(1.0)
         dih = (dih*180.0)/pi

         return
         end
c---------------------------------------------------------
      subroutine orthonorm ( v1, v2 )
c
      real v1 ( 3 ), v2 ( 3 )
c
      rv1m1 = 1.0 / veclen ( v1 )
c
      v1 ( 1 ) = v1 ( 1 ) * rv1m1
      v1 ( 2 ) = v1 ( 2 ) * rv1m1
      v1 ( 3 ) = v1 ( 3 ) * rv1m1
c
        v1dv2 = dot ( v1, v2 )
c
      v2 ( 1 ) =  v2 ( 1 ) - v1dv2 * v1 ( 1 )
      v2 ( 2 ) =  v2 ( 2 ) - v1dv2 * v1 ( 2 )
      v2 ( 3 ) =  v2 ( 3 ) - v1dv2 * v1 ( 3 )
      rv2m1 = 1.0 / veclen ( v2 )
c
      v2 ( 1 ) = v2 ( 1 ) * rv2m1
      v2 ( 2 ) = v2 ( 2 ) * rv2m1
      v2 ( 3 ) = v2 ( 3 ) * rv2m1
       return
      end
c----------------------------------------------------------
      subroutine xcross ( v1, v2, v3 )
      real v1(3), v2(3), v3(3)
      v3 ( 1 ) = v1 ( 2 ) * v2 ( 3 ) - v1 ( 3 ) * v2 ( 2 )
      v3 ( 2 ) = v1 ( 3 ) * v2 ( 1 ) - v1 ( 1 ) * v2 ( 3 )
      v3 ( 3 ) = v1 ( 1 ) * v2 ( 2 ) - v1 ( 2 ) * v2 ( 1 )

      return
      end
c---------------------------------------------------------
      function dot ( v1, v2 )
c
      real v1 ( 3 ), v2 ( 3 )
      dot = v1 ( 1 ) * v2 ( 1 ) + v1 ( 2 ) * v2 ( 2 )
     $ + v1 ( 3 ) * v2 ( 3 )
c
      return
      end
c---------------------------------------------------------
c
       function veclen ( v )
```

```
c
      real v ( 3 )
c
      veclen = sqrt ( v ( 1 ) ** 2 + v ( 2 ) ** 2 + v ( 3 ) ** 2 )
c
      return
      end
c-------------------------------------------------------

      subroutine getdis (xi, yi, zi, a,b,c,dis)
       real xi(2), yi(2), zi(2), v(3)
       v(1) = xi(2) - xi(1)
       v(1) = v(1) - a*anint(v(1)/a)
       v(2) = yi(2) - yi(1)
       v(2) = v(2) - b*anint(v(2)/b)
       v(3) = zi(2) - zi(1)
       v(3) = v(3) - c*anint(v(3)/c)
       dis = veclen(v)
      return
      end
c-------------------------------------------------------
      subroutine getang (xi, yi, zi, a,b,c,ang)
       real xi(3), yi(3), zi(3), v1(3), v2(3)
       v1(1) = xi(1) - xi(2)
       v1(2) = yi(1) - yi(2)
       v1(3) = zi(1) - zi(2)
       v2(1) = xi(3) - xi(2)
       v2(2) = yi(3) - yi(2)
       v2(3) = zi(3) - zi(2)
       r1 = veclen(v1)
       r2 = veclen(v2)
       dotp = dot(v1,v2)
       theta = acos(dotp/(r1*r2))
        pi = 4.0*atan(1.0)
       ang = (theta*180.0)/pi

      return
      end
c---------------------------------------------------------
c                         END
c---------------------------------------------------------
```

The resulting CV values are normalized using MATLAB:

```
load -ASCII CV_all.mat
N1 = normalize(CV_all,1);
```

Next, the likelihood maximization script is shown below. It should be executed once. This script was originally created by Dr. Sanjib Paul.

```
%To select different CVs, vary the value in the CV_all_nor(:,12)
```

```
function l=log_lik(alpha,CV_all_nor)
r=CV_all_nor(:,12)*alpha(1);   %Single variable model
% r=CV_all_nor(:,[1 4])*alpha([1;2]);   %Two variable model
% r=CV_all_nor(:,[10 11 12])*alpha([1;2;3]); %Three variable model
% r=CV_all_nor(:,[11 12 13 14])*alpha([1;2;3;4]);
% r=CV_all_nor(:,[1 9 10 12 16])*alpha([1;2;3;4;5]);
p=(1+tanh(r))/2;
l=0;
for i=1:252     %%%Outward simulations
    l=l+log(p(i));
end
for i=253:1000  %%%Inward simulations
    l=l+log(1-p(i));
end
l=-l;
end
```

Then, the following lines of code should be typed in the MATLAB console:

```
1 load -ASCII CV_all_nor.mat
2 options = optimset('Display','iter','MaxIter',10000,'TolX',10^-6
3 ,'TolFun',10^-6);
4 alpha0=[0.05];
5 [alpha,fval,exitflag,output,grad,hessian]=fminunc('log_lik'
6 ,alpha0,options,CV_all);
```

The first line loads the data file, the second and third lines define the options for maximizing the likelihood function, the fourth line defines the initial a_0 parameter which is then optimized to produce the maximum value of the likelihood function, and the last two lines execute the likelihood maximization script. The number of CVs can be varied by varying r model in the likelihood maximization script. Another a_0 value needs to be added to the fourth line when an additional CV is added to the model. After execution of the script, the MATLAB program outputs likelihood scores and a_n values. In case if the script output states that the results were not converged, one needs to vary the a_0 values until the convergence is reached.

APPENDIX I

FORCE-FIELD FILES

I.1 Preparation of Parameter Files for Small Molecules Using GAFF

The first step is to use the Antechamber program to generate files (**.PREPI** and **.FRCMOD**) which will contain geometry and charge information of the modeled ligand and will be read as input by the TLEAP later.

```
$AMBERHOME/bin/antechamber -i my.pdb -fi pdb -o my.prepi -fo prepi -nc 0
-pf y -c bcc

-help print these instructions and other that are not included here
-i       input file name
-fi     input file format
-o     output file name
-fo    output file format
-c      charge method
-nc    net molecular charge
-pf     remove the intermediate files: can be yes (y) and no (n, default)
-rn     residue name, if not available in the input file

List of file formats that can be used as output files:

$AMBERHOME/bin/parmchk -i my.prepi -f prepi -o my.frcmod
```

I.2 Parameter Files for Small Molecules

I am providing the parameters from the **.PREPI** and **.FRCMOD** files that were generated in studies presented in chapters 3, 4, and 7. Each file is labeled by the PDB name.

I.2.1 PDB Code: 1LVJ

The **.PREPI** file is shown below:

```
    0    0    2

This is a remark line
molecule.res
PMZ   INT  0
CORRECT     OMIT DU   BEG
  0.0000
   1  DUMM  DU    M    0  -1  -2     0.000       .0        .0     .00000
   2  DUMM  DU    M    1   0  -1     1.449       .0        .0     .00000
   3  DUMM  DU    M    2   1   0     1.523    111.21        .0     .00000
   4  CE1   c3    M    3   2   1     1.540   111.208  -180.000  0.157100
   5  HE11  h1    E    4   3   2     1.079    83.882    19.719  0.032200
   6  HE12  h1    E    4   3   2     1.080    79.216   -91.729  0.032200
   7  HE13  h1    E    4   3   2     1.084   159.299   150.960  0.032200
   8  ND    n3    M    4   3   2     1.486    50.453   141.450 -0.738600
   9  CE2   c3    3    8   4   3     1.504   111.659   -88.694  0.157100
  10  HE21  h1    E    9   8   4     1.081   109.840   154.724  0.032200
  11  HE22  h1    E    9   8   4     1.080   109.887    34.388  0.032200
  12  HE23  h1    E    9   8   4     1.082   109.268   -85.594  0.032200
  13  CG    c3    M    8   4   3     1.485   111.469   147.025  0.172800
  14  HG1   h1    E   13   8   4     1.076   107.235    89.377  0.050200
  15  HG2   h1    E   13   8   4     1.069   108.057  -154.698  0.050200
  16  CB1   c3    M   13   8   4     1.567   117.660   -32.040 -0.155400
  17  HB11  hc    E   16  13   8     1.078   108.543   -69.134  0.054700
  18  HB12  hc    E   16  13   8     1.069   109.120    48.655  0.054700
  19  CA1   c3    M   16  13   8     1.565   113.474   170.739  0.042300
  20  HA11  h1    E   19  16  13     1.074   107.640     6.642  0.060700
  21  HA12  h1    E   19  16  13     1.069   107.996   123.054  0.060700
  22  N1    na    M   19  16  13     1.484   117.127  -115.427 -0.168100
  23  C6    ca    M   22  19  16     1.366   120.197   -67.022  0.060700
  24  C1    ca    M   23  22  19     1.402   121.164    17.607 -0.139000
  25  H1    ha    E   24  23  22     1.078   120.219     0.537  0.140000
  26  C2    ca    M   24  23  22     1.401   119.813  -179.615 -0.150600
  27  CA2   c     B   26  24  23     1.502   120.296  -179.606  0.575700
  28  CB2   c3    3   27  26  24     1.510   121.948    14.107 -0.198100
  29  HB21  hc    E   28  27  26     1.078   109.599  -169.614  0.064367
  30  HB22  hc    E   28  27  26     1.083   108.898    70.557  0.064367
  31  HB23  hc    E   28  27  26     1.081   109.437   -49.285  0.064367
  32  OB3   o     E   27  26  24     1.221   119.381  -169.241 -0.532100
  33  C3    ca    M   26  24  23     1.392   120.239     0.210 -0.092000
  34  H3    ha    E   33  26  24     1.080   120.029   179.968  0.160000
  35  C4    ca    M   33  26  24     1.391   119.917    -0.222 -0.107000
  36  H4    ha    E   35  33  26     1.072   119.790   179.911  0.148000
  37  C5    ca    M   35  33  26     1.393   120.340     0.100  0.002900
  38  S5    ss    M   37  35  33     1.767   119.583  -179.717 -0.147800
  39  C7    ca    M   38  37  35     1.765    91.786   140.255 -0.031100
  40  C8    ca    M   39  38  37     1.398   120.091  -140.324 -0.083000
  41  H8    ha    E   40  39  38     1.080   120.090     0.476  0.143000
  42  C9    ca    M   40  39  38     1.398   119.911  -179.760 -0.151000
  43  H9    ha    E   42  40  39     1.081   120.043  -179.985  0.138000
  44  C10   ca    M   42  40  39     1.399   119.865     0.199 -0.097000
  45  H10   ha    E   44  42  40     1.079   120.043   179.971  0.134000
  46  C11   ca    M   44  42  40     1.395   119.929    -0.163 -0.176000
  47  H11   ha    E   46  44  42     1.069   120.117   179.558  0.138000
  48  C12   ca    M   46  44  42     1.388   120.373    -0.050  0.079700


LOOP
  C12   N1
   C5   C6
  C12   C7

IMPROPER
  CA1   C6   N1  C12
   C1   C5   C6   N1
   C6   C2   C1   H1
  CA2   C1   C2   C3
  CB2   C2  CA2  OB3
   C2   C4   C3   H3
   C3   C5   C4   H4
   C6   C4   C5   S5
   C8  C12   C7   S5
   C7   C9   C8   H8
   C8  C10   C9   H9
   C9  C11  C10  H10
  C10  C12  C11  H11
   C7  C11  C12   N1

DONE
STOP
```

The **.FRCMOD** file is shown below:

```
1LVJ
remark goes here
MASS

BOND

ANGLE

DIHE

IMPROPER
ca-ca-ca-na         1.1          180.0          2.0  Using default value
ca-ca-ca-ha         1.1          180.0          2.0  General improper
torsional angle (2 general atom types)
c -ca-ca-ca         1.1          180.0          2.0  Using default value
c3-ca-c -o         10.5          180.0          2.0  General improper
torsional angle (2 general atom types)
ca-ca-ca-ss         1.1          180.0          2.0  Using default value

NONBON
```

I.2.2 PDB Code: 2L8H

The **.PREPI** file is shown below:

```
    0    0    2

This is a remark line
molecule.res
L8H   INT  0
CORRECT     OMIT DU   BEG
  0.0000
   1  DUMM  DU    M    0  -1  -2      0.000        .0        .0    .00000
   2  DUMM  DU    M    1   0  -1      1.449        .0        .0    .00000
   3  DUMM  DU    M    2   1   0      1.523    111.21        .0    .00000
   4  NH1   N2    M    3   2   1      1.540   111.208  -180.000 -0.493200
   5  HH11  H     E    4   3   2      1.000    32.436   173.577  0.333450
   6  HH12  H     E    4   3   2      1.000   146.080  -138.341  0.333450
   7  CZ    CA    M    4   3   2      1.306    90.790    17.880  0.526300
   8  NH2   N2    B    7   4   3      1.305   120.210  -165.542 -0.493200
   9  HH21  H     E    8   7   4      1.001   120.040  -179.713  0.333450
  10  HH22  H     E    8   7   4      0.999   120.026     0.576  0.333450
  11  NE    N2    M    7   4   3      1.328   119.983    14.793 -0.415100
  12  HE    H     E   11   7   4      0.979   119.522     0.151  0.329700
  13  CD    CT    M   11   7   4      1.456   121.113  -179.946  0.011300
  14  HD2   H1    E   13  11   7      1.079   109.352   -44.021  0.100200
  15  HD3   H1    E   13  11   7      1.081   109.482  -163.649  0.100200
  16  CG    CT    M   13  11   7      1.527   110.204    76.224 -0.112400
  17  HG2   HC    E   16  13  11      1.079   109.477   -50.677  0.066200
  18  HG3   HC    E   16  13  11      1.079   109.407  -170.687  0.066200
  19  CB    CT    M   16  13  11      1.530   110.070    69.276 -0.105400
  20  HB2   HC    E   19  16  13      1.083   109.777    83.177  0.096700
  21  HB3   HC    E   19  16  13      1.081   109.528   -37.159  0.096700
  22  CA    CT    M   19  16  13      1.534   108.550  -156.875  0.014500
  23  N     N3    3   22  19  16      1.491   110.097    62.857 -0.838600
  24  H2    H     E   23  22  19      1.041   109.400   129.921  0.469133
  25  H     H     E   23  22  19      1.039   109.679     9.660  0.469133
  26  HXT   H     E   23  22  19      1.040   109.343  -110.285  0.469133
  27  HA    HP    E   22  19  16      1.079   109.195  -177.318  0.151700
  28  C     C     M   22  19  16      1.529   109.568   -57.350  0.641100
  29  O     O     E   28  22  19      1.232   120.832  -104.047 -0.536100
  30  NA    N2    M   28  22  19      1.329   115.866    75.174 -0.446100
  31  HN    H     E   30  28  22      0.980   119.536    -0.319  0.304500
  32  CAT   CA    M   30  28  22      1.334   121.059  -179.398  0.032600
  33  CAJ   CA    M   32  30  28      1.395   120.597    -1.587 -0.252000
  34  HAJ   HA    E   33  32  30      1.002   119.887   -10.676  0.158000
  35  CAU   CA    M   33  32  30      1.393   119.530  -178.959  0.197100
  36  OAQ   OS    S   35  33  32      1.356   119.943   179.593 -0.306900
  37  CAA   CT    3   36  35  33      1.429   108.777   -85.805  0.107700
  38  HAA   H1    E   37  36  35      1.000   116.058    60.456  0.057033
  39  HAAA  H1    E   37  36  35      1.000   105.755   177.810  0.057033
  40  HAAB  H1    E   37  36  35      0.999   116.165   -64.806  0.057033
  41  CAW   CA    M   35  33  32      1.390   120.456    -0.052 -0.053000
  42  CAI   CA    M   41  35  33      1.393   120.218   177.999 -0.085000
  43  HAI   HA    E   42  41  35      1.000   119.848     6.293  0.166000
  44  CAG   CA    M   42  41  35      1.391   120.258  -178.313 -0.115000
  45  HAG   HA    E   44  42  41      0.999   119.743  -168.464  0.159000
  46  CAF   CA    M   44  42  41      1.393   119.767    -0.770 -0.085000
  47  HAF   HA    E   46  44  42      1.000   119.505  -167.730  0.154000
  48  CAH   CA    M   46  44  42      1.392   119.901     0.892 -0.140000
  49  HAH   HA    E   48  46  44      1.004   119.761   176.668  0.137000
  50  CAV   CA    M   48  46  44      1.391   120.246    -0.221 -0.003000
  51  CAK   CA    M   50  48  46      1.391   120.203   177.918 -0.174000
  52  HAK   HA    E   51  50  48      1.001   119.731   -15.308  0.128000


LOOP
  CAK  CAT
  CAV  CAW

IMPROPER
    NE  NH1   CZ  NH2
    CA   NA    C    O
     C  CAT   NA   HN
  CAJ  CAK  CAT   NA
  CAT  CAU  CAJ  HAJ
  CAJ  CAW  CAU  OAQ
  CAU  CAI  CAW  CAV
  CAW  CAG  CAI  HAI
  CAI  CAF  CAG  HAG
  CAG  CAH  CAF  HAF
  CAF  CAV  CAH  HAH
  CAW  CAH  CAV  CAK
  CAT  CAV  CAK  HAK

DONE
STOP
```

The **.FRCMOD** file is shown below:

```
2L8H
remark goes here
MASS
N2 14.010        0.530               same as n3
H  1.008         0.161               same as hn
CA 12.010        0.360               same as c2
CT 12.010        0.878               same as c3
H1 1.008         0.135               same as hc
HC 1.008         0.135               same as hc
N3 14.010        0.530               same as n4
HP 1.008         0.135               same as hc
C  12.010        0.616               same as c
O  16.000        0.434               same as o
HA 1.008         0.135               same as hc
OS 16.000        0.465               same as os

```

```
BOND
N2-H   392.40   1.019       same as hn-n3
N2-CA  417.90   1.386       same as ca-nh
N2-CT  325.90   1.465       same as c3-n3
CT-H1  330.60   1.097       same as c3-hc
CT-CT  300.90   1.538       same as c3-c3
CT-HC  330.60   1.097       same as c3-hc
CT-N3  283.30   1.511       same as c3-n4
CT-HP  330.60   1.097       same as c3-hc
CT-C   313.00   1.524       same as c -c3
N3-H   373.20   1.030       same as hn-n4
C -O   637.70   1.218       same as c -o
C -N2  490.00   1.335
CA-CA  461.10   1.398       same as ca-ca
CA-HA  344.30   1.087       same as c2-hc
CA-OS  389.20   1.360       same as c2-os
OS-CT  308.60   1.432       same as c3-os

ANGLE
N2-CA-N2   70.270     120.980   same as nh-ca-nh
H -N2-H    41.400     106.400   same as hn-n3-hn
H -N2-CA   49.100     119.380   same as c2-n3-hn
CA-N2-CT   64.646     118.515   Calculated with empirical approach
N2-CT-H1   49.550     109.800   same as hc-c3-n3
N2-CT-CT   66.020     111.040   same as c3-c3-n3
H -N2-CT   47.420     109.290   same as c3-n3-hn
CT-CT-HC   46.340     109.800   same as c3-c3-hc
CT-CT-CT   62.860     111.510   same as c3-c3-c3
H1-CT-H1   39.400     107.580   same as hc-c3-hc
H1-CT-CT   46.391     109.545   Calculated with empirical approach
HC-CT-HC   39.400     107.580   same as hc-c3-hc
CT-CT-N3   64.180     114.210   same as c3-c3-n4
CT-CT-HP   46.340     109.800   same as c3-c3-hc
CT-CT-C    63.270     111.040   same as c -c3-c3
CT-N3-H    45.850     110.110   same as c3-n4-hn
CT-C -O    67.400     123.200   same as c3-c -o
CT-C -N2   70.000     116.000
N3-CT-HP   48.640     107.900   same as hc-c3-n4
N3-CT-C    65.470     110.730   same as c -c3-n4
H -N3-H    40.580     108.300   same as hn-n4-hn
HP-CT-C    46.930     108.770   same as c -c3-hc
C -N2-H    48.330     117.550
C -N2-CA   63.82        123.71
O -C -N2   74.22        123.05
N2-CA-CA   68.290     120.950   same as ca-ca-nh
CA-CA-HA   50.010     119.700   same as c2-c2-hc
CA-CA-CA   66.620     120.020   same as ca-ca-ca
CA-CA-OS   70.710     121.870   same as c2-c2-os
CA-OS-CT   63.360     115.590   same as c2-os-c3
OS-CT-H1   51.050     108.700   same as hc-c3-os

DIHE
N2-CA-N2-H    1    0.300       180.000       2.000  same as X -c2-n3-X
N2-CA-N2-CT   1    0.300       180.000       2.000  same as X -c2-n3-X
```

```
CA-N2-CT-H1   1    0.300         0.000       3.000  same as X -c3-n3-X
CA-N2-CT-CT   1    0.300         0.000       3.000  same as X -c3-n3-X
N2-CT-CT-HC   1    0.156         0.000       3.000  same as X -c3-c3-X
N2-CT-CT-CT   1    0.156         0.000       3.000  same as X -c3-c3-X
H -N2-CT-H1   1    0.300         0.000       3.000  same as X -c3-n3-X
H -N2-CT-CT   1    0.300         0.000       3.000  same as X -c3-n3-X
CT-CT-CT-HC   1    0.160         0.000       3.000  same as hc-c3-c3-c3
CT-CT-CT-CT   1    0.180         0.000      -3.000  same as c3-c3-c3-c3
CT-CT-CT-CT   1    0.250       180.000      -2.000  same as c3-c3-c3-c3
CT-CT-CT-CT   1    0.200       180.000       1.000  same as c3-c3-c3-c3
H1-CT-CT-HC   1    0.150         0.000       3.000  same as hc-c3-c3-hc
H1-CT-CT-CT   1    0.160         0.000       3.000  same as hc-c3-c3-c3
CT-CT-CT-N3   1    0.156         0.000       3.000  same as X -c3-c3-X
CT-CT-CT-HP   1    0.160         0.000       3.000  same as hc-c3-c3-c3
CT-CT-CT-C    1    0.156         0.000       3.000  same as X -c3-c3-X
HC-CT-CT-HC   1    0.150         0.000       3.000  same as hc-c3-c3-hc
CT-CT-N3-H    1    0.156         0.000       3.000  same as X -c3-n4-X
CT-CT-C -O    1    0.000       180.000       2.000  same as X -c -c3-X
CT-CT-C -N2   1    0.000       180.000       2.000  same as X -c -c3-X
HC-CT-CT-N3   1    0.156         0.000       3.000  same as X -c3-c3-X
HC-CT-CT-HP   1    0.150         0.000       3.000  same as hc-c3-c3-hc
HC-CT-CT-C    1    0.156         0.000       3.000  same as X -c3-c3-X
CT-C -N2-H    1    10.00        180.00       2.000  same as X -C -N -X
CT-C -N2-CA   1    10.00        180.00       2.000  same as X -C -N -X
N3-CT-C -O    1    0.000       180.000       2.000  same as X -c -c3-X
N3-CT-C -N2   1    0.000       180.000       2.000  same as X -c -c3-X
H -N3-CT-HP   1    0.156         0.000       3.000  same as X -c3-n4-X
H -N3-CT-C    1    0.156         0.000       3.000  same as X -c3-n4-X
HP-CT-C -O    1    0.800         0.000      -1.000  same as hc-c3-c -o
HP-CT-C -O    1    0.000         0.000      -2.000  same as hc-c3-c -o
HP-CT-C -O    1    0.080       180.000       3.000  same as hc-c3-c -o
HP-CT-C -N2   1    0.000       180.000       2.000  same as X -c -c3-X
C -N2-CA-CA   1    0.300       180.000       2.000  same as X -c2-n3-X
O -C -N2-H    1    10.00       180.000       2.000  ATTN, need revision
O -C -N2-CA   1    10.00       180.000       2.000  ATTN, need revision
N2-CA-CA-HA   1    6.650       180.000       2.000  same as X -c2-c2-X
N2-CA-CA-CA   1    3.625       180.000       2.000  same as X -ca-ca-X
H -N2-CA-CA   1    0.300       180.000       2.000  same as X -c2-n3-X
CA-CA-CA-OS   1    6.650       180.000       2.000  same as X -c2-c2-X
CA-CA-CA-CA   1    3.625       180.000       2.000  same as X -ca-ca-X
CA-CA-CA-HA   1    6.650       180.000       2.000  same as X -c2-c2-X
CA-CA-OS-CT   1    1.050       180.000       2.000  same as X -c2-os-X
HA-CA-CA-OS   1    6.650       180.000       2.000  same as X -c2-c2-X
CA-OS-CT-H1   1    0.383         0.000       3.000  same as X -c3-os-X
HA-CA-CA-HA   1    6.650       180.000       2.000  same as X -c2-c2-X

IMPROPER
N2-N2-CT-N2         1.1          180.0         2.0   Using default value
CT-N2-C -O          1.1          180.0         2.0   Using default value
C -CA-N2-H          1.1          180.0         2.0   Using default value
CA-CA-CA-N2         1.1          180.0         2.0   Using default value
CA-CA-CA-HA         1.1          180.0         2.0   Using default value
CA-CA-CA-OS         1.1          180.0         2.0   Using default value
CA-CA-CA-CA         1.1          180.0         2.0   Using default value
```

```

NONBON
  N2          1.8240  0.1700             same as nh
  H           0.6000  0.0157             same as hn
  CA          1.9080  0.0860             same as ca
  CT          1.9080  0.1094             same as c3
  H1          1.4870  0.0157             same as hc
  HC          1.4870  0.0157             same as hc
  N3          1.8240  0.1700             same as n4
  HP          1.4870  0.0157             same as hc
  C           1.9080  0.0860             same as c
  O           1.6612  0.2100             same as o
  HA          1.4870  0.0157             same as hc
  OS          1.6837  0.1700             same as os
```

I.2.3 PDB Code: 1UTS

The **.PREPI** file is shown below:

```
    0    0    2

This is a remark line
molecule.res
P13   INT  0
CORRECT     OMIT DU   BEG
  0.0000
   1  DUMM  DU    M    0  -1  -2     0.000      .0        .0      .00000
   2  DUMM  DU    M    1   0  -1     1.449      .0        .0      .00000
   3  DUMM  DU    M    2   1   0     1.523   111.21       .0      .00000
   4  C91   ca    M    3   2   1     1.540   111.208  -180.000 -0.097800
   5  C41   ca    S    4   3   2     1.381   139.869   -99.505 -0.042000
   6  H4    ha    E    5   4   3     1.076   118.839    26.688  0.136000
   7  C31   cd    M    4   3   2     1.429    19.539   -21.089 -0.183200
   8  H31   ha    E    7   4   3     1.081   125.436  -119.422  0.178000
   9  C21   cc    M    7   4   3     1.361   108.352    60.738 -0.062100
  10  H2    h4    E    9   7   4     1.081   128.214   178.179  0.200000
  11  N1    na    M    9   7   4     1.387   107.020    -0.925 -0.168400
  12  HN1   hn    E   11   9   7     1.012   124.700   179.543  0.322700
  13  C81   ca    M   11   9   7     1.371   111.223     0.501 -0.021200
  14  C71   ca    M   13  11   9     1.382   133.001   179.000 -0.126000
  15  H7    ha    E   14  13  11     1.073   119.856     2.476  0.154000
  16  C61   ca    M   14  13  11     1.398   119.215  -177.184 -0.113000
  17  H61   ha    E   16  14  13     1.079   119.178   178.532  0.115000
  18  C51   cp    M   16  14  13     1.413   120.726    -1.839 -0.158000
  19  C2    cp    M   18  16  14     1.508   121.273   178.253  0.032000
  20  C3    ca    S   19  18  16     1.402   119.055   -56.029 -0.120000
  21  H3    ha    E   20  19  18     1.089   116.062     3.355  0.132000
  22  C1    ca    M   19  18  16     1.404   121.857   126.236 -0.022000
  23  H1    ha    E   22  19  18     1.085   118.473    -4.137  0.187000
  24  C6    ca    M   22  19  18     1.403   120.399   176.263 -0.197000
  25  H6    ha    E   24  22  19     1.079   115.993   177.871  0.154000
  26  C5    ca    M   24  22  19     1.427   121.474    -2.698  0.157100
  27  O     os    S   26  24  22     1.458   122.896  -174.348 -0.350900
  28  C     c3    3   27  26  24     1.447   118.788    31.527  0.137400
  29  CA    c3    3   28  27  26     1.564   114.749  -136.166 -0.149400
  30  CB    c3    3   29  28  27     1.556   117.500   -61.301  0.102800
  31  N     n4    3   30  29  28     1.509   111.337  -175.607 -0.836600
  32  HN1A  hn    E   31  30  29     1.042   111.899  -178.994  0.474467
  33  HN2   hn    E   31  30  29     1.041   110.864   -58.962  0.474467
  34  HN3   hn    E   31  30  29     1.041   111.559    60.190  0.474467
  35  HB1   hx    E   30  29  28     1.103   111.447   -55.560  0.121200
  36  HB2   hx    E   30  29  28     1.103   110.282    64.669  0.121200
  37  HA1   hc    E   29  28  27     1.118   106.626   176.459  0.086200
  38  HA2   hc    E   29  28  27     1.116   106.809    61.986  0.086200
  39  HC1   h1    E   28  27  26     1.113   112.644   -16.321  0.085700
  40  HC2   h1    E   28  27  26     1.111   108.878    98.903  0.085700
  41  C4    ca    M   26  24  22     1.442   117.474     5.561 -0.239300
  42  CA1   c3    M   41  26  24     1.534   126.532   178.547  0.219100
  43  HA11  hx    E   42  41  26     1.113   108.479  -102.843  0.115200
  44  HA12  hx    E   42  41  26     1.111   109.529   136.341  0.115200
  45  NB    n4    M   42  41  26     1.543   113.067    16.716 -0.785000
  46  HB11  hn    E   45  42  41     1.013   103.579    25.935  0.463300
  47  HB12  hn    E   45  42  41     1.014   100.360   -79.803  0.463300
  48  CG    c3    M   45  42  41     1.566   122.078   158.993  0.079800
  49  HG1   hx    E   48  45  42     1.105   103.928   150.997  0.124700
  50  HG2   hx    E   48  45  42     1.114   102.394    39.157  0.124700
  51  CD    c3    M   48  45  42     1.604   120.280   -82.888  0.081800
  52  HD1   hx    E   51  48  45     1.105   110.229   133.343  0.135200
  53  HD2   hx    E   51  48  45     1.103   108.253    14.815  0.135200
  54  NE    n4    M   51  48  45     1.612   117.890  -106.236 -0.694400
  55  HE    hn    E   54  51  48     1.044   105.228    45.772  0.467800
  56  CH1   c3    M   54  51  48     1.584   116.475   -70.340  0.080300
  57  HH11  hx    E   56  54  51     1.105   104.895   -59.391  0.151950
  58  HH12  hx    E   56  54  51     1.099   106.042    57.577  0.151950
  59  CI1   c3    M   56  54  51     1.591   114.176   179.032  0.072800
  60  HI11  hx    E   59  56  54     1.109   110.504   -70.270  0.168450
  61  HI12  hx    E   59  56  54     1.109   111.535   167.457  0.168450
  62  NJ    n4    M   59  56  54     1.548   114.508    47.835 -0.769000
  63  HJ1   hn    E   62  59  56     1.025   108.862  -169.641  0.501800
  64  HJ2   hn    E   62  59  56     1.023   112.035    80.025  0.501800
  65  CI2   c3    M   62  59  56     1.552   113.907   -47.830  0.072800
  66  HI21  hx    E   65  62  59     1.108   105.581   171.773  0.168450
  67  HI22  hx    E   65  62  59     1.110   104.444   -72.886  0.168450
  68  CH2   c3    M   65  62  59     1.592   113.708    48.005  0.080300
  69  HH21  hx    E   68  65  62     1.105   108.849  -167.984  0.151950
  70  HH22  hx    E   68  65  62     1.108   110.845    70.849  0.151950


LOOP
  C81  C91
  C51  C41
   C4   C3
  CH2   NE

IMPROPER
```

```
  C81  C41  C91  C31
  C91  C51  C41   H4
  C91  C21  C31  H31
  C31   H2  C21   N1
  C81  C21   N1  HN1
  C71  C91  C81   N1
  C81  C61  C71   H7
  C71  C51  C61  H61
  C61  C41  C51   C2
   C3   C1   C2  C51
   C4   C2   C3   H3
   C6   C2   C1   H1
   C1   C5   C6   H6
   C6   C4   C5    O
  CA1   C3   C4   C5

DONE
STOP
```

The **.FRCMOD** file is shown below:

```
1UTS

remark goes here
MASS

BOND

ANGLE

DIHE
ca-ca-cd-ha   1    0.700       180.000       2.000      same as X -c2-ca-X
ca-ca-cd-cc   1    0.700       180.000       2.000      same as X -c2-ca-X

IMPROPER
ca-ca-ca-cd         1.1          180.0         2.0   Using default value
ca-cp-ca-ha         1.1          180.0         2.0   General improper
torsional angle (2 general atom types)
ca-cc-cd-ha         1.1          180.0         2.0   Using default value
cd-h4-cc-na         1.1          180.0         2.0   Using default value
ca-cc-na-hn         1.1          180.0         2.0   General improper
torsional angle (2 general atom types)
ca-ca-ca-na         1.1          180.0         2.0   Using default value
ca-ca-ca-ha         1.1          180.0         2.0   General improper
torsional angle (2 general atom types)
ca-ca-cp-cp         1.1          180.0         2.0   Using default value
ca-ca-ca-os         1.1          180.0         2.0   Using default value

NONBON
```

I.2.4 PDB Code: 1UUD

The **.PREPI** file is shown below:

```
    0    0    2

This is a remark line
molecule.res
P14   INT  0
CORRECT     OMIT DU   BEG
  0.0000
   1  DUMM  DU    M    0  -1  -2     0.000      .0        .0      .00000
   2  DUMM  DU    M    1   0  -1     1.449      .0        .0      .00000
   3  DUMM  DU    M    2   1   0     1.523   111.21       .0      .00000
   4  NZ1   nh    M    3   2   1     1.540   111.208  -180.000 -0.493700
   5  HZ11  hn    E    4   3   2     0.979    63.110  -100.487  0.334450
   6  HZ12  hn    E    4   3   2     0.980   137.616     5.210  0.334450
   7  CE    cz    M    4   3   2     1.368    73.651   121.668  0.529300
   8  NZ2   nh    B    7   4   3     1.358   119.260  -136.077 -0.493700
   9  HZ21  hn    E    8   7   4     0.981   119.988     0.056  0.334450
  10  HZ22  hn    E    8   7   4     0.980   120.016   179.980  0.334450
  11  ND    nh    M    7   4   3     1.320   121.505    43.991 -0.419100
  12  HD    hn    E   11   7   4     0.980   119.162   179.899  0.326700
  13  CG    c3    M   11   7   4     1.312   121.520    -0.648 -0.006700
  14  HG1   h1    E   13  11   7     1.081   109.437  -164.767  0.080700
  15  HG2   h1    E   13  11   7     1.080   109.400   -45.074  0.080700
  16  CB    c3    M   13  11   7     1.535   109.914    75.092  0.114400
  17  HB1   h1    E   16  13  11     1.080   109.286   -57.373  0.087700
  18  HB2   h1    E   16  13  11     1.079   109.198  -176.880  0.087700
  19  OA    os    M   16  13  11     1.435   110.180    62.826 -0.365900
  20  C1    ca    M   19  16  13     1.433   110.091    72.571  0.068100
  21  C6    ca    B   20  19  16     1.423   119.673  -106.474 -0.116000
  22  C5    ca    B   21  20  19     1.421   119.899   179.068 -0.088000
  23  C4    ca    B   22  21  20     1.421   120.061     0.672  0.169100
  24  C3    ca    S   23  22  21     1.420   120.071     0.079 -0.205000
  25  H3    ha    E   24  23  22     1.080   119.919   179.913  0.148000
  26  O1    os    S   23  22  21     1.430   119.983   179.975 -0.288900
  27  C11   c3    3   26  23  22     1.431   109.570  -111.118  0.102700
  28  H11   h1    E   27  26  23     1.080   109.471   172.948  0.066033
  29  H12   h1    E   27  26  23     1.080   109.426    52.964  0.066033
  30  H13   h1    E   27  26  23     1.079   109.438   -67.034  0.066033
  31  H5    ha    E   22  21  20     1.078   119.999  -179.298  0.194000
  32  H6    ha    E   21  20  19     1.083   120.121    -0.344  0.153000
  33  C2    ca    M   20  19  16     1.425   120.342    74.186 -0.153300
  34  CA    c3    M   33  20  19     1.543   120.486     0.478  0.178100
  35  HA1   hx    E   34  33  20     1.080   109.615   -78.172  0.108700
  36  HA2   hx    E   34  33  20     1.080   109.208   162.154  0.108700
  37  NB    n4    M   34  33  20     1.484   109.712    42.108 -0.767000
  38  HB11  hn    E   37  34  33     1.030   109.460    33.693  0.461300
  39  HB12  hn    E   37  34  33     1.029   109.485   -86.218  0.461300
  40  CG1   c3    M   37  34  33     1.482   109.799   153.747  0.100800
  41  HG11  hx    E   40  37  34     1.080   109.423   -48.688  0.112200
  42  HG12  hx    E   40  37  34     1.081   109.357  -168.481  0.112200
  43  CD    c3    M   40  37  34     1.533   109.537    71.232 -0.100400
  44  HD1   hc    E   43  40  37     1.080   109.342  -108.569  0.073200
  45  HD2   hc    E   43  40  37     1.080   109.362    11.170  0.073200
  46  CE1   c3    M   43  40  37     1.533   109.881   131.432 -0.113400
  47  HE1   hc    E   46  43  40     1.079   109.477    23.433  0.072200
  48  HE2   hc    E   46  43  40     1.081   109.443   143.408  0.072200
  49  CZ    c3    M   46  43  40     1.531   109.519   -96.639  0.031300
  50  HZ1   h1    E   49  46  43     1.081   109.405   -56.803  0.085200
  51  HZ2   h1    E   49  46  43     1.080   109.371    62.967  0.085200
  52  NH    nh    M   49  46  43     1.312   109.846  -176.957 -0.418100
  53  HH    hn    E   52  49  46     0.981   119.285  -105.358  0.329700
  54  CI    cz    M   52  49  46     1.319   121.502    75.570  0.530300
  55  NJ2   nh    B   54  52  49     1.366   121.543    -1.284 -0.490700
  56  HJ21  hn    E   55  54  52     0.980   119.765   179.943  0.334200
  57  HJ22  hn    E   55  54  52     0.980   120.486    -0.079  0.334200
  58  NJ1   nh    M   54  52  49     1.358   119.191   178.206 -0.490700
  59  HJ11  hn    E   58  54  52     0.979   120.035   179.993  0.334200
  60  HJ12  hn    E   58  54  52     0.980   119.946    -0.131  0.334200


LOOP
   C2   C3

IMPROPER
   ND  NZ1   CE  NZ2
   C6   C2   C1   OA
   C1   C5   C6   H6
   C6   C4   C5   H5
   C5   C3   C4   O1
   C4   C2   C3   H3
   CA   C1   C2   C3
   NH  NJ2   CI  NJ1

DONE
STOP
```

The **.FRCMOD** file is shown below:

```
1UUD

remark goes here
MASS

BOND

ANGLE

DIHE
nh-cz-nh-hn   1    0.675       180.000     2.000      same as X -c2-nh-X
nh-cz-nh-c3   1    0.675       180.000     2.000      same as X -c2-nh-X

IMPROPER
```

```
nh-nh-cz-nh          1.1          180.0     2.0       Using default value
ca-ca-ca-os          1.1          180.0     2.0       Using default value
ca-ca-ca-ha          1.1          180.0     2.0       General improper
torsional angle (2 general atom types)

NONBON
```

I.2.5 PDB Code: 1UUI

The **.PREPI** file is shown below:

```
    0    0    2

This is a remark line
molecule.res
P12   INT  0
CORRECT     OMIT DU   BEG
  0.0000
   1  DUMM  DU    M    0  -1  -2     0.000      .0        .0      .00000
   2  DUMM  DU    M    1   0  -1     1.449      .0        .0      .00000
   3  DUMM  DU    M    2   1   0     1.523   111.21       .0      .00000
   4  CG1   c3    M    3   2   1     1.540   111.208  -180.000  0.099300
   5  HG11  hx    E    4   3   2     1.080     5.145    78.690  0.129450
   6  HG12  hx    E    4   3   2     1.079   110.911   146.783  0.129450
   7  CD1   c3    M    4   3   2     1.538   103.842   -96.459  0.000300
   8  HD11  h1    E    7   4   3     1.077   108.291  -133.130  0.098700
   9  HD12  h1    E    7   4   3     1.080   109.046   -14.556  0.098700
  10  NE1   nh    M    7   4   3     1.497   112.100   106.363 -0.325100
  11  CZ    cz    B   10   7   4     1.407   118.810   159.635  0.529300
  12  NH1   nh    B   11  10   7     1.363   120.751    -9.200 -0.463200
  13  HH11  hn    E   12  11  10     0.981   119.622  -179.886  0.342700
  14  HH12  hn    E   12  11  10     0.975   120.670     0.393  0.342700
  15  NH2   nh    B   11  10   7     1.364   120.693   177.444 -0.463200
  16  HH21  hn    E   15  11  10     0.980   119.721   179.847  0.342700
  17  HH22  hn    E   15  11  10     0.976   120.597    -0.299  0.342700
  18  CD2   c3    M   10   7   4     1.496   121.331   -32.407  0.000300
  19  HD21  h1    E   18  10   7     1.077   109.075   150.975  0.098700
  20  HD22  h1    E   18  10   7     1.079   109.089   -89.920  0.098700
  21  CG2   c3    M   18  10   7     1.538   112.057    31.015  0.099300
  22  HG21  hx    E   21  18  10     1.079   109.156  -103.336  0.129450
  23  HG22  hx    E   21  18  10     1.081   108.913   137.661  0.129450
  24  NB    n4    M   21  18  10     1.487   111.639    17.336 -0.692400
  25  HB    hn    E   24  21  18     1.029   108.120   178.600  0.479800
  26  CA    c3    M   24  21  18     1.500   110.933    59.612  0.186100
  27  HA1   hx    E   26  24  21     1.082   109.398    19.346  0.106700
  28  HA2   hx    E   26  24  21     1.080   109.774  -100.432  0.106700
  29  C2    ca    M   26  24  21     1.546   110.110   139.248 -0.159300
  30  C3    ca    M   29  26  24     1.423   120.406    33.458 -0.185000
  31  H3    ha    E   30  29  26     1.078   120.091    -0.090  0.149000
  32  C4    ca    M   30  29  26     1.421   120.029  -179.855  0.159100
  33  OA1   os    S   32  30  29     1.430   119.952   179.914 -0.288900
  34  CB1   c3    3   33  32  30     1.430   109.502    71.638  0.103700
  35  HB11  h1    E   34  33  32     1.080   109.485   -55.153  0.065700
  36  HB12  h1    E   34  33  32     1.080   109.461    64.790  0.065700
  37  HB3   h1    E   34  33  32     1.080   109.479  -175.152  0.065700
  38  C5    ca    M   32  30  29     1.420   120.044    -0.059 -0.073000
  39  H5    ha    E   38  32  30     1.080   119.976   179.924  0.196000
  40  C6    ca    M   38  32  30     1.420   119.977     0.111 -0.134000
  41  H6    ha    E   40  38  32     1.080   119.957   179.844  0.150000
  42  C1    ca    M   40  38  32     1.420   120.060    -0.102  0.073100
  43  OA    os    M   42  40  38     1.430   119.908   179.841 -0.378900
  44  CB    c3    M   43  42  40     1.433   110.036    70.070  0.137400
  45  HB1   h1    E   44  43  42     1.081   109.895   -40.718  0.080200
  46  HB2   h1    E   44  43  42     1.080   109.400    79.536  0.080200
  47  CG    c3    M   44  43  42     1.529   109.074  -160.623 -0.140400
  48  HG1   hc    E   47  44  43     1.078   109.202   -39.696  0.082200
  49  HG2   hc    E   47  44  43     1.080   109.368    79.936  0.082200
  50  CD    c3    M   47  44  43     1.532   109.988  -159.890  0.100800
  51  HD1   hx    E   50  47  44     1.082   109.570   140.705  0.121200
  52  HD2   hx    E   50  47  44     1.080   109.546    20.746  0.121200
  53  NE    n4    M   50  47  44     1.480   109.493   -99.287 -0.836600
  54  HE1   hn    E   53  50  47     1.030   109.488   162.398  0.472800
  55  HE2   hn    E   53  50  47     1.030   109.439   -77.546  0.472800
  56  HE3   hn    E   53  50  47     1.030   109.455    42.425  0.472800


LOOP
   NB  CG1
   C1   C2

IMPROPER
  NH1  NH2   CZ  NE1
   CA   C3   C2   C1
   C2   C4   C3   H3
   C3   C5   C4  OA1
   C4   C6   C5   H5
   C5   C1   C6   H6
   C2   C6   C1   OA

DONE
STOP
```

The **.FRCMOD** file is shown below:

```
1UUI

remark goes here
MASS

BOND

ANGLE

DIHE
```

```
c3-nh-cz-nh   1    0.675       180.000    2.000       same as X -c2-nh-X
nh-cz-nh-hn   1    0.675       180.000    2.000       same as X -c2-nh-X

IMPROPER
nh-nh-cz-nh         1.1          180.0     2.0      Using default value
ca-ca-ca-ha         1.1          180.0     2.0      General improper
torsional angle (2 general atom types)
ca-ca-ca-os         1.1          180.0     2.0      Using default value

NONBON
```

I.2.6 Porphyrin: DPD Molecule

The **.PREPI** file is shown below:

```
    0    0    2

This is a remark line
molecule.res
LIG   INT  0
CORRECT     OMIT DU   BEG
  0.0000
   1  DUMM  DU    M    0  -1  -2     0.000      .0        .0      .00000
   2  DUMM  DU    M    1   0  -1     1.449      .0        .0      .00000
   3  DUMM  DU    M    2   1   0     1.523   111.21       .0      .00000
   4  N7    nb    M    3   2   1     1.540   111.208  -180.000 -0.846500
   5  C26   ca    B    4   3   2     1.321    90.651    54.397  0.794700
   6  C25   ca    S    5   4   3     1.404   119.142   -89.871 -0.268200
   7  N5    nc    E    6   5   4     1.373   133.232  -179.506 -0.500600
   8  N9    nh    B    5   4   3     1.343   118.893    89.202 -0.958500
   9  H12   hn    E    8   5   4     0.994   118.370    10.576  0.447550
  10  H11   hn    E    8   5   4     0.993   119.527   170.196  0.447550
  11  C27   ca    M    4   3   2     1.339    89.315   -64.181  0.926400
  12  N10   nh    B   11   4   3     1.359   115.257   -91.536 -0.928500
  13  H13   hn    E   12  11   4     0.994   116.619    18.053  0.428050
  14  H14   hn    E   12  11   4     0.994   116.666   163.807  0.428050
  15  N8    nb    M   11   4   3     1.323   127.956    89.820 -0.796500
  16  C24   ca    M   15  11   4     1.327   111.971     0.916  0.502300
  17  N6    na    M   16  15  11     1.357   127.964   179.515 -0.253500
  18  C28   c3    3   17  16  15     1.447   126.497    -0.206  0.035100
  19  H15   h1    E   18  17  16     1.082   110.560  -118.652  0.058200
  20  H16   h1    E   18  17  16     1.082   110.545   120.087  0.058200
  21  H17   h1    E   18  17  16     1.079   107.623     0.684  0.058200
  22  C23   cd    M   17  16  15     1.383   105.781  -179.918  0.444200
  23  C22   ch    M   22  17  16     1.430   121.641  -179.931 -0.119200
  24  C21   cg    M   23  22  17     1.190   178.709   170.287 -0.095300
  25  C10   ce    M   24  23  22     1.439   178.421     5.300  0.105600
  26  C8    cd    S   25  24  23     1.362   118.022     1.967  0.132600
  27  N2    nd    S   26  25  24     1.383   125.261  -179.908 -0.631800
  28  C6    cc    B   27  26  25     1.303   106.114   179.852  0.437900
  29  C7    cc    B   28  27  26     1.466   112.002     0.100 -0.251100
  30  C9    cd   S   29  28  27     1.334  106.119    -0.105 -0.147500
  31  H5    ha   E   30  29  28     1.070  128.555   179.984  0.168500
  32  H4    ha   E   29  28  27     1.071  125.368  -179.992  0.156000
  33  C5    ce   B   28  27  26     1.435  126.134   179.998 -0.170600
  34  C2    cd   S   33  28  27     1.348  126.991     0.013  0.094000
  35  N1    na   B   34  33  28     1.368  128.313    -0.103 -0.534800
  36  C1    cc   S   35  34  33     1.360  111.514  -179.933  0.117000
  37  C3    cc   B   36  35  34     1.462  105.987     0.053 -0.147500
  38  C4    cd   S   37  36  35     1.332  108.292    -0.079 -0.147500
  39  H2    ha   E   38  37  36     1.072  127.779  -179.970  0.161000
  40  H1    ha   E   37  36  35     1.070  123.701  -179.875  0.160500
  41  H21   hn   E   35  34  33     0.996  125.056     0.141  0.500800
  42  H3    ha   E   33  28  27     1.074  116.422   179.957  0.148500
  43  C11   cc   M   25  24  23     1.443  115.984  -177.964 -0.001100
  44  C12   cd   B   43  25  24     1.389  128.040     0.139 -0.147500
  45  C14   cd   S   44  43  25     1.391  107.337   179.863 -0.147500
  46  H7    ha   E   45  44  43     1.071  127.200   179.945  0.161000
  47  H6    ha   E   44  43  25     1.069  125.389    -0.223  0.160500
  48  N3    na   M   43  25  24     1.355  124.444  -179.996 -0.167100
  49  H22   hn   E   48  43  25     0.995  124.306    -0.052  0.345700
  50  C13   cc   M   48  43  25     1.355  110.194  -179.848 -0.024100
  51  C15   ce   M   50  48  43     1.430  125.777   179.997 -0.072200
  52  H8    ha   E   51  50  48     1.076  114.074   179.896  0.148500
  53  C16   cd   M   51  50  48     1.348  129.266    -0.070  0.117600
  54  C17   cd   B   53  51  50     1.458  124.053  -179.977 -0.164500
  55  C18   cc   S   54  53  51     1.334  106.705   179.959 -0.234100
  56  H10   ha   E   55  54  53     1.070  128.857  -179.953  0.168500
  57  H9    ha   E   54  53  51     1.072  125.108     0.014  0.156000
  58  N4    nd   M   53  51  50     1.388  126.668     0.021 -0.631800
  59  C19   cd   M   58  53  51     1.300  106.118  -179.981  0.452900
  60  C20   cf   M   59  58  53     1.452  124.579   179.902  0.007200
  61  C29   ch   M   60  59  58     1.434  117.649   179.721 -0.095300
  62  C30   cg   M   61  60  59     1.192  177.422    -0.649 -0.119200
  63  C31   cc   M   62  61  60     1.429  178.948     0.185  0.444200
  64  N11   na   S   63  62  61     1.384  121.466  -173.951 -0.253500
  65  C34   c3   3   64  63  62     1.446  127.726    -0.085  0.035100
  66  H18   h1   E   65  64  63     1.082  110.499    57.260  0.058200
  67  H19   h1   E   65  64  63     1.083  110.657   -64.069  0.058200
  68  H20   h1   E   65  64  63     1.079  107.688   176.635  0.058200
  69  N12   nd   M   63  62  61     1.286  125.153     5.825 -0.500600
  70  C32   ca   M   69  63  62     1.376  104.354  -179.788 -0.268200
  71  C33   ca   M   70  69  63     1.378  110.970    -0.012  0.502300
  72  N13   nb   M   71  70  69     1.326  126.654   179.889 -0.796500
  73  C35   ca   M   72  71  70     1.323  111.928    -0.495  0.926400
  74  N16   nh   B   73  72  71     1.360  116.728  -177.674 -0.928500
  75  H23   hn   E   74  73  72     0.995  116.421  -162.603  0.428050
  76  H24   hn   E   74  73  72     0.994  116.527   -17.955  0.428050
  77  N14   nb   M   73  72  71     1.338  127.904     1.016 -0.846500
  78  C36   ca   M   77  73  72     1.322  118.571    -0.675  0.794700
  79  N15   nh   M   78  77  73     1.344  118.743   178.706 -0.958500
  80  H25   hn   E   79  78  77     0.994  118.262    11.424  0.447550
  81  H26   hn   E   79  78  77     0.994  119.326   169.650  0.447550


```

```
LOOP
  C24  C25
  C23   N5
   C9   C8
   C4   C2
  C20   C1
  C13  C14
  C19  C18
  C33  N11
  C36  C32

IMPROPER
  C25   N7  C26   N9
  C24  C26  C25   N5
  C26  H12   N9  H11
   N8   N7  C27  N10
  C27  H13  N10  H14
  C25   N6  C24   N8
  C28  C24   N6  C23
  C22   N6  C23   N5
  C11   C8  C10  C21
   C9  C10   C8   N2
   C7   C5   C6   N2
   C6   C9   C7   H4
   C7   C8   C9   H5
   C6   C2   C5   H3
   C4   C5   C2   N1
   C1   C2   N1  H21
   C3  C20   C1   N1
   C1   C4   C3   H1
   C3   C2   C4   H2
  C12  C10  C11   N3
  C11  C14  C12   H6
  C13  C12  C14   H7
  C11  C13   N3  H22
  C14  C15  C13   N3
  C13  C16  C15   H8
  C17  C15  C16   N4
  C18  C16  C17   H9
  C17  C19  C18  H10
  C18  C20  C19   N4
   C1  C19  C20  C29
  C30  N11  C31  N12
  C34  C33  N11  C31
  C33  C36  C32  N12
  C32  N11  C33  N13
  N13  N14  C35  N16
  C35  H23  N16  H24
  C32  N14  C36  N15
  C36  H25  N15  H26

DONE
STOP
```

The **.FRCMOD** file is shown below:

```
DPD

remark goes here
MASS

BOND

ANGLE
nc-cd-ch   75.218     115.295   Calculated with empirical approach
cg-ce-cc   66.080     114.640   same as ce-ce-cg
cd-cf-ch   66.120     123.130   same as ce-cf-ch
cg-cc-nd   75.218     115.295   Calculated with empirical approach

DIHE
nc-cd-ch-cg   1    0.000       180.000    2.000       same as X -c1-cd-X
na-cd-ch-cg   1    0.000       180.000    2.000       same as X -c1-cd-X
ch-cg-ce-cd   1    0.000       180.000    2.000       same as X -c1-ce-X
ch-cg-ce-cc   1    0.000       180.000    2.000       same as X -c1-ce-X
cg-ce-cd-nd   1    1.000       180.000    2.000       same as X -ce-ce-X
cg-ce-cd-cd   1    1.000       180.000    2.000       same as X -ce-ce-X
cg-ce-cc-cd   1    1.000       180.000    2.000       same as X -ce-ce-X
cg-ce-cc-na   1    1.000       180.000    2.000       same as X -ce-ce-X
cd-ce-cc-cd   1    1.000       180.000    2.000       same as X -ce-ce-X
cd-ce-cc-na   1    1.000       180.000    2.000       same as X -ce-ce-X
nd-cd-ce-cc   1    1.000       180.000    2.000       same as X -ce-ce-X
nd-cc-ce-cd   1    1.000       180.000    2.000       same as X -ce-ce-X
nd-cc-ce-ha   1    1.000       180.000    2.000       same as X -ce-ce-X
cc-ce-cd-na   1    1.000       180.000    2.000       same as X -ce-ce-X
cc-ce-cd-cd   1    1.000       180.000    2.000       same as X -ce-ce-X
cc-cc-ce-cd   1    1.000       180.000    2.000       same as X -ce-ce-X
cc-cc-ce-ha   1    1.000       180.000    2.000       same as X -ce-ce-X
na-cd-ce-ha   1    1.000       180.000    2.000       same as X -ce-ce-X
na-cc-cf-cd   1    6.650       180.000    2.000       same as X -ce-cf-X
na-cc-cf-ch   1    6.650       180.000    2.000       same as X -ce-cf-X
cc-cf-cd-cc   1    6.650       180.000    2.000       same as X -ce-cf-X
cc-cf-cd-nd   1    6.650       180.000    2.000       same as X -ce-cf-X
cc-cf-ch-cg   1    0.000       180.000    2.000       same as X -c1-cf-X
cc-cc-cf-cd   1    6.650       180.000    2.000       same as X -ce-cf-X
cc-cc-cf-ch   1    6.650       180.000    2.000       same as X -ce-cf-X
cd-cd-ce-ha   1    1.000       180.000    2.000       same as X -ce-ce-X
cd-cc-ce-ha   1    1.000       180.000    2.000       same as X -ce-ce-X
na-cc-ce-ha   1    1.000       180.000    2.000       same as X -ce-ce-X
ha-ce-cd-nd   1    1.000       180.000    2.000       same as X -ce-ce-X
cc-cd-cf-ch   1    6.650       180.000    2.000       same as X -ce-cf-X
nd-cd-cf-ch   1    6.650       180.000    2.000       same as X -ce-cf-X
cd-cf-ch-cg   1    0.000       180.000    2.000       same as X -c1-cf-X
ch-cg-cc-na   1    0.000       180.000    2.000       same as X -c1-cc-X
ch-cg-cc-nd   1    0.000       180.000    2.000       same as X -c1-cc-X

IMPROPER
ca-nb-ca-nh         1.1          180.0    2.0          Using default value
```

```
ca-ca-ca-nc         1.1          180.0    2.0           Using default value
ca-hn-nh-hn         1.1          180.0    2.0           Using default value
nb-nb-ca-nh         1.1          180.0    2.0           Using default value
ca-na-ca-nb         1.1          180.0    2.0           Using default value
c3-ca-na-cd         1.1          180.0    2.0           Using default value
ch-na-cd-nc         1.1          180.0    2.0           Using default value
cc-cd-ce-cg         1.1          180.0    2.0           Using default value
cd-ce-cd-nd         1.1          180.0    2.0           Using default value
cc-ce-cc-nd         1.1          180.0    2.0           Using default value
cc-cd-cc-ha         1.1          180.0    2.0           Using default value
cc-cd-cd-ha         1.1          180.0    2.0           Using default value
cc-cd-ce-ha         1.1          180.0    2.0           Using default value
cd-ce-cd-na         1.1          180.0    2.0           Using default value
cc-cd-na-hn         1.1          180.0    2.0           General improper
torsional angle (2 general atom types)
cc-cf-cc-na         1.1          180.0    2.0           Using default value
cd-ce-cc-na         1.1          180.0    2.0           Using default value
cc-cc-na-hn         1.1          180.0    2.0           General improper
torsional angle (2 general atom types)
cd-cd-cc-ha         1.1          180.0    2.0           Using default value
cc-cf-cd-nd         1.1          180.0    2.0           Using default value
cc-cd-cf-ch         1.1          180.0    2.0           Using default value
cg-na-cc-nd         1.1          180.0    2.0           Using default value
c3-ca-na-cc         1.1          180.0    2.0           Using default value
ca-ca-ca-nd         1.1          180.0    2.0           Using default value

NONBON
```

APPENDIX J
MEDIA COVERAGE OF PUBLISHED WORK

The work on ligand recognition in RNA which is presented in chapter 4 has been highlighted by multiple media sources. In this appendix, I provide the links to these media sources.

UNH NEWS

`https://www.unh.edu/unhtoday/news/release/2020/07/21/unh-researchers-discover-new-pathways-could-help-treat-rna-viruses`

UNH TODAY

`https://www.unh.edu/unhtoday/2020/07/new-pathways-could-help-treat-rna-viruses-discovered`

FOSTERS.COM

`https://www.fosters.com/story/news/coronavirus/2020/07/21/unh-researchers-discover-pathways-that-could-help-treat-rna-viruses/113776498/`

SEACOASTONLINE

`https://www.seacoastonline.com/story/news/coronavirus/2020/07/21/unh-researchers-discover-pathways-that-could-help-treat-rna-viruses/113776498/`

GRANITE-GEEK

`https://granitegeek.concordmonitor.com/2020/07/20/new-pathways-could-help-treat-rna-viruses/`

NEWS-MED-LIFESCI

`https://www.news-medical.net/news/20201217/Supercomputer-simulations-lead-to-an-important-viral-inhibitor-discovery.aspx`

XSEDE

https://www.xsede.org/-/supercomputers-simulate-new-pathways-for-potential-rna-virus-treatment

UCSD

https://ucsdnews.ucsd.edu/pressrelease/supercomputers-simulate-new-pathways-for-potential-rna-virus-treatment

APPENDIX K
CURRICULUM VITAE

K.1 Education

University of New Hampshire, BS, Chemical Engineering, May 2016

K.2 Publications

Tannir, S., **Levintov, L.**, Townley, M. A., Leonard, B. M., Kubelka, J., Vashisth, H., Varga, K., and Balaz, M. (2020). “Functional nanoassemblies with mirror-image chiroptical properties templated by a single homochiral DNA strand.” *Chem. Mater.*, 32(6), 22722281. (PDF)

Levintov, L., and Vashisth, H. (2020). “Ligand recognition in viral RNA necessitates rare conformational transitions.” *J. Phys. Chem. Lett.*, 11(14), 54265432. (PDF)

Levintov, L., Paul, S., and Vashisth, H. (2021) “Reaction coordinate and thermodynamics of base flipping in RNA.” *J. Chem. Theory Comput.*, 17(3), 19141921. (PDF)

Levintov, L., and Vashisth, H. (2021). “Role of conformational heterogeneity in ligand recognition by viral RNA molecules.” *Phys. Chem. Chem. Phys.*, *DOI: 10.1039/D1CP00679G.* (PDF)

Levintov, L., and Vashisth, H. (2021). “Role of salt-bridging interactions in recognition of viral RNA by arginine-rich peptides.” *Biophys. J.*, (Under review).

K.3 Presentations

Lev Levintov, Harish Vashisth. Simulation Study of Supramolecular Nanoassembly

2018 Department of Chemical Engineering Spring 2018 Seminar Series, Durham, NH 30 March 2018.

2018 Graduate Research Conference, Durham, NH 9 April 2018.

2018 UNH Bioengineering Symposium, Durham, NH 8 May 2018.

Lev Levintov, Harish Vashisth. Atomistic simulation studies of DNA-Porphyrin nanoassemblies

78th Physical Electronics Conference, Durham, NH 25 Jun 2018.

AIChE Annual Meeting, Pittsburgh, PA 30 Oct 2018.

Lev Levintov, Harish Vashisth. Atomically resolved simulation studies of RNA/small-molecule interactions

Annual Fall Meeting of the American Chemical Society, Boston, MA 21 August 2018.

Lev Levintov, Harish Vashisth. Conformational Mapping of Viral RNA Elements Using Atomistic Simulations

AIChE Annual Meeting, Pittsburgh, PA 29 October, 2018.

Lev Levintov, Harish Vashisth. Conformational mapping of HIV-1 TAR RNA

Graduate Research Conference, Durham, NH 1 April, 2019.

UNH Bioengineering Symposium, Durham, NH 8 May, 2019.

Lev Levintov, Harish Vashisth. Conformational heterogeneity and its role in ligand recognition by RNA molecules

Molecular Biophysics in the Northeast, Boston, MA 9 November, 2019.

Lev Levintov, Harish Vashisth. Long Time-Scale Atomistic Simulations of HIV-1 TAR RNA

Annual Meeting, Orlando, FL 11 November, 2019.

Lev Levintov, Harish Vashisth. Studies on Conformational Transitions in RNA upon

Ligand Binding

2020 Departmental Seminar Series, UNH Chemical Engineering, Durham, NH 30 October 2020.

Lev Levintov, Sanjib Paul, Harish Vashisth. Transition Path Sampling Simulations of Base Flipping in RNA

2020 Virtual AIChE Annual Meeting, 18 November 2020.

Lev Levintov, Harish Vashisth. Rare conformational transition in viral RNA upon ligand binding

65th Biophysical Society Virtual Annual Meeting, 26 February 2021.