DISSERTATION

MOLECULAR DYNAMICS SIMULATIONS OF PEPTIDE AND PROTEIN SYSTEMS

Submitted by

Ryan Nicholas Weber

Department of Chemistry

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Summer 2021

Doctoral Committee:

Advisor: Martin McCullagh

Grzegorz Szamel
Richard Finke
Qiang Wang

ABSTRACT

MOLECULAR DYNAMICS SIMULATIONS OF PEPTIDE AND PROTEIN SYSTEMS

Molecular systems composed of amino acids play an important role in biological systems and have numerous functions and applications due to their enormous chemical versatility. These systems are usually divided into peptides and proteins based on the number of amino acids that compose each molecule. Molecular dynamics simulations can provide molecular-level insights into the self-assembly of peptide systems and the function of protein systems where experimental methods fail. Peptides are utilized for their switchable and self-assembling properties for the engineering of novel biomaterials which are responsive to external stimuli. Often, peptides are paired with aromatic molecules to incorporate interesting optoelectronic properties into the material. Chapter 2 discusses a molecular dynamics simulation study on the self-assembling properties of the self-complimentary $(RXDX)_4$ sequence paired with an unnatural coumarin amino acid for the design of a pH-switchable, optoelectronic, self-assembling biomaterial. Specifically, it is found that the hydrophobicity of the peptide sequence plays a significant role in the stability and pH-switchability of $(RXDX)_4$ and coumarin-$(RXDX)_4$ β-sheet fibers. Proteins are essential to all known life and participate in nearly every cellular process. There are many varieties of proteins with important diverse functions. Helicase proteins hydrolyze NTP to catalyze the translocation and unwinding of double-stranded nucleic acids such as RNA and DNA and play a critical and extensive role in viral replication. Nsp13 is a helicase protein that is an important component of the viral replication machinery of the severe acute respiratory syndrome coronavirus-2 and remains a promising target for antiviral drugs. Chapter 3 presents a molecular dynamics simulation study on the ATP-dependent translocation mechanism of the SARS-CoV-2 nsp13 helicase. Specifically, the results from the study suggest that nsp13 may translocate using an inchworm stepping mechanism and that the binding of ATP may cause the first step in the translocation cycle. Motifs Ia, IV, and V are identified as

key motifs in the translocation mechanism of nsp13 and as potential targets for the development of antiviral drugs against SARS-CoV-2. Although molecular dynamics simulation is a powerful approach to investigate condensed phase molecular phenomenon such as protein folding, allostery, and self-assembly, molecular dynamics is limited in the size and length of simulations that can be performed. Implicit solvent simulation methods, such as Implicit Solvation using the Superposition Approximation (IS-SPA), were developed to address these issues in solvated systems. The goal of IS-SPA is to improve the efficiency of molecular dynamics simulations by removing the solvent from the system, but still include the effect of the solvent on the solute. Chapter 4 presents the development and optimization of an IS-SPA molecular dynamics code on a GPU using CUDA. Specifically, the performance of three different IS-SPA CUDA algorithms are compared. The future studies of the self-assembly of peptide systems for the design of biomaterials, the ATP-dependent translocation mechanism of the SARS-CoV-2 nsp13, and the optimization of the GPU-capable IS-SPA molecular dynamics code in CUDA are discussed in the final chapter.

ACKNOWLEDGEMENTS

I would like to thank Martin McCullagh for his guidance during my time in graduate school at Colorado State University. I would also like to thank the members of the McCullagh group: Jake Anderson, Kelly Du Pont, Russell Davidson, Peter Lake (Rex), Max Mattson, Mortaza Derakhashani Molayousefi, Heidi Klem, and Kevin Vatow for their constant support and friendships which has continued beyond my time at Colorado State University. I want to specifically express my gratitude to Peter Lake for all of the insightful conversations and regular guidance that he provided during my time in graduate school.

I would like to wholeheartedly thank Jake Anderson, Rae Anderson, Heidi Klem, Russel Davidson, Max Mattson, Peter Lake, Kelly Du Pont, Matthew Lacroix, and Jeffery Ma for the nonstop laughter, entertainment, distractions, patience, love, and compassion that you have provided me.

Lastly, I am grateful for my family; my parents, Dan and Joni Weber, my siblings and their spouses, David and Kristi Weber, Gina and Sam Lodhi, Jenny and Chad Martie, and Katie and Mike Ellis, as well as all of my nieces and nephews, for their unceasing love, wisdom, and support.

DEDICATION

I would like to dedicate this dissertation to my friends and family for their endless love, support, and encouragement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

Amino acids are organic compounds composed of 3 parts: a basic amino group, an acidic carboxyl group, and an organic sidechain unique to each amino acid. There are many naturally occurring amino acids, yet only 20 are used in the genetic code of all life.[1] Amino acids are diverse and have a range of varying chemical properties including size, charge, polarity, hydrophobicity, and aromaticity.[2] They form polymer chains, through peptide bonds, of varying size with enormous chemical versatility.[1,3–5] Due to the countless possible combinations and the diverse molecular properties of those combinations, amino acids polymer chains have many potential applications both biological and as a source of molecular engineering.

Molecules composed of amino acids are ubiquitous in biological systems and are fundamental to life. This fact has led to a substantial interest in categorizing and gaining a fundamental understanding of these molecules by the scientific community. In the scientific literature there is an abundance of studies focusing on two major types of systems composed of amino acids: peptides and proteins. The main difference between peptides and proteins are their size. Although the cut-off number of peptides is somewhat arbitrary, peptides are smaller than proteins and generally contain between 2 and 50 amino acids. On the other hand, proteins are much longer and usually consist of several hundred amino acids.

Beyond understanding natural peptides and proteins there are many studies that report the synthesis of new unnatural amino acids with distinct properties, further expanding the versatility of proteins and peptides.[4,6–10] Unnatural amino acids are engineered for specific applications, such as peptidomimetic drugs used by the pharmaceutical industry,[4,11–13] fluorescent probes for biological imaging,[8,9,12,14–16] and, in conjunction with natural amino acids, material design.[4,10–12]

Molecular dynamics (MD) simulations are a computational tool that provide atomistic insight into the behavior of individual molecules and play an important role in understanding protein and peptide systems. MD simulations are performed by numerically solving Newton's equations of

motion following the time evolution of a system.[17,18] The interactions between atoms are defined by an input model and force field which describe the interatomic forces. Each step performed gives a configuration of the system, representing one point in phase space, and is a single frame in a trajectory. Making use of statistical mechanics, thermodynamic quantities of the system can be measured from the trajectories by calculating an arithmetic average over the instantaneous value of the quantity at each frame. In the limit of infinite frames, these averages converge to the value of the thermodynamic property.[18]

MD can be used to study a wide variety of important processes occurring on nanometer-length and microsecond-time scales, without making any assumptions about the process beforehand. Furthermore, it can provide an understanding of peptide and protein systems where experimental methods fail. For example, crystal structures provide us with only a single static structure of a protein, whereas MD simulations can give us information about the larger phase space available to the protein providing details about conformational changes and structural fluctuations.[17–20] The work presented in Chapter 2 and Chapter 3 utilize MD simulations to provide an understanding of two particular peptide and protein systems, while Chapter 4 discusses the development of an MD model to accelerate those studies.

1.1 Peptide Systems

The bottom-up engineering of self-assembling peptide-based nanostructures has attracted a lot of interest from researchers due to the potential applications in biomedicine and biotechnology.[3,6,12,21–24] Self-assembling peptides rely on spontaneous diffusion and the formation of a large number of weak non-covalent and reversible interactions that lead to a highly stable and often ordered assembly.[21] Certain peptides have been labeled as "smart" or switchable as they are responsive to external stimuli leading to changes in structure or function of the material. These stimuli can be environmental, biological, or optical and can be controlled by changes in pH, temperature, ionic strength, or enzymatic manipulation.[7,21,23,25,26]

One example of a class of switchable peptides are ionic self-complimentary peptides. These peptides are typically composed of 16 amino acids comprised of alternating oppositely charge hydrophilic residues and neutral hydrophobic residues. This repeating pattern leads to the formation of β-sheets that are hydrophobic on one side and have alternating positive and negative charges on the other.[26–29] The formation of this β-sheet structure is pH-sensitive as the charges on the hydrophilic residues change under acidic and basic conditions.[28,30]

Switchable materials are of particular interest for the design of self-assembling optoelectronic materials. Peptides have been combined with a variety of aromatic molecules in efforts to combine the self-assembling properties of peptides with the optoelectronic properties of dye molecules.[1,2,8,9,14,31–44] Often the dye molecules are quite large, making it difficult to balance the hydrophobic and π-stacking interactions of the dye with the weak, self-assembling interactions of the peptides. Unnatural amino acids with small aromatic sidechains are hypothesized to make balancing these interactions easier. Chapter 2 seeks to answer whether or not a switchable, self-assembling, ionic, self-complimentary peptide can be combined with an unnatural amino acid with a small aromatic sidechain to create a pH-responsive, optoelectronic material. Furthermore, it addresses if the switchable properties of this material can be further tuned by altering the weak non-covalent interactions between the peptides through specific residue mutation. To answer these questions, Chapter 2 presents a study of $(RXDX)_4$ peptides and coumarin-$(RXDX)_4$ conjugates through molecular dynamic simulations.

1.2 Protein Systems

Proteins are essential to all known life and participate in nearly every cellular process. There are many types of proteins with a variety of roles such as enzymatic proteins that catalyze biochemical reactions,[45,46] antibody proteins that fight diseases,[45] glycosylases that repair DNA,[47] transport proteins that move molecules throughout the body,[48] structural and nonstructural proteins that contribute to the replication and packaging of viral genome,[19,49] and a myriad of other protein types

with crucial functions. Understanding the mechanisms by which these proteins function is of vital importance for understanding biological processes and to be able to treat viruses and diseases.

One particularly interesting class of proteins are helicases, which use the energy of adenosine triphosphate (ATP) hydrolysis to catalyze the unwinding of double-stranded nucleic acids such as ribonucleic acid (RNA) and deoxyribonucleic acid (DNA).[50–54] Helicases are encoded by all cellular life as well as many viruses and are involved in virtually every step of DNA and RNA metabolism.[52,53] As a result of the critical and extensive role helicase proteins play in viral replication, viral helicases are prominent targets for antiviral drug development.[55]

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) responsible for the COVID-19 pandemic has infected over 150 million people and has caused more than 3 million deaths worldwide.[56] The need to understand the mechanism by which this virus replicates and spreads cannot be overstated. The SARS-CoV-2 nonstructural protein 13 (nsp13) is a helicase protein that plays a critical role in the replication of the virus and is highly conserved across SARS viruses.[57–59] For example, SARS-CoV-2 nsp13 has a 99.8% sequence identity with SARS-CoV-1 nsp13.[57,59] Studies have shown that suppression of nsp13 in SARS-CoV-1 leads to inhibition of viral replication.[59,60] This suggests that SARS-CoV-2 nsp13 helicase is a promising target for antiviral drugs.[61] Furthermore, understanding the mechanism by which this helicase operates would aid in the development of antiviral drugs of possible future SARS viruses as well as provide insight into the ATP-dependent translocation mechanism of other superfamily 1 (SF1) helicase proteins. Chapter 3 presents work characterizing the structure–function relationships utilized by SARS-CoV-2 nsp13 during ATP-dependent RNA translocation using extensive Gaussian accelerated molecular dynamics (GaMD) simulations.

1.3 Development of Implicit Solvation Models

MD is limited in the system size and time scale of simulations that can be performed. Some processes require large length- and time-scales that are computationally unfeasible. With continual advances in code efficiency and the computational capabilities of hardware, the limits of MD

simulation length- and time-scales are steadily expanding.[17] Still, larger system sizes and longer simulations are needed to model many processes such as the macroscopic self-assembly of peptides. Implicit solvent models have been developed to address these issues in solvated systems.[62]

In all-atom molecular dynamics (aaMD) simulations of solvated systems, explicit solvent molecules constitute a majority of the atoms in the system. This means a bulk of the calculations are being performed on the solvent, however, typically only the solute is considered during analysis of the system. For this reason implicit solvation models have been developed to reduce the computational cost of simulating solvated systems by removing the solvent degrees of freedom from the simulation.[62,63] One challenge associated with developing implicit solvent models is finding a good balance between computational cost and accuracy of the model.[62,64] For example, models such as generalized Born (GB)[65] or constant density dielectric are computationally efficient, but fail to accurately capture aggregation features. Inversely, models such as the Reference Interaction Site Model (RISM)[66,67] are accurate, but extremely computationally demanding making their use for MD simulations impractical.[64]

The Implicit Solvation Using the Superposition Approximation (IS-SPA) implicit solvent model was developed to maintain a high degree of accuracy relative to other implicit solvent models and still be computationally feasible.[20,64] IS-SPA has out performed other implicit solvent approaches, such as the GB or constant density dielectric models.[20] The IS-SPA MD code was initially implemented in Fortran and was only capable of running on central processing units (CPUs). Due to the significantly better performance of graphics processing units (GPUs) over CPUs, development of an efficient GPU-enabled IS-SPA algorithm is critical for increasing IS-SPA MD performance. Chapter 4 presents work on the development of an efficient IS-SPA algorithm in CUDA resulting in an efficient GPU-capable IS-SPA MD code.

Chapter 2

The Role of Hydrophobicity in the Stability and pH-Switchability of (RXDX) $_4$ and Coumarin-(RXDX)$_4$ Conjugate β-Sheets[a]

2.1 Overview

pH-switchable, self-assembling materials are of interest in biological imaging and sensing applications. Here I propose that combining the pH-switchability of RXDX (X=Ala, Val, Leu, Ile, Phe) peptides and the optical properties of coumarin creates an ideal candidate for these materials. This suggestion is tested with a thorough set of all-atom molecular dynamics simulations. I first investigate the dependence of pH-switchability on the identity of the hydrophobic residue, X, in the bare (RXDX)$_4$ systems. Increasing the hydrophobicity stabilizes the fiber which, in turn, reduces the pH-switchability of the system. This behavior is found to be somewhat transferable to systems in which a single hydrophobic residue is replaced with a coumarin containing amino acid. In this case, conjugates with X=Ala are found to be unstable at both pHs while conjugates with X=Val, Leu, Ile and Phe are found to form stable β-sheets at least at neutral pH. The (RFDF)$_4$-coumarin conjugate is found to have the largest change in the ordering of the coumarin sidechains with pH change. Thus, I posit that coumarin-(RFDF)$_4$ containing peptide sequences are ideal candidates for pH-sensing bioelectronic materials.

[a]Reproduced with permission from Weber, R., & McCullagh, M. (2020). The Role of Hydrophobicity in the Stability and pH-Switchability of (RXDX)$_4$ and Coumarin-(RXDX)$_4$ Conjugate β-Sheets. Journal of Physical Chemistry B, 124(9), 1723–1732. https://doi.org/10.1021/acs.jpcb.0c00048. Copyright 2021 American Chemical Society

2.2 Introduction

Materials that are electronically conductive, tunable, and biocompatible are of great interest for bioelectronic applications. Peptides have been combined with a variety of optoelectronic materials for use in photovoltaic cells[68] and other electronic applications[14] to attempt to attain all of these properties.[33,69–73] These materials combine the self-assembly, biocompatibility, and tunable behavior of peptides[26,74] with the optoelectronic properties of a dye molecule.[75] Stacking of highly π-conjugated dye molecules can lead to electronic delocalization in the self-assembled aggregate and provide the electronic property of interest. Thus, it necessary to predict macroscopic structure from molecular details as the optoelectronic properties of the material depend on the assembled structure.

Peptides are inexpensive, easily synthesized and provide a rich diversity of self-assembled structures, making them ideal candidates for scaffolding optoelectronic materials.[76] To date, peptides have been used to create molecular wires,[77] switches[21] and surfactants[78] among other materials.[12,76,79,80] Due to the diverse set of chemical building blocks and resulting assembled structures, peptides continue to be an active area of research in bottom-up materials engineering.[5] The amino acid sequence $(\text{RADA})_n$ (Arg-Ala-Asp-Ala) has received a lot of attention over the past decade for its self-assembling properties and applications in tissue engineering.[81] $(\text{RADA})_n$ is a self-complimentary peptide sequence comprised of alternating oppositely charged hydrophilic residues (Arg and Asp) and neutral hydrophobic residues (Ala) allowing it to form highly organized β-sheet structures leading to nanofiber and hydrogel formation.[7,22,27,28,30,76,81–87] Potentially, the most interesting property of $(\text{RADA})_4$ is the transformation in the self-assembled structure when the pH is changed.[21,22]

Many of the proposed applications for small self-assembling peptides require the inclusion of additional functional components. Peptides have been tethered to aliphatic[3,88–91] and aromatic groups[1,92] to achieve control over assembly and additional functionality.[26] Aromatic amphiphiles are of particular interest here due to their potential use as bioelectronic and biosensing materials.[1] The peptide is used as a solubilizing and scaffolding agent for the aromatic groups which act as both

agents of hydrophobic assembly and electron rich or deficient semiconducting materials. A variety of aromatic groups have been linked to peptides include OPV3,[31–33] napthalene[34,35] , perylene diimide,[2,36–43] coumarin,[8,9,14,44] and others.[1] Here I focus on the use of coumarin due to its high quantum yield, extended spectroscopic range, photostability, and general solubility in a variety of solvents.[8,9,14] Furthermore, many coumarin derivatives have been synthesized that could be used to further tune the optical properties of the material.[8,9,14] Coumarins alone suffer a lack of tunability and biocompatibility making them less than ideal candidates for bioelectronic properties, but coumarin derived unnatural amino acids have been synthesized for biological imaging applications allowing it to be easily incorporated into a peptide sequence.[14]

The relative importance of hydrophobicity and aromaticity in the formation of β-sheet fibrils from self-complimentary amphipathic peptide sequences is debated.[23,93–96] Generally, a hydrophobic threshold must be reached to have fibrilization in these types of systems.[23,96,97] Some of these studies suggest that aromatic interactions are not important for determining fibrilization rates, but do have an impact on the self-assembly of these materials, specifically the morphology of the fibers.[94,95] It is even less clear how all of these factors will affect the self-assembly and pH switchability of a peptide-dye conjugate.

In this chapter, I describe results of a systematic computational investigation of the self-assembly and pH switchability of $(RXDX)_4$ (X = Ala, Val, Leu, Ile, and Phe) and $(RXDX)_4$-coumarin peptides. The set of five amino acids chosen for the X position have incremental change in hydrophobicity, based on water-octanol partition coefficients,[98] allowing us to probe the effect of hydrophobicity on the properties of the materials. Phenylalanine was added to this set to study the effect of adding a residue which is capable of π-stacking. Lastly, a single hydrophobic residue in each $(RXDX)_4$ fiber is mutated to a coumarin derived amino acid to study the stability and pH-switchability of coumarin-$(RXDX)_4$ conjugates.

2.3 Computational Methods

2.3.1 System Setup

Simulations are performed on five separate systems of $(RXDX)_4$ with X = Ala (A), Val (V), Leu (L), Ile (I), and Phe (F). β-strands of each $(RXDX)_4$ sequence are created using the sequence command in tleap within the AmberTools 18 package.[99] Each system contains twenty molecules of the $(RXDX)_4$ sequence arranged in two β-sheets, forming a fiber that is ten β-strands long as shown in Figure 2.1. The initial structure for each system is based on spacings between β-strands provided by Cormier *et al.*[100] The distance between the two β-sheets are varied based on the identity of the nonpolar residue. For X = Ala, Val, Leu, Ile, and Phe the initial distance for $(RXDX)_4$ fibers are 6.0, 7.0, 9.0, 9.0, and 11.0 Å, respectively. Each fiber is solvated with TIP3P water and 0.1 M NaCl plus neutralizing ions in a 143 Å cubic box with periodic boundary conditions.

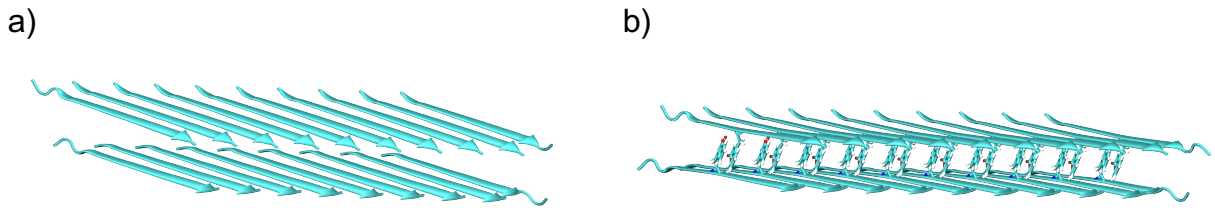

Figure 2.1: Initial $(RXDX)_4$ β-sheet sandwich fiber structure (a) without and (b) with a single hydrophobic amino acid (X) mutated to an unnatural coumarin amino acid.

To determine the effect of pH on the stability of the fibers, I simulate each system at neutral and acidic pH in triplicate for a total of 30 systems. The pH of the system is controlled by the protonation state of the Asp residues; Acidic pH (2-4) is achieved by protonating the functional group of the Asp residues while neutral pH leaves the Asp residues deprotonated giving each β-strand a total charge of 0 and +4, respectively. The C-termini remain deprotonated at both acidic and neutral pH.

I simulate five coumarin-peptide systems to ascertain the stability and optoelectronic properties of coumarin fibers with varying hydrophobicity. An unnatural coumarin amino acid is substituted for the fourth hydrophobic residue in the $(RXDX)_4$ sequence ((RXDX)-RXD-Coumarin-$(RXDX)_2$)

and is shown in Figure 2.2. Analogous to the $(RXDX)_4$ fiber simulations, each coumarin-RXDX fiber is simulated at both neutral and acidic pH in triplicate.

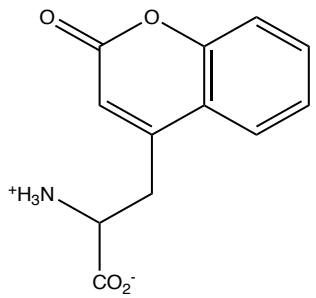

Figure 2.2: Molecular structure of an unnatural coumarin amino acid.

2.3.2 Simulation Details

Coumarin and an unnatural amino acid coumarin are parameterized in ff15ipq using the IPolQ method[101,102] with the ωB97X-D3 functional[103,104] and the cc-pVTZ basis set.[105] Atom types and charges are provided in the SI. QM calculations were run using ORCA 3.1.3.[106]

All MD simulations are run using the AMBER18 software[99] and modeled using AMBER's ff15ipq[101,102] and GAFF force fields.[107] Comparison of fiber thickness with experimental values provided in Table 2.1 is evidence that the force fields accurately reproduce experimental results as both the $(RADA)_4$ and $(RLDL)_4$ systems are found to have thicknesses within error of experiment. Furthermore, the calculated solubility of our coumarin model is 0.014 ± 0.006 $\frac{mol}{L}$, in agreement with the solubility value of 0.01706 $\frac{mol}{L}$ reported by Yalkowsky *et al.*[108] Hydrogen atoms are constrained using the SHAKE algorithm.[109] Direct nonbonding interactions are cutoff at 12 Å and long range electrostatic interactions are modeled using the particle mesh Ewald (PME) treatment.[110] An integration time step of 2 fs is used. MD simulations are performed in the NPT-ensemble with a Monte Carlo (MC) barostat set to 1 atm and a Langevin thermostat set to 298 K.

The simulation protocol used for all systems is the same. Solvated systems are minimized in two steps: 8,000 steepest descent minimization steps with harmonic constraints (force constant of 100 kcal mol^{-1} $Å^{-2}$) placed on all peptide atoms, followed by 12,000 conjugate gradient steps with

Table 2.1: Height of the RXDX β-sheet sandwich.

System	Height (Å)				
	RADA	RVDV	RLDL	RIDI	RFDF
Calculated	18.357(3)	20.718(3)	23.771(4)	22.170(5)	22.34(2)
Experimental	19(1)[28]	-	23(4)[28]	-	-

no constraints. The system is heated to a temperature of 298K over 50 ps with harmonic restraints (force constant of 5 kcal mol^{-1} Å^{-2}) on all peptide atoms. The water in the system relaxes with a harmonic restraint (force constant of 5 kcal mol^{-1} Å^{-2}) on all solute atoms for 40 ps. Following the heating protocol, the system is simulated at 298K for 10 ns allowing the pressure to equilibrate. For the $(RXDX)_4$ and coumarin-$(RXDX)_4$ simulations I use simulated annealing molecular dynamics (SAMD) due to the slow time dynamics of the fiber systems. Thirty simulated annealing cycles with a period of 10.05 ns were performed. In each cycle, the temperature is increased from 298 to 450 K in 50 ps, then kept constant for 1 ns, subsequently lowered back to 298K in 8.0 ns, and finally kept constant for 1 ns. The system is simulated at 298K for 200 ns from which ensemble averages were measured. The combined simulation time for this study is 31.25 μs.

2.4 Results and Discussion

Previous studies of the RXDX peptide sequence (X=A,L) have shown that the RXDX peptides form β-sheet sandwich fibers at neutral pH that dissociate in acidic and basic environments.[22,30,84–86,100] Experimental studies suggest that hydrophobicity is the main driving force for fibril formation rates and stability. It was proposed that aromaticity is not necessary for fibril formation, but does affect self-assembly in these materials as changes in peptide and ion concentrations lead to different fiber morphologies that are not seen when other non-aromatic non-polar residue were used. Here, I present a set of simulations that investigate the effect hydrophobicity and aromaticity have on the stability and switchability of $(RXDX)_4$ fibers.

I present a similar set of simulations with the addition of an unnatural coumarin amino acid to the $(RXDX)_4$ sequence as a model system for the design of switchable self-assembling biomaterials with

interesting optoelectronic properties. These simulations probe the stability and pH-switchability of the the β-sheet sandwich structure, as well as characterize the crucial role hydrophobicity plays in the stability of the fiber. Furthermore, I study the pH-switchability of the coumarin order which determines the optical and electronic properties of coumarin-$(RXDX)_4$ fibers.

2.4.1 pH-Switchability of $(RXDX)_4$ Fibers

Changing the pH of the $(RXDX)_4$ systems from neutral to acidic leads to instability within the β-sheet fibers. Representative snapshots of the $(RADA)_4$ simulations at neutral and acidic pH are shown in the first column of Figure 2.3a. At neutral pH, the system retains the β-sheet sandwich structure for the entire simulation other than some end fraying. Under acidic pH, the system completely loses all β-sheet structure leading to all of the peptides being dispersed in the solvent. This behavior is quantified for the entire simulation by measuring the percent of residues in an anti-parallel β-sheet, Figure 2.3b, determined by the define secondary structure protein (DSSP) analysis as a measure of the stability of each of the fibers.[111] At neutral pH, 51% of residues in $(RADA)_4$ are still in a β-sheet after 500 ns of simulation while at acidic pH this percentage falls to zero as the fiber dissociates. This demonstrates the pH-switchability of $(RADA)_4$ fibers.

Mutation of the RADA peptide sequence at the alanine position with increasingly hydrophobic residues enhances fiber stability and diminishes the pH-switchability. Representative snapshots of the $(RXDX)_4$ (X = V, L, I, F) simulations are provided in Figure 2.3a. Generally, the fibers at neutral pH seem relatively unaffected by the identity of the hydrophobic residue in the RXDX sequence. This is supported by the 50-70% of residues residing in a β-sheet for the course of these simulations as quantified in Figure 2.3b. Under acidic pH, increasing the hydrophobicity of the sequence stabilizes the fiber, leading to a smaller variation in stability with pH change. $(RFDF)_4$ fibers show little change with pH as only the ends of the fiber become more disordered at acidic pH. Generally, $(RFDF)_4$ fibers are more structured than $(RADA)_4$ fibers. Again this is further supported by the small changes in % β-sheet content at the two different pHs for the RFDF system in Figure 2.3b.

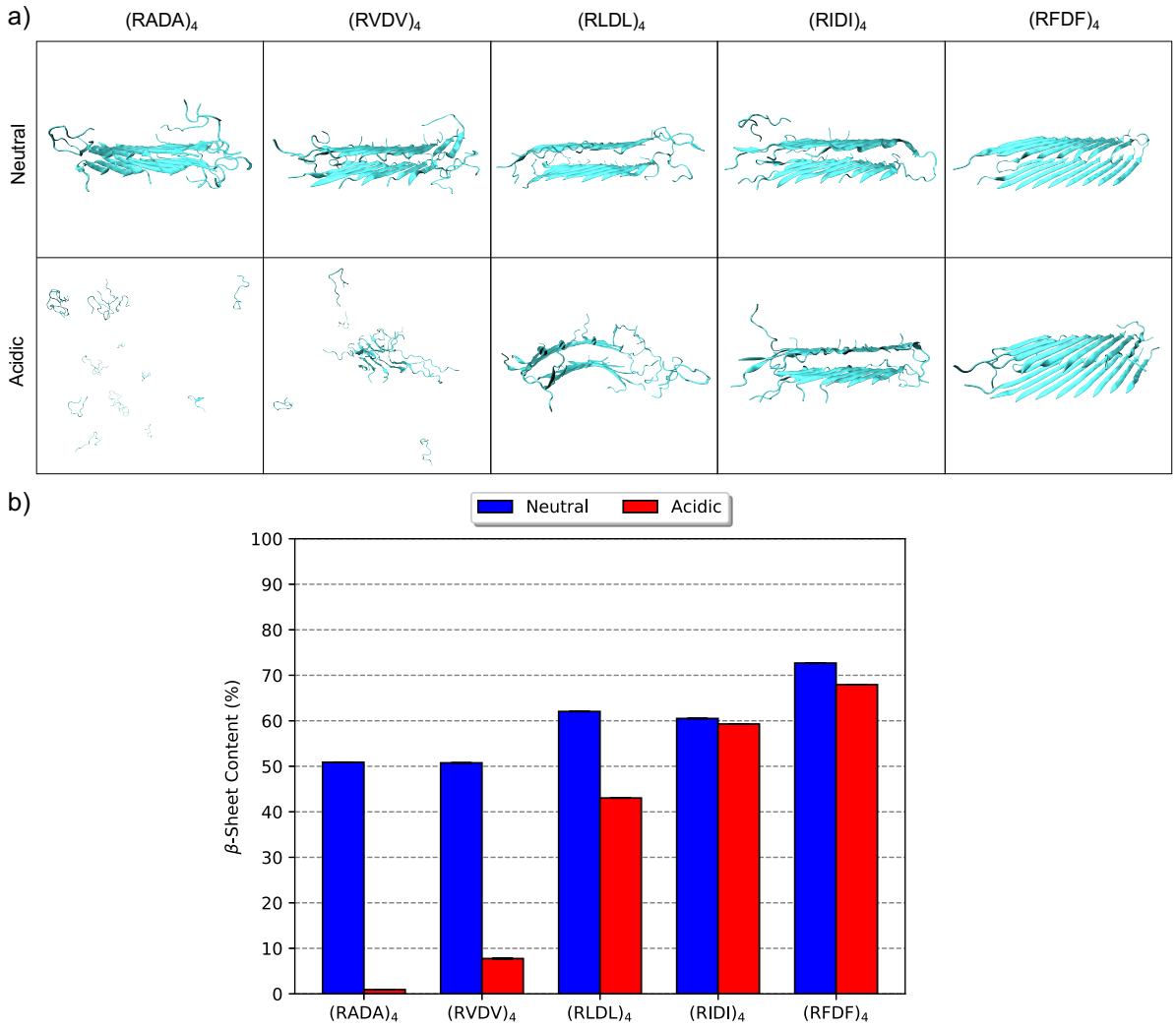

Figure 2.3: Hydrophobicity and pH control $(RXDX)_4$ fiber stability. (a) Representative snapshots of the structure of each $(RXDX)_4$ β-sheet fiber after 500 ns of simulation. The top and bottom rows show the fiber structure at neutral and acidic pH, respectively. (b) Percent of amino acids in a β-sheet secondary structure at the end of the simulations for each $(RXDX)_4$ fiber at both neutral and acidic pH. Secondary structure was calculated using the DSSP method as a measure of fiber stability. Fibers become more stable with increasing hydrophobicity at both pH leading to a reduction in the pH-switchability of the fiber structure.

Aromaticity of the nonpolar residues plays a minor roll in β-sheet fiber stability. It has been argued that aromaticity is unnecessary for fibrillization, instead, it is primarily hydrophobicity that influences fibrilization; a peptide sequence needs to reach a hydrophobic threshold to fibrilize.[94] Figure 2.3 shows that fibers containing Phe are more stable than fibers containing Ile, even though Phe and Ile have the same hydrophobicity and Ile has a higher propensity to form β-sheets (Table A.1). To demonstrate that aromaticity further stabilizes $(RXDX)_4$ fibers due to π-stacking interactions between aromatic sidechains I calculate the free energy as a function of the angle between neighboring phenyl rings separated into inter-sheet and intra-sheet pairs for the $(RFDF)_4$ systems. As shown in Figure 2.4, there is a preference for the phenyl rings to π-stack within a β-sheet as there is an approximately 3 kcal/mol well at an angle of 0° for intra-sheet pairs. Interestingly, there is no intercalation of the phenyl rings between β-sheets. The phenyl rings sit end-to-end in a T-shape configuration with an angle of 58° as shown by the inter-sheet free energy curve. These data support the conclusion that hydrophobicity is an important factor in determining fiber stability, but additionally suggest that the planar geometry of the phenyl ring further stabilizes $(RXDX)_4$ fibers at both neutral and acidic pH. The increase in stability for the $(RFDF)_4$ fiber could originate from aromatic interactions between phenyl rings, the packing of the phenyl rings within the fiber, or a more favorable change in entropy during self-assembly due to the more restricted range of motion of the Phe sidechain structure relative to the Ile sidechain structure before self-assembly.

2.4.2 The Interplay of Hydrophobic and Coulombic Interactions.

The coulombic interactions between residues within the fiber are responsible for the pH-switchability of the $(RXDX)_4$ fibers. The only difference between the neutral and acidic pH systems is the protonation state of the Asp residue leading to a change in the coulombic interactions stabilizing the fiber. Figure 2.5a contains the coulombic interactions of $(RADA)_4$ averaged over the first 10 ns of the simulation broken down in to residue type pairs for the initial self-assembled structure. There are a mix of stabilizing and destabilizing electrostatic interactions. The Arg–Arg, Asp–Asp, and X–X interactions are large and destabilize the fiber's β-sheet sandwich structure.

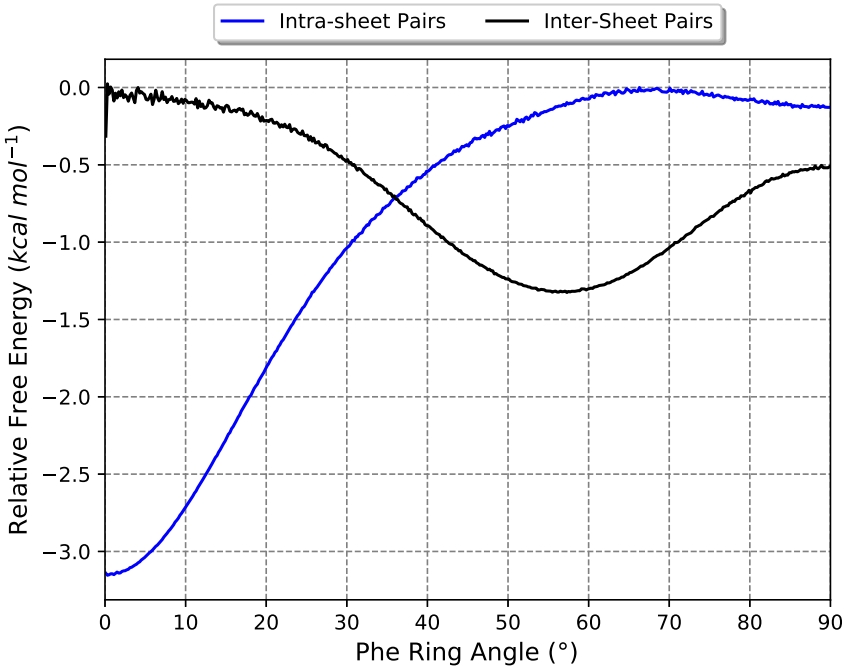

Figure 2.4: Free energy of the dihedral between phenyl rings in the $(RFDF)_4$ fiber separated into intra- and inter-sheet pairs. The free energy minima at 0° for intra-sheet pair demonstrate π-stacking interactions within each β-sheet while inter-sheet phenyl ring pairs form a T-shape configuration with a dihedral of 58°.

These interactions are counterbalanced by the Arg–Asp, X–Asp, and X–Arg stabilizing interactions. At neutral pH, the total coulombic interactions are large and negative, stabilizing the fiber. When the pH is lowered both the Asp–Asp and Arg–Asp coulombic interactions are significantly reduced due to the protonation of the aspartates to aspartic acids. Since the Arg-Asp interactions are 3-4 times larger than the Asp–Asp interactions this leads to an overall 9,870 kcal/mol increase in the total coulombic interaction energy. It is the large reduction of the Arg–Asp stabilizing interaction energy which leads to a destabilization of the $(RADA)_4$ fiber at acidic pH. Furthermore, the large change in coulombic interaction energy is consistent for all 5 $(RXDX)_4$ fibers and is independent of the identity of the hydrophobic residue.

Increasing the hydrophobicity of $(RXDX)_4$ peptides leads to an enhanced fiber stability and a reduction in pH-switchability due to an increase in hydrophobic interactions within the fiber. Hydrophobic molecules aggregate in water because this maximizes the number of hydrogen bonds between water molecules and minimizes the contact area between the hydrophobic and water

molecules. Therefore, the free energy of solvation of hydrophobic molecules is negatively correlated to the molecule's solvent accessible surface area (SASA). The change in hydrophobic SASA of $(RXDX)_4$ molecules during self-assembly is used as a measure of the hydrophobic interactions within each fiber. The hydrophobic SASA is calculated as the total SASA of the hydrophobic residues in the $(RXDX)_4$ peptides. Figure 2.5b shows the total change in SASA of the hydrophobic residues due to the self-assembly of $(RXDX)_4$ peptides into β-sheet fibers. As expected, there is a larger change in hydrophobic SASA as the hydrophobicity of the peptide is increased at both pHs. This suggests there is a subtle balance between the coulombic and hydrophobic interactions controlling the stability of $(RXDX)_4$ fibers. For $(RADA)_4$ the hydrophobic interactions are small enough that the reduction in stabilizing coulombic interactions with pH change lead to instability within the fiber. As the non-polar residue is mutated to a more hydrophobic residue, the balance between coulombic and hydrophobic interactions shift such that the hydrophobic interactions are large enough to stabilize the fiber even at acidic pH.

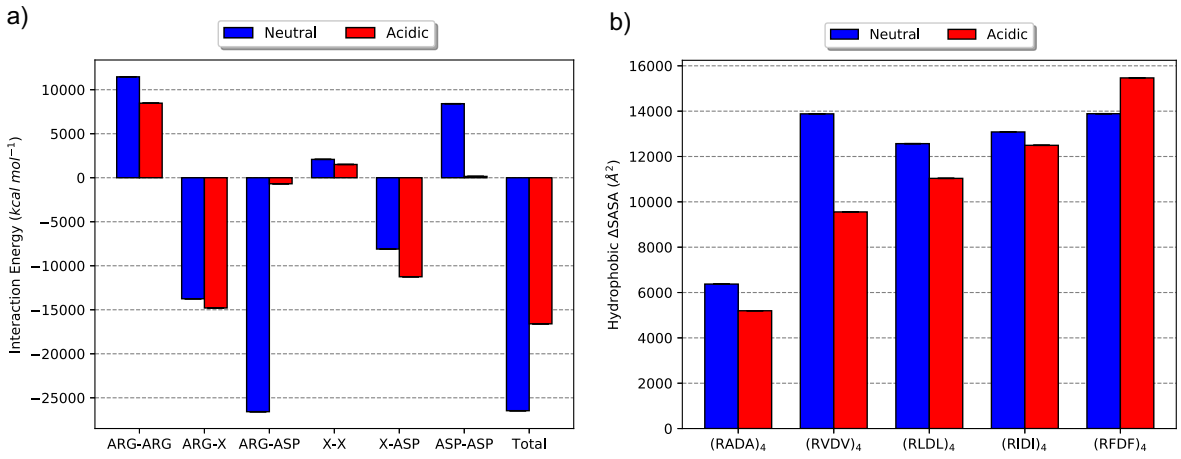

Figure 2.5: Coulombic and hydrophobic interactions control the stability and pH-switchability of $(RXDX)_4$ fibers. (a) Coulombic interaction energy decomposed into residue type pairs for $(RADA)_4$ fibers at neutral and acidic pH. The total coulombic interactions are large and attractive stabilizing the fiber. Under acidic conditions there is a 9870 kcal/mol increase in coulombic interaction leading to instability in the fiber structure. (b) The change in hydrophobic SASA (ΔSASA) during the self-assembly of each $(RXDX)_4$ fiber. ΔSASA is negatively correlated with the free energy of solvation of hydrophobic molecules. Increasing the hydrophobicity of the peptide leads to a larger ΔSASA and therefore, larger hydrophobic interactions stabilizing the β-sheet fiber structure at both neutral and acidic pH.

2.4.3 pH-Switchability of Coumarin-$(RXDX)_4$ Fibers

The mutation of a single hydrophobic residue within the $(RXDX)_4$ sequence to coumarin leads to changes in fiber stability, pH-switchability, and structure. Addition of coumarin leads to an induced curvature and a destabilization of the β-sheet fiber due to differences in the size of the hydrophobic residue and the coumarin side chain. Figure 2.6b shows the percent of residues in a β-sheet. $(RADA)_4$ fibers are completely unstable with the addition of coumarin making it a poor candidate for the design of switchable electronic biomaterials. The other systems are able to retain 40-70% of the β-sheet structure. Figure 2.6a shows each of the coumarin-$(RXDX)_4$ fiber structures. Even though the systems retain most of their β-sheet content with pH change, the overall fiber structure is unstable either breaking into smaller β-sheets containing only a few peptide strands or the two β-sheets that make up the fiber separate exposing the hydrophobic layers to the solvent. Although the pH-switchability of the fibers are not shown in the β-sheet metric, there is a change in stability and structure of the fiber with pH. The fibers containing coumarin are more stable with increasing hydrophobicity as the hydrophobic residues become similar in size to the coumarin sidechain.

Increasing the hydrophobicity and size of the nonpolar residue decreases the curvature of the fiber structure and increases fiber stability. Figure 2.6d and Figure 2.6e show the distribution of the bend angle for each system at neutral and acidic pH, respectively. The bend angle is a measure of the curvature of the fiber in the direction of fibrilization and is calculated using the method described by Fujiwara *et al.*[112] At neutral pH the $(RFDF)_4$ fiber has very little curvature as the bend angles are close to zero. As the hydrophobicity of the peptide decreases, the bend angle distribution broadens out and shifts to the right corresponding to less order in the fiber and an increase in fiber curvature. Furthermore, lowering the pH leads to increased fiber curvature and more disorder in the fiber as the distributions broaden and shift to higher angles.

The total SASA of the coumarin residues within each fiber provide further evidence of an enhanced fiber stability with increasing hydrophobicity of the coumarin-$(RXDX)_4$ sequence. For an ordered fiber, the coumarin SASA is small as only the coumarins at the end of the fiber will be

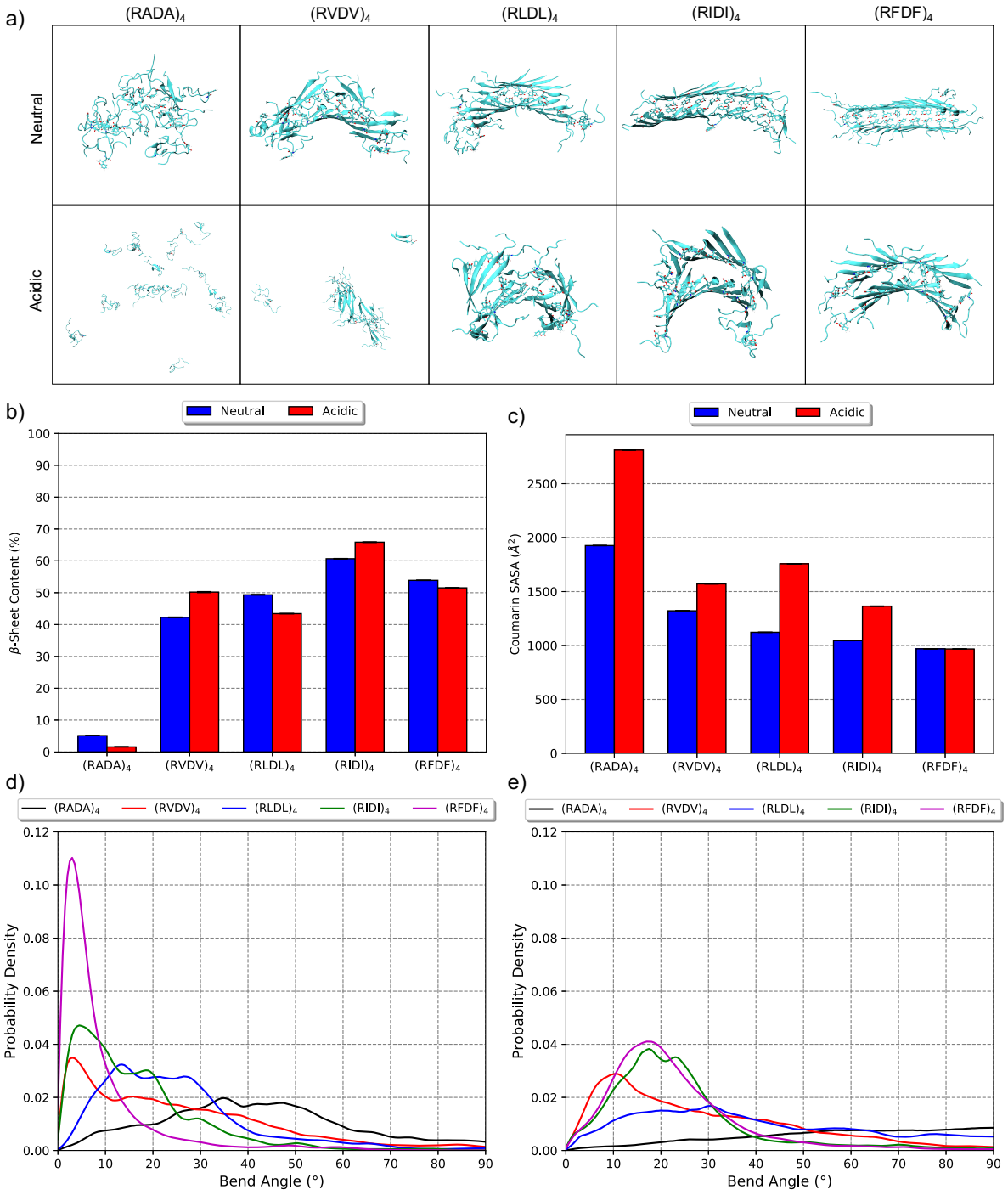

Figure 2.6: Hydrophobicity and pH control coumarin-$(RXDX)_4$ fiber stability. (a) Representative snapshots of the structure of each coumarin-$(RXDX)_4$ β-sheet fiber after 500 ns of simulation at neutral and acidic pH. (b) Percent of amino acids in a β-sheet secondary structure. $(RADA)_4$ fiber is unstable with the addition of coumarin, but increasing the hydrophobicity and size of the hydrophobic amino acid stabilizes coumarin-$(RXDX)_4$ fibers. (c) Average total SASA of coumarin amino acids for each coumarin-$(RXDX)_4$ fiber. The distribution of bend angles as a measure of fiber curvature and stability along the fibrilization axis at (d) neutral pH and (e) acidic pH.

solvent exposed, but as the fiber dissociates, more coumarins become solvent exposed. Figure 2.6c shows the average coumarin SASA for each fiber at both neutral and acidic pH. In agreement with the trends observed in the bend angle analysis, there is a decrease in coumarin SASA and therefore, an increase in fiber stability as the hydrophobicity of the peptide increases. When the pH is lowered the coumarin SASA increases corresponding to a decrease in fiber stability.

2.4.4 Order of Coumarin Sidechains within Coumarin-$(RXDX)_4$ Fibers

As part of designing a switchable material, it is important to not only have pH-switchability of the coumarin-$(RXDX)_4$ fiber structure, but also in the the optical properties of the fiber. For this reason it is of interest to analyze the coumarin order within the fibers because it is the arrangement of the coumarin sidechains that will dictate the optical and electronic properties of the fiber. The addition of coumarin to the $(RADA)_4$ sequence leads to instability and dissociation of the fiber making coumarin-$(RADA)_4$ a poor candidate for the design of switchable electronic biomaterials. Therefore, I analyze the pH-switchability of the coumarin order within $(RVDV)_4$, $(RLDL)_4$, $(RIDI)_4$, and $(RFDF)_4$ fibers to discern which of these sequences are the most optimal for these materials. The joint probability density as a function of both the coumarin-coumarin separation distance and the angle between the dipole moments of the two coumarin is used to measure the organization of coumarin within coumarin-$(RXDX)_4$ fibers. Furthermore, I use relative entropy to quantify the pH-switchability of the structure of the coumarin network. Relative entropy is a measure of the difference between two probability densities. The relative entropy between the joint probability densities of the neutral and acidic systems for each fiber is calculated as

$$S_{rel} = \int_0^{\pi} \int_0^{\infty} P_n(r,\theta) \ln \frac{P_n(r,\theta)}{P_a(r,\theta)} r^2 \sin\theta dr d\theta \tag{2.1}$$

where $P_n(r,\theta)$ and $P_a(r,\theta)$ are the joint probability densities of the neutral and acidic system, respectively.

Coumarin sidechains become more ordered within the β-sheet fibers with increasing peptide hydrophobicity. Figure 2.7 shows the joint probability distribution for $(RLDL)_4$, $(RIDI)_4$, and

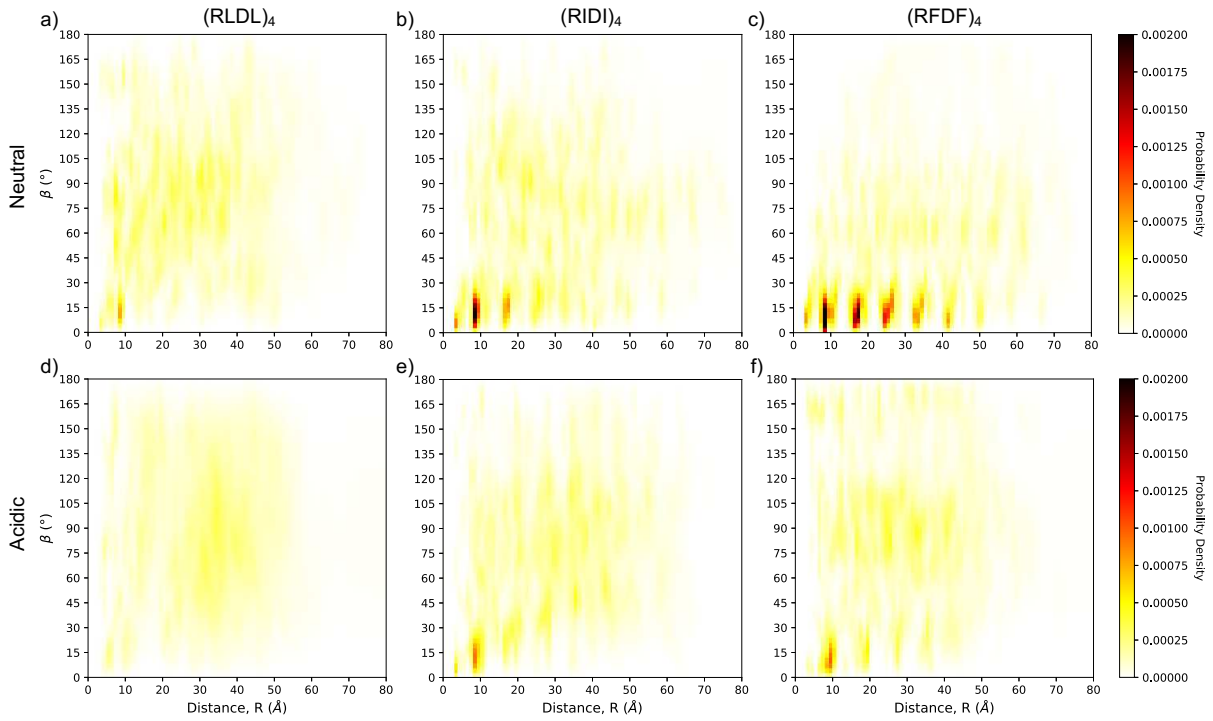

Figure 2.7: Joint probability densities of all coumarin sidechain dimers as a function of separation distance and angle between their respective dipole moments for (a,d) $(RLDL)_4$, (b,e) $(RIDI)_4$, and (c,f) $(RFDF)_4$ at neutral and acidic pH, respectively.

$(RFDF)_4$ systems at neutral and acidic pH. At neutral pH there is a high degree of coumarin order for the $(RFDF)_4$ fibers as there are large peaks near 0° that repeat out to long distances corresponding to a bricklayer structure composed of a majority of the coumarins within the fiber. $(RLDL)_4$ shows little coumarin order as there is a broad distribution of angles between coumarin sidechain and little density at large distances. Generally, as the hydrophobicity, size, and aromaticity of the peptide increases there is a shift to a parallel arrangement of the coumarins and an increase in density at larger distances corresponding to larger coumarin aggregates within the fiber. When the pH is lowered there is a reduction in long range structure and the distribution of angles between coumarin dimers broadens. This is most apparent in the coumarin-$(RFDF)_4$ fibers where there is a significant increase in sampling of angles near 90° and a reduction in sampling of angles near 0°.

Table 2.2: Coumarin-$(RXDX)_4$ peptide sequences and relative entropy.

Sequence	X	S_{rel}
$(RVDV)_4$	Val	0.3047(1)
$(RLDL)_4$	Leu	0.28047(2)
$(RIDI)_4$	Ile	0.583(2)
$(RFDF)_4$	Phe	0.884(1)

Modifying the hydrophobicity of the nonpolar residue alters the pH-switchability of coumarin order within coumarin-$(RXDX)_4$ fibers. Table 2.2 shows the relative entropy between the neutral and acidic joint probability densities for each fiber system. Contrary to the trends in pH-switchability of $(RXDX)_4$ and coumarin-$(RXDX)_4$ fiber structures, there is an increase in pH-switchability of coumarin ordering with increasing hydrophobicity of the peptide sequence, where coumarin-$(RFDF)_4$ fibers have the largest relative entropy between the neutral and acidic probability densities. These results suggest that for the development of switchable self-assembling biomaterials using Coumarin-$(RXDX)_4$ sequences, coumarin-$(RFDF)_4$ is the most ideal as it shows the largest pH-switchability of the coumarin sidechain order and therefore, the largest change in the optical and electronic properties of the fiber.

2.5 Conclusion

I simulate a series of $(RXDX)_4$ and coumarin-$(RXDX)_4$ β-sheet sandwich fibers to investigate the role of hydrophobicity in the stability and pH-switchability of the fibers for the development of switchable self-assembling biomaterials with interesting optical and electronic properties. Generally, the fibers are shown to be pH switchable as they became destabilized at acidic pH. Increasing the hydrophobicity of the peptide sequence leads to an increase in fiber stability at both neutral and acidic pH resulting in a decrease in pH-switchability. Addition of coumarin to the $(RXDX)_4$ sequence results in fiber instability and induces a curvature in the fiber structure due to differences in the size between the coumarin sidechains and the hydrophobic amino acid sidechains. These effects are reduced at neutral pH relative to acidic conditions and with larger and more hydrophobic sidechains. Coumarin sidechains become more ordered and showed the largest pH-switchability with increasing hydrophobicity. Of the systems I studied the coumarin-$(RFDF)_4$ sequence was the most stable fiber with the addition of coumarin into the sequence and the fiber was most stable and had the largest change in the organization of coumarin with pH change making it the most ideal for these types of materials. Future work will measure the optical and electron transfer properties of these fibers as well as investigate the self-assembly mechanism of these materials.

2.6 Funding

I gratefully acknowledge the Army Research Office for financial support (ARO Fund number W911NF-17-1-0383). This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562. I would specifically like to acknowledge the San Diego Supercomputer Center Comet and Pittsburgh Super Computing Bridges resources used under XSEDE allocation CHE160008 awarded to Martin McCullagh.

Chapter 3

The Role of ATP in the RNA Translocation Mechanism of SARS-CoV-2 NSP13 Helicase[b]

3.1 Overview

The COVID-19 pandemic has demonstrated the need to develop potent and transferable therapeutics to treat coronavirus infections. Numerous antiviral targets are being investigated, but non-structural protein 13 stands out as a highly conserved and yet under studied target. Nsp13 is a superfamily 1 helicase that translocates along and unwinds viral RNA in an ATP dependent manner. Currently, there are no available structures of nsp13 from SARS-CoV-1 or SARS-CoV-2 with either ATP or RNA bound presenting a significant hurdle to the rational design of therapeutics. To address this knowledge gap, I have built models of SARS-CoV-2 nsp13 in Apo, ATP, ssRNA and ssRNA+ATP substrate states. Using 30 μs of GaMD simulation (at least 6 μs per substrate state), these models were confirmed to maintain substrate binding poses that are similar to other SF1 helicases. A Gaussian mixture model and linear discriminant analysis structural clustering protocol was used to identify key aspects of the ATP-dependent RNA translocation mechanism. Namely, four RNA-nsp13 structures are identified that exhibit ATP-dependent populations and support the inch-worm mechanism for translocation. These four states are characterized by different RNA-binding poses for motifs **Ia**, **IV** and **V** and suggest a power stroke–like motion of domain 2A relative to domain 1A . This structural and mechanistic insight of nsp13 RNA translocation presents novel targets for the further development of antivirals.

[b]Reproduced with permission from Journal of Physical Chemistry, submitted for publication. Unpublished work copyright 2021 American Chemical Society.

3.2 Introduction

Severe acute respiratory syndrome coronavirus 2, responsible for the COVID-19 pandemic, has infected over a 150 million people and caused more than 3 million deaths worldwide as of May 2021.[56] While antigen-based vaccines have demonstrated significant success at mitigating severe disease and spread, the need to treat infected patients as well as the evolution of potentially vaccine resistant mutants make the development of potent antivirals a pressing concern. Additionally, it has been suggested that there is the potential for new coronaviruses to become infectious to humans necessitating the development of alternative and possibly general therapeutics. To that end, the characterization of structure-function relationship of vital SARS-CoV-2 proteins is necessary to aid in the development of antivirals.

A promising target for antiviral drug development against SARS-CoV-2 is the nsp13, one of 16 nonstructural proteins (nsps),[113] because it plays a critical role in viral replication and the inhibition of which in SARS-CoV-1 has been demonstrated to lead to inhibition of viral replication.[114,115] Nsp13 is a helicase protein that is highly conserved across SARS viruses[13,59,116] and is hypothesized to be a component of the RNA replication complex with the RNA polymerase, nsp12, and other nsps.[59,113,117–119] Nsp13 has also been implicated in viral RNA capping activity[59,60] and as an interferon antagonist.[120] Viral helicases have been targets for antiviral development in SARS,[13,113,121] flaviviruses[116,122–132] and other positive-sense RNA viruses. Further characterization of the structure–function relationships of SARS-CoV-2 nsp13 will allow for clarification of its role in viral replication and aid in the development of antivirals.

SARS-COV-2 nsp13 is classified as a SF1 helicase allowing for the prediction of the ATP-pocket and RNA-binding cleft within its structure. While there is no SARS-CoV-2 nsp13 crystal structure available in the literature, the very close homolog from SARS-CoV-1 (PDB: 6JYT),[57] has a 99.8% sequence identity with SARS-CoV-2 nsp13.[59] The SARS-CoV-1 nsp13 structure, depicted with subdomain coloring in Figure 3.1(b), is composed of five domains: zinc binding domain (ZBD)(red), stalk domain (blue), domain 1B (pink), domain 1A (green), and domain 2A (cyan). Nsp13 has been classified as a SF1 Upf1-like helicase, a family of enzymes with similar

sequence characteristics, such as distinct structural features, specificity for ATP, and unwinding polarity, that are found to interact with DNA/RNA in both eukaryotes and viruses.[53,57,133–135] As found in all SF1 and SF2 helicases, domains 1A and 2A, also known as the RecA-like domains, comprise the conserved helicase region. The classification of nsp13 as a SF1 helicase allows for the prediction of the ATP binding site in-between domains 1A and 2A as well as the RNA-binding cleft separating the RecA-like domains from domain 1B. ATP is depicted in its predicted binding pocket in Figure 3.1(c). Verification of these structural inferences is necessary for developing well-founded structure–function hypotheses as well as rationally designing therapeutics.

Aspects of the structure–function relationship for nsp13 can be inferred through comparison to sequence-similar enzymes. Upf1-like helicases utilize an NTPase cycle to provide the free energy to unwind dsRNA and translocate along the nucleic acid substrate in a 5′ to 3′ direction.[53,57] These enzymes also exhibit RNA-dependent NTPase activity.[136] A set of highly conserved motifs including nucleoside triphosphate (NTP) binding and hydrolysis motifs (**I**, **II**), RNA binding and unwinding motifs (**Ia**, **Ib**, **IV**), and motifs connecting the two binding regions (**III**, **V** and **VI**) have been found to be important for the function of SF1 helicases.[53,133,134,137] The motifs found in SARS-CoV-2 nsp13 are indicated in Figure 3.1(c). How all of these motifs work in concert to allow for NTP-dependent RNA translocation remains unknown for SARS-CoV-2 nsp13, yet this information is critical for rational design of inhibitors.

SF1 helicases are thought to translocate by either an inchworm stepping or Brownian ratchet mechanism.[138] The inchworm mechanism has two sites that alternate between strongly and weakly bound states such that one site is always strongly bound to RNA. The weakly bound site performs a power stroke before strongly binding RNA one basepair forward. This behavior is dependent on the ATP substrate state and can lead to unidirectional translocation along an oligonucleotide substrate. The Brownian ratchet is a simpler two state model in which RNA is either strongly or weakly bound to the protein. RNA is translocated through a power stroke when the protein enters the short lived, weakly bound state. These mechanisms are distinguished by their ATP-dependent RNA-binding activity and, as such, are testable based on ATP-dependent RNA-bound structures.

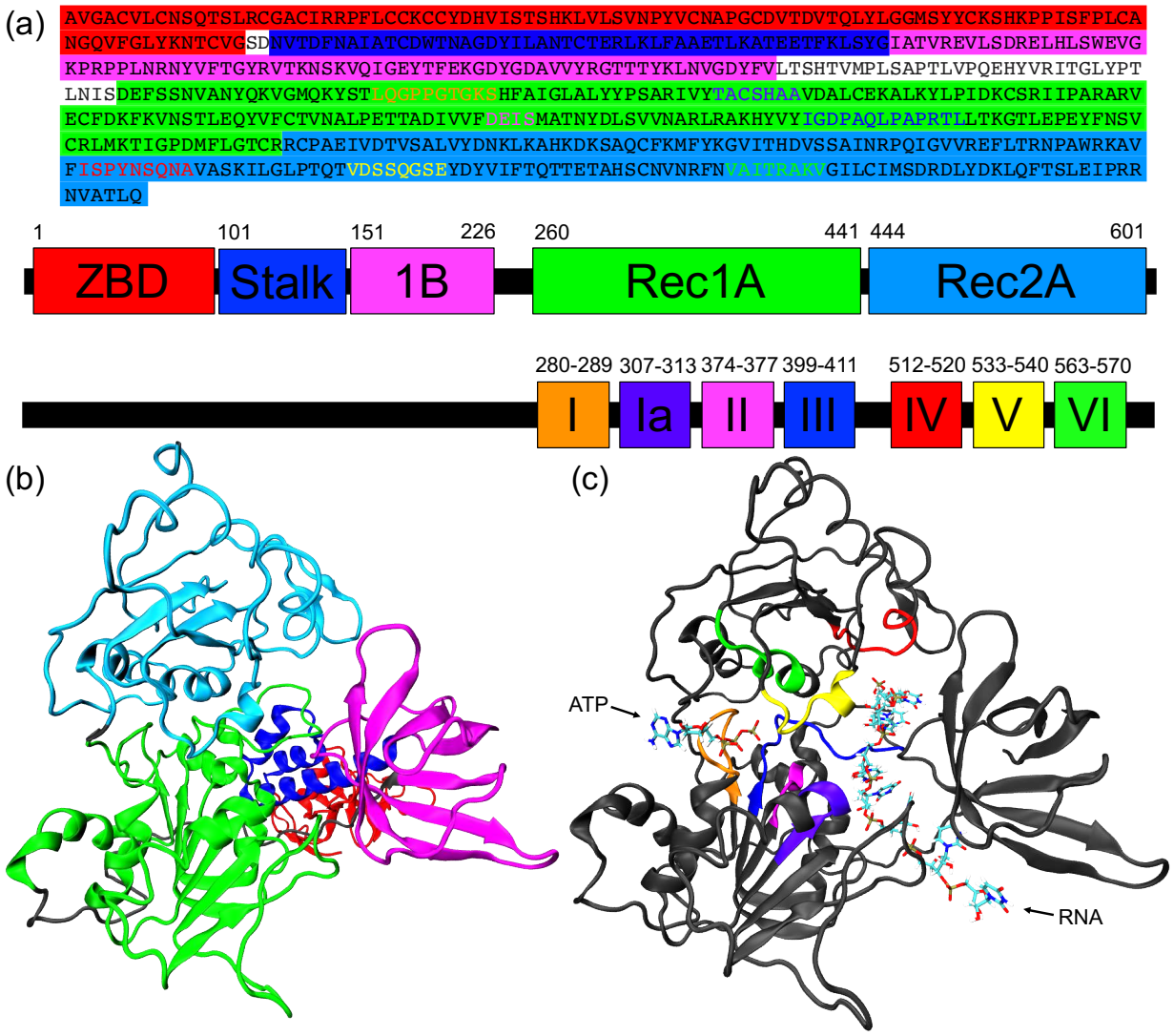

Figure 3.1: SARS-CoV-2 Nsp13 helicase structure. (a) Sequence, domain structure, and motif structure of nsp13. Nsp13 model based on the I570V mutation of SARS-CoV-1 nsp13 (6JYT), colored by (b) domain and (c) motif.[54] The ZBD and Stalk Domains were removed in the motif structure for clarity.

The mechanism of RNA-translocation by SARS-CoV-1 nsp13 has been examined at an ensemble kinetics level[113] and single-molecule kinetics level[139] to identify the rate of ATP hydrolysis and translocation. H/D mass exchange data suggest the presence of at least two RNA-bound states depending on the presence of an ATP substrate.[57] These data are insufficient, however, to distinguish between the two proposed mechanisms and do not provide enough structural information to inform antiviral development.

The work presented here examines the structure–function relationships utilized by SARS-CoV-2 nsp13 during ATP-dependent RNA translocation. Specifically, I focus on the effect of the presence of ATP and RNA on the structural ensemble of the protein and elucidate the role of ATP in the translocation mechanism. I perform extensive GaMD simulations[140] in combination with Gaussian Mixture Model (GMM) structural clustering and characterization by Linear discriminant analysis (LDA). From these analyses I identify four states in the RNA-binding cleft which is indicative of nsp13 utilizing an inchworm stepping translocation mechanism. The role of ATP binding in the translocation mechanism is also elucidated. Furthermore, I analyze the ATP-pocket of the four states identifying key motifs that allosterically connect the ATP-pocket to the RNA-binding cleft.

3.3 Computational Methods

3.3.1 System Setup

Simulations were performed for four ligand-bound states of the nsp13 helicase: Apo, single-stranded RNA (ssRNA), ATP, and ssRNA+ATP. The initial structure for the Apo state is based on the I570V mutation of SARS-CoV-1 (PDB: 6JYT).[57] Due to the lack of ligand-bound crystal structures of SARS nsp13s, RNA and ATP substrates were extracted from Upf1-like helicases aligned to the mutated SARS-CoV-1 nsp13 crystal structure. For the ssRNA and ssRNA+ATP states, polyuracil ssRNA was extracted from the RNA-bound Upf1 helicase crystal structure (PDB: 2XZL)[141] after the two crystal structures were aligned using a sequence–based maximum likelihood protocol as implement in THESEUS.[142] For the ssRNA+ATP bound state, the ATP-analog AMP-PNP and the coordinated Mg^{2+} ion were extracted from the AMP-PNP-bound Upf1 helicase crystal structure

(PDB: 2GJK)[143] after the two crystal structures were aligned using the P-loop, motif **II**, motif **VI**, and motif **V** as alignment landmarks. The amide group between the β- and γ-phosphates in AMP-PNP were replaced by an oxygen atom to form ATP. Furthermore, residues Gly282-Gly287 from the P-loop region in domain 2A were replaced by residues Gly430-Gly435 from the P-loop region in the AMP-PNP-bound Upf1 helicase. The ATP state was created by taking the ssRNA+ATP state and removing the ssRNA.

The protein is modeled using ff14SB parameters,[144] RNA is modeled using ff99bsc0$_{\chi OL3}$ parameters,[145,146] and the parameterization files for ATP[147] are obtained from the AMBER parameter database. Additionally, nsp13 has a ZBD with three non-standard zinc binding pockets. The three zincs are parameterized in Cys-Cys-Cys-Cys, Cys-Cys-Cys-HID, and Cys-Cys-HID-HIE environments, respectively, using the MCPB tool in AMBER.[148] Crystallographic waters are maintained for each state and TIP3P water was added to each system with at least a 12 Å buffer yielding a cubic box of linear dimension of at least 130 Å and a total of at least ~ 215K atoms. See Table A.7 in the supporting information for specific details of the size of each system. Na^+ and Cl^- ions were added to neutralize charge and provide a 0.1 M salt concentration.

3.3.2 Simulation Details

All-atom, explicit solvent GaMD simulations for the Apo, ssRNA, ATP, and ssRNA+ATP states of nsp13 are performed using the GPU-enabled AMBER18 software.[99] Hydrogen atoms are constrained using the SHAKE algorithm.[109] Direct nonbonding interactions are cut off at 12 Å, and long-range electrostatic interactions are modeled using the PME treatment.[110] An integration time step of 2 fs is used. GaMD simulations are performed in the NPT ensemble with a MC barostat set to 1 atm and a Langevin thermostat set to 300 K.

The simulation protocol used for all systems are the same. The energy of the systems are minimized in ten stages. In all minimization stages 2000 steepest descent minimization steps are performed with varying harmonic restraints. In the first stage there is a 500 kcal mol^{-1} Å^{-2} restraint on all protein and ligand atoms. In the next four stages the restraints on the protein sidechains are

reduced to 10.0, 1.0, 0.1, and 0.0 kcal mol^{-1} Å^{-2}, respectively. Finally, in the last five stages there are diminishing restraints on ATP+RNA the protein backbone and all ligands of 50.0, 5.0, 1.0, 0.1 and 0.0 kcal mol^{-1} Å^{-2}. The system is heated to a temperature of 300K over 1 ns with a harmonic restraint of 40 kcal mol^{-1} Å^{-2} on all protein and ligand atoms. Pressure equilibration is performed in six stages. First 1 ns of NVT simulation is performed maintaining the harmonic restraint from the heating. In the next five pressure equilibration stages the restraint was reduced to 20.0, 10.0, 5.0, 1.0, and 0.1 kcal mol^{-1} Å^{-2} for 200 ps each. A conventional molecular dynamics (cMD) NPT simulation is then performed for 10 ns.

Following the cMD simulation each substrate state is simulated using GaMD. For GaMD simulations, the threshold energy for applying boost potential is set to E = Vmax and the default 6 kcal mol^{-1} is used for both σ_{0P} and σ_{0D}. The maximum, minimum, average, and standard deviation values of the system potential are obtained from an initial 10 ns cMD simulation with no boost potential. Then GaMD simulations are performed with boost potential applied to both the dihedral and total potential energy terms. Each GaMD simulation is proceeded with a 40 ns equilibration run after adding the boost potential, followed by 2 μs of production runs. The GaMD simulations of all substrate states are performed in triplicate except for the ssRNA bound state in which six replicates were performed yielding a total of 30 μs of GaMD simulation to be analyzed.

3.3.3 Model Corroboration

Due to the lack of ATP-bound and ssRNA-bound crystal structures, the ATP, ssRNA, and ssRNA+ATP starting structures were created from combining ATP-bound and ssRNA-bound Upf1 helicase crystal structures with a I570V mutated SARS-CoV-1 nsp13 apo helicase crystal structure. The contacts between the SARS-CoV-2 nsp13 protein and the bound ATP and ssRNA ligands for the ATP, ssRNA, and ssRNA+ATP systems are compared to similar contacts in other SF1 helicase proteins, including the Upf1 and IGHMBP2 helicases, to show that these are suitable initial structures that are stable during simulation. For the RNA-bound systems, contacts between motifs **Ia**, **IV**, and **V** with ssRNA phosphates are determined for each frame as these motifs are highly

conserved across SF1 helicase proteins. Similarly, the contacts between motifs **I**, **II**, **III**, **V** and **VI** with ATP and Mg^{2+} are calculated in the ATP-bound systems. A residue and ligand were considered to be in contact if any atom of the residue is within 5 Å of any atom in the ligand. The residue identities and the percentage of frames that each residue is in contact with the ligand are shown in the supporting information for both the ssRNA and ATP contacts in Table A.8 and Table A.9, respectively. These tables show that a majority of the contacts formed in the Upf1 and IGHMBP2 crystal structures are maintained for 60-100% of the frames in the simulation, although, some of the contacts are formed or broken depending on whether both or only a single ligand are bound to the protein.

3.3.4 Gaussian Mixture Model Clustering and Linear Discriminant Analysis

Variational Bayesian Gaussian mixture model[149] is a probabilistic model that effectively fits a given set of data to a specified number of Gaussian distributions with unknown parameters assigning each data point to a cluster. GMM is used to identify structural states in the nsp13 protein by clustering a set of protein–protein distances calculated from the nsp13 simulations using a GMM tolerance of 10^{-6}. The number of clusters that the data is fit to is determined by calculating the silhouette,[150] Calinski-Harabasz (CH),[151] and Davies-Bouldin (DB)[152] scores for a range of cluster sizes from two clusters to ten clusters. Then based on maximums in the silhouette and CH scores and minimums in the DB score the number of clusters is chosen.

Linear discriminant analysis is a classification and dimensionality reduction tool. LDA is a supervised algorithm (data must already be clustered) that finds the linear combination of features that maximize cluster separation. LDA is used to identify the protein–protein distances that best differentiate the four states identified in the RNA-binding cleft.

The GMM clustering and characterization using LDA is performed iteratively. Each iteration uses the projection of the distance data onto the previous iterations LD eigenvectors as the input data for the GMM clustering. The distance data is still used as the input data for LDA. The cycle

is iterated until the distance between the current iterations projected distances and the previous iterations projected distances is below a threshold value of 10^{-3}. Both GMM and LDA are performed using the machine-learning Python library scikit-learn.[153]

3.4 Results and Discussion:

The translocation mechanism of the SARS-CoV-2 nsp13 helicase is likely an important component of the viral lifecycle and yet structural details regarding this mechanism are lacking.[119] Here, I present a set of simulations for the Apo, ATP, ssRNA, and ssRNA+ATP bound states to provide insight into the translocation mechanism of nsp13 along ssRNA. First, I discuss large scale changes in domains 1A, 2A, and 1B between all four systems. I identify changes in the RNA-binding cleft due to the binding of ATP by analyzing differences in the ssRNA and ssRNA+ATP systems which provide insight into the translocation mechanism of SARS-CoV-2 along ssRNA. Finally, I discuss allostery between the ATP-pocket and the RNA-binding cleft focusing on how the presence of ATP changes the ATP-pocket and how those changes affect the RNA-binding cleft.

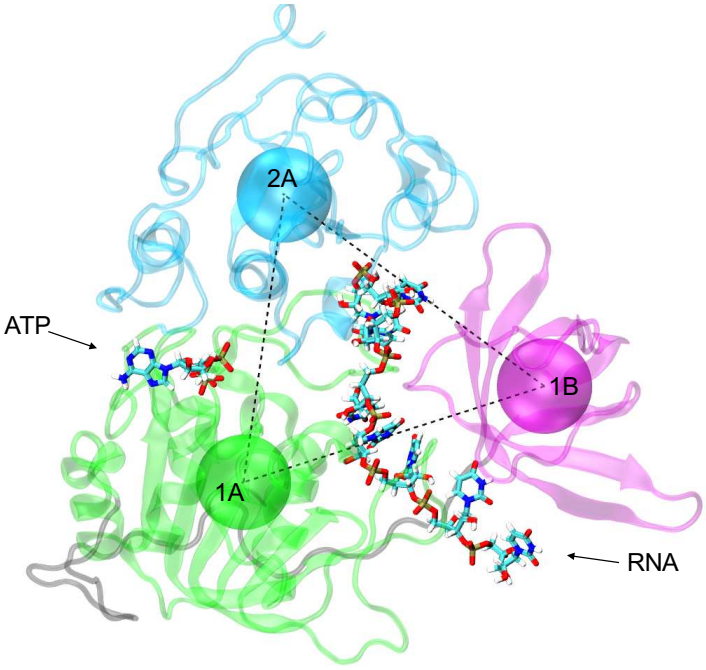

Figure 3.2: Structural depiction of the distances between the center-of-mass of domains 1B (magenta), 1A (green), and 2A (cyan). The stalk and ZBD domains are not depicted for clarity.

3.4.1 Large-Scale Changes to Protein Structure

The interdomain distances between domains 1A, 2A, and 1B, depicted in Figure 3.2, are calculated to investigate large-scale changes within the nsp13 protein structure due to the presence of ssRNA and ATP. The center-of-mass of the β-sheets for each domain are used to calculate the interdomain distances due to their rigidity within each domain. The ATP-pocket sits at the boundary between domains 1A and 2A. The average 1A–2A domain distance remains constant around 31 Å, independent of the presence of ATP and ssRNA as demonstrated by the average separation distances shown in Table 3.1. The standard deviation of all distances are calculated from the set of average reweighted distance values of each replicate of the system.

Although there are no large scale changes to the ATP-pocket, the presence of ATP leads to large scale changes to the RNA-binding cleft. Domain 1B runs along the edge of both domains 1A and 2A forming the RNA-binding cleft. The interplay of these two boundaries are important in the translocation mechanism of nsp13. The 1A–1B and 2A–1B domain distances provide insight into how these boundaries change with ssRNA and ATP binding. The 1A–1B domain distance increases when ssRNA is bound to nsp13. Furthermore, the binding of ATP into the ssRNA-bound state leads to additional widening of the 1A–1B distance from 36.6 Å in the ssRNA system to 38.4 Å in the ssRNA+ATP state. Similar behavior is observed between domains 2A and 1B where the presence of ATP leads to an increase in the 2A–1B distance by 3.4 Å relative to the RNA-bound state. There is more fluctuation in the 2A–1B distance relative to 1A-1B distance as demonstrated by the larger standard deviation value of 2.0 Å for the 2A–1B distance compared to 0.8 Å for the 1A–1B distance for the ssRNA-bound system, which can be attributed to large fluctuations in the tertiary structure of the 2A domain. The widening of the RNA-binding cleft for both the 1A–1B and 2A–1B boundaries suggests a possible reduction in the binding strength of RNA within the RNA-binding cleft due to the presence of ATP. The linear interaction energy (LIE) between each phosphate and the surrounding protein residues and the RMSF of each phosphate were calculated to determine the binding strength between nsp13 and ssRNA and can be found in Table A.10 and Table A.11 of the supporting information. In both analyses the fluctuations in these values were

too large to differentiate between the ssRNA and ssRNA+ATP bound states. One exception is phosphate P5, labeled in Figure 3.3, in which the interaction energy is higher at -63 kcal/mol for the ssRNA+ATP system relative to -112 kcal/mol for the ssRNA system. The lower LIE for P5 in the ssRNA system can be attributed to motif **Ia** binding phosphate P5 in one state of the ssRNA simulations. The LIE values are phosphate specific but, due to the homogeneous nature of the RNA sequence, the nature of the nsp13-RNA state is relatively agnostic to specific phosphate interactions. In the subsequent section, we use a minimum distance analysis to overcome this concern.

Table 3.1: Average center-of-mass separation distance between domains 1A, 2A, and 1B of the nsp13 helicase Apo, ATO, ssRNA, and ssRNA+ATP ligand bound states. Error in the last digit is provided in parentheses.

Separation Distance (Å)				
Domains	APO	ATP	ssRNA	ssRNA+ATP
1A-2A	31.1(6)	31.6(2)	31(1)	31.3(4)
1A-1B	34.6(6)	34(1)	36.6(8)	38.4(4)
2A-1B	30(1)	31.1(4)	33(2)	36.4(1)

3.4.2 Structural Changes of the RNA-Binding Cleft due to the Presence of ATP

The presence of ATP leads to large scale changes in the RNA-binding cleft as shown by the increase in the 1A–1B and 2A–1B interdomain distances. Motifs **Ia** and **IV** are both highly conserved regions in SF1 helicases and it has been suggested that these motifs are involved in RNA binding.[54,154] The structure of the RNA-binding cleft from each frame of these simulations are clustered based on the distances between motif **Ia**, motif **IV**, and the closest RNA phosphates (**P**) using a GMM-LDA approach. Table A.12 in the supporting information contains the residues used as the position of each motif when calculating these distances. Based on the three distances, the

GMM analysis separated the ssRNA and ssRNA+ATP structures into four states: **S1**, **S2**, **S3**, and **S4**. A representative structure of the RNA-binding cleft for each state is shown in Figure 3.3.

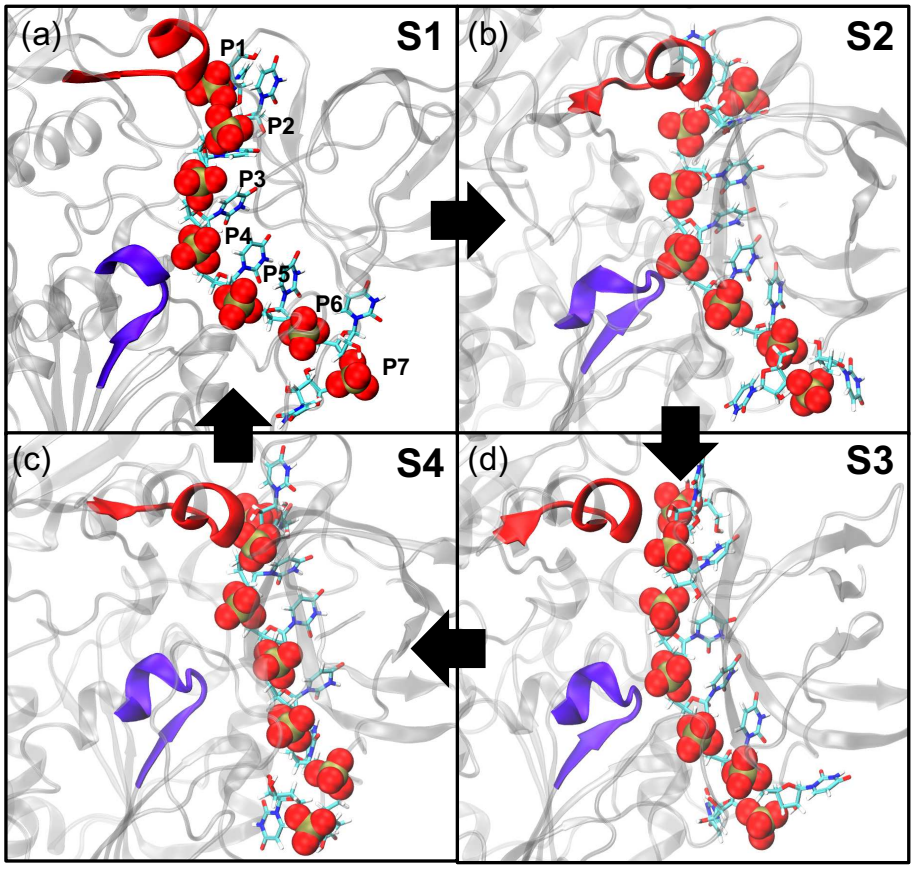

Figure 3.3: Representative structures of the four states identified by the GMM-LDA clustering analysis performed on distances between motif **Ia** (purple), motif **IV** (red), and the closest ssRNA phosphates. The phosphates along the ssRNA have been highlighted. (a) In state **S1** motif **Ia** and **IV** are both strongly bound to ssRNA phosphates and are separated by two unbound phosphates. (b) In state **S2** motif **Ia** and **IV** are separated by a two phosphate gap, but motif **Ia** is strongly bound to a phosphate while motif **IV** is weakly or unbound to the ssRNA phosphates. (c) In state **S4** both motif **Ia** and motif **IV** are strongly bound to ssRNA phosphates, but are separated by one unbound phosphate. (d) In state **S3** motifs **Ia** and **IV** are separated by a one phosphate gap, but motif **IV** is strongly bound to a phosphate while motif **IV** is weakly or unbound to the phosphates. The four states are evidence of a possible 4-step inchworm stepping translocation mechanism. The arrows illustrate the 4-step cycle possibly utilized by nsp13 during translocation.

Analysis of the structure of the RNA-binding cleft for **S1**, **S2**, **S3**, and **S4** reveals that motif **Ia** and motif **IV** independently bind and release ssRNA at different phosphate positions. To determine how the structure of RNA-binding cleft varies for each of the states, the three distances were projected onto the LD eigenvectors calculated by the LDA. The coefficients of LD1 and LD2 are

shown in the supporting information in Table A.13. Figure 3.4(a,b) show the projection of the three distances on to the LD1 and LD2 eigenvectors for the ssRNA and ssRNA+ATP systems, respectively. LD1 separates **S4** from the other three states and is dominated by the distance between motif **Ia** and the RNA phosphates. Table 3.2 shows the average of each of the distances used in the clustering analysis for all clusters. **S1**, **S2**, and **S3** have an average **Ia – P** distance around 5.5 Å, while in **S4** this distance increases to 9.4 Å. This can be seen in Figure 3.3(c) where the ssRNA has separated from motif **Ia**, causing ssRNA to become more linear in the RNA-binding cleft. LD2 distinguishes between **S1**, **S2**, and **S3** and is dominated by both the **Ia – IV** and **IV – P** distances. The distance between motif **Ia** and **IV** is smallest for **S3** at 14.1 Å and largest for **S2** and **S1** at 18.2 and 20.2 Å, respectively. The change in the **Ia – IV** distance between **S1** and **S3** can be explained by a change in the number of phosphates between the phosphates bound by motifs **Ia** and **IV**. In the representative structure for **S3** (Figure 3.3(d)), motif **Ia** is bound to P4 and motif **IV** is bound to P2 leaving only a single phosphate gap (P3) between them. On the other hand, in the representative structure for **S1** (Figure 3.3(a)), motif **Ia** is bound to P4 and motif **IV** is bound to P1, leaving a 2 phosphate gap (P2 and P3) between them. This is qualitatively consistent with the linear interaction energy of P5 as the state where motif **Ia** binds P5 is an example of a conformation where there is a two phosphate gap between motif **Ia** and **IV** as motif **Ia** is bound to P5 and motif **IV** is bound to P2. **S1** and **S2** are not well separated by the **Ia – IV** distance, but are separated by the distance between motif **IV** and the RNA phosphates. In **S1**, the average **IV – P** distance is 5.0 Å, while in **S2** the average distance increases to 8.1 Å as the ssRNA bends away from motif **IV**.

Table 3.2: Average separation distances and standard deviations between motif **IV**, motif **Ia**, and the closest ssRNA phosphates (**P**) for states **S1**, **S2**, **S3**, and **S4**.

	Average Distance (Å)			
Residues	**S1**	**S2**	**S3**	**S4**
IV – P	5.0(8)	8(1)	6(1)	4.7(4)
IV – Ia	20(1)	18(2)	14.2(9)	16.3(7)
Ia – P	5.5(5)	5.4(4)	5.4(2)	9.4(9)

The four states identified in the GMM-LDA analysis suggest a 4-step inchworm stepping translocation mechanism of nsp13 along ssRNA. In this inchworm mechanism, motif **Ia** and motif **IV** act as binding sites that independently bind and release the phosphates of ssRNA. There is always one of the binding sites that is strongly bound to ssRNA, unlike in a Brownian ratchet mechanism in which the protein as a whole, strongly or weakly binds ssRNA. The mechanism follows the cycle shown in Figure 3.3. Let's assume the cycle starts in **S1** where motif **IV** is bound to P_n and motif **Ia** is bound to P_{n+3} leaving a two phosphate gap between them. In the first step of the mechanism motif **IV** and P_n unbind as the protein transitions to **S2**. In the second step motif **IV** does a power stroke moving down one base of ssRNA and binds P_{n+1} resulting in the protein being in **S3**. This leaves a single phosphate gap between the phosphates bound by motifs **Ia** and **IV**. In the third step the protein transitions to **S4** as motif **Ia** and P_{n+3} unbind. Finally, in the fourth step, **Ia** performs a power stroke moving down one base of ssRNA and binds to P_{n+4} and the protein ends up back in **S1**. Each 4-step cycle performed by nsp13 results in a one base pair translocation along ssRNA.

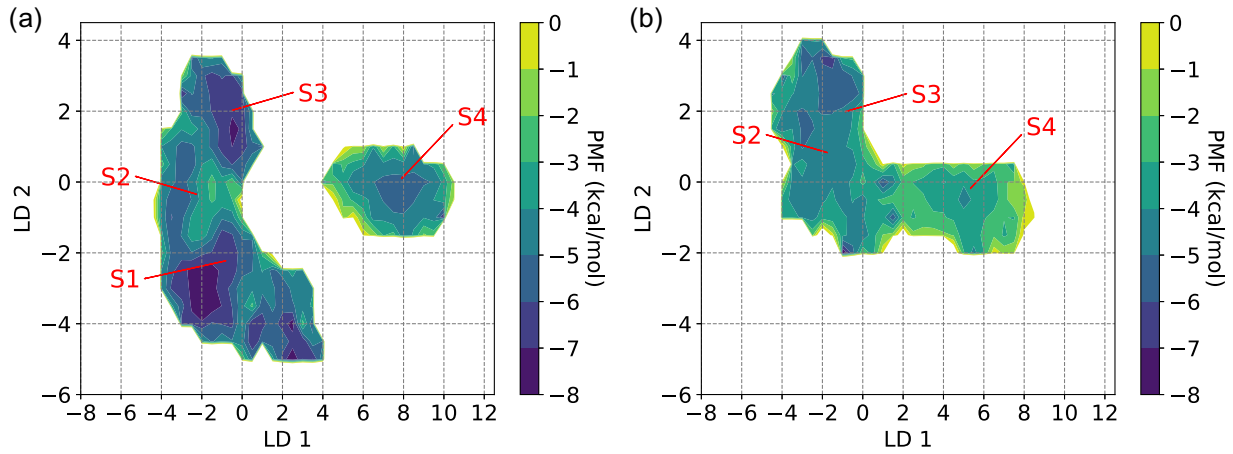

Figure 3.4: TThe projection of the distances between motif **Ia**, motif **IV**, and the closest ssRNA phosphates on to LD1 and LD2 eigenvectors from the GMM-LDA clustering analysis for the (a) ssRNA and (b) ssRNA+ATP systems.

Changes in the sampling of **S1**, **S2**, **S3**, and **S4** when ATP is present in the ATP-pocket suggests that the binding of ATP leads to the release of ssRNA by motif **IV**. In Table 3.3, the probability

of being in each of the four state for the ssRNA and ssRNA+ATP systems is calculated to provide insight into which of the four states the protein is in when RNA is bound and how the sampling of the four states change when ATP binds into the RNA-bound protein. In the ssRNA system, both **S1** and **S2** are each sampled 33% of the time. With ATP bound, **S1** is no longer sampled and the sampling of **S2** and **S3** both increase to 51% and 33%, respectively. These probabilities suggest that the cycle begins with a two phosphate gap between the phosphates bound by motifs **Ia** and **IV**. The binding of ATP then increases the probability of states where motif **IV** unbinds (**S2**) and performs a power stroke leading to a single phosphate gap between motifs **Ia** and **IV** (**S3**). These probabilities provide evidence that the binding of ATP into the ssRNA-bound state of nsp13 causes the first step in the inchworm stepping translocation mechanism. The remaining translocation steps most likely occur in the later steps of the hydrolysis cycle such as ATP hydrolysis and release of adenosine diphosphate (ADP) and the inorganic phosphate (Pi). This is supported by several studies that suggest the hydrolysis of ATP to ADP+Pi and the release of these products are responsible for the unwinding of dsRNA and leads to the power stroke resulting in the translocation of RNA helicase proteins.[135,138,155,156] It is necessary to perform simulations of ssRNA+ADP+Pi and ssRNA+ADP ligand bound systems to further elucidate the translocation mechanism of nsp13 along the ATP-hydrolysis cycle.

Table 3.3: The probability of being in each of the states identified in the GMM-LDA clustering analysis for both the ssRNA and ssRNA+ATP systems.

	Probability			
Residues	**S1**	**S2**	**S3**	**S4**
RNA	33.01%	35.39%	15.14%	16.45%
RNA+ATP	0.02%	50.78%	33.22%	15.99%

3.4.3 Structural Changes of the ATP-pocket due to the Presence of ATP

The binding of ATP to nsp13 leads to a change in the structure of the RNA-binding cleft as shown by the change in sampling of the four states identified in the RNA-pocket and the increase

in the 1A–1B and 2A–1B interdomain distances when ATP is bound into the ssRNA state. To identify how the presence of ATP changes the ATP-pocket and how these changes allosterically alter the RNA-binding cleft, LDA was performed on states **S1**, **S2**, **S3**, and **S4** using distances between ATP-binding motifs (**I**,**II**,**VI**), RNA-binding motifs (**IV**), and motifs that bind both ATP and RNA (**Ia**,**V**,**III**). Initially, all residues in all conserved motifs were included in the LDA analysis. Iteratively, LDA was performed on these distances and the distances that received low coefficients or contained similar information to other distances in LD1 and LD2 were removed until the eight distances shown in Table 3.4 were chosen.

Table 3.4: Average separation distances and standard deviations between motifs **I**, motif **Ia**, motif **II**, motif **IV**, motif **V**, motif **VI**, and the closest ssRNA phosphates for states **S1**, **S2**, **S3**, and **S4**.

	Average Distance (Å)			
Distance	**S1**	**S2**	**S3**	**S4**
I – Ia	21.4(5)	21.7(6)	21.8(6)	20.0(6)
V – VI	11.9(6)	11.7(7)	11.4(4)	10.9(8)
I – V	21.4(8)	19(1)	18(1)	19.1(7)
II – V	16.0(7)	13(2)	13(2)	13.7(6)
V – P	6.2(6)	6(1)	7.5(8)	8.2(6)
Ia – V	14.0(8)	11(2)	10(1)	9.5(8)
IV – V	8.5(5)	10(1)	9.3(6)	9.5(9)
II – VI	18(2)	17(2)	18(1)	17.9(5)

Motif **V** is identified as a key motif for allosteric communication between the ATP-pocket and the RNA-binding cleft due to changes in its positioning between them for the four states identified in the GMM-LDA analysis of the RNA-binding cleft. To determine how the structure of the ATP-pocket varies in the four states, the eight distances are projected onto the LD1 and LD2 eigenvectors for the ssRNA and ssRNA+ATP systems as shown in Figure 3.5(a,b), respectively. The coefficients of LD1 and LD2 are shown in the supporting information in Table A.14. LD1 separates **S4** from **S1**, **S2**, and **S3**. The average distances shown in Table 3.4 are ordered based on the LD1 coefficients. **S4** has a smaller **I – Ia** distance relative to the other states as domain **Ia** unbinds ssRNA and sits close to the ATP-pocket. Similarly, motif **V** is closer to both the ATP-pocket and motif **VI** for **S4**,

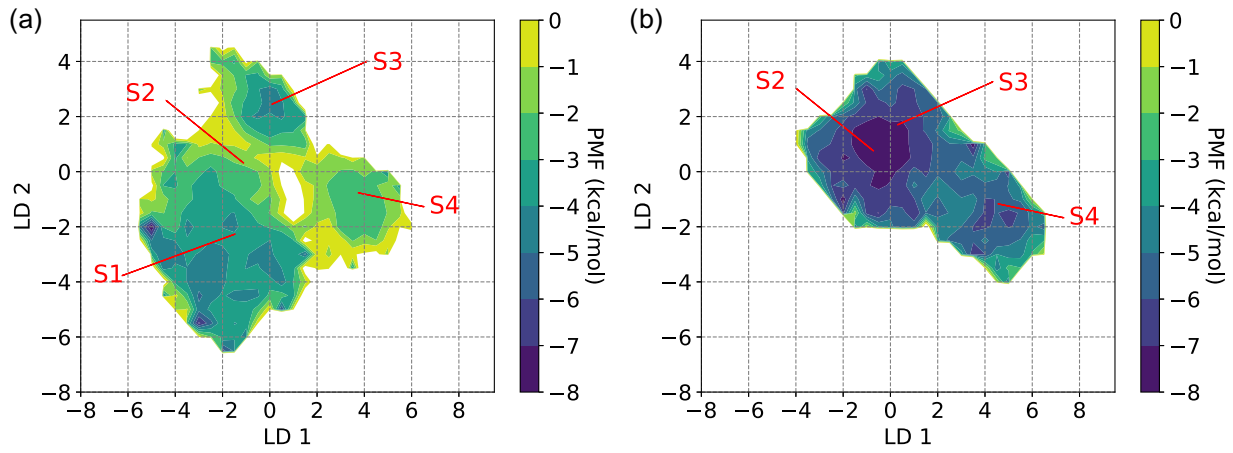

Figure 3.5: The projection of select distances between motif **I**, motif **Ia**, motif **II**, motif **IV**, motif **V**, motif **VI**, and the closest ssRNA phosphates on to LD1 and LD2 eigenvectors from the GMM-LDA clustering analysis for the (a) ssRNA and (b) ssRNA+ATP systems.

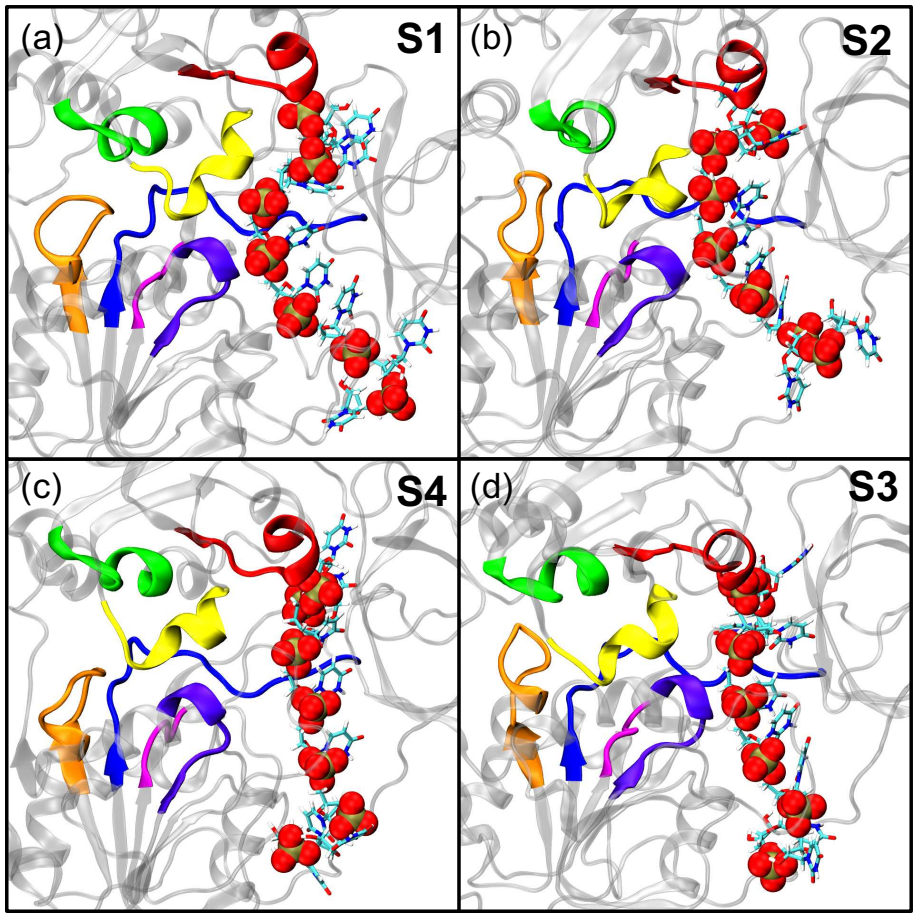

Figure 3.6: Representative structures of the ATP-pocket for states (a) **S1**, (b) **S2**, (c) **S4**, and (d) **S3** four states identified by the GMM-LDA clustering analysis performed on select distances between motif **I** (orange), motif **Ia** (purple), motif **II** (pink), motif **IV** (red), motif **V** (yellow), motif **VI** (green), and the closest ssRNA phosphates

leading to the largest **V – P** distance and smallest **V – VI** distance of all four states. Furthermore, motif **Ia** and **V** sit much closer together relative to the other states. Overall, **S4** has a much more compact ATP-pocket which leads to motifs **V** and **Ia** not binding or weakly binding ssRNA. LD2 differentiates between **S1**, **S2**, and **S3** and has the largest contribution from the and **I – V** distance. The **I – V** distance is smallest for **S3** at 18 Å as motif **V** sits closer to the ATP-pocket as shown in Figure 3.6(c). The **I – V** distance increases for **S2** to 19 Å and increases again for **S1** to 21.4 Å as motif **V** moves closer to the RNA-binding cleft and binds to the ssRNA phosphates. This is further shown by the decrease in the **V – P** distance for **S1** and **S2**. Not only does motif **V** sit closer to the RNA-binding cleft for these two states, but it rotates up away from motif **Ia** becoming more parallel with the ssRNA backbone as it binds one or two of the phosphates as shown in Figure 3.6(a,b). This is evidenced by the increasing **Ia – V** and **II – V** distances.

Table 3.5: Average separation distance and standard deviation between all residues in motif **V** with ATP and Mg^{2+} for states **S1**, **S2**, **S3**, and **S4**.

	Average Distance (Å)			
Residues	**S1**	**S2**	**S3**	**S4**
Val 533	5.3(8)	4(2)	3(1)	4(1)
ASP 534	8(1)	5(1)	6.0(4)	6(1)
SER 535	8(1)	3(1)	4.7(5)	5(2)
SER 536	6.6(7)	4.3(8)	4.2(8)	6.9(7)
GLN 537	9.2(8)	6(1)	7.2(5)	8.5(8)
GLY 538	12(1)	9(1)	9(1)	12.3(8)
SER 539	12.2(7)	10.1(8)	10(1)	12.3(8)
GLU 540	11.0(4)	10.1(5)	10.0(6)	11.0(7)

The analysis of the ATP-pocket allows us to form a more complete picture as to how the presence of ATP leads to a change in sampling of **S1**, **S2**, **S3**, and **S4**. As motif **V** moves farther from the ATP-pocket it binds to ssRNA phosphates closer to motif **IV** stabilizing the two phosphate gap between motifs **Ia** and **IV** and, therefore, stabilizing **S1**. As nsp13 binds ATP there is an increase in the sampling of **S2** and **S3** where motif **V** sits closer to the ATP-pocket due to the formation of contacts between ATP and motif **V**. Table 3.5 shows the average distance between all residues in

motif **V** and ATP or Mg^{2+} for the ssRNA+ATP system. In **S2**, **S3**, and **S4** Val 533, Asp 534, Ser 535, and Ser 536 form contacts with ATP and Mg^{2+} with an average separation distance of 3-6 Å. These contacts make it energetically unfavorable for motif **V** to move close enough to ssRNA to stabilize the two phosphate gap between motif **Ia** and motif **IV** preventing the protein from sampling **S1** when ATP is bound.

These results are consistent with other studies in the literature that identify the importance of motif **V** in the communication between the ATP-pocket and the RNA-binding cleft of other SF1 and SF2 helicases. Molecular dynamics simulations of the flavivirus NS3 helicase protein revealed motif **V** as an allosteric link between the ATP-binding pocket and the RNA-binding cleft due to strong correlations between motif **V** and both binding pockets.[157] Furthermore, these results were supported by mutagenesis studies both *in vitro* and *in silico*.[49] SARS-CoV-2 motif V and subdomain 2A are more dynamic than their flaviviral homologs but the mechanistic role of motif V is conserved. This further reflects the importance of motif V in the translocation mechanism making it an interesting target for antiviral development.

3.5 Conclusions

To provide insight into the translocation mechanism utilized by nsp13 and the role ATP-binding plays in the translocation mechanism I performed simulations of Apo, ATP, ssRNA, and ssRNA+ATP ligand-bound states of the nsp13 helicase. Our models were verified by comparing substrate binding poses to crystal structure of other SF1 helicases that contain these substrates.

Interdomain distances revealed that the binding of ATP leads to an increase in the 1A–1B and 2A–1B domain distances corresponding to a widening of the RNA-binding cleft. A GMM-LDA approach revealed the presence of 4 states in the RNA-binding cleft. These four states represent a 4-step inchworm stepping translocation cycle where motif **Ia** and motif **IV** alternate in releasing a ssRNA phosphate before performing a power stroke and binding a phosphate one basepair forward along ssRNA. The change in sampling of the four states in the ssRNA and ssRNA+ATP systems suggests that the first step in the cycle occurs due to the binding of ATP by nsp13. Analysis of the

ATP-pocket of the 4 states reveal that motif **V** plays an important role in stabilizing ssRNA in states where motifs **Ia** and **IV** are largely separated. When ATP is present, motif **V** forms contacts with ATP reducing its interaction with ssRNA. Based on the simulations that I presented here motifs **Ia**, **IV**, and **V** play a crucial role in the translocation mechanism of nsp13. These motifs would be ideal targets for antiviral drugs to inhibit the function of nsp13.

Future work will focus on performing simulations of ssRNA+ADP+Pi and ssRNA+ADP ligand-bound states of nsp13 to further investigate the translocation mechanism of nsp13 in the later stages of the hydrolysis cycle. Specifically, the role of ATP-hydrolysis, release of the inorganic phosphate, and release of ADP in the inchworm stepping mechanism will be elucidated. Also, enhanced sampling simulations of motif **V** would further clarify the ATP dependence of the structural conformations of motif **V** and its role in stabilizing the two phosphate gap in states where motif **Ia** and **IV** are largely separated.

3.6 Funding

The computing for this project was performed at the High Performance Computing Center at Oklahoma State University supported in part through the National Science Foundation grant OAC-1531128. This work also used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562. I would specifically like to acknowledge the San Diego Supercomputer Center Comet used under XSEDE allocation CHE160008 awarded to Martin McCullagh.

Chapter 4

Implicit Solvation Using the Superposition Approximation (IS-SPA) Simulations on GPUs

4.1 Overview

Molecular dynamics simulation is an important tool that can provide molecular-level insight into condensed phase molecular phenomenon. Often the process of interest occurs on length- and time-scales that is computationally unfeasible to approach with MD. Implicit solvent models were developed to reduce the computational demand of simulating solvated systems by removing the solvent degrees of freedom. Implicit Solvation using the Superposition Approximation is an implicit solvent model that was developed to reduce the computational demand of simulating solvated systems while maintaining a high degree of accuracy relative to explicit solvent simulations. Advancements in the performance of molecular dynamics simulations have recently focused on the use of graphics processing units due to their parallel computing power. I present a GPU implementation of the IS-SPA molecular dynamics code. Three different parallelizations of the IS-SPA algorithm are discussed including a highly parallelized algorithm, a tiling method algorithm, and a modified tiling method algorithm that removes the use of atomic operations at the expense of the level of parallelization of the IS-SPA calculations. The tiling method algorithm outscales the other two algorithms with increasing system size and the number of Monte Carlo integration points used per solute atom. Our results demonstrate that the GPU-implementation of IS-SPA is more computationally efficient than AMBER's explicit solvent all-atom molecular dynamics simulations at concentrations below 25 mM for a system of 50 alanine dipeptide molecules performed on a GPU.

4.2 Introduction

Molecular simulation is a powerful approach to investigate condensed phase molecular phenomenon such as protein folding, allostery, and self-assembly. Many of these processes, however, occur on time- or length-scales that are still challenging to approach using explicit solvent all-atom simulations. A variety of methods have been developed to address this issue including enhanced sampling protocols, particle coarse-graining, and implicit solvation. Of these methods, implicit solvation is the most computationally appealing since it offers the possibility for drastic computational savings; for many biomolecular simulations, the solvent represents $\sim$90% of the computational cost and yet the macroscopic properties of interest typically only depend on the solute positions. Despite its appeal, implicit solvation has not become mainstream due to a lack of models that provide both computational efficiency and reasonable thermodynamic consistency.

Implicit solvation is achieved by removing explicit representation of the solvent particles but including their effect in the forces felt by the solute. A variety of theoretical approaches have been used to determine the forms of the forces that should be applied. While certainly a simplification of the plethora of available methods, I separate implicit solvent models into two types: those meant for single point calculations and those meant for simulation. The single point calculation methods are largely based on classical density functional theory and include methods such as RISM[158–160] and variational implicit solvation.[161] These methods can provide highly accurate free energies of solvation but come at a significant computational cost and are thus not typically applicable for large-scale biomolecular structural sampling. One common component of these accurate methods is the consistent treatment of the polar and non-polar components of solvation. Methods designed for molecular simulation include GB[65] for polar solvation and solvent-accessible surface area (SASA) for non-polar solvation.[162–164] While these methods can be utilized in a simulation their accuracy is application, implementation, and parameterization dependent.[165] It is suggested that these methods are insufficient for molecular aggregation phenomena.[20,64]

IS-SPA is a recently developed implicit solvent model that accurately captures polar[20] and non-polar[64] components to solvation. This method has been shown to be applicable to aggregation

phenomena.[20,64] The non-polar model demonstrated significant computational speed-up compared to both explicit solvent and RISM.[64] The treatment of both polar and non-polar solvation is more computationally expensive but is still found to outperform explicit solvent simulations at solute concentrations well above typical experimental concentrations for self-assembly applications.[20] While the algorithm has demonstrated impressive accuracy and performance, the applications have been limited to relatively small systems. It is of interest, therefore, to further develop this method for large-scale applications.

Recent performance advancements in molecular simulation have focused on GPU implementation of explicit solvent aaMD methods. The recent advancements in the AMBER implementation of GB and SASA, in particular, demonstrate the importance of considering GPU implementation for simulation-based implicit solvation methods.[166] The AMBER implementation is based on a previous implementation of GB on a GPU[167,168] that used a tiling particle decomposition to distribute the pairwise computation workload efficiently across the GPU. Recent performance results have been impressive suggesting the importance of considering this scheme in the GPU algorithm for IS-SPA.

In this chapter, I present a GPU implementation of the IS-SPA algorithm. The entire code was written from scratch and each force calculation is performed on the GPU. Three different GPU parallelization algorithms are compared and contrasted. I start by briefly outlining the theory behind IS-SPA, next I describe the three algorithms, and finally I present the results of these algorithms as applied to systems of alanine dipeptide (AP) molecules.

4.3 Theory

The goal of IS-SPA is to accurately reproduce the mean solvent force on a given solute configuration. The IS-SPA theory is described in full detail by Lake *et al.*[20,64] For polar solutes in an arbitrary solvent, the mean solvent force on an atom j, given the position $\boldsymbol{R}^N$ of the N solute atoms, in the case of pairwise additive interactions is given by

$$\langle \boldsymbol{f}_j \rangle_{solv} = \rho \iint \boldsymbol{f}_{j,solv}(\boldsymbol{r} - \boldsymbol{R}_j, \boldsymbol{\Omega}) g(\boldsymbol{r}, \boldsymbol{\Omega}; \boldsymbol{R}^N)\, d\boldsymbol{\Omega}\, d\boldsymbol{r} \tag{4.1}$$

where ρ is the solvent density, $\boldsymbol{f}_{j,solv}$ is the solvent force, g is the solvent distribution function, and $\boldsymbol{\Omega}$ represents the internal coordinates of the solvent molecule. IS-SPA makes use of several approximations to analytically integrate over $\boldsymbol{\Omega}$ such that only the position of the solvent, $\boldsymbol{r}$, needs to be sampled by MC sampling.

The first approximation presumes that the only important degrees of freedom of a solvent molecule is the alignment of its dipole moment and that the molecule is axially symmetric. This reduces the 12 degrees of freedom in $\boldsymbol{\Omega}$ for a chloroform molecule to two, represented by $\hat{\boldsymbol{p}}$. The second approximation used is the the Kirkwood superposition approximation, where the many-body distribution function is reduced to a product of two-body distribution functions g_k that also take into account the orientation of the dipole moment. Specifically,

$$g(\boldsymbol{r}, \hat{\boldsymbol{p}}; \boldsymbol{R}^N) \simeq \frac{\prod_{k=1}^{N} g_k(|\boldsymbol{r} - \boldsymbol{R}_k|) P_k(\hat{\boldsymbol{p}}; \boldsymbol{r} - \boldsymbol{R}_k)}{\int \prod_{k=1}^{N} P_k(\hat{\boldsymbol{p}}; \boldsymbol{r} - \boldsymbol{R}_k)\, d\hat{\boldsymbol{p}}} \tag{4.2}$$

where P_k is the probability distribution of the dipole moment of the solvent given the distance vector $\boldsymbol{r} - \boldsymbol{R}_k$ from atom k.

The probability distribution of the dipole moment is approximated by the probability distribution of a thermal ideal dipole in a constant electric field and is given as

$$P_k(\hat{\boldsymbol{p}}; \boldsymbol{r} - \boldsymbol{R}_k) \simeq \exp\left[-\frac{p\hat{\boldsymbol{p}} \cdot (\boldsymbol{r} - \boldsymbol{R}_k)}{T|\boldsymbol{r} - \boldsymbol{R}_k|} E_k(|\boldsymbol{r} - \boldsymbol{R}_k|) \right] \tag{4.3}$$

where p is the static dipole moment of a chloroform molecule, T is the temperature in units of energy, and E_k is the radially symmetric effective mean field along its separation vector produced by each atom.

Finally, a MC integration is used to approximate the integral of Equation 4.1. Namely,

$$\langle \boldsymbol{f}_j \rangle_{solv} \simeq \frac{1}{M} \sum_{i=1}^{M} \frac{\rho}{\Gamma(\boldsymbol{r}_i)} \boldsymbol{f}_{j,solv}(\boldsymbol{r}_i - \boldsymbol{R}_j, \boldsymbol{E}_i) \prod_{k=1}^{N} g_k(|\boldsymbol{r}_i - \boldsymbol{R}_k|) \tag{4.4}$$

where Γ represents the probability distribution from which the MC points are sampled, M is the total number of MC points for all atoms, and $\boldsymbol{E}_i$ is the sum of the electric fields produced by each atom, E_k, which comes out of the product of the probability distributions of the dipole moments defined in Equation 4.3. It is important to note that the MC integration is only performed inside a defined interaction volume around each atom. The long ranged electrostatics outside of the interaction volume are approximated by a constant density dielectric. The treatment of the long ranged electrostatics are explained in greater detail by Lake *et al.*[20]

4.4 GPU Architecture and CUDA Software

To better understand the IS-SPA CUDA implementations discussed in this chapter it is import to understand the differences between GPUs and CPUs, the architecture of GPUs, and how algorithms are implemented on GPUs using the CUDA platform. GPUs and CPUs are designed to be efficient for different types of computational tasks. CPUs have latency-optimized cores and can handle a wide-range of complex tasks quickly, but these tasks must be run in serial. Alternatively, GPUs contain thousands of throughput-optimized cores making them more efficient than CPUs for computational algorithms that process large sets of data in parallel such as in the case of IS-SPA. The NVIDIA GPU architecture is built around a scalable array of multithreaded Streaming Multiprocessors (SMs) that each consist of a variable number of CUDA cores or Stream Processors (SPs) which are the most basic processing units. Both the number of SMs and SPs per SM is GPU dependent, for example a GeForce GTX 1080 GPU card contains 20 SMs each containing 128 SPs per SM for a total of 2560 CUDA cores.

CUDA is a general purpose parallel computing platform and programming model that utilizes the parallel compute engine of GPUs. A function executed on a GPU is called a kernel. A kernel is executed in parallel by an array of threads. All threads run the same code, but each thread has a unique ID that is used to select which data to work on. A block is a group of threads which are executed together. Blocks are dynamically assigned by the GPU scheduler to SMs giving GPUs their scalability. An SM can execute more than one block concurrently, but the number of blocks

that can simultaneously execute on an SM is limited by both hardware and the amount of resources each block uses including number of threads, shared memory, and number of registers. A grid is a set of blocks which together execute the GPU operation. The number of threads per block and the total number of blocks can be changed to fit the needs of a kernel, but there are limits for each that depends on the specific GPU card. Furthermore, threads are bound together in groups of 32 called warps. All threads in a single warp can only run a single set of instructions at once. For example if one thread within a warp follows one condition of an if-else statement and another thread in the same warp follows the other condition than all 32 threads in the warp will go down both sides of the statement. This does not create a problem functionally as the threads that do not satisfy the condition become inactive when following that branch of the statement, but this could cause a reduction in performance if both sides of the statement are long. Additionally, all threads within a warp fetch data from the memory together. Therefore, it is important to consider warps when trying to optimize the efficiency of the code.

Another important factor to consider when creating an efficient application is the use of memory. Due to the hierarchical nature of the GPU architecture there is also a hierarchy of memory types in size, access speeds, and lifetimes. The smallest and fastest accessible memory type on the SMs are registers. Registers are only accessible by and have the lifetime of a thread. The next fastest memory type is shared memory which can be as fast to access as registers depending on a few factors such as bank conflicts. Each block has its own shared memory that is accessible by any thread within that block and has the lifetime of the block. Shared memory is an important factor to consider when choosing how to divide calculations into blocks. The largest and slowest form of memory is global memory. Global memory can be 150x slower than registers or shared memory and has the lifetime of the application. Global memory can be accessed by both the host CPU and the GPU device and is persistent between kernel launches. Reducing the number of calls to global memory will typically improve code performance.

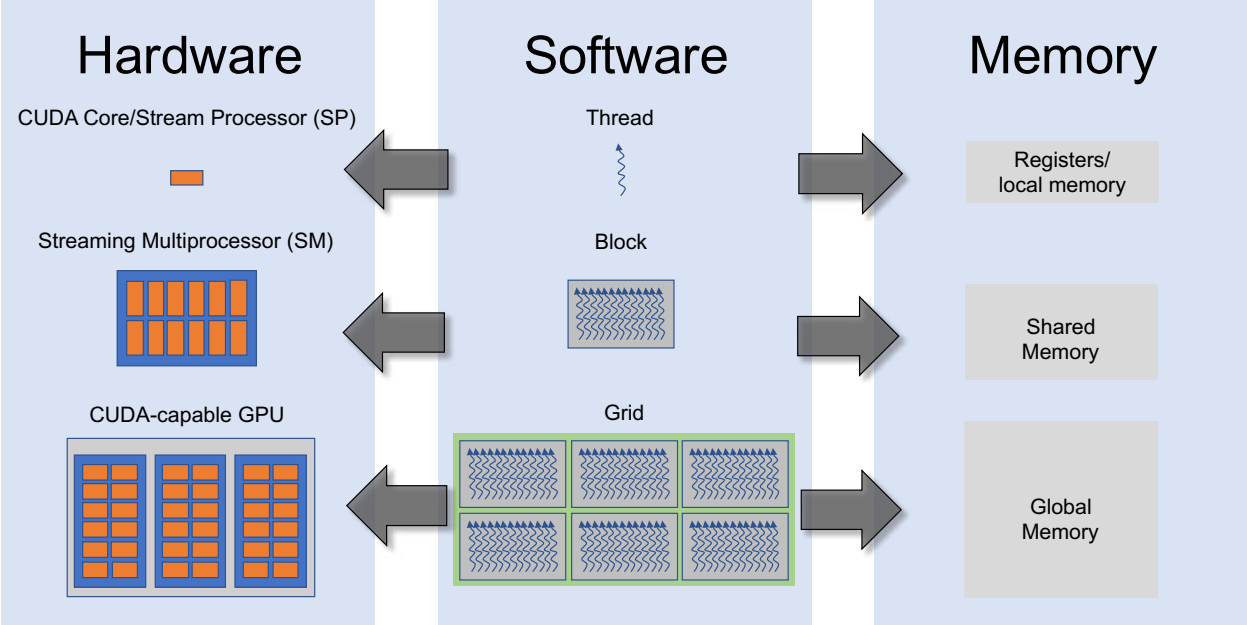

Figure 4.1: Schematic representation of the GPU architecture, the logical hierarchy of the CUDA software, and the GPU memory hierarchy.

4.5 CUDA Implementations

The CUDA algorithms that I use to perform IS-SPA implicit solvent simulations divide the calculation of the IS-SPA forces into three kernels, delineated as the MC kernel, the field kernel, and the force kernel. The MC kernel generates the positions of all of the MC points used to calculate the mean solvent force in Equation 4.4. The user defines the number of MC points per atom, N_{MC}, in the input file. The MC kernel generates N_{MC} MC points within the interaction volume of each atom producing a total of M MC points ($M = N \cdot N_{MC}$). At each MC point i, the field kernel calculates the product of the densities and the sum of the electric field, $\boldsymbol{E}_i$, produced by each atom as in Equation 4.2. Look-up tables are used to determine the density and electric field values produced at an MC point by each atom. Finally, the density and electric field are used by the force kernel to calculate the mean solvent force on each atom j by each MC point i using Equation 4.4. Both the field and force kernels calculate $N \cdot M$ atom–MC point interactions per step of the simulation making them both computationally demanding. There are numerous ways in which these kernels can be written, each with significantly differing computational efficiencies. I present three versions of the CUDA algorithm.

4.5.1 Algorithm 1: Highly Parallelized

The main goal of algorithm 1 is to fully parallelize the IS-SPA calculations such that each thread is only calculating the interactions between a single atom–MC point pair so that no loops are used. The high degree of parallelization of the IS-SPA calculations should make the CUDA algorithm quite efficient. This parallelization is done in both the field and force kernels.

The field kernel parallelizes the calculations of the electric field and density at each MC point such that each thread is only calculating the partial electric field and density that a specific atom is placing onto a specific MC point. For a particular MC point, N/W warps are created, where W is the total number of threads in a warp. If N is not a multiple of W than this leads to some warps having inactive threads. Since all threads within a warp are calculating the electric field and solvent density at a single MC point, the partial electric field and solvent densities are reduced using a warp reduction. Finally, each warp uses atomic operations to push their partial electric field and density to the global total solvent density and electric field for each MC point. Atomic operations are used to prevent race conditions where multiple threads may try to read and write to the same variable simultaneously which could cause a value to be computed incorrectly. A total of $M \cdot (N/W)$ warps are generated filling up $M \cdot (N/W)/S$ blocks, where S is the max number of warps per block.

Similarly, the force kernel parallelizes the calculation of the mean solvent forces on each atom such that each thread only calculates the partial mean solvent force from one specific MC point onto one atom. The electric field must be converted into polarization by each thread instead of being calculated one time in the field kernel due to how the field kernel is organized. For a particular atom, M/W warps are created; if M is not a multiple of W than some warps will have inactive threads. Since each warp is calculating the forces from various MC points onto one atom, the forces from each thread in the warp can be reduced using a warp reduction. Atomic operations are then used to calculate the total mean solvent force on each atom. The force kernel generates $N \cdot (M/W)$ warps across $N \cdot (M/W)/S$ blocks.

4.5.2 Algorithm 2: Tiling Method

In algorithm 2, I use a tiling method approach to calculate the mean solvent forces that is similar to the tiling method used by AMBER and Friedrich *et al.* to calculate the non-bonded forces between atoms in GB implicit solvent simulations in which the calculation of the interaction between each pair of atoms is divided into warp sized tiles that the GPU scheduler distributes across the GPU into individual blocks.[167–169] Albeit, there are a few major differences between the two methods due to differences in how the two forces are calculated. The pairwise interactions between MC point i and atom j can be represented by the matrix shown in Figure 4.2, where the atom–MC point interactions are grouped into tiles of dimension $W \times W$. The reason for the division of the calculations in this fashion is because the GPU scheduler works in terms of warps, where each warp performs the same mathematical operation on W values simultaneously.[169] Unlike calculating the non-bonded interactions between all pairs of atoms, there is no symmetry that can be exploited in the interactions between MC point–atom pairs to reduce the number of calculations needed to be performed. In determining the non-bonded forces, the diagonal tiles must be treated differently then the off-diagonal tiles and only the upper triangle of the matrix needs to be calculated due to Newton's third law of motion. In the tiling method used in calculating the mean solvent forces, every tile in the matrix must be calculated and every tile is calculated in the same way. Both the field and force kernel generate $(N/W) \cdot (M/W)$ blocks, assigning each tile to a block.

In the field kernel a tile assigns each thread in the block an MC points between i to $i + W - 1$. Each thread then loops over atoms, from atom j to $j + W - 1$, calculating the electric field and density produced by the atom at the MC point assigned to that thread as shown by the blue tile in Figure 4.2. Since this kernel is calculating a per MC point quantity, this setup avoids any race conditions within a tile. Race conditions between tiles must still be considered, therefore, atomic operations are used to update the global electric field and densities calculated from each tile. Also, the charge and position of atoms j to $j + W - 1$ are stored in shared memory.

Conversely, a tile in the force kernel assigns each thread in the block an atom between j to $j + W - 1$. Each thread loops over MC points between i and $i + W - 1$, calculating the partial

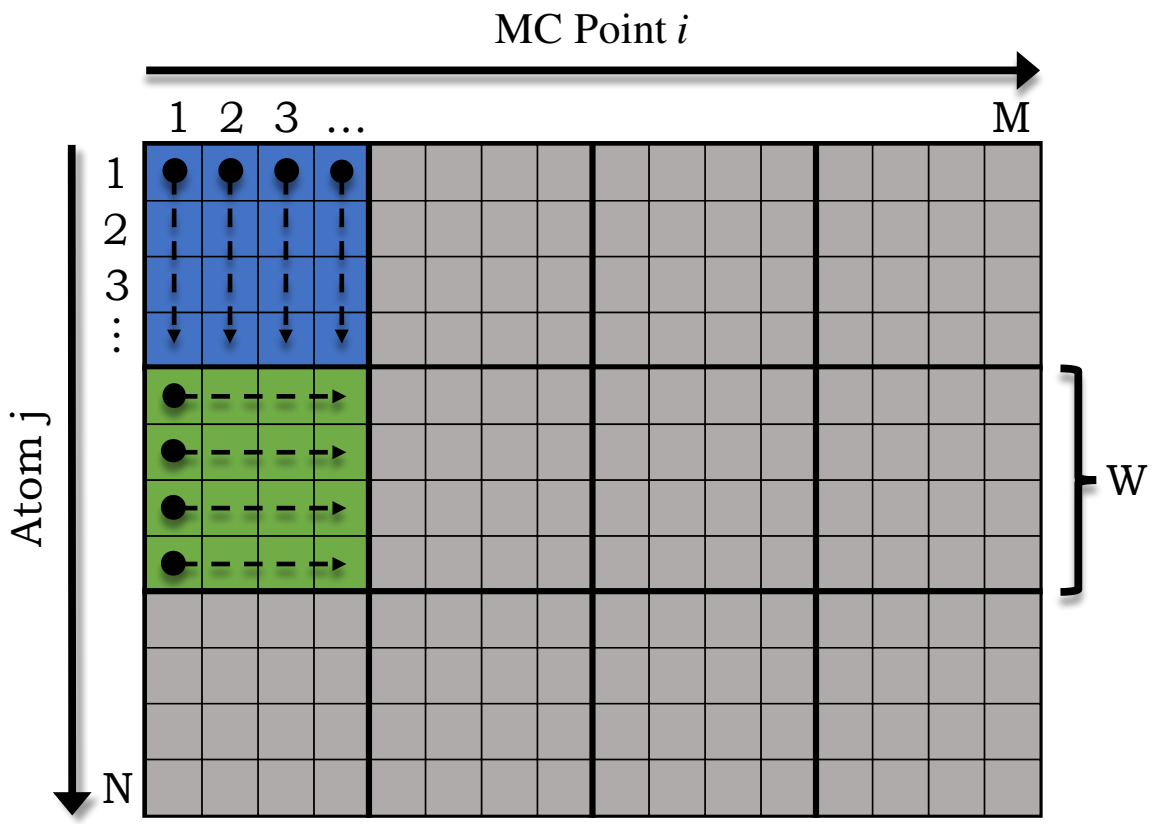

Figure 4.2: Schematic representation of the of the work-load distribution for the calculation of the IS-SPA forces using algorithm 2. Each box represents the interaction between an MC point i with an atom j. These interactions are grouped into tiles of size $W \times W$ that are each assigned to an independent warp. A tile in the IS-SPA field kernel assigns each thread an MC point i and loops over W atoms as illustrated by the blue tile. A tile in the IS-SPA force kernel assigns each thread an atom j and loops over MC points as demonstrated by the green tile.

mean solvent force that the MC point produces on the atom assigned to that thread as shown by green tile in Figure 4.2. This setup allows race conditions to be avoided within a tile because the force kernel is calculating a per atom quantity. Similar to the field kernel, atomic operations must be used to update the global IS-SPA forces due to possible race conditions between tiles. In this kernel, the electric field and density of the MC points i to $i + W - 1$ are stored in shared memory.

4.5.3 Algorithm 3: Tiling Method with no Atomic Operations

In algorithm 3, I use an approach similar to the tiling method with changes in the size of the tile such that all race conditions are avoided. It is suggested that atomic operations are a bottleneck in

terms of performance in the tiling method.[169] Therefore, this approach removes all race conditions between tiles by increasing the size of the loops performed by each tile. Tiles are still organized such that there are no race conditions within a tile.

In algorithm 2, the field kernel assigns a warp-sized block to each MC point for a total of M blocks. Each block then calculates the total electric field and density for its MC point i by looping over all atoms as shown in Figure 4.3. The loop over all atoms is divide among W threads within the warp, such that atoms $j = 0$ to $j = W - 1$ are assigned to a thread. Then the the loop increments j for each thread by W until all atoms are considered for MC point i. At the end of the loop the total electric field and density at the MC point assigned to the warp is calculated by warp reducing the W partial electric fields and densities of each thread. Since each warp is considering a different MC point, the total electric field is converted into a polarization and is then pushed to the global variable without any race conditions.

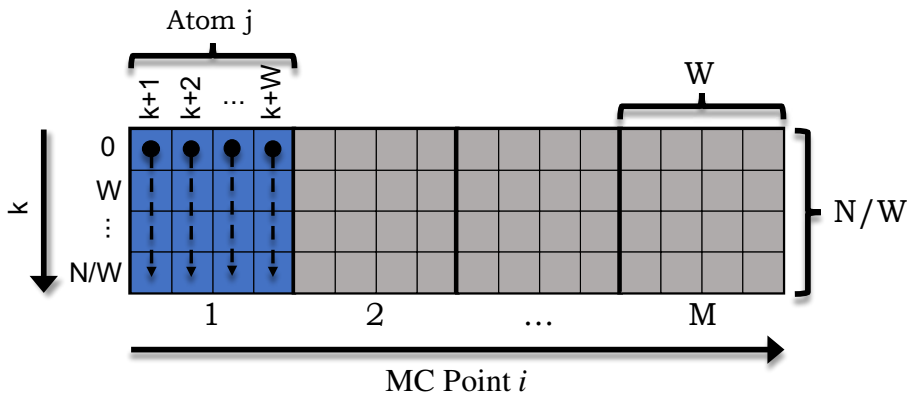

Figure 4.3: Schematic representation of the of the work-load distribution for the calculation of the electric fields and densities by the field kernel using algorithm 3. Each box represents the interaction between an MC point i with an atom j. These interactions are grouped into tiles of size $W \times N/W$ that are each assigned to an independent warp. A tile in the IS-SPA field kernel assigns each warp am MC point i and loops over all atoms as illustrated by the blue tile.

Conversely, the force kernel assigns a warp-size block to each atom for a total of N blocks. Each block calculates the mean solvent force for its atom j by looping over all MC points as shown in Figure 4.4. The loop over all MC points is divided among all W threads within the warp. The

total mean solvent force of the atom is calculated by warp reducing the W partial mean solvent forces of each thread. Since each warp is calculating the mean solvent force on a single atom, no race conditions are encountered when pushing the total force to the global variable.

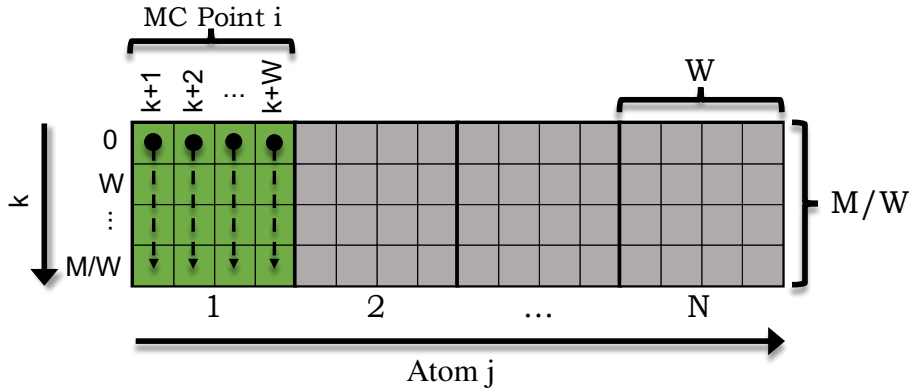

Figure 4.4: Schematic representation of the of the work-load distribution for the calculation of the IS-SPA forces by the force kernel using algorithm 3. Each box represents the interaction between an MC point i with an atom j. These interactions are grouped into tiles of size $W \times M/W$ that are each assigned to an independent warp. A tile in the IS-SPA field kernel assigns each warp an atom j and loops over all MC points as illustrated by the green tile.

4.6 Validation

IS-SPA simulations of both Lennard-Jones ions and AP molecules are performed to validate the accuracy of the IS-SPA CUDA code. The charged Lennard-Jones spheres are used as a simple test case for the IS-SPA CUDA procedure and AP is chosen as a model molecular system as it is a typical test system for solvent models.[161,170] In particular, these two systems were chosen so that the results of IS-SPA CUDA code could be compared to the results presented by Lake *et al.*[20] The IS-SPA simulations are performed in the NVT ensemble with an Andersen thermostat with a collision frequency of 33.33 ps^{-1}. The mass of the hydrogen atoms is set to 12 u to reduce the frequency of those bonds and allow the use of a 2 fs integration time step. No cutoff is used in the calculation of the non-bonded and solvent forces in the IS-SPA simulations. N_{MC} is set to 100 MC points per atom for the dimer IS-SPA simulations of both systems. A radial distance of 12 Å defines

the spherical interaction volume around each atom and is chosen based on when variations in the density and polarization tend to the bulk values.[20] The external dielectric constant of chloroform is set to 2.3473 as measured from the bulk chloroform model.[171,172] The same IS-SPA parameters are used for both the Lennard-Jones ions and the AP molecules as the IS-SPA parameters used by Lake *et al*.

4.6.1 Lennard-Jones Sphere

To validate the accuracy of the IS-SPA CUDA code, the dimerization potential of mean force (PMF) of a pair of Lennard-Jones particles in chloroform is calculated and compared to the Lennard-Jones dimer PMF presented by Lake *et al*.[20] To parallel the simulations presented in that study, the Lennard-Jones ions are given charges of $q = +1$ and $q = -1$ and Lennard-Jones parameters of $\epsilon = 0.152$ kcal/mol and $r_{min} = 7$ Å. The dimer PMF calculated by Lake *et al.* relies on grid integration to numerically solve the integral in Equation 4.1, where as the IS-SPA CUDA code utilizes MC integration to the solve the integral. In both cases, the total force on the positive and negative ions are calculated at separation distances from 3.5 Å to 15.0 Å at intervals of 0.1 Å. The red and blue IS-SPA PMF curves presented in Figure 4.5 are the integration of the positive and negative ion forces, respectively. There is complete agreement between the curves produced by the IS-SPA CUDA code and those presented by Lake *et al* demonstrating the accuracy of the IS-SPA CUDA code. There is deviation from the PMF produced by explicit solvent simulations, but that is a statement on the accuracy of IS-SPA as an implicit solvent model and not on the accuracy of the IS-SPA CUDA code.

4.6.2 Alanine Dipeptide

Further validation of the accuracy of the IS-SPA CUDA code is performed by comparison of the dimerization PMFs of AP molecules in chloroform presented in Figure 4.6. The dimerization behavior along the center-of-mass separation distance is investigated using umbrella sampling simulations performed by the IS-SPA CUDA code. Similarly, umbrella sampling simulations are performed by the IS-SPA Fortran CPU code developed by Lake *et al*.[20] These simulations are

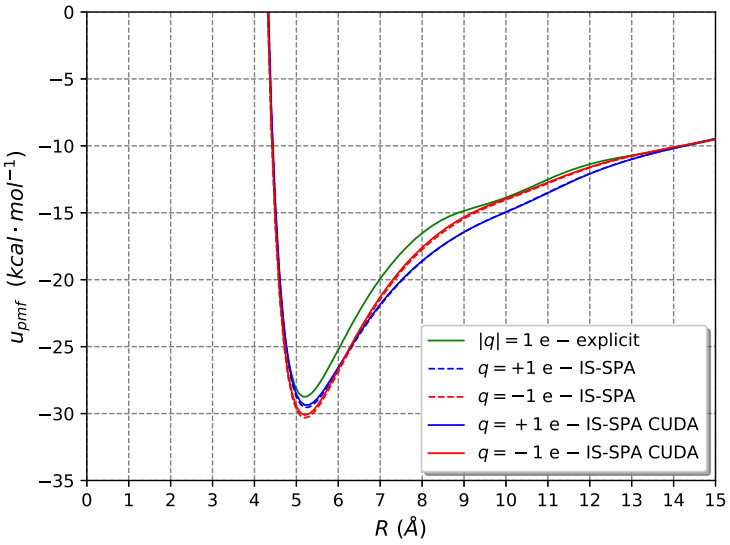

Figure 4.5: Dimerization PMFs of two Lennard-Jones ions with a charge of $q = +1$ (blue curves) and $q = -1$ (red curve) produced by the IS-SPA CUDA code (solid lines) and presented by Lake *et al.* (dotted lines). Furthermore, the explicit solvent dimerization PMF is shown by the green curve.

performed every 0.5 Å from 3.5 Å to 16.0 Å using a force constant of 20 kcal/mol/Å^2 and each window is simulated for 150 ns. There is complete agreement between the two PMFs as shown by the red and blue curves in Figure 4.6 further demonstrating the accuracy of the IS-SPA CUDA code. Again, the discrepancy between the IS-SPA PMFs and the explicit solvent PMF is due to the accuracy of IS-SPA as an implicit solvent model and not the IS-SPA CUDA code, specifically.

4.7 Performance

Although I have developed a full MD CUDA code to perform IS-SPA simulations, the work presented here focuses on optimizing the efficiency of calculating the IS-SPA solvent forces. Other parts of the code, such as the non-bonded interactions, are not highly optimized and, therefore, could provide additional increase in code performance. There are an abundance of studies in the literature that discuss the optimization of cMD simulations performed using CUDA, with a large focus on the non-bonded interactions which constitutes the largest bottleneck in cMD code performance.[161,166,168,169,171,173–177] In the IS-SPA simulations the largest bottleneck lies in

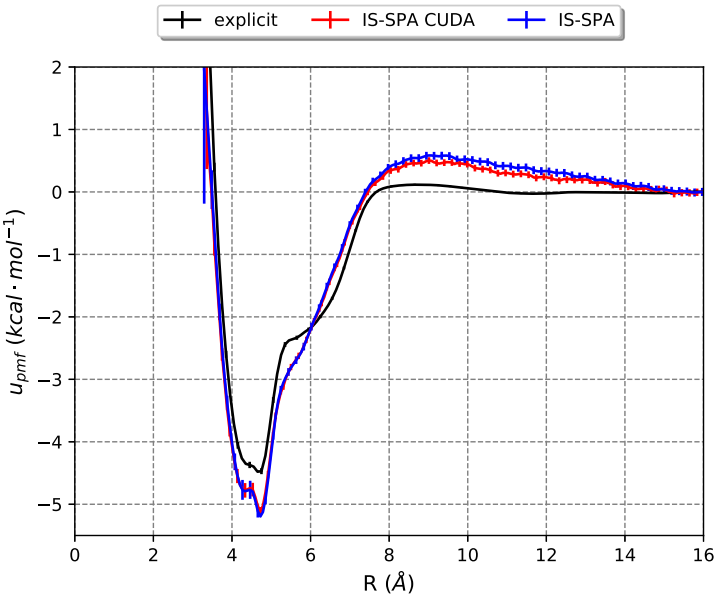

Figure 4.6: Dimerization PMFs of AP along the center-of-mass separation distance using explicit solvent, IS-SPA performed by the IS-SPA Fortran CPU code, and IS-SPA performed by the IS-SPA CUDA GPU code.

the calculation of the IS-SPA solvent forces on the solutes. Therefore, it is not our goal to fully optimize the entire MD code, rather, our goal is to highly optimize the calculation of the IS-SPA solvent forces such that it could be adapted into a highly-developed MD software. For this reason I discuss the relative code efficiency between IS-SPA algorithms, but do not compare the overall code performance with other MD codes. For the three algorithms, I compare how their performance scales with both the number of solute atoms and the number of MC points per solute atom. All performance calculations were performed on a GeForce GTX 1080 GPU card.

To determine how the performance of each algorithm scales with the total number of solute atoms I performed a series of timescaling simulations for systems of 2, 10, 25, and 50 AP molecules in chloroform. Each system was performed with $N_{MC} = 96$ for 1000 steps. To obtain each timescaling point the performance was averaged over 100 individual runs. The timescaling as a function of the number of solute atoms is shown in Figure 4.7. For all system sizes, the tiling method, algorithm 2, outperforms the other algorithms. As the system size increases the performance of algorithms 1 and 3 declines more quickly than in algorithm 2, suggesting that algorithm 2 will continue to outscale algorithms 1 and 3 at continually increasing system sizes. Algorithms 1 and

3 perform similarly for the dimer AP system, but as the system size increases the performance of algorithm 3 surpasses that of algorithm 1.

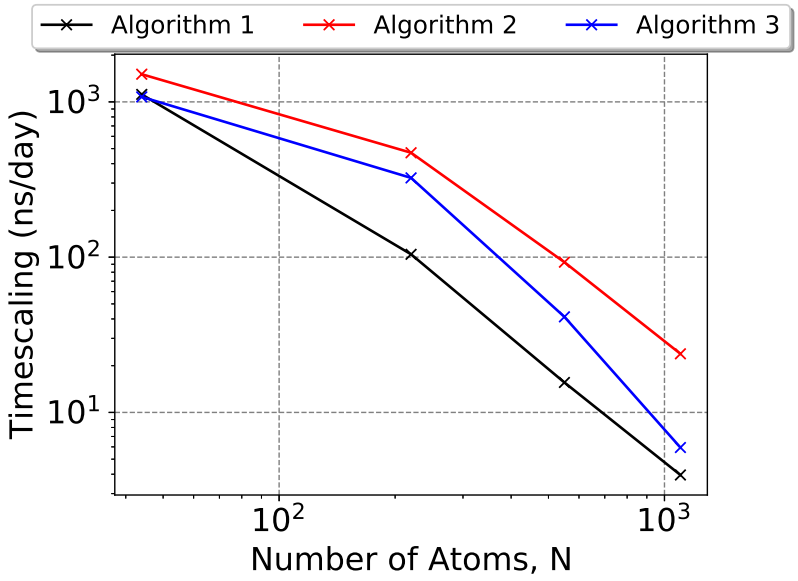

Figure 4.7: Timescaling as a function of the number of atoms, N, for $N_{MC} = 96$.

To determine how the performance of the three algorithms scale with the number of MC points per solute atom I performed timescaling simulations of AP systems with 2 and 50 AP molecules in chloroform with varying N_{MC} values. Figure 4.8(a) presents the performance of the dimer AP system with N_{MC} ranging from 8 to 256 MC points. Figure 4.8(b) shows the performance of the system with 50 AP molecules with N_{MC} ranging from 8 to 152 MC points. In both cases N_{MC} was increased incrementally by 8 MC points. For the dimer system, at low N_{MC} values algorithm 1 and 3 outperform algorithm 2. Once N_{MC} reaches a value of 40 MC points or higher algorithm 2 begins to outscale the other two, as the performance of algorithms 1 and 3 decreases sharply with increasing N_{MC}. Algorithms 1 and 3 scale similarly for the dimer system until N_{MC} is greater than 96 MC points at which algorithm 1 performs slightly better than algorithm 2. For the system of 50 AP molecules, algorithm 2 greatly outperforms algorithms 1 and 3 for all N_{MC} values, especially when N_{MC} is small. For small N_{MC}, algorithm 3 performs slightly better than algorithm 1, but the difference between them diminishes as N_{MC} increases. In general for both system sizes algorithm

2 performs significantly better than algorithm 1 and 3, which both perform similarly. Also, the performance of all three algorithms declines significantly as N_{MC} increases.

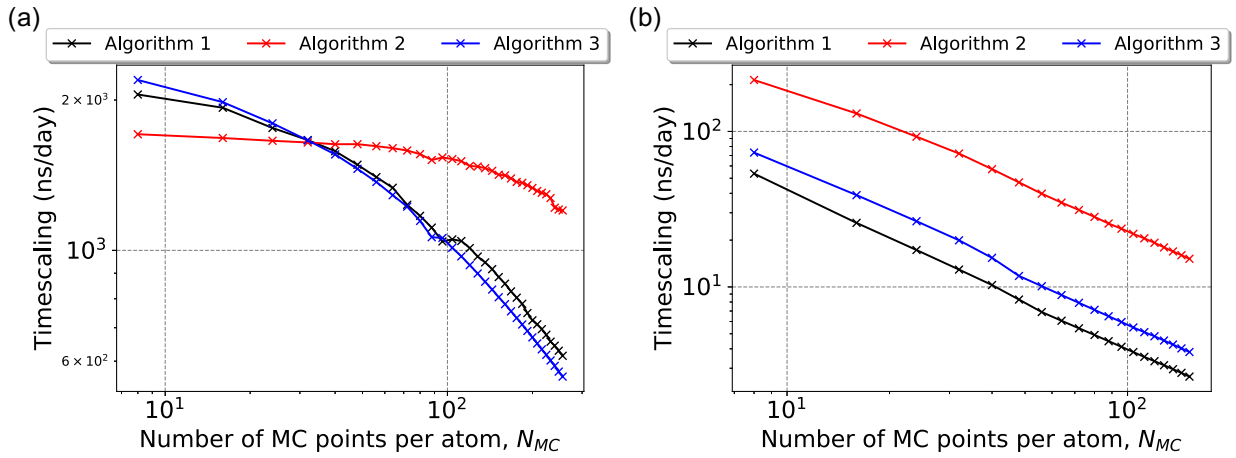

Figure 4.8: Timescaling as function of the number of MC points per atom, N_{MC}, for (a) $N = 44$ and (b) $N = 1100$.

These results demonstrate that fully parallelizing the calculation of the IS-SPA solvent forces is not optimal and that grouping things into tiles leads to better code performance. Although it is suggested that atomic operations can lead to a significant loss in performance, I show here that removing atomic operations at the expense of higher parallelism leads to lower performance of the code. Our results suggest that the code must strike a delicate balance between the use of expensive operations such as atomic operations and the level of parallelization of the calculations such as in the tiling method used by algorithm 2.

The current utility of IS-SPA lies in simulating systems of small molecules at low concentrations due to the current requirement of a fixed number of MC points per atom.[20] Table 4.1 shows the performance of explicit solvent aaMD of a system of 50 AP molecules in chloroform for a variety of concentrations and, therefore, system sizes performed on a GeForce GTX 1080 card using AMBER's CUDA MD code. It was demonstrated that algorithm 2 is the highest performing implementation of IS-SPA on a GPU for a system of 50 AP molecules, producing 24 ns of simulation a day. Comparison of the performance of the algorithm 2 implementation of IS-SPA to the performance of

the explicit solvent aaMD simulations reveals that IS-SPA begins to outperform explicit solvent simulation at concentrations below 25 mM. At concentrations below 25 mM the systems size becomes to large to be simulated efficiently. While this demonstrates where the usage of the current IS-SPA CUDA MD code becomes beneficial, the performance of the IS-SPA MD code can still be improved, further extending the utility of IS-SPA.

Explicit Solvent Performance		
Concentration (mM)	Number of Atoms	Performance (ns/day)
5	613920	9.02
10	301125	16.33
25	118665	39.78
50	57450	84.82
100	31765	156.16
500	7215	561.00
1000	4200	744.41

Table 4.1: The performance of explicit solvent aaMD simulations of a system of 50 AP molecules in chloroform for a variety of concentrations and system sizes. All simulations were performed using AMBER's highly optimized CUDA MD software on a GeForce GTX 1080 GPU card.

4.8 Conclusion

IS-SPA was developed to reduce the computational cost of simulating solvated systems by removing the solvent degrees of freedom while still capturing the force of the solvent on the solute. Until now, IS-SPA has only been implemented in a parallelizable CPU Fortran code. It has been shown that the parallel computing power of GPUs can provide significant speed up in the performance of MD simulations over CPU MD codes. Therefore, an implementation of IS-SPA in a GPU-capable CUDA code is needed to further improve the IS-SPA simulation performance. I present a full IS-SPA MD code with a focus on optimizing the calculation of the IS-SPA solvent forces. Three different algorithms are developed focusing on different methods to enhance performance. Based on the scaling of the performance of the three algorithms with the number of solute atoms

and the number of MC points per atom, algorithm 2, the tiling method, outperforms the other two algorithms as it strikes a balance between the use of atomic operations and parallelization of the IS-SPA calculations.

Future work will aim at further optimizing the IS-SPA CUDA MD code. There are at least three methods that could be utilized to improve code performance. The first is to further optimize the current version of the CUDA code. One way to enhance the performance of the code is to limit the use of atomic operations in a different manner than that done by algorithm 3 that still allows for a high level of parallelization of the IS-SPA calculations. For example, this was done in the tiling method used by AMBER in which they used buffers to avoid race conditions between tiles and, therefore, avoided the use of atomic operations.[169] Such a method could be implemented into algorithm 2 and may provide additional increase in computational performance of the IS-SPA CUDA code. A second way in which the IS-SPA MD code may be improved is to introduce a cutoff for the IS-SPA interactions, although, it is not clear if this could be implemented in a way to enhance code performance. For all three algorithms the performance decreases quickly with an increasing N_{MC} value. This suggests that third possible source of optimization is reducing the number of MC points per atom that is needed to sufficiently sample the IS-SPA solvent forces. In the current implementation of IS-SPA, 96 MC points are uniformly chosen in a spherical volume around a solute to estimate the integral in Equation 4.1. To reduce the value of N_{MC}, various sampling distributions of the MC points could be tested to find a distribution that reduces the variance of the mean force on a particular test molecule allowing for smaller values of N_{MC} to be used but still maintain a relatively low degree of error in solving the integral in Equation 4.1. This was done by Lake *et al.* in the implementation of IS-SPA for the nonpolar component of solvation, but has not been done for the extension of IS-SPA to polar solvents.[20,64]

Chapter 5

Conclusion

Amino acids have a tremendous chemical versatility and are the building blocks of peptides and proteins. Both peptides and proteins play a significant role in biological systems and have a wide array of important functions and applications. The previous three chapters have focused on the study of peptide systems, protein systems, and the development of a simulation method. The work that I report here has forwarded our understanding of the interplay of forces within $(RXDX)_4$ and coumarin-$(RXDX)_4$ β-sheet fibers, the role ATP-binding plays in the translocation mechanism of SARS-CoV-2 nsp13, and the development of an efficient GPU-capable IS-SPA MD code in CUDA.

5.1 Design of Switchable Optoelectronic Materials

Peptides are prime candidates for functional material design due their large chemical diversity and potential for switchable, self-assembling properties. For the application of peptide systems to optoelectronic biomaterials, other molecules with interesting optoelectronic properties are often incorporated into the peptides. To this end, Chapter 2 investigates the pairing of an unnatural coumarin amino acid with pH-switchable $(RXDX)_4$ β-sheet fibers elucidating the role hydrophobicity of the peptide sequence plays in fiber stability and pH-switchability. Specifically, I perform a set of aaMD simulations of $(RXDX)_4$ fibers where the hydrophobic residue, X, was mutated to increasingly hydrophobic amino acids (X=Ala,Val,Leu,Ile,Phe). From these simulations I find that increasing the hydrophobicity of the sequence leads to increased fiber stability, but an overall decrease in pH-switchability of the fiber. When these simulations are compared to a set of aaMD simulations of coumarin-$(RXDX)_4$ fibers I find that the addition of coumarin into the $(RXDX)_4$ sequence leads to less fiber stability and an increased curvature of the fiber structure, although, this effect is minimized by larger hydrophobic residues. Furthermore, increasing hydrophobicity in the coumarin-$(RXDX)_4$ conjugates leads to higher order and pH-switchability within the coumarin sidechains. This data

suggests that coumarin can be coupled with more hydrophobic versions of $(RXDX)_4$ β-sheet fibers for the design of switchable biomaterials with potentially interesting optoelectronic properties.

In this study I provide novel insight into the role hydrophobicity plays in the stability of $(RXDX)_4$ fibers and in the integration of an unnatural coumarin amino acid into these β-sheet structures. Beyond this study, there are still many questions that remain surrounding this system and other related peptide systems. Further work measuring the change in optical properties are vital from a material design perspective. Although this study shows clear changes in the layering and ordering of coumarin within the β-sheet fibers, it is still not apparent what impact this would have on the optoelectronic properties of the fiber. Computational approaches to address this question could utilize quantum mechanics (QM) calculations to compute UV-Vis or Fluorescence spectra or to analyze the electron transfer process within the fibers under various pH. Furthermore, spectroscopic techniques could be utilized to measure and verify the effect pH has on the optical properties of coumarin-$(RXDX)_4$ fibers.

Additionally, experiments that examine the effect of changing the molecule associated with $(RXDX)_4$ and other peptide sequences would shed light on the types of dye-peptide conjugates that would be ideal for optoelectronic biomaterial design. There are many studies available that look at the self-assembly of π-conjugated peptides, but in most of these studies the aromatic core molecule is quite large relative to the peptides and significantly control the self-assembling properties of the molecules.[2,8,9,14,31–44] As shown in the research presented in Chapter 2, even incorporating a small aromatic molecule, coumarin, has a significant effect on peptide self-assembly. Although, this can be limited by changing properties of the peptide such as the size and hydrophobicity of certain residues. An investigation of other unnatural amino acids with a small aromatic sidechain would be crucial from a material design perspective to find peptides which do not significantly alter the self-assembling properties as well as to tune the optoelectronic properties of the resulting nanostructures. A similar aaMD approach exploring pH and hydrophobic effects on fiber stability and switchability could be used to investigate a variety of small aromatic unnatural amino acids coupled to the $(RXDX)_4$ peptide.

In addition to investigating different aromatic molecules, studying various peptides would also be important for material design. In Chapter 2, the hydrophobic residues were mutated to tune the stability and switchability of the β-sheet fibers. One could consider a similar study where the charged residues could be mutated to other similarly charged residues to probe the effect of the distance between the charges and the β-sheet plane or the relative heights of the positive and negative charges on fiber stability and pH-switchability. For example the arginine and aspartic acid residues in $(RXDX)_4$ could be mutated to histidine and glutamic acid. These mutations would maintain the repeating positive and negative charges that are important for fiber stability and pH-switchability but would change the height of the charges relative to the plane of the β-sheet.

Along with investigating mutations of the $(RXDX)_4$ sequence, aromatic molecules such as coumarin could be coupled with other peptide sequences that form different types of nanostructures besides β-sheet fibers or are responsive to stimuli other than changes in pH. One example is the coupling of the unnatural coumarin amino acid with a tetratricopeptide repeat protein (TPR). TPR is one of the most studied repeat proteins along with the ankyrin repeat, leucine-rich repeat, and transcription activator-like proteins.[178] TPR consists of a 34 residue repeat and forms two antiparallel α-helices.[178,178–180] These helical motifs are modular and can stack on top of each other forming a superhelical structure. Several studies have investigated the formation of hydrogels, films, and fibrils using TPR proteins.[178–181] The formation and deformation of the hydrogel could be controlled by the ionic strength of the solution.[181] One could envision performing aaMD studies investigating the self-assembly of coumarin-TPR conjugates and their response to ionic strength. One interesting aspect of using the TPR protein in particular is its larger size relative to $(RXDX)_4$ which may allow for larger aromatic molecules to be incorporated without disrupting the self-assembling and switchable behavior of the peptide.

5.2 ATP-dependent Translocation Mechanism of SARS-CoV-2 Nsp13 Helicase

Viral helicase proteins play a critical role in viral replication making them ideal targets for the development of antiviral drugs. Due to the emergence of the COVID-19 global pandemic caused by SARS-CoV-2, understanding the mechanism by which the SARS-CoV-2 nsp13 helicase functions is of high importance. The ATP-dependent RNA translocation mechanism of nsp13 was previously unknown, although, SF1 helicases are thought to translocate by either an inchworm stepping mechanism or a Brownian ratchet mechanism. Chapter 3 characterizes the structure-function relationship utilized by nsp13 during ATP-dependent RNA translocation. In particular, through the comparison of simulations of the Apo, ATP, ssRNA, and ssRNA+ATP ligand bound states of nsp13, I gain insight into the translocation mechanism utilized by nsp13 and the role ATP-binding plays in this process. Using a GMM-LDA approach I find the presence of four states in the RNA-binding cleft which supports nsp13 utilizing an inchworm stepping translocation mechanism. In this mechanism motif **Ia** and motif **IV** alternate between strongly and weakly bound states, such that one site is always strongly bound to ssRNA while the other enters a weakly bound state and performs a power stroke before strongly rebinding ssRNA. The presence of ATP in the ATP-pocket changes the relative sampling of the four states such that motif **IV** enters into a weakly bound state with ssRNA. This suggests that the first step in the translocation cycle occurs due to the binding of ATP by nsp13. When the ATP-pocket of the 4 states are analyzed it is found that motif **V** is strongly interacting with ssRNA in states where motifs **Ia** and **IV** are largely separated. In the presence of ATP, motif **V** forms contacts with ATP reducing its interaction with RNA. The weakened interaction between motif **V** and ssRNA explains the reduction in the sampling of the states where motif **IV** is strongly bound to ssRNA when motif **Ia** and **IV** are largely separated.

In this study I provide novel insight into the ATP-dependent translocation mechanism of nsp13, yet, additional work must be conducted to fully understand it. Specifically, the later stages of the ATP-hydrolysis cycle, including ATP-hydrolysis, the release of the inorganic phosphate, and the

release of ADP must be analyzed to elucidate the ATP-dependence of the translocation mechanism. GaMD simulations of the ssRNA+ADP+Pi and ssRNA+ADP ligand bound states of nsp13 could be performed in combination with GMM clustering and characterization by linear discriminant analysis to find any new states sampled by the RNA-binding cleft in these systems. Furthermore, the populations of the four states already identified in the RNA-binding cleft could be measured in these new systems and would reveal information about the ATP-dependence of the inchworm stepping translocation mechanism during ATP-hydrolysis, release of the inorganic phosphate, and release of ADP. Furthermore, the interactions in the ATP-pocket which allosterically cause the change in states of the RNA-binding cleft could be elucidated in a similar fashion. Simulation and analysis of the ssRNA+ADP+Pi and ssRNA+ADP ligand bound states in combination with the results discussed in Chapter 3 would provide a deep understanding of the ATP-dependent translocation mechanism of nsp13. Furthermore, the knowledge gained about the translocation mechanism of nsp13 could be used to understand how other SF1 helicases translocate along RNA or DNA, due to the highly conserved nature of the translocation machinery of nsp13 between SF1 helicases.

Beyond understanding the translocation mechanism of nsp13, the role of specific key motifs could be elucidated further. One example of this is motif **V**; based on the simulations presented in Chapter 3, motif **V** plays a role in stabilizing specific states in the RNA-binding cleft. Performing enhanced sampling simulations such as umbrella sampling would allow us to better understand the change in the binding affinity of motif **V** to ssRNA when ATP is present. Also, umbrella sampling simulations could be used to determine to what extent motif **V** stabilizes the states where motif **Ia** and motif **IV** are largely separated. Similarly, enhanced sampling simulations could also be performed for other important motifs identified in the ssRNA+ADP+Pi and ssRNA+ADP ligand bound systems of nsp13.

In addition to understanding the translocation function of nsp13 it is important to discern its role in the RNA replication complex. Nsp13 is hypothesized to be a component of the RNA replication complex with the RNA polymerase, nsp8, nsp12, and other nsps[59,113,117–119] This inference is due, in part, to experiments demonstrating that nsp12 enhances the function of nsp13.[57,113] MD simulations

of nsp13 with other nsps, such as nsp12, would clarify its role in the RNA replication complex. The specific interactions between nsp12 and nsp13 could be elucidated to provide insight into how nsp12 enhances the function of nsp13.

After developing a more complete understanding of the ATP-dependent translocation mechanism of nsp13 and its role in the RNA replication complex, drug binding experiments targeting the key motifs utilized by nsp13 for RNA translocation could be performed. These experiments would identify molecules to be used as potential antiviral drugs that would target the inhibition of the nsp13 helicase function. Potential computational and experimental methods that could be utilized for these experiments are MD drug docking simulations and ligand binding assays, respectively.

5.3 Development and Optimization of a GPU-enable IS-SPA algorithm

Molecular dynamic simulations are a powerful tool for studying protein and peptide systems as it provides atomistic insight into the behavior of the systems for which experimental methods are inadequate. MD simulations are used to provide novel insight into the systems discussed in Chapter 2 and Chapter 3. Often, as encountered many times within my own research, a process of interest occurs on length- and time-scales outside of the current capabilities of conventional aaMD. For this reason, implicit solvent models, such as IS-SPA, are developed to reduce the computational cost of solvated systems by removing the solvent degrees of freedom. IS-SPA was originally implemented in a parallelizable CPU-capable Fortran MD code. To further optimize the efficiency of IS-SPA, an efficient GPU-enabled IS-SPA algorithm must be developed in CUDA. Accordingly, Chapter 4 discusses the development of an efficient IS-SPA algorithm in CUDA focusing on the optimization of calculating the IS-SPA solvent forces. Specifically, three algorithms are discussed including a highly-parallelized algorithm, a tiling method algorithm, and a modified tiling algorithm that removes the use of atomic operations by decreasing the level of parallelization in the code. Comparison of the performance of each algorithm and how they scale with the number of solute atoms and the number of MC points per solute atom reveals that a balance between the level of

parallelization of the IS-SPA force calculations and the use of expensive operations, such as atomic operations, must be met. An example of this is the tiling method algorithm which had the best performance for a majority of system sizes and choice of N_{MC}.

In this study I present and compare the performance of three novel algorithms designed to perform IS-SPA simulations on a GPU. Still the efficiency of the IS-SPA algorithm in CUDA could be improved further. As shown by the results in Chapter 4, the tiling method had the best performance of any algorithm I have developed so far. This algorithm was based on the tiling method used by AMBER and Friedrich et al. to calculate the non-bonded forces between atoms in GB implicit solvent simulations in which they suggest that the use of atomic operations are a serial bottleneck that would be unacceptable in terms of performance.[167–169] Therefore, it is a reasonable goal to modify the tiling algorithm used to calculate the IS-SPA solvent forces to avoid the use of atomic operations while still maintaining a high level of parallelization within the force and field kernels. One approach that could be utilized to this end would be to introduce output buffers to which each warp could write to its own global array removing the need for atomic operations. Following the calculation of the partial values by each tile, the partial values in the output buffers could be reduced to obtain the final total values. This is just one of many possible methods that could be tested to further enhance the efficiency of the IS-SPA algorithm as it is currently implemented.

Besides making improvements to the current implementation of the IS-SPA algorithm in CUDA, the performance of the code could be improved by modifying the distribution from which MC points are chosen. When using MC integration to solve an integral, such as the one in Equation 4.1, MC points can be chosen from any probability distribution as long as the domain contains the entire domain of the integrand function.[64] Sampling from a different probability distribution results in different efficiencies of converging the average. In the application of IS-SPA to nonpolar solvation, the variance of the mean solvent force for a fixed number of sample points were measured for three distributions. The distribution with the smallest variance, a parabolic distribution, was chosen to reduce the number of MC points required to converge the solvent forces.[64] A similar process could be performed for IS-SPA, applied to both polar and nonpolar solvation, to find a probability

distribution that more quickly converges the mean solvent force on each atom compared to the current uniform probability density selected from a spherical volume. If the value of N_{MC} could be reduced from the 100 MC points per solute atom as used by Lake et al. to less than 32 MC points, it would provide a 3-10x increase in performance of the current implementation of IS-SPA in CUDA.

5.4 A Broad Perspective

The research presented in this dissertation demonstrates the utility of MD simulation applied to biological processes. My research has provided a detailed understanding of how the self-assembling properties of RXDX peptides can be tuned to allow for aromatic molecules to be incorporated for biomaterial design and has laid the groundwork for understanding the ATP-dependent translocation mechanism of nsp13. With continually improving MD techniques and computational capabilities of hardware, the length- and time-scales that are computationally feasible continues to grow. This growth allows us to form better connections between the behavior of individual molecules and the macroscopic properties or larger functionality of a system. As system sizes and time-scales of MD simulations expand, so does the utility and insight that MD can provide for designing biomaterials with desirable properties and understanding the structure-function relationship of proteins. Precisely for this reason, part of my research focuses on the development of a GPU implementation of the IS-SPA implicit solvent model, working towards expanding the systems and biological processes approachable by MD simulations and narrowing the gap between the length- and time-scales accessible by MD and experimental techniques.

Bibliography

1. Fleming, S.; Ulijn, R. V. Design of nanostructures based on aromatic peptide amphiphiles. *Chem. Soc. Rev.* **2014**, *43*, 8150–8177.

2. Eakins, G. L.; Gallaher, J. K.; Keyzers, R. A.; Falber, A.; Webb, J. E. A.; Laos, A.; Tidhar, Y.; Weissman, H.; Rybtchinski, B.; Thordarson, P.; Hodgkiss, J. M. Thermodynamic factors impacting the peptide-driven self-assembly of perylene diimide nanofibers. *J. Phys. Chem. B* **2014**, *118*, 8642–8651.

3. Matson, J. B.; Zha, R. H.; Stupp, S. I. Peptide self-assembly for crafting functional biological materials. *Curr. Opin. Solid State Mater. Sci.* **2011**, *15*, 225–235.

4. Cui, H.; Webber, M. J.; Stupp, S. I. Self-assembly of peptide amphiphiles: from molecules to nanostructures to biomaterials. *Biopolymers* **2010**, *94*, 1–18.

5. Frederix, P. W.; Scott, G. G.; Abul-Haija, Y. M.; Kalafatovic, D.; Pappas, C. G.; Javid, N.; Hunt, N. T.; Ulijn, R. V.; Tuttle, T. Exploring the sequence space for (tri-)peptide self-assembly to design and discover new hydrogels. *Nat. Chem.* **2015**, *7*, 30–37.

6. Cavalli, S.; Albericio, F.; Kros, A. Amphiphilic peptides and their cross-disciplinary role as building blocks for nanoscience. *Chem. Soc. Rev.* **2010**, *39*, 241–263.

7. Lowik, D. W.; Leunissen, E. H. P.; Heuvel, M. V. D.; Hansen, M. B.; Hest, J. C.; Löwik, D. W. P. M.; Leunissen, E. H. P.; van den Heuvel, M.; Hansen, M. B.; van Hest, J. C. M. Stimulus responsive peptide based materials. *Chem. Soc. Rev.* **2010**, *39*, 3394–3412.

8. Katritzky, A. R.; Narindoshvili, T. Fluorescent amino acids: Advances in protein-extrinsic fluorophores. *Org. Biomol. Chem.* **2009**, *7*, 627–634.

9. Sameiro, M.; Gonçalves, T. Fluorescent labeling of biomolecules with organic probes. *Chem. Rev.* **2009**, *109*, 190–212.

10. Truex, N. L.; Nowick, J. S. Assembly of Peptides Derived from β-Sheet Regions of β-Amyloid. *J. Am. Chem. Soc.* **2016**, *138*, 13891–13900.

11. Hamley, I. W. Self-assembly of amphiphilic peptides. *Soft Matter* **2011**, *7*, 4122–4138.

12. Chow, D.; Nunalee, M. L.; Lim, D. W.; Simnick, A. J.; Chilkoti, A. Peptide-based biopolymers in biomedicine and biotechnology. *Mater. Sci. Eng. R Reports* **2008**, *62*, 125–155.

13. Kumar, K.; Lupoli, T. J. Exploiting Existing Molecular Scaffolds for Long-Term COVID Treatment. *ACS Med. Chem. Lett.* **2020**, *11*, 1357–1360.

14. Harkiss, A. H.; Sutherland, A. Recent advances in the synthesis and application of fluorescent α-amino acids. *Org. Biomol. Chem.* **2016**, *14*, 8911–8921.

15. Wang, J.; Xie, J.; Schultz, P. G. A genetically encoded fluorescent amino acid. *J. Am. Chem. Soc.* **2006**, *128*, 8738–8739.

16. Taylor, P. P.; Pantaleone, D. P.; Senkpeil, R. F.; Fotheringham, I. G. Novel biosynthetic approaches to the production of unnatural amino acids using transaminases. *Trends Biotechnol.* **1998**, *16*, 412–418.

17. Hollingsworth, S. A.; Dror, R. O. Molecular Dynamics Simulation for All. *Neuron* **2018**, *99*, 1129–1143.

18. Cheng, X.; Ivanov, I. Molecular dynamics. *Methods Mol. Biol.* **2012**, *929*, 243–285.

19. Davidson, R. B.; Hendrix, J.; Geiss, B. J.; McCullagh, M. RNA-Dependent Structures of the RNA-Binding Loop in the Flavivirus NS3 Helicase. *J. Phys. Chem. B* **2020**, *124*, 2371–2381.

20. Lake, P. T.; Mattson, M. A.; McCullagh, M. Implicit Solvation Using the Superposition Approximation (IS-SPA): Extension to Peptides in a Polar Solvent. *J. Chem. Theory Comput.* **2021**, *17*, 703–713.

21. Fairman, R.; Åkerfeldt, K. S. Peptides as novel smart materials. *Curr. Opin. Struct. Biol.* **2005**, *15*, 453–463.

22. Kuczer, M.; Konopińska, D.; Rosiński, G. Insect gonadotropic peptide hormones : some recent. *J. Pept. Sci.* **2007**, *14*, 16–26.

23. Bowerman, C. J.; Nilsson, B. L. Self-assembly of amphipathic β-sheet peptides: insights and applications. *Biopolymers* **2012**, *98*, 169–184.

24. Anderson, J.; Lake, P. T.; McCullagh, M. Initial Aggregation and Ordering Mechanism of Diphenylalanine from Microsecond All-Atom Molecular Dynamics Simulations. *J. Phys. Chem. B* **2018**, *122*, 12331–12341.

25. Sun, K.; Xiao, C.; Liu, C.; Fu, W.; Wang, Z.; Li, Z. Thermally Sensitive Self-Assembly of Glucose-Functionalized Tetrachloro-Perylene Bisimides: From Twisted Ribbons to Microplates. *Langmuir* **2014**, *30*, 11040–11045.

26. Ulijn, R. V.; Smith, A. M. Designing peptide based nanomaterials. *Chem. Soc. Rev.* **2008**, *37*, 664–675.

27. Paradís-Bas, M.; Tulla-Puche, J.; Zompra, A. A.; Albericio, F. RADA-16: A tough peptide - Strategies for synthesis and purification. *European J. Org. Chem.* **2013**, *2013*, 5871–5878.

28. Bagrov, D.; Gazizova, Y.; Podgorsky, V.; Udovichenko, I.; Danilkovich, A.; Prusakov, K.; Klinov, D. Morphology and aggregation of RADA-16-I peptide studied by AFM, NMR and molecular dynamics simulations. *Biopolymers* **2016**, *106*, 72–81.

29. Wang, T.; Zhong, X.; Wang, S.; Lv, F.; Zhao, X. Molecular mechanisms of RADA16-1 peptide on fast stop bleeding in rat models. *Int. J. Mol. Sci.* **2012**, *13*, 15279–15290.

30. Arosio, P.; Owczarz, M.; Wu, H.; Butté, A.; Morbidelli, M. End-to-end self-assembly of RADA 16-I nanofibrils in aqueous solutions. *Biophys. J.* **2012**, *102*, 1617–1626.

31. Mansbach, R. A.; Ferguson, A. L. Control of the hierarchical assembly of π-conjugated optoelectronic peptides by pH and flow. *Org. Biomol. Chem.* **2017**, *15*, 5484–5502.

32. Ardona, H. A. M.; Tovar, J. D. Peptide π-electron conjugates: organic electronics for biology? *Bioconjug. Chem.* **2015**, *26*, 2290–2302.

33. Wall, B. D.; Zacca, A. E.; Sanders, A. M.; Wilson, W. L.; Ferguson, A. L.; Tovar, J. D. Supramolecular polymorphism: Tunable electronic interactions within π-conjugated peptide nanostructures dictated by primary amino acid sequence. *Langmuir* **2014**, *30*, 5946–5956.

34. Chen, L.; Morris, K.; Laybourn, A.; Elias, D.; Hicks, M. R.; Rodger, A.; Serpell, L.; Adams, D. J. Self-assembly mechanism for a naphthalene-dipeptide leading to hydrogelation. *Langmuir* **2010**, *26*, 5232–5242.

35. Pepe-Mooney, B. J.; Fairman, R. Peptides as materials. *Curr. Opin. Struct. Biol.* **2009**, *19*, 483–494.

36. Datar, A.; Balakrishnan, K.; Zang, L. One-dimensional self-assembly of a water soluble perylene diimide molecule by pH triggered hydrogelation. *Chem. Commun.* **2013**, *49*, 6894–6896.

37. Zhou, Y.; Li, B.; Li, S.; Ardona, H. A. M.; Wilson, W. L.; Tovar, J. D.; Schroeder, C. M. Concentration-Driven Assembly and Sol-Gel Transition of π-Conjugated Oligopeptides. *ACS Cent. Sci.* **2017**, *3*, 986–994.

38. Chen, S.; Slattum, P.; Wang, C.; Zang, L. Self-Aassembly of perylene imide molecules into 1D nanostructures: methods, morphologies, and applications. *Chem. Rev.* **2015**, *115*, 11967–11998.

39. Bai, S.; Debnath, S.; Javid, N.; Frederix, P. W. J. M.; Fleming, S.; Pappas, C.; Ulijn, R. V. Differential self-assembly and tunable emission of aromatic peptide bola -amphiphiles containing perylene bisimide in polar solvents including water. *Langmuir* **2014**, *30*, 7576–7584.

40. Castilla, A. M.; Draper, E. R.; Nolan, M. C.; Brasnett, C.; Seddon, A.; Mears, L. L.; Cowieson, N.; Adams, D. J. Self-sorted oligophenylvinylene and perylene bisimide hydrogels. *Sci. Rep.* **2017**, *7*, 8380–8390.

41. Lemouchi, C.; Simonov, S.; Zorina, L.; Gautier, C.; Hudhomme, P.; Batail, P. Amino acid derivatives of perylenediimide and their N-H...O peptide bond dipoles-templated solid state assembly into stacks. *Org. Biomol. Chem.* **2011**, *9*, 8096–8101.

42. Eakins, G. L.; Pandey, R.; Wojciechowski, J. P.; Zheng, H. Y.; Webb, J. E. A.; Valéry, C.; Thordarson, P.; Plank, N. O. V.; Gerrard, J. A.; Hodgkiss, J. M. Functional organic semiconductors assembled via natural aggregating peptides. *Adv. Funct. Mater.* **2015**, *25*, 5640–5649.

43. Walsh, J. J.; Lee, J. R.; Draper, E. R.; King, S. M.; Jäckel, F.; Zwijnenburg, M. A.; Adams, D. J.; Cowan, A. J. Controlling Visible Light Driven Photoconductivity in Self-Assembled Perylene Bisimide Structures. *J. Phys. Chem. C* **2016**, *120*, 18479–18486.

44. Seo, H. J.; Kim, J. C. Self-assembly of coumarin-conjugated acidic proteinoids1. *Polym. Sci. - Ser. A* **2012**, *54*, 358–363.

45. Agarwal, P. K. Enzymes: An integrated view of structure, dynamics and function. *Microb. Cell Fact.* **2006**, *5*, 1–12.

46. Bolon, D. N.; Mayo, S. L. Enzyme-like proteins by computational design. *Proc. Natl. Acad. Sci. U. S. A.* **2001**, *98*, 14274–14279.

47. Votaw, K. A.; McCullagh, M. Characterization of the Search Complex and Recognition Mechanism of the AlkD-DNA Glycosylase. *J. Phys. Chem. B* **2019**, *123*, 95–105.

48. Hediger, M. A.; Romero, M. F.; Peng, J. B.; Rolfs, A.; Takanaga, H.; Bruford, E. A. The ABCs of solute carriers: Physiological, pathological and therapeutic implications of human membrane transport proteins. *Pflugers Arch. Eur. J. Physiol.* **2004**, *447*, 465–468.

49. Du Pont, K. E.; Davidson, R. B.; McCullagh, M.; Geiss, B. J. Motif v regulates energy transduction between the flavivirus NS3 ATPase and RNA-binding cleft. *J. Biol. Chem.* **2020**, *295*, 1551–1564.

50. Gilhooly, N. S.; Gwynn, E. J.; Dillingham, M. S. Superfamily 1 helicases. *Front. Biosci. - Sch.* **2013**, *5 S*, 206–216.

51. Singleton, M. R.; Dillingham, M. S.; Wigley, D. B. Structure and mechanism of helicases and nucleic acid translocases. *Annu. Rev. Biochem.* **2007**, *76*, 23–50.

52. Byrd, A. K.; Raney, K. D. Superfamily 2 helicases. *Front. Biosci.* **2012**, *17*, 2070–2088.

53. Fairman-Williams, M. E.; Guenther, U. P.; Jankowsky, E. SF1 and SF2 helicases: Family matters. *Curr. Opin. Struct. Biol.* **2010**, *20*, 313–324.

54. Raney, K. D.; Byrd, A. K.; Aarattuthodiyil, S. Structure and mechanisms of SF1 DNA helicases. *Adv. Exp. Med. Biol.* **2013**, *767*, 17–46.

55. Adam, M. K.; Jarrett-Wilkins, C.; Beards, M.; Staykov, E.; MacFarlane, L. R.; Bell, T. D.; Matthews, J. M.; Manners, I.; Faul, C. F.; Moens, P. D.; Ben, R. N.; Wilkinson, B. L. 1D Self-Assembly and Ice Recrystallization Inhibition Activity of Antifreeze Glycopeptide-Functionalized Perylene Bisimides. *Chem. - A Eur. J.* **2018**, *24*, 7834–7839.

56. WHO COVID-19 Explorer. Geneva: World Health Organization. 2020; https://worldhealthorg.shinyapps.io/covid/.

57. Jia, Z.; Yan, L.; Ren, Z.; Wu, L.; Wang, J.; Guo, J.; Zheng, L.; Ming, Z.; Zhang, L.; Lou, Z.; Rao, Z. Delicate structural coordination of the Severe Acute Respiratory Syndrome coronavirus Nsp13 upon ATP hydrolysis. *Nucleic Acids Res.* **2019**, *47*, 6538–6550.

58. Ugurel, O. M.; Mutlu, O.; Sariyer, E.; Kocer, S.; Ugurel, E.; Inci, T. G.; Ata, O.; Turgut-Balik, D. Evaluation of the potency of FDA-approved drugs on wild type and mutant SARS-CoV-2 helicase (Nsp13). *Int. J. Biol. Macromol.* **2020**, *163*, 1687–1696.

59. Romano, M.; Ruggiero, A.; Squeglia, F.; Maga, G.; Berisio, R. A Structural View of SARS-CoV-2 RNA Replication Machinery: RNA Synthesis, Proofreading and Final Capping. *Cells* **2020**, *9*.

60. Ivanov, K. A.; Thiel, V.; Dobbe, J. C.; van der Meer, Y.; Snijder, E. J.; Ziebuhr, J. Multiple Enzymatic Activities Associated with Severe Acute Respiratory Syndrome Coronavirus Helicase. *J. Virol.* **2004**, *78*, 5619–5632.

61. Mirza, M. U.; Froeyen, M. Structural elucidation of SARS-CoV-2 vital proteins: Computational methods reveal potential drug candidates against main protease, Nsp12 polymerase and Nsp13 helicase. *J. Pharm. Anal.* **2020**, *10*, 320–328.

62. Orozco, M.; Luque, F. J. Theoretical methods for the description of the solvent effect in biomolecular systems. *Chem. Rev.* **2000**, *100*, 4187–4225.

63. Zhang, J.; Zhang, H.; Wu, T.; Wang, Q.; Van Der Spoel, D. Comparison of Implicit and Explicit Solvent Models for the Calculation of Solvation Free Energy in Organic Solvents. *J. Chem. Theory Comput.* **2017**, *13*, 1034–1043.

64. Lake, P. T.; Mccullagh, M. Implicit Solvation Using the Superposition Approximation (ISSPA): An Implicit Treatment of the Nonpolar Component to Solvation for Simulating Molecular Aggregation. *J. Chem. Theory Comput.* **2017**, *13*, 5911–5924.

65. Onufriev, A.; Bashford, D.; Case, D. A. Modification of the generalized born model suitable for macromolecules. *J. Phys. Chem. B* **2000**, *104*, 3712–3720.

66. Beglov, D.; Roux, B. Solvation of complex molecules in a polar liquid: An integral equation theory. *J. Chem. Phys.* **1996**, *104*, 8678–8689.

67. Chandler, D.; Andersen, H. C. Optimized cluster expansions for classical fluids. II. Theory of molecular liquids. *J. Chem. Phys.* **1972**, *57*, 1918–1929.

68. Bian, L.; Zhu, E.; Tang, J.; Tang, W.; Zhang, F. Recent progress in the design of narrow bandgap conjugated polymers for high-efficiency organic solar cells. *Prog. Polym. Sci.* **2012**, *37*, 1292–1331.

69. Hoeben, F. J.; Jonkheijm, P.; Meijer, E. W.; Schenning, A. P. About supramolecular assemblies of π-conjugated systems. *Chem. Rev.* **2005**, *105*, 1491–1546.

70. Vadehra, G. S.; Wall, B. D.; Diegelmann, S. R.; Tovar, J. D. On-resin dimerization incorporates a diverse array of π-conjugated functionality within aqueous self-assembling peptide backbones. *Chem. Commun.* **2010**, *46*, 3947–3949.

71. Kim, S. H.; Parquette, J. R. A model for the controlled assembly of semiconductor peptides. *Nanoscale* **2012**, *4*, 6940–6947.

72. Wall, B. D.; Tovar, J. D. Synthesis and characterization of p-conjugated peptide-based supramolecular materials. *Pure Appl. Chem.* **2012**, *84*, 1039–1045.

73. Guo, X.; Baumgarten, M.; Müllen, K. Designing π-conjugated polymers for organic electronics. *Prog. Polym. Sci.* **2013**, *38*, 1832–1908.

74. Zelzer, M.; Ulijn, R. V. Next-generation peptide nanomaterials: molecular networks, interfaces and supramolecular functionality. *Chem. Soc. Rev.* **2010**, *39*, 3351–3357.

75. Lampel, A.; Ulijn, R. V.; Tuttle, T. Guiding principles for peptide nanotechnology through directed discovery. *Chem. Soc. Rev.* **2018**, *47*, 3737–3758.

76. Zhang, S. Emerging biological materials through molecular self-assembly. *Biotechnol. Adv.* **2002**, *20*, 321–339.

77. Scheibel, T.; Parthasarathy, R.; Sawicki, G.; Lin, X.-M.; Jaeger, H.; Lindquist, S. L. Conducting nanowires built by controlled self-assembly of amyloid fibers and selective metal deposition. *Proc. Natl. Acad. Sci. U. S. A.* **2003**, *100*, 4527–32.

78. Vauthey, S.; Santoso, S.; Gong, H.; Watson, N.; Zhang, S. Molecular self-assembly of surfactant-like peptides to form nanotubes and nanovesicles. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 5355–5360.

79. Burgess, N. C.; Sharp, T. H.; Thomas, F.; Wood, C. W.; Thomson, A. R.; Zaccai, N. R.; Brady, R. L.; Serpell, L. C.; Woolfson, D. N. Modular design of self-assembling peptide-based nanotubes. *J. Am. Chem. Soc.* **2015**, *137*, 10554–10562.

80. Ma, Y.; Zhang, F.; Zhang, J.; Jiang, T. A water-soluble perylene derivative for live-cell imaging. *Turkish J. Chem.* **2015**, *39*, 835–842.

81. Apostolovic, B.; Danial, M.; Klok, H.-A. A. Coiled coils: attractive protein folding motifs for the fabrication of self-assembled, responsive and bioactive materials. *Chem. Soc. Rev.* **2010**, *39*, 3536–3541.

82. Koutsopoulos, S.; Unsworth, L. D.; Nagai, Y.; Zhang, S. Controlled release of functional proteins through designer self-assembling peptide nanofiber hydrogel scaffold. *Proc. Natl. Acad. Sci.* **2009**, *106*, 4623–4628.

83. Zhang, S.; Altman, M. Peptide self-assembly in functional polymer science and engineering. *React. Funct. Polym.* **1999**, *41*, 91–102.

84. Yokoi, H.; Kinoshita, T.; Zhang, S. Dynamic reassembly of peptide RADA16 nanofiber scaffold. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 8414–8419.

85. Ye, Z.; Zhang, H.; Luo, H.; Wang, S.; Zhou, Q.; Du, X.; Tang, C.; Chen, L.; Liu, J.; Shi, Y. K.; Zhang, E. Y.; Ellis-Behnke, R.; Zhao, X. Temperature and pH effects on biophysical and morphological properties of self-assembling peptide RADA 16-1. *J. Pept. Sci.* **2008**, *14*, 152–162.

86. Sun, Y.; Zhang, Y.; Tian, L.; Zhao, Y.; Wu, D.; Xue, W.; Ramakrishna, S.; Wu, W.; He, L. Self-assembly behaviors of molecular designer functional RADA16-I peptides: Influence of motifs, pH, and assembly time. *Biomed. Mater.* **2017**, *12*, 015007.

87. Tsutsumi, H.; Mihara, H. Soft materials based on designed self-assembling peptides: from design to application. *Mol. Biosyst.* **2013**, *9*, 609–617.

88. Hartgerink, J. D.; Beniash, E.; Stupp, S. I. Self-Assembly and Mineralization of Peptide-Amphiphile Nanofibers. *Science (80-.).* **2001**, *294*, 1684–1688.

89. Stendahl, J. C.; Rao, M. S.; Guler, M. O.; Stupp, S. I. Intermolecular forces in the self-assembly of peptide amphiphile nanofibers. *Adv. Funct. Mater.* **2006**, *16*, 499–508.

90. Greenfield, M. A.; Hoffman, J. R.; De La Cruz, M. O.; Stupp, S. I. Tunable mechanics of peptide nanofiber gels. *Langmuir* **2010**, *26*, 3641–3647.

91. Ruff, Y.; Moyer, T.; Newcomb, C. J.; Demeler, B.; Stupp, S. I. Precision templating with DNA of a virus-like particle with peptide nanostructures. *J. Am. Chem. Soc.* **2013**, *135*, 6211–6219.

92. Shao, H.; Parquette, J. R. A π-conjugated hydrogel based on an Fmoc-dipeptide naphthalene diimide semiconductor. *Chem. Commun.* **2010**, *46*, 4285–4287.

93. Doran, T. M.; Kamens, A. J.; Byrnes, N. K.; Nilsson, B. L. Role of amino acid hydrophobicity, aromaticity, and molecular volume on IAPP(20-29) amyloid self-assembly. *Proteins Struct. Funct. Bioinforma.* **2012**, *80*, 1053–1065.

94. Bowerman, C. J.; Ryan, D. M.; Nissan, D. A.; Nilsson, B. L. The effect of increasing hydrophobicity on the self-assembly of amphipathic β-sheet peptides. *Mol. Biosyst.* **2009**, *5*, 1058–1069.

95. Lakshmanan, A.; Cheong, D. W.; Accardo, A.; Di Fabrizio, E.; Riekel, C.; Hauser, C. A. Aliphatic peptides show similar self-assembly to amyloid core sequences, challenging the importance of aromatic interactions in amyloidosis. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, 519–524.

96. Bowerman, C. J.; Liyanage, W.; Federation, A. J.; Nilsson, B. L. Tuning β-sheet peptide self-assembly and hydrogelation behavior by modification of sequence hydrophobicity and aromaticity. *Biomacromolecules* **2011**, *12*, 2735–2745.

97. Betush, R. J.; Urban, J. M.; Nilsson, B. L. Balancing hydrophobicity and sequence pattern to influence self-assembly of amphipathic peptides. *Pept. Sci.* **2018**, *110*, e23099.

98. Fauchere, J.-L.-L.; Charton, M.; Kier, L. B.; Verloop, A.; Pliska, V. Amino acid side chain parameters for correlation studies in biology and pharmacology. *Int. J. Pept. Protein Res.* **1988**, *32*, 269–278.

99. Case, D.; Betz, R.; Cerutti, D.; Cheatham III, T.; Darden, T.; Duke, R.; Giese, T.; Gohlke, H.; Goetz, A.; Homeyer, N.; Izadi, S.; Janowski, P.; Kaus, J.; Kovalenko, A.; Lee, T.; LeGrand, S.; Li, P.; Lin, C.; Luchko, T.; Luo, R.; Madej, B.; Mermelstein, D.; Merz, K.; Monard, G.; Nguyen, H.; Nguyen, H.; Omelyan, I.; Onufriev, A.; Roe, D.; Roitberg, A.; Sagui, C.; Simmerling, C.; Botello-Smith, W.; Swails, J.; Walker, R.; Wang, J.; Wolf, R.; Wu, X.; Xiao, L.; Kollman, P. AMBER 2018. 2016.

100. Cormier, A. R.; Pang, X.; Zimmerman, M. I.; Zhou, H. X.; Paravastu, A. K. Molecular structure of RADA16-I designer self-assembling peptide nanofibers. *ACS Nano* **2013**, *7*, 7562–7572.

101. Cerutti, D. S.; Swope, W. C.; Rice, J. E.; Case, D. A. Ff14ipq: A self-consistent force field for condensed-phase simulations of proteins. *J. Chem. Theory Comput.* **2014**, *10*, 4515–4534.

102. Debiec, K. T.; Cerutti, D. S.; Baker, L. R.; Gronenborn, A. M.; Case, D. A.; Chong, L. T. Further along the road less traveled: AMBER ff15ipq, an original protein force field built on a self-consistent physical model. *J. Chem. Theory Comput.* **2016**, *12*, 3926–3947.

103. Chai, J. D.; Head-Gordon, M. Long-range corrected hybrid density functionals with damped atom-atom dispersion corrections. *Phys. Chem. Chem. Phys.* **2008**, *10*, 6615–6620.

104. Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys.* **2010**, *132*, 154104.

105. Kendall, R. A.; Dunning, T. H.; Harrison, R. J. Electron affinities of the first-row atoms revisited. Systematic basis sets and wave functions. *J. Chem. Phys.* **1992**, *96*, 6796–6806.

106. Thiede, E. H.; Van Koten, B.; Weare, J.; Dinner, A. R. Eigenvector method for umbrella sampling enables error analysis. *J. Chem. Phys.* **2016**, *145*, 084115.

107. Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general Amber force field. *J. Comput. Chem.* **2004**, *25*, 1157–1174.

108. Yalkowsk, S. H.; He, Y. *Handb. Aqueous Solubility Data*, 2nd ed.; Taylor and Francis Group, LLC: Boca Raton, FL, 2003; pp 1–1496.

109. Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* **1977**, *23*, 327–341.

110. Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98*, 10089–10092.

111. Kabsch, W.; Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983**, *22*, 2577–2637.

112. Fujiwara, K.; Ebisawa, S.; Watanabe, Y.; Toda, H.; Ikeguchi, M. Local sequence of protein β-strands influences twist and bend angles. *Proteins Struct. Funct. Bioinforma.* **2014**, *82*, 1484–1493.

113. Adedeji, A. O.; Marchand, B.; te Velthuis, A. J. W.; Snijder, E. J.; Weiss, S.; Eoff, R. L.; Singh, K.; Sarafianos, S. G. Mechanism of nucleic acid unwinding by SARS-CoV helicase. *PLoS One* **2012**, *7*, e36521.

114. Kim, M. K.; Yu, M. S.; Park, H. R.; Kim, K. B.; Lee, C.; Cho, S. Y.; Kang, J.; Yoon, H.; Kim, D. E.; Choo, H.; Jeong, Y. J.; Chong, Y. 2,6-Bis-arylmethyloxy-5-hydroxychromones with antiviral activity against both hepatitis C virus (HCV) and SARS-associated coronavirus (SCV). *Eur. J. Med. Chem.* **2011**, *46*, 5698–5704.

115. Adedeji, A. O.; Singh, K.; Calcaterra, N. E.; DeDiego, M. L.; Enjuanes, L.; Weiss, S.; Sarafianos, S. G. Severe acute respiratory syndrome coronavirus replication inhibitor that interferes with the nucleic acid unwinding of the viral helicase. *Antimicrob. Agents Chemother.* **2012**, *56*, 4718–4728.

116. Ndjomou, J.; Corby, M. J.; Sweeney, N. L.; Hanson, A. M.; Aydin, C.; Ali, A.; Schiffer, C. A.; Li, K.; Frankowski, K. J.; Schoenen, F. J.; Frick, D. N. Simultaneously Targeting the NS3 Protease and Helicase Activities for More Effective Hepatitis C Virus Therapy. *ACS Chem. Biol.* **2015**, *10*, 1887–1896.

117. Yan, L.; Zhang, Y.; Ge, J.; Zheng, L.; Gao, Y.; Wang, T.; Jia, Z.; Wang, H.; Huang, Y.; Li, M.; Wang, Q.; Rao, Z.; Lou, Z. Architecture of a SARS-CoV-2 mini replication and transcription complex. *Nat. Commun.* **2020**, *11*, 5874.

118. Yan, L.; Ge, J.; Zheng, L.; Zhang, Y.; Gao, Y.; Wang, T.; Huang, Y.; Yang, Y.; Gao, S.; Li, M.; Liu, Z.; Wang, H.; Li, Y.; Chen, Y.; Guddat, L. W.; Wang, Q.; Rao, Z.; Lou, Z. Cryo-EM Structure of an Extended SARS-CoV-2 Replication and Transcription Complex Reveals an Intermediate State in Cap Synthesis. *Cell* **2021**, *184*, 184–193.e10.

119. Chen, J.; Malone, B.; Llewellyn, E.; Grasso, M.; Shelton, P. M.; Olinares, P. D. B.; Maruthi, K.; Eng, E. T.; Vatandaslar, H.; Chait, B. T.; Kapoor, T. M.; Darst, S. A.; Campbell, E. A. Structural Basis for Helicase-Polymerase Coupling in the SARS-CoV-2 Replication-Transcription Complex. *Cell* **2020**, *182*, 1560–1573.e13.

120. Yuen, C. K.; Lam, J. Y.; Wong, W. M.; Mak, L. F.; Wang, X.; Chu, H.; Cai, J. P.; Jin, D. Y.; To, K. K. W.; Chan, J. F. W.; Yuen, K. Y.; Kok, K. H. SARS-CoV-2 nsp13, nsp14, nsp15 and orf6 function as potent interferon antagonists. *Emerg. Microbes Infect.* **2020**, *9*, 1418–1428.

121. Adedeji, A. O.; Singh, K.; Kassim, A.; Coleman, C. M.; Elliott, R.; Weiss, S. R.; Frieman, M. B.; Sarafianos, S. G. Evaluation of SSYA10-001 as a replication inhibitor of severe acute respiratory syndrome, mouse hepatitis, and Middle East respiratory syndrome coronaviruses. *Antimicrob. Agents Chemother.* **2014**, *58*, 4894–4898.

122. Kadaré, G.; Haenni, A. L. Virus-encoded RNA helicases. *J. Virol.* **1997**, *71*, 2583–2590.

123. Borowski, P.; Niebuhr, A.; Schmitz, H.; Hosmane, R. S.; Bretner, M.; Siwecka, M. A.; Kulikowski, T. NTPase/helicase of Flaviviridae: Inhibitors and inhibition of the enzyme. *Acta Biochim. Pol.* **2002**, *49*, 597–614.

124. Raney, K. D.; Sharma, S. D.; Moustafa, I. M.; Cameron, C. E. Hepatitis C virus non-structural protein 3 (HCV NS3): A multifunctional antiviral target. *J. Biol. Chem.* **2010**, *285*, 22725–22731.

125. Leung, D.; Schroder, K.; White, H.; Fang, N. X.; Stoermer, M. J.; Abbenante, G.; Martin, J. L.; Young, P. R.; Fairlie, D. P. Activity of Recombinant Dengue 2 Virus NS3 Protease in the Presence of a Truncated NS2B Co-factor, Small Peptide Substrates, and Inhibitors. *J. Biol. Chem.* **2001**, *276*, 45762–45771.

126. Byrd, C. M.; Grosenbach, D. W.; Berhanu, A.; Dai, D.; Jones, K. F.; Cardwell, K. B.; Schneider, C.; Yang, G.; Tyavanagimatt, S.; Harver, C.; Wineinger, K. A.; Page, J.; Stavale, E.; Stone, M. A.; Fuller, K. P.; Lovejoy, C.; Leeds, J. M.; Hruby, D. E.; Jordan, R. Novel benzoxazole inhibitor of dengue virus replication that targets the NS3 helicase. *Antimicrob. Agents Chemother.* **2013**, *57*, 1902–1912.

127. Sweeney, N. L.; Hanson, A. M.; Mukherjee, S.; Ndjomou, J.; Geiss, B. J.; Steel, J. J.; Frankowski, K. J.; Li, K.; Schoenen, F. J.; Frick, D. N. Benzothiazole and Pyrrolone Flavivirus Inhibitors Targeting the Viral Helicase. *ACS Infect. Dis.* **2015**, *1*, 140–148.

128. Lee, H.; Ren, J.; Nocadello, S.; Rice, A. J.; Ojeda, I.; Light, S.; Minasov, G.; Vargas, J.; Nagarathnam, D.; Anderson, W. F.; Johnson, M. E. Identification of novel small molecule inhibitors against NS2B/NS3 serine protease from Zika virus. *Antiviral Res.* **2017**, *139*, 49–58.

129. Mastrangelo, E.; Pezzullo, M.; De burghgraeve, T.; Kaptein, S.; Pastorino, B.; Dallmeier, K.; De lamballerie, X.; Neyts, J.; Hanson, A. M.; Frick, D. N.; Bolognesi, M.; Milani, M. Ivermectin is a potent inhibitor of flavivirus replication specifically targeting NS3 helicase activity: New prospects for an old drug. *J. Antimicrob. Chemother.* **2012**, *67*, 1884–1894.

130. Basavannacharya, C.; Vasudevan, S. G. Suramin inhibits helicase activity of NS3 protein of dengue virus in a fluorescence-based high throughput assay format. *Biochem. Biophys. Res. Commun.* **2014**, *453*, 539–544.

131. Shadrick, W. R.; Mukherjee, S.; Hanson, A. M.; Sweeney, N. L.; Frick, D. N. Aurintricarboxylic acid modulates the affinity of hepatitis C virus NS3 helicase for both nucleic acid and ATP. *Biochemistry* **2013**, *52*, 6151–6159.

132. Drummer, H. E.; Boo, I.; Maerz, A. L.; Poumbourios, P. A Conserved Gly436-Trp-Leu-Ala-Gly-Leu-Phe-Tyr Motif in Hepatitis C Virus Glycoprotein E2 Is a Determinant of CD81 Binding and Viral Entry. *J. Virol.* **2006**, *80*, 7844–7853.

133. Jankowsky, E.; Fairman-Williams, M. E. In *RSC Biomol. Sci.*; Jankowsky, E., Ed.; Royal Society of Chemistry Publishing: Cambridge (UK), 2010; Chapter 1, pp 1–31.

134. Ranji, A.; Boris-Lawrie, K. RNA helicases: Emerging roles in viral replication and the host innate response. *RNA Biol.* **2010**, *7*, 775–787.

135. Pyle, A. M. RNA helicases and remodeling proteins. *Curr. Opin. Chem. Biol.* **2011**, *15*, 636–642.

136. Seybert, A.; Hegyi, A.; Siddell, S. G.; Ziebuhr, J. The human coronavirus 229E superfamily 1 helicase has RNA and DNA duplex-unwinding activities with 5'-to-3' polarity. *RNA* **2000**, *6*, 1056–1068.

137. Marchat, L. A.; Arzola-Rodríguez, S. I.; Hernandez-de la Cruz, O.; Lopez-Rosas, I.; Lopez-Camarillo, C. DEAD/DExH-Box RNA helicases in selected human parasites. *Korean J. Parasitol.* **2015**, *53*, 583–595.

138. Patel, S. S.; Donmez, I. Mechanisms of helicases. *J. Biol. Chem.* **2006**, *281*, 18265–18268.

139. Mickolajczyk, K. J.; Shelton, P. M.; Grasso, M.; Cao, X.; Warrington, S. E.; Aher, A.; Liu, S.; Kapoor, T. M. Force-dependent stimulation of RNA unwinding by SARS-CoV-2 nsp13 helicase. *Biophys. J.* **2021**, *120*, 1020–1030.

140. Miao, Y.; McCammon, J. A. Gaussian Accelerated Molecular Dynamics: Theory, Implementation, and Applications. *Annu. Rep. Comput. Chem.* **2017**, *13*, 231–278.

141. Chakrabarti, S.; Jayachandran, U.; Bonneau, F.; Fiorini, F.; Basquin, C.; Domcke, S.; Le Hir, H.; Conti, E. Molecular Mechanisms for the RNA-Dependent ATPase Activity of Upf1 and Its Regulation by Upf2. *Mol. Cell* **2011**, *41*, 693–703.

142. Theobald, D. L.; Wuttke, D. S. THESEUS: Maximum likelihood superpositioning and analysis of macromolecular structures. *Bioinformatics* **2006**, *22*, 2171–2172.

143. Cheng, Z.; Muhlrad, D.; Lim, M. K.; Parker, R.; Song, H. Structural and functional insights into the human Upf1 helicase core. *EMBO J.* **2007**, *26*, 253–264.

144. Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **2015**, *11*, 3696–3713.

145. Banáš, P.; Hollas, D.; Zgarbová, M.; Jurečka, P.; Orozco, M.; Cheatham, T. E.; Šponer, J.; Otyepka, M. Performance of molecular mechanics force fields for RNA simulations: Stability of UUCG and GNRA hairpins. *J. Chem. Theory Comput.* **2010**, *6*, 3836–3849.

146. Zgarbová, M.; Otyepka, M.; Šponer, J.; Mládek, A.; Banáš, P.; Cheatham, T. E.; Jurečka, P. Refinement of the Cornell et al. Nucleic acids force field based on reference quantum chemical calculations of glycosidic torsion profiles. *J. Chem. Theory Comput.* **2011**, *7*, 2886–2902.

147. Meagher, K. L.; Redman, L. T.; Carlson, H. A. Development of polyphosphate parameters for use with the AMBER force field. *J. Comput. Chem.* **2003**, *24*, 1016–1025.

148. Li, P.; Merz, K. M. MCPB.py: A Python Based Metal Center Parameter Builder. *J. Chem. Inf. Model.* **2016**, *56*, 599–604.

149. Huynh, V.; Phung, D.; Venkatesh, S. Streaming variational inference for dirichlet process mixtures. *ACML 2015 - 7th Asian Conf. Mach. Learn.* **2015**, 237–252.

150. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53–65.

151. Caliński, T.; Harabasz, J. A Dendrite Method Foe Cluster Analysis. *Commun. Stat.* **1974**, *3*, 1–27.

152. Davies, D. L.; Bouldin, D. W. A Cluster Separation Measure. *IEEE Trans. Pattern Anal. Mach. Intell.* **1979**, *PAMI-1*, 224–227.

153. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

154. Hall, M. C.; Matson, S. W. Helicase motifs: The engine that powers DNA unwinding. *Mol. Microbiol.* **1999**, *34*, 867–877.

155. Wang, Q.; Arnold, J. J.; Uchida, A.; Raney, K. D.; Cameron, C. E. Phosphate release contributes to the rate-limiting step for unwinding by an RNA helicase. *Nucleic Acids Res.* **2009**, *38*, 1312–1324.

156. Garai, A.; Chowdhury, D.; Betterton, M. D. Two-state model for helicase translocation and unwinding of nucleic acids. *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.* **2008**, *77*, 1–9.

157. Davidson, R. B.; Hendrix, J.; Geiss, B. J.; McCullagh, M. Allostery in the dengue virus NS3 helicase: Insights into the NTPase cycle from molecular simulations. *PLoS Comput. Biol.* **2018**, *14*, e1006103.

158. Chandler, D.; McCoy, J. D.; Singer, S. J. Density functional theory of nonuniform polyatomic systems. I. General formulation. *J. Chem. Phys.* **1986**, *85*, 5971–5976.

159. Beglov, D.; Roux, B. An integral equation to describe the solvation of polar molecules in liquid water. *J. Phys. Chem. B* **1997**, *101*, 7821–7826.

160. Kovalenko, A.; Hirata, F. Three-dimensional density profiles of water in contact with a solute of arbitrary shape: A RISM approach. *Chem. Phys. Lett.* **1998**, *290*, 237–244.

161. Ishizuka, R.; Huber, G. A.; McCammon, J. A. Solvation effect on the conformations of alanine dipeptide: Integral equation approach. *J. Phys. Chem. Lett.* **2010**, *1*, 2279–2283.

162. Chen, J.; Brooks, C. L. Implicit modeling of nonpolar solvation for simulating protein folding and conformational transitions. *Phys. Chem. Chem. Phys.* **2008**, *10*, 471–481.

163. Harris, R. C.; Pettitt, B. M.; Debenedetti, P. G. Effects of geometry and chemistry on hydrophobic salvation. *Proc. Natl. Acad. Sci. U. S. A.* **2014**, *111*, 14681–14686.

164. Harris, R. C.; Pettitt, B. M. Reconciling the understanding of 'hydrophobicity' with physics-based models of proteins. *J. Phys. Condens. Matter* **2016**, *28*, 83003.

165. Shell, M. S. The relative entropy is fundamental to multiscale and inverse thermodynamic problems. *J. Chem. Phys.* **2008**, *129*, 144108.

166. Salomon-Ferrer, R.; Götz, A. W.; Poole, D.; Le Grand, S.; Walker, R. C. Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. *J. Chem. Theory Comput.* **2013**, *9*, 3878–3888.

167. Friedrichs, M. S.; Eastman, P.; Vaidyanathan, V.; Houston, M.; LeGrand, S.; Beberg, A. L.; Ensign, D. L.; Bruns, C. M.; Pande, V. S. Accelerating Molecular Dynamic Simulation on Graphics Processing Units. *J. Comput. Chem.* **2009**, *30*, 864–872.

168. Eastman, P.; Pande, V. S. OpenMM: A Hardware-Independent Framework for Molecular Simulations. *Comput. Sci. Eng.* **2010**, *12*, 34–39.

169. Götz, A. W.; Williamson, M. J.; Xu, D.; Poole, D.; Le Grand, S.; Walker, R. C. Routine microsecond molecular dynamics simulations with AMBER on GPUs. 1. generalized born. *J. Chem. Theory Comput.* **2012**, *8*, 1542–1555.

170. Wu, W.; Xing, L.; Zhou, B.; Lin, Z. Active protein aggregates induced by terminally attached self-assembling peptide ELK16 in Escherichia coli. *Microb. Cell Fact.* **2011**, *10*, 9.

171. Cieplak, P.; Caldwell, J.; Kollman, P. Molecular mechanical models for organic and biological systems going beyond the atom centered two body additive approximation: Aqueous solution free energies of methanol and N-methyl acetamide, nucleic acid base, and amide hydrogen bonding and chloroform/. *J. Comput. Chem.* **2001**, *22*, 1048–1057.

172. Fox, T.; Kollman, P. A. Application of the RESP methodology in the parametrization of organic solvents. *J. Phys. Chem. B* **1998**, *102*, 8070–8079.

173. Páll, S.; Hess, B. A flexible algorithm for calculating pair interactions on SIMD architectures. *Comput. Phys. Commun.* **2013**, *184*, 2641–2650.

174. Harvey, M. J.; De Fabritiis, G. An implementation of the smooth particle mesh Ewald method on GPU hardware. *J. Chem. Theory Comput.* **2009**, *5*, 2371–2377.

175. Glaser, J.; Nguyen, T. D.; Anderson, J. A.; Lui, P.; Spiga, F.; Millan, J. A.; Morse, D. C.; Glotzer, S. C. Strong scaling of general-purpose molecular dynamics simulations on GPUs. *Comput. Phys. Commun.* **2015**, *192*, 97–107.

176. Anthopoulos, A.; Grimstead, I.; Brancale, A. GPU-accelerated molecular mechanics computations. *J. Comput. Chem.* **2013**, *34*, 2249–2260.

177. Anderson, J. A.; Lorenz, C. D.; Travesset, A. General purpose molecular dynamics simulations fully implemented on graphics processing units. *J. Comput. Phys.* **2008**, *227*, 5342–5359.

178. Main, E. R.; Phillips, J. J.; Millership, C. Repeat protein engineering: Creating functional nanostructures/biomaterials from modular building blocks. *Biochem. Soc. Trans.* **2013**, *41*, 1152–1158.

179. Grove, T. Z.; Regan, L.; Cortajarena, A. L. Nanostructured functional films from engineered repeat proteins. *J. R. Soc. Interface* **2013**, *10*.

180. Phillips, J. J.; Millership, C.; Main, E. R. Fibrous nanostructures from the self-assembly of designed repeat protein modules. *Angew. Chemie - Int. Ed.* **2012**, *51*, 13132–13135.

181. Grove, T. Z.; Osuji, C. O.; Forster, J. D.; Dufresne, E. R.; Regan, L. Stimuli-responsive smart gels realized via modular protein design. *J. Am. Chem. Soc.* **2010**, *132*, 14024–14026.

182. Costantini, S.; Colonna, G.; Facchiano, A. M. Amino acid propensities for secondary structures are influenced by the protein structural class. *Biochem. Biophys. Res. Commun.* **2006**, *342*, 441–451.

183. Lim, S. C.; Bowler, M. W.; Lai, T. F.; Song, H. The Ighmbp2 helicase structure reveals the molecular basis for disease-causing mutations in DMSA1. *Nucleic Acids Res.* **2012**, *40*, 11009–11022.

Appendix A

Supporting Information

A.1 Chapter 2 SI: The Role of Hydrophobicity in the Stability and pH-Switchability of $(RXDX)_4$ and Coumarin-$(RXDX)_4$ Conjugate β-Sheets

A.1.1 Hydrophobic Residues

I generalize the amino acid sequence to $(RXDX)_4$, where five different hydrophobic amino acids (Ala, Val, Leu, Ile and Phe) are substituted at the X position to investigate the effect of changing the hydrophobic group on both stability and pH-switchability of $(RXDX)_4$ β-sheet sandwich fibers. This set of five residues is chosen due to their incremental change in hydrophobicity, based on water-octanol partition coefficients,[98] without adding additional functional groups which could affect the self-assembled behavior of the material. These values along with β-sheet propensity for these amino acids are provided in Table A.1.

Table A.1: $(RXDX)_4$ peptide sequences, hydrophobicity, and secondary structure propensity.

Sequence	X	Hydrophobicity (π)[a]	β-sheet Propensity (P_β)[b]
$(RADA)_4$	Ala	0.31	0.75
$(RVDV)_4$	Val	1.22	1.86
$(RLDL)_4$	Leu	1.70	1.32
$(RIDI)_4$	Ile	1.80	1.71
$(RFDF)_4$	Phe	1.79	1.43

[a] Amino Acid hydrophobicity based on water-octanol partition coefficients relative to glycine.[98]

[b] Propensity to occur in β-sheet secondary structures from the PDBselect dataset.[182]

A.1.2 Force Field Selection

To support my choice in force fields I compare the thickness of each each $(RXDX)_4$ fiber at neutral pH calculated from simulation with atomic force microscopy measurements published by Bagrov *et al.*,[28] as shown in Table 2.1. The fiber in the simulation is aligned such that the principal component vector that is perpendicular to the plane of the 2 β-sheets is parallel to the x-axis of the simulation box. The thickness is calculated as the distance from the end of each central arginine sidechain on one β-sheet to the end of each central arginine on the other β-sheet projected along the x-axis. The center of mass of the NH1 and NH2 atoms is used as the position of the end of the arginine sidechain. Arginine residues on the outer two β-strands on both ends of each β-sheet and all N-termini arginines are excluded from these calculations as these parts of the β-sheet sandwich are less stable. The thickness is calculated every 100th frame and averaged over the entire 250 ns of aaMD simulation. The calculated thickness of $(RXDX)_4$ fibers with X=Ala and Leu are in good agreement with experimental results.

Additionally, I simulated 100 coumarins in water for 400 ns. From these simulations the solubility of coumarin in water is calculated to be 0.014 ± 0.006 $\frac{mol}{L}$, in agreement with the solubility value of 0.01706 $\frac{mol}{L}$ reported by Yalkowsky *et al.*[108]

A.1.3 Coulombic Interaction Energies

The coulombic interaction energies broken down into residue type pairs was calculated for each fiber to understand the interactions which lead to the pH-switchability of the $(RXDX)_4$ fibers. Below are the coulombic interaction energies for the $(RVDV)_4$, $(RLDL)_4$, $(RIDI)_4$, and $(RFDF)_4$ fibers.

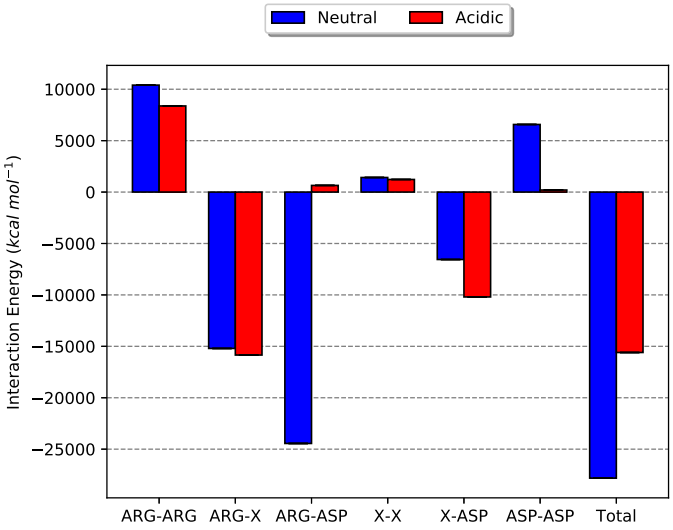

Figure A.1: Coulombic interaction energy decomposed into residue type pairs for $(RVDV)_4$ fibers at neutral and acidic pH. The total coulombic interactions are large and attractive stabilizing the fiber. Under acidic conditions there is a 11967 kcal/mol increase in the total coulombic interaction energy.

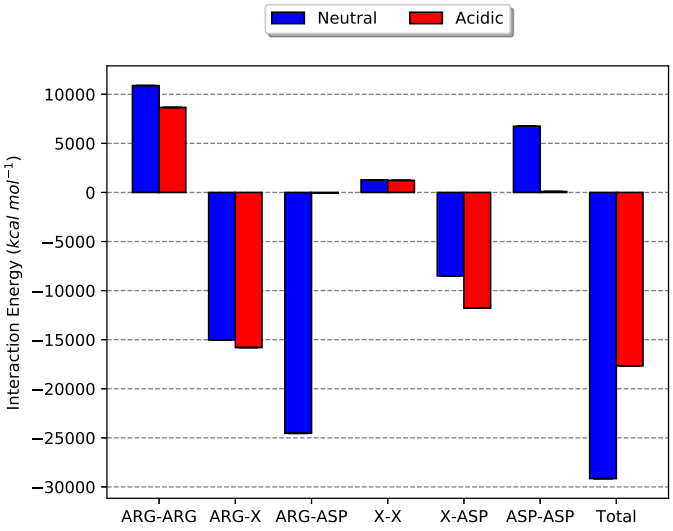

Figure A.2: Coulombic interaction energy decomposed into residue type pairs for $(RLDL)_4$ fibers at neutral and acidic pH. The total coulombic interactions are large and attractive stabilizing the fiber. Under acidic conditions there is a 11480 kcal/mol increase in the total coulombic interaction energy.

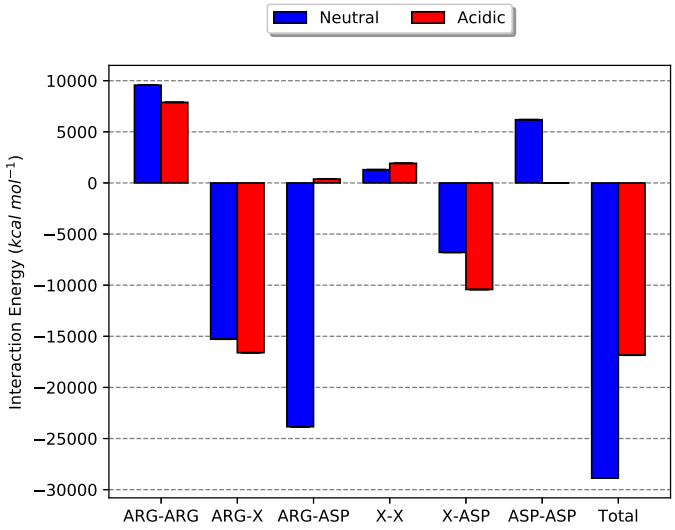

Figure A.3: Coulombic interaction energy decomposed into residue type pairs for $(RIDI)_4$ fibers at neutral and acidic pH. The total coulombic interactions are large and attractive stabilizing the fiber. Under acidic conditions there is a 12023 kcal/mol increase in the total coulombic interaction energy.

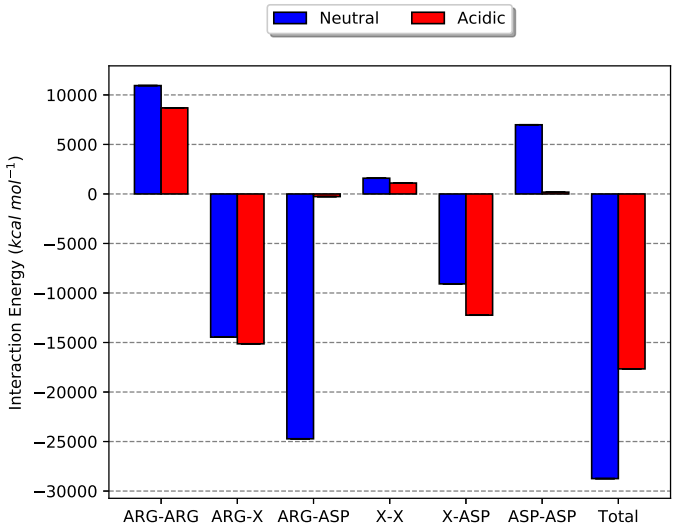

Figure A.4: Coulombic interaction energy decomposed into residue type pairs for $(RFDF)_4$ fibers at neutral and acidic pH. The total coulombic interactions are large and attractive stabilizing the fiber. Under acidic conditions there is a 11077 kcal/mol increase in the total coulombic interaction energy.

A.1.4 Coumarin Ordering

The joint probability density as a function of both the coumarin-coumarin separation distance and the angle between the dipole moments of the two coumarin is used to measure the organization of coumarin within coumarin-$(RXDX)_4$ fibers. Below are the joint probability densities for coumarin-$(RADA)_4$ and coumarin-$(RVDV)_4$ fibers.

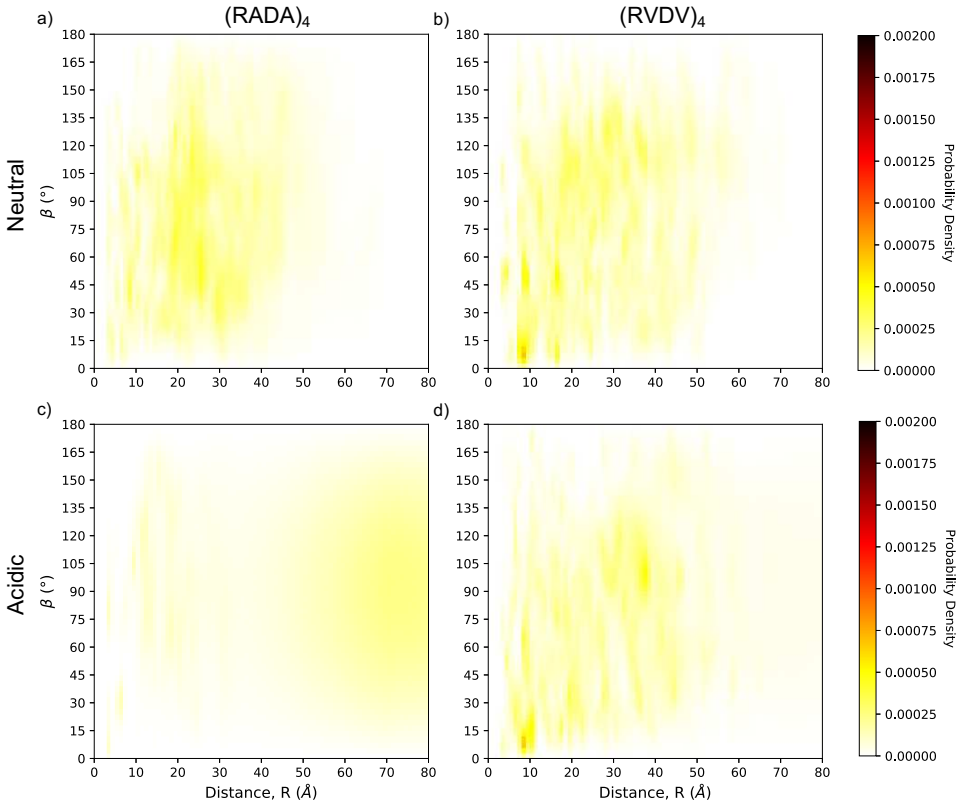

Figure A.5: Joint probability distributions of all coumarin sidechain dimers as a function of separation distance and angle between their respective dipole moments for (a,c) $(RADA)_4$ and (b,d) $(RVDV)_4$ at neutral and acidic pH, respectively.

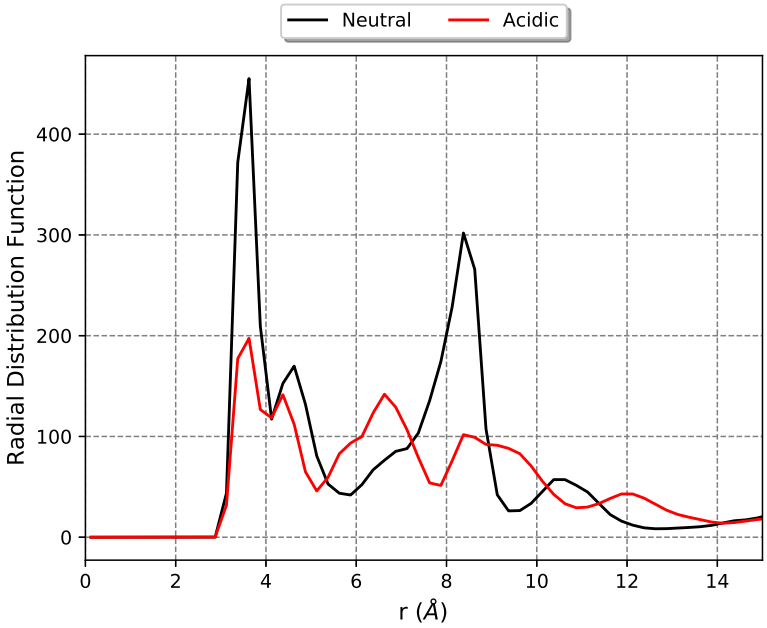

Figure A.6: Radial distribution function of coumarin sidechain dimers as a function of separation distance at neutral and acidic pH.

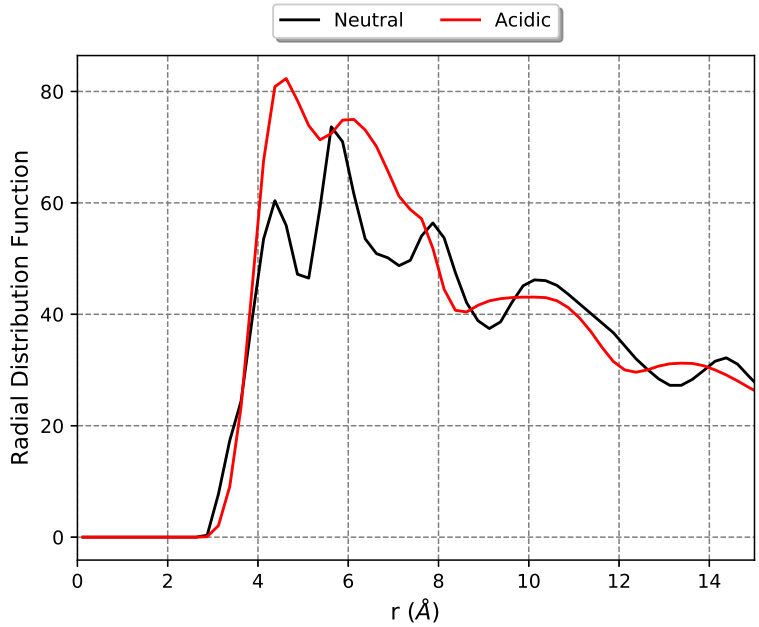

Figure A.7: Radial distribution function of coumarin and phenylalanine sidechains as a function of separation distance at neutral and acidic pH.

A.1.5 Pressure Equilibration

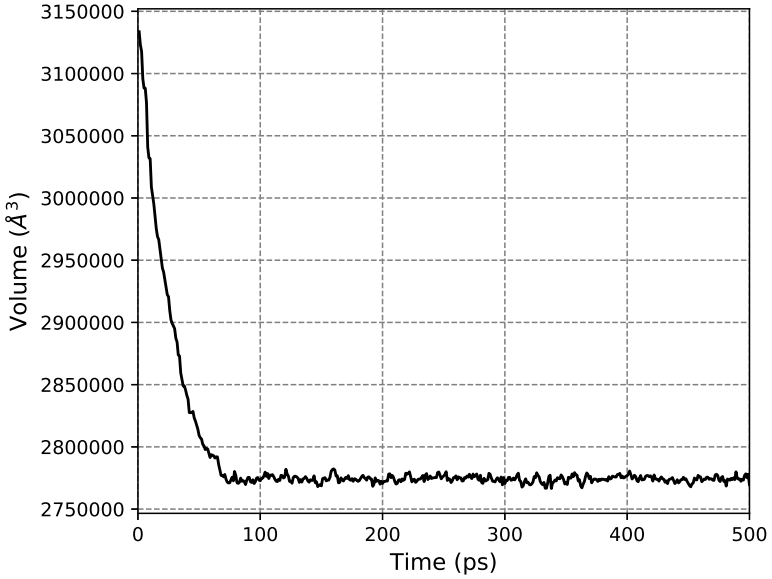

Figure A.8: Timetrace of the volume of the simulation box for the coumarin-$(RADA)_4$ system (replica 1). The volume and, therefore, pressure become equilibrated after 80 ps of simulation following the heating step.

A.1.6 Atom Parameters

Below are the atom and bonded parameters used for the unnatural coumarin amino acid. The bonds, angles, and dihedrals shown below are the parameters used for interacting atoms that are not within the same forcefield. All other parameters are provided by the ff15ipq and GAFF forcefields.

A.2 Chapter 3 SI: The Role of ATP in the RNA Translocation Mechanism of SARS-CoV-2 NSP13 Helicase

A.2.1 System Setup

The extent of the largest principle component of the protein was calculated to confirm that the size of the simulation box is large enough to prevent interactions between periodic images of

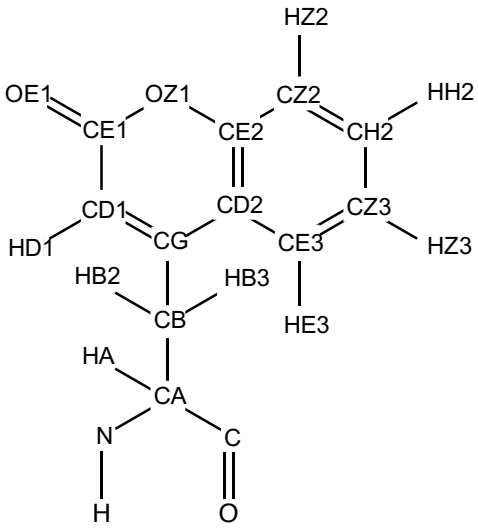

Figure A.9: Atom names of unnatural coumarin amino acid

Table A.2: Parameters for unnatural coumarin amino acid atoms.

Name	Type	Forcefield	Charge
O	OD	ff15ipq	-0.610490
C	C	ff15ipq	0.701560
CA	CX	ff15ipq	-0.146170
N	N	ff15ipq	-0.417020
H	H	ff15ipq	0.316160
HA	H1	ff15ipq	0.156000
CB	TA	ff15ipq	-0.240850
HB2	HC	ff15ipq	0.138730
HB3	HC	ff15ipq	0.138730
CG	c2	GAFF	0.239000
CD1	c2	GAFF	-0.520650
HD1	ha	GAFF	0.200630
CE1	c2	GAFF	0.895610
OE1	o	GAFF	-0.674080
OZ1	os	GAFF	-0.60182
CE2	ca	GAFF	0.473750
CZ2	ca	GAFF	-0.322010
HZ2	ha	GAFF	0.197470
CH2	ca	GAFF	-0.074560
HH2	ha	GAFF	0.316160
CZ3	ca	GAFF	-0.171280
HZ3	ha	GAFF	0.161480
CE3	ca	GAFF	-0.160930
HE3	ha	GAFF	0.159950
CD2	ca	GAFF	-0.155440

Table A.3: Bond parameters for unnatural coumarin amino acid. The parameters used in an AMBER frcmod file.

Bond Atom Types	Force Constant ($\frac{kcal}{mol \cdot A^2}$)	Bond Length (Å)
TA-c2	17.0000	1.5100

Table A.4: Angle parameters for unnatural coumarin amino acid. The parameters below are used in an AMBER frcmod file.

Angle Atom Types	Force Constant ($\frac{kcal}{mol \cdot rad^2}$)	Equilibrium Angle (°)
CX-TA-c2	63.0000	114.00
TA-c2-c2	70.0000	120.00
TA-c2-ca	70.0000	120.00
HC-TA-c2	50.0000	109.50
o -c2-os	76.662	118.370
os-c2-o	76.662	118.370
c2-TA-HC	50.0000	109.50

Table A.5: Dihedral parameters for unnatural coumarin amino acid. The parameters below are used in an AMBER frcmod file.

Dihedral Atom Types	Barrier Height	Phase Shift Angle(°)	Periodicity
C -CX-TA-c2	-0.22262	0.0	-4.0
C -CX-TA-c2	-1.19867	0.0	-3.0
C -CX-TA-c2	-0.03841	0.0	-2.0
C -CX-TA-c2	-0.17107	0.0	1.0
CX-TA-CA-ca	0.00000	0.0	-4.0
CX-TA-CA-ca	0.11558	0.0	-3.0
CX-TA-CA-ca	-0.39107	0.0	-2.0
CX-TA-CA-ca	-0.50022	0.0	1.0
CX-TA-CA-c2	0.00000	0.0	-4.0
CX-TA-CA-c2	0.11558	0.0	-3.0
CX-TA-CA-c2	-0.39107	0.0	-2.0
CX-TA-CA-c2	-0.50022	0.0	1.0
N -CX-TA-c2	-0.22082	0.0	-4.0
N -CX-TA-c2	0.13006	0.0	-3.0
N -CX-TA-c2	0.01776	0.0	-2.0
N -CX-TA-c2	-0.62013	0.0	1.0
H1-CX-TA-c2	0.38033	0.0	3.0
HC-TA-c2-c2	-0.46048	0.0	2.0
HC-TA-c2-ca	-0.46048	0.0	2.0
CX-TA-c2-c2	0.00000	0.0	-4.0
CX-TA-c2-c2	0.11558	0.0	-3.0
CX-TA-c2-c2	-0.39107	0.0	-2.0
CX-TA-c2-c2	-0.50022	0.0	1.0
CX-TA-c2-ca	0.00000	0.0	-4.0
CX-TA-c2-ca	0.11558	0.0	-3.0
CX-TA-c2-ca	-0.39107	0.0	-2.0
CX-TA-c2-ca	-0.50022	0.0	1.0

the protein. As shown in Table A.7, the extent of the protein is 104(2) Å maintaining a 16 Å gap between periodic images for the smallest simulation box size providing a large enough buffer as a 12 Å interaction cutoff is used.

A.2.2 Model Corroboration

To provide support that the initial structures of the ATP, ssRNA, and ssRNA+ATP ligand-bound states of nsp13 are suitable I compare the contacts in the simulations to the contacts in the crystal structures of other SF1 helicases with ssRNA and ATP bound. The ssRNA contacts are shown in Table A.8 and the ATP contacts are shown in Table A.9.

A.2.3 ssRNA Binding Strength

Inter-domain distance analysis of the Apo, ATP, ssRNA, and ssRNA+ATP states show that when nsp13 binds ATP there is a widening of the RNA-binding cleft. To measure the change in binding strength between nsp13 and ssRNA the linear interaction energy (Table A.10) and root-mean-square fluctuation of the RNA phosphates (Table A.11) are calculated for the ssRNA and ssRNA+ATP systems. The error in both analyses are too large to differentiate between the two systems. Figure A.10 shows the labeling of the RNA phosphates and the highly conserved motifs of nsp13.

A.2.4 Inter-domain Distances

The inter-domain distances between domains 1A, 2A, and 1B were calculated for the Apo, ATP, ssRNA, and ssRNA+ATP ligand-bound states of nsp13. The distributions of the 1A–1B, 2A–1B, and 1A–2A distances for each ligand-bound state are shown in Figure A.11(a-c), respectively.

A.2.5 Gaussian Mixture Model and Linear Discriminant Analysis

Figure A.12 shows the Silhouette, CH, and DB scores for cluster sizes ranging from two clusters to ten clusters for the RNA-binding cleft distances. Based on the maximums of the Silhouette and CH scores and minimums of the DB score a cluster size of four was chosen. Linear discriminant

Table A.6: Dihedral parameters for unnatural coumarin amino acid. The parameters below are used in an AMBER frcmod file.

Dihedral Atom Types	Barrier Height (÷2)	Phase Shift Angle(°)	Periodicity
TA-c2-c2-ca	1.1	180.0	2.0
c2-c2-c2-ha	1.1	180.0	2.0
c2-o -c2-os	1.1	180.0	2.0
ca-ca-ca-os	1.1	180.0	2.0
ca-ca-ca-ha	1.1	180.0	2.0

Table A.7: The average extent of the largest principle axis of the nsp13 protein and the size of the simulation box for each ligand bound state.

System	Length (Å)	Box Length (Å)	Number of Water Molecules
Apo	104(2)	131.0(1)	212502
ATP	110(3)	130.6(2)	211356
ssRNA	105(2)	130.9(1)	212184
ssRNA+ATP	104(2)	120.33(8)	162858

Table A.8: Residues from motifs **Ia**, **IV**, and **V** in contact with RNA phosphates (≤ 5.0 Å) for various SF1 RNA-bound helicase protein crystal structures and the percentage of frames where the corresponding nsp13 residues are in contact with RNA phosphates for both the RNA and RNA+ATP systems.

motif	Upf1 (2XZL)[141]	IGHMBP2 (4B3G)[183]	nsp13	RNA	RNA+ATP
Ia	SER 461	SER 244	SER 310	83.15%	84.25%
Ia	ASN 462	ASN 245	HIE 311	83.68%	83.61%
IV	PRO 731	PRO 540	PRO 514	17.82%	1.34%
IV	TYR 732	TYR 541	TYR 515	97.35%	55.16%
IV	GLU 793	ASN 542	ASN 516	66.49%	54.30%
V	SER 761	SER 563	SER 535	0.27%	0.00%
V	ALA 764	ASP 565	VAL 533	44.14%	0.12%

Table A.9: Residues from motifs **I**, **II**, **III**, **V** and **VI** in contact with ATP or MG^{2+} (≤ 5.0 Å) for various SF1 ATP-bound helicase protein crystal structures and the percentage of frames where the corresponding nsp13 residues are in contact with ATP or MG^{2+} for both the ATP and RNA+ATP systems.

motif	Upf1 (2GJK)[143]	nsp13	ATP	RNA+ATP
I	GLY 492	GLY 282	8.93%	29.54%
I	PRO 493	PRO 283	97.18%	99.99%
I	PRO 494	PRO 284	98.28%	100.00%
I	GLY 495	GLY 285	100.00%	100.00%
I	THR 496	THR 286	100.00%	100.00%
I	GLY 497	GLY 287	100.00%	100.00%
I	LYS 498	LYS 288	100.00%	100.00%
I	THR 499	SER 289	100.00%	100.00%
I	VAL 500	HIE 290	100.00%	100.00%
II	ASP 636	ASP 374	60.07%	50.49%
II	GLU 637	GLU 275	0.46%	30.70%
III	GLN 665	GLN 404	15.39%	73.33%
V	GLY 831	GLY 538	63.19%	80.30%
V	ARG 832	SER 539	36.77%	19.22%
V	GLU 833	GLU 540	99.27%	84.46%
VI	ARG 865	ARG 567	100.00%	99.95%
VI	ARG 867	LYS 569	80.17%	45.12%

Table A.10: Average linear interaction energy between each phosphate and protein residues within 12 Å for the ssRNA and ssRNA+ATP systems.

Linear Interaction Energy ($kcal \cdot mol^{-1}$)		
Phosphates	ssRNA	ssRNA+ATP
P0	-97(18)	-94(12)
P1	-122(21)	-126(17)
P2	-103(23)	-143(28)
P3	-130(16)	-144(57)
P4	-124(24)	-95(44)
P5	-112(24)	-63(26)
P6	-48(26)	-39(37)

Table A.11: RMSF of each phosphate for the ssRNA and ssRNA+ATP systems.

	RMSF (Å)	
Phosphates	ssRNA	ssRNA+ATP
P0	1.316(0.465)	2.495(1.028)
P1	1.211(0.220)	1.857(0.458)
P2	1.042(0.191)	1.520(0.606)
P3	0.849(0.178)	1.177(0.523)
P4	0.879(0.205)	1.012(0.038)
P5	1.328(0.268)	1.322(0.214)
P6	2.427(0.812)	2.204(0.994)

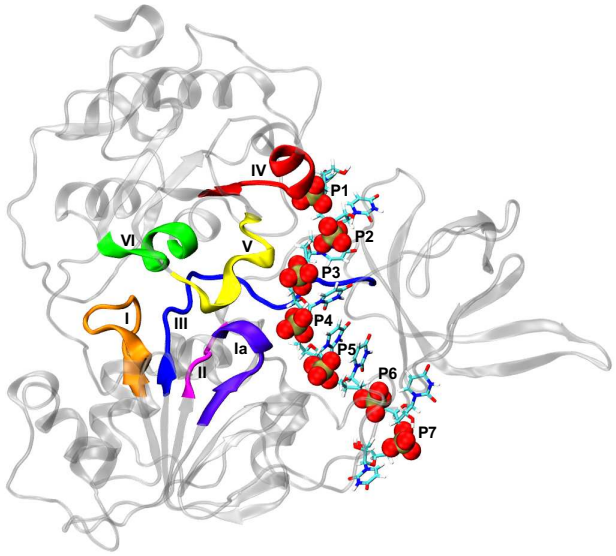

Figure A.10: Representative structure of nsp13 with ssRNA bound. Motifs **I** (orange), **Ia** (violet), **II** (magenta), **III** (blue), **IV** (red), **V** (yellow), **VI** (green), and each phosphate in the ssRNA backbone is highlighted and labeled. The ZBD and Stalk domain are removed for clarity.

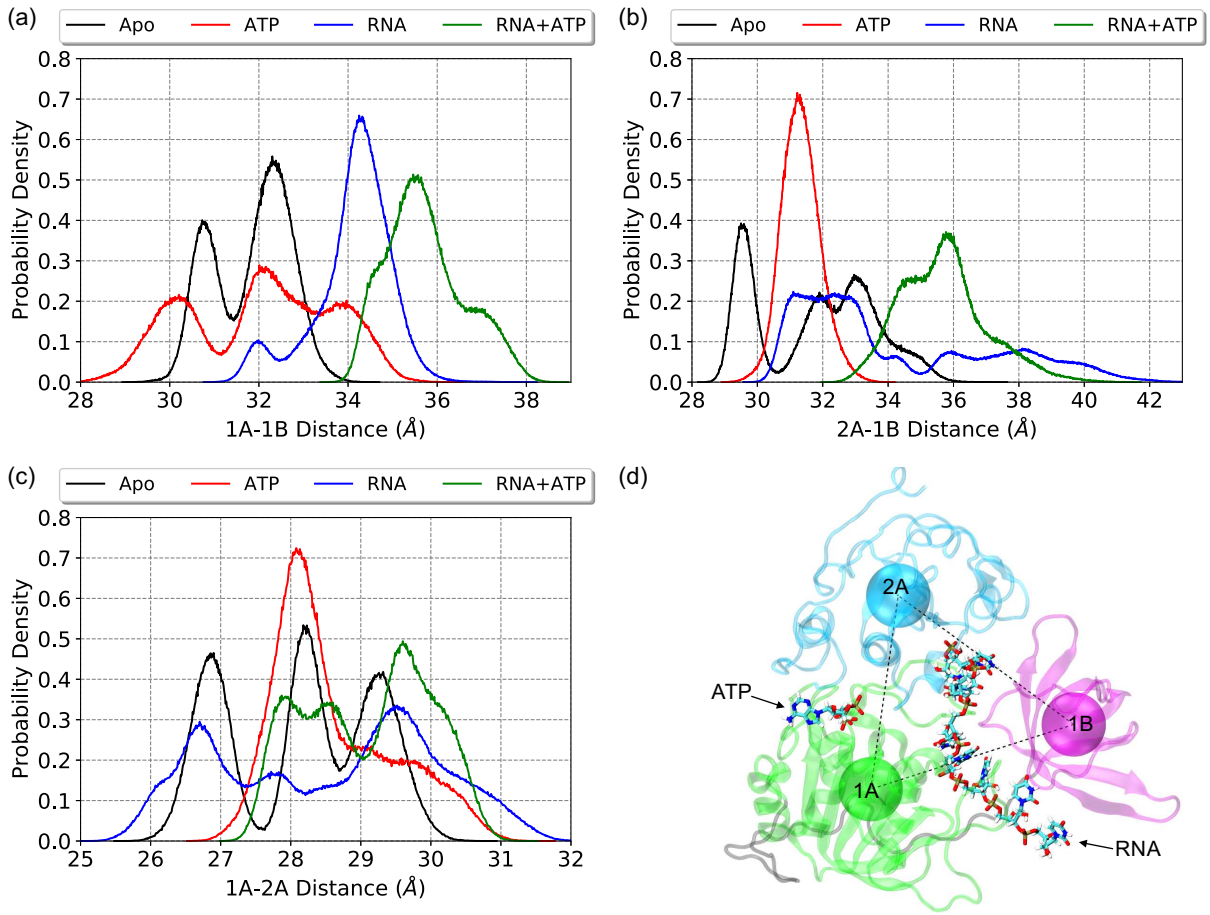

Figure A.11: Probability density of the center-of-mass separation distance between domains (a) 1A–1B, (b) 1A–2A, and (c) 2A–1B of the nsp13 Apo, ATP, ssRNA, and ssRNA+ATP ligand-bound states. (d) Structural depiction of the center-of-mass of domains 1B (magenta), 1A (green), and 2A (cyan)

analysis (LDA) was utilized to differentiate between the 4 states in the RNA-binding cleft and the ATP-pocket. Table A.12 shows the α-carbon of the residues used to represent the position of each motif used in the LDA. The LD1 and LD2 vectors for the RNA-binding cleft and the ATP pocket are shown in Table A.13 and Table A.14, respectively.

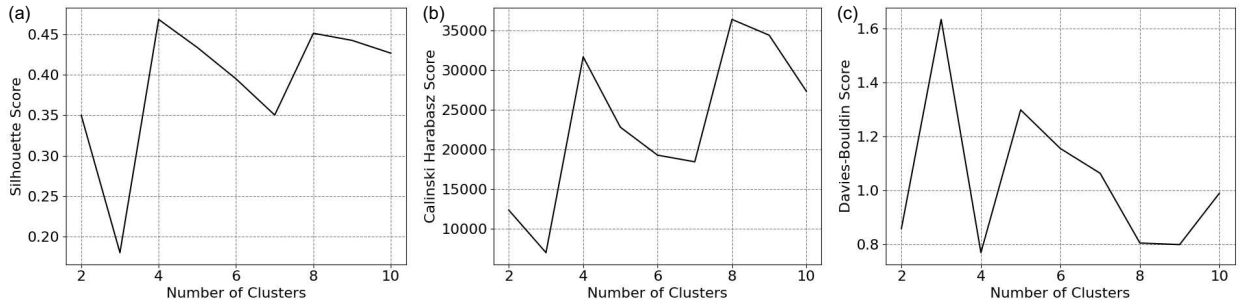

Figure A.12: (a) Silhouette, (b) Calinski-Harabasz, and (c) Davies-Bouldin scores for various number of clusters from GMM clustering of the RNA-binding cleft.

Table A.12: Residues used as the position of each motif utilized by the linear discriminant analysis in calculating the difference between states **S1**, **S2**, **S3**, and **S4**.

Motif	Residue
I	GLN 281
Ia	HID 311
II	ILE 375
IV	ASN 516
V	ASP 534
VI	ARG 567

A.2.6 Motif V–ssRNA Contacts

Table A.15 shows the percentage of frames that motif **V** was in contact with each ssRNA phosphate. If any residue of motif **V** was within 5 Å of an ssRNA phosphate than it was considered a contact. The phosphates are labeled relative to the phosphate bound by motif **Ia**. Table A.16 shows the average separation distance between each residue in motif **V** and the closest ssRNA phosphate.

Table A.13: Coefficients for each distance used in the linear discriminant analysis to describe the RNA-binding cleft for LD1 and LD2.

LDA Coefficients		
Residues	LD1	LD2
IV – P	-0.424	0.335
IV – Ia	-0.095	-0.657
Ia – P	1.859	-0.050

Table A.14: Coefficients for each distance used in the linear discriminant analysis to describe the ATP-pocket for LD1 and LD2.

LDA Coefficients		
Residues	LD1	LD2
I – V	0.931	-1.080
I – Ia	-1.173	1.427
Ia – V	-0.520	-0.318
II – V	-0.765	0.207
II – VI	-0.096	0.020
IV – V	-0.373	-0.067
V – VI	-1.024	0.509
V – P	0.646	0.244

Table A.15: Percentage of frames where motif **V** is bound (≤ 5.0 Å) to each ssRNA phosphates. Phosphates are labeled relative to the phosphate motif **Ia** is binding, where motif **Ia** is binding P_n.

Residues	**S1**	**S2**	**S3**	**S4**
P_n	0.88%	0.46%	0.10%	7.02%
P_{n-1}	53.60%	77.07%	62.50%	44.14%
P_{n-2}	43.92%	12.35%	1.87%	0.39%
P_{n-3}	0.56%	0.01%	0.00%	0.00%

Table A.16: Average separation distance and standard deviation between all residues in motif **V** with ssRNA phosphates for states **S1**, **S2**, **S3**, and **S4**.

Average Distance (Å)				
Residues	**S1**	**S2**	**S3**	**S4**
Val 533	16(2)	17(2)	18.6(6)	19.2(8)
ASP 534	12(2)	13(2)	14.6(7)	15(1)
SER 535	11(1)	12(2)	12.7(8)	15(1)
SER 536	6(2)	8(1)	9.9(9)	10(1)
GLN 537	7(1)	8(2)	10.1(5)	10.8(6)
GLY 538	4(2)	6(3)	7.4(7)	7.9(8)
SER 539	4.3(6)	4.6(9)	4.8(5)	5.1(6)
GLU 540	7.1(9)	7(2)	7.9(9)	8.6(5)

List of Abbreviations

aaMD	all-atom molecular dynamics	5
ADP	adenosine diphosphate	36
AP	alanine dipeptide	45
ATP	adenosine triphosphate	4
CH	Calinski-Harabasz	30
cMD	conventional molecular dynamics	28
CPU	central processing unit	5
DB	Davies-Bouldin	30
DNA	deoxyribonucleic acid	4
DSSP	define secondary structure protein	12
GaMD	Gaussian accelerated molecular dynamics	4
GB	generalized Born	5
GMM	Gaussian Mixture Model	27
GPU	graphics processing unit	5
IS-SPA	Implicit Solvation Using the Superposition Approximation	5
LDA	linear discriminant analysis	27
LIE	linear interaction energy	32
MC	Monte Carlo	10
MD	molecular dynamics	1
nsp13	nonstructural protein 13	4
nsps	nonstructural proteins	24
NTP	nucleoside triphosphate	25
Pi	inorganic phosphate	36
PME	particle mesh Ewald	10
PMF	potential of mean force	55
QM	quantum mechanics	63
RISM	Reference Interaction Site Model	5
RNA	ribonucleic acid	4
SAMD	simulated annealing molecular dynamics	10
SARS-CoV-2	severe acute respiratory syndrome coronavirus 2	4
SASA	solvent accessible surface area	15
SF1	superfamily 1	4
SM	Streaming Multiprocessor	47
SP	Stream Processor	47
ssRNA	single-stranded RNA	27
TPR	tetratricopeptide repeat protein	64
ZBD	zinc binding domain	24