DISSERTATION

# ALLOSTERY OF THE FLAVIVIRUS NS3 HELICASE AND BACTERIAL IGPS STUDIED WITH MOLECULAR DYNAMICS SIMULATIONS

Submitted by

Russell Bruce Davidson

Department of Chemistry

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Spring 2020

Doctoral Committee:

Advisor: Martin McCullagh

Elliot Bernstein
George Barisas
Brian Geiss

## ABSTRACT

## ALLOSTERY OF THE FLAVIVIRUS NS3 HELICASE AND BACTERIAL IGPS STUDIED WITH MOLECULAR DYNAMICS SIMULATIONS

Allostery is a biochemical phenomenon where the binding of a molecule at one site in a biological macromolecule (e.g. a protein) results in a perturbation of activity or function at another distinct active site in the macromolecule's structure. Allosteric mechanisms are seen throughout biology and play important functions during cell signaling, enzyme activation, and metabolism regulation as well as genome transcription and replication processes. Biochemical studies have identified allosteric effects for numerous proteins, yet our understanding of the molecular mechanisms underlying allostery is still lacking. Molecular-level insights obtained from all-atom molecular dynamics simulations can drive our understanding and further experimentation on the allosteric mechanisms at play in a protein. This dissertation reports three such studies of allostery using molecular dynamics simulations in conjunction with other methods. Specifically, the first chapter introduces allostery and how computational simulation of proteins can provide insight into the mechanisms of allosteric enzymes. The second and third chapters are foundational studies of the flavivirus non-structural 3 (NS3) helicase. This enzyme hydrolyzes nucleoside triphosphate molecules to power the translocation of the enzyme along single-stranded RNA as well as the unwinding of double-stranded RNA; both the hydrolysis and helicase functions (translocation and unwinding) have allosteric mechanisms where the hydrolysis active site's ligand affects the protein-RNA interactions and bound RNA enhances the hydrolysis activity. Specifically, a bound RNA oligomer is seen to affect the behavior and positioning of waters within the hydrolysis active site, which is hypothesized to originate, in part, from the RNA-dependent conformational states of the RNA-binding loop. Additionally, the substrate states of the NTP hydrolysis reaction cycle are seen to affect protein-RNA interactions, which is hypothesized to drive unidirectional translocation of the enzyme along the RNA polymer. Finally, chapter four introduces a novel method to study the biophysical coupling between two active sites in a protein. The short-ranged residue-residue interactions within the protein's three dimensional structure are used to identify paths that connect the two active sites. This method is used to highlight the paths and residue-residue interactions that are important to the allosteric enhancement observed for the *Thermatoga maritima* imidazole glycerol phosphate synthase (IGPS) protein. Results from this new quantitative analysis have provided novel insights into the allosteric paths of IGPS. For both the NS3 and IGPS proteins, results presented in this dissertation have highlighted structural regions that may be targeted for small-molecule inhibition or mutagenesis studies. Towards this end, the future studies of both allosteric proteins as well as broader impacts of the presented research are discussed in the final chapter.

# ACKNOWLEDGEMENTS

I would like to thank Professor Martin McCullagh. He took me on as a student on his first day working at Colorado State University. The following years under his guidance were formative to me as a scientist, much of which is due to Martin's tenacity and approach to science. I thank him for all that he has done for me.

The McCullagh group is also an origin for much of my development as a scientist. Dr. Peter Lake, Kevin Votaw, Mortaza Derakhshani Molayousefi, Kelly Du Pont, Ryan Weber, Jakob Anderson, and Heidi Klem have been exceptional coworkers who pushed my scientific thinking in numerous directions. Additionally, these people have been some of my closest friends. I'd like to specifically acknowledge Ryan and Jake for putting up with me and my idiosyncrasies on a regular basis. Additionally, Rex was often either the source or the fuel for innumerable and insightful conversations. These discussions covered any and all topics, yet his words often drove me to introspection and critical thinking.

I would also like to thank friends from outside the McCullagh group. Specifically and in no specific order: I appreciate the support that Dan Agocs provided throughout my time at Elon University as well as CSU. To Matthew Lacroix, my first DM, I will never forget our mythical adventures and I thank him for introducing me to my new hobby. I thank Rachel Anderson, Max Mattson, and Jeffrey Ma for the fellowship and support they provided these past years.

Lastly and most importantly, I emphatically thank my parents, John and Hillary, and my sister, Kelly. My family has been a never-ending font of support, wisdom, and laughter. I cannot describe in words my appreciation for all that they have done for me.

# DEDICATION

*I would like to dedicate this dissertation to the many creators of art, literature, and music whose works inspired me to live my own style of creative life.*

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# Chapter 1

# Introduction

## 1.1 What Is Allostery?

Allostery is a biochemical phenomenon where the binding of a molecule at one site in a biological macromolecule (e.g. a protein) results in a perturbation of activity or function at another distinct active site in the macromolecule's structure. A marvel of eons-worth of evolution, allostery is described as "the second secret of life" – second only to the genetic code – because allosteric mechanisms allow biological systems to adapt to ever-changing chemical environments. Proteins that have some form of allosteric control are ubiquitous in all biology, from the largest eukaryotes to the smallest prokaryotes as well as viruses. Allosteric mechanisms are found in cell signaling, enzyme activation, and metabolism regulation as well as genome transcription and replication processes.[1–5] It is exactly for these reasons that the categorization and fundamental understanding of this phenomenon is extremely compelling yet difficult to study.

Allostery, as defined above, requires at least two active sites in the protein's three dimensional structure. The first, termed the orthosteric site, is associated with the native function(s) of the protein. Within this active site, the protein's ligand binds and undergoes some biophysical or chemical reaction that is catalyzed by the enzyme. The second site, termed the allosteric site, is the structural location where a second molecule binds and perturbs the orthosteric site's functionality. The binding of this molecule, termed the effector molecule, can lead to the enhancement or inhibition or regulation, in some fashion, of the protein's activity at the other site. Examples of effector molecules range from molecular oxygen ($O_2$) for hemoglobin[6] to peptidic- or nucleic acid (NA) oligomers, as observed in kinase-peptide complexes[4] and DNA- or RNA-helicases[7].

A classic example of allostery is the hemoglobin protein, which reversibly-binds and transports $O_2$ throughout the body. The basal function of hemoglobin is binding and unbinding of $O_2$ molecules in the four heme groups located in the protein structure. The initial binding of an $O_2$

molecule induces structural changes in the hemoglobin protein that enhances binding of subsequent $O_2$ molecules. This form of cooperative allostery allows for increased transport capabilities of the wild-type hemoglobin enzyme. Breakdown of this cooperativity via inhibition (e.g. by carbon monoxide) or mutation of the protein (e.g. sickle cell hemoglobin) leads to issues with oxygen transport and stability of the native enzyme.

Hemoglobin has been studied for over 50 years; it is a good case study for this introduction to allostery due to its historical significance as well as its complexity. The observed allosteric cooperativity in hemoglobin was the spark for the development of numerous enzymatic models, all of which attempt to generalize a quantitative description of an allosteric enzyme's response to the effector molecule as measured by biochemical assays.[8–10] These models are quantitative narratives describing hypothesized steps in the mechanism for allosteric regulation, yet they lack specific insight into the physical interactions between the effector molecule and the protein. As one of the first crystal structures of a biological macromolecule obtained using X-ray crystallography, the structures of hemoglobin in deoxy- and oxygenated conformations (Protein Data Bank - PDB - IDs 2HHB and 1HHO) were used to identify how a bound-$O_2$ molecule affects the protein structure.[11,12] The combination of biochemical assays, enzymatic models, and structural insights have resulted in an extremely detailed understanding of the structure and function of hemoglobin.

Since the initial studies of hemoglobin's allostery, the number of identified allosteric enzymes has exploded. The biochemical studies of an allosteric enzyme can provide detailed understanding of an effector's influence on the enzyme's kinetics. Yet, structural insights are needed to understand how allostery is chemically or biophysically manifested in the enzyme of interest. Specific questions of interest in this regard are:

1. What is the allosteric effect? In other words, what are effector-induced changes in the orthosteric site that lead to the observed biochemical allostery?
2. What are the interactions between the enzyme and effector that lead to this allosteric effect?
3. How are the orthosteric and allosteric sites coupled?

X-ray crystallography studies of proteins can provide information about the enzyme in the presence and absence of the effector and ligand molecules. These static, structural details provide critical information that can be used to preliminarily answer these questions, such as where the allosteric site is located in the protein's structure or the large-scale structural changes that occur upon effector binding. Yet, proteins naturally function in dynamic environments, which might be poorly described by the static structures obtained from crystallography studies.

## 1.2 Using Molecular Dynamics to Study Allostery

The structural fluctuations and conformational changes of proteins have become increasingly more important to our understanding of the ever growing field of allostery.[13,14] Interactions between a protein and its ligands (including the effector molecule), as seen in crystal structures, may be incomplete or poor descriptions of the interactions important to allostery. Additionally, these static structures represent a single conformation in the highly complex phase space of the protein-ligand complex.

Molecular dynamics (MD) is a theoretical chemistry method that can be used to sample the ensemble of conformations for the protein-ligand structures, thereby providing a much more detailed description of the biophysically important phase space. This method numerically solves Newton's equations of motion to propagate the atomic positions of the modeled structure in time, where interatomic forces are defined by the molecular mechanics Hamiltonian and force-field parameters. The Hamiltonian used to describe the potential energy between all atoms in the system is

$$\begin{aligned} V_{MM} = &\sum_{i}^{n_{bonds}} b_i \left(r_i - r_{i,eq}\right)^2 + \sum_{i}^{n_{angles}} a_i \left(\theta_i - \theta_{i,eq}\right)^2 + \sum_{i}^{n_{dihedrals}} \sum_{n}^{n_{i,max}} \frac{V_{i,n}}{2} \left[1 + \cos(n\phi_i - \gamma_{i,n})\right] \\ &+ \sum_{i<j}^{n_{atoms}} \left(\frac{A_{i,j}}{r_{i,j}^{12}} - \frac{B_{i,j}}{r_{i,j}^{6}}\right) + \sum_{i<j}^{n_{atoms}} \frac{q_i q_j}{4\pi\epsilon_0 r_{i,j}} \end{aligned}$$

where $b_i$ and $r_{i,eq}$ are parameters describing the harmonic bonding potential, $a_i$ and $\theta_{i,eq}$ are parameters describing the harmonic angle potential, $V_{i,n}$ and $\gamma_{i,n}$ are parameters to describe the torsional rotation around a central bond, $A_{i,j}$ and $B_{i.j}$ are parameters to describe the Lennard-Jones potential between non-bonded atoms, and $q_i$ and $q_j$ are atomic charge parameters used to quantify the pairwise electrostatic interaction potential. Over two decades of research has been focused on the development of these force field parameters to accurately model biomolecule structures and dynamics.[15]

Computational resources are used to perform these numerical calculations, where the modeled system is propagated for a large number of time steps. This results in a trajectory where each frame represents a conformation of the system that is hypothesized to be biophysically plausible. Therefore, a large set of frames serves as a description of the accessible phase space for the system.

From such a set of frames, one can begin to answer the scientific questions posed above from an atomistic perspective. Comparative analyses between modeled systems in the presence and absence of ligands can identify the allosteric effect caused by the effector molecule. Frames that strongly represent the average structure of the modeled system can be used to visualize the structural interactions between the protein and ligands as well as initiate further studies of the protein-ligand complexes. Using enhanced sampling methods, MD simulations can provide enough sampling of protein-ligand interactions to quantify the relative free energies of conformations within the phase space of the system. Additionally, MD trajectories can be used to study the short-range, residue-residue interactions that couple the allosteric and orthosteric active sites. These results, obtained from sets of MD simulations, can be validated and contextualized with a combination of bioinformatics and experimental insights. Hypotheses developed from the study of MD trajectories provides new research avenues for experimental collaborators.

## 1.3 Chapter Overview

My research, which relies heavily on the use of MD simulations to study allostery, will be presented in the subsequent chapters of this dissertation. The chapters are organized to mirror the scientific questions presented above. Chapters two and three present work highlighting the allosteric effect and protein-ligand interactions (questions 1 and 2) for the flavivirus NS3 helicase. Chapter four introduces a new method to analyze the couplings between the allosteric and orthosteric sites (question 3) for imidazole glycerol phosphate synthase (IGPS), which is a model system for the study of allostery. The concluding chapter highlights potential research avenues for these specific proteins as well as provides a forward-looking perspective on the scientific questions of interest.

### 1.3.1 Flavivirus NS3 helicase

The NS3 protein is a helicase that hydrolyzes nucleoside triphosphate (NTP) molecules to translocate the enzyme along a NA polymer. Often referred to as motor proteins, helicases are analogous to a motor in a car: the fuel (an NTP molecule) is burned (hydrolyzed), resulting in the release of energy that the motor (the protein) converts into mechanical work to move along the road (the NA polymer). Allostery in the flaviviral NS3 helicase is observed for both the NTP hydrolysis reaction as well as the translocation process: the hydrolysis of NTP is seen to be enhanced by the bound-NA polymer and translocation along the polymer is dependent on the fuel-burning reaction cycle.

For the flaviviruses (Family *Flaviviridae*), the NS3 helicase plays a pivotal role in the replication of the viral RNA genome. This viral helicase utilizes energy released from the hydrolysis reaction to translocate along and unwind double-stranded RNA, thereby resolving the double-stranded replication intermediate into single-stranded, positive sense RNAs. An understanding of the natural workings of this enzyme, including the allosteric mechanisms underlying the helicase functions, could aid the development of antiviral therapeutics against flaviviruses such as dengue, Zika, and West Nile. The work presented in chapters two and three represent foundational research on the allosteric mechanisms of the flavivirus NS3 helicase. Specifically, chapter two reports a set of all-atom, explicit solvent MD simulations modeling the dengue NS3 helicase, from which the allosteric effects of both the RNA and NTP-hydrolysis ligands were studied. Chapter three highlights RNA-dependent conformations of a secondary structure in the Zika virus NS3 that is hypothesized to be an origin site of the RNA-induced allosteric effect.

### 1.3.2 IGPS

The IGPS protein has been a model enzyme for the study of allostery, similar to hemoglobin. It plays an important role in the purine and histidine biosynthesis pathways in plants, fungi, archaea, and bacteria. Allostery in IGPS is seen between two active sites where the binding of a ligand at the allosteric site induces a 4,900-fold enhancement of the reactivity at the orthosteric site. The specific chemical reactions at these active sites as well as the biophysical importance of IGPS are of little importance to this dissertation. Rather, the extreme allosteric enhancement observed for the IGPS system has driven the field to use it as the proving ground for new methods to study allostery. As presented here, chapter four uses the IGPS system in this way.

Methods to study the interactions that couple allosteric and orthosteric sites have used graph theoretical frameworks to describe the protein as a network of interacting, coupled nodes. The correlated fluctuations of these nodes have been used to highlight the coupled motions in the protein between the two active sites. Our new method, presented in chapter four, continues in this vein of research yet highlights a new, more physically-relevant quantity to describe the protein network. Additionally, we present two new metrics to quantify the “importance” or “centrality” of nodes to the protein’s allosteric mechanism. This new quantitative analysis of MD simulations is used to study the short-ranged interactions observed in the IGPS system that build up into the coupling between the two active sites of the protein. Results from this analysis provide novel insights into the biophysics of the protein and are validated with experimental results.

# Chapter 2

# Allostery in the Dengue Virus NS3 Helicase: Insights into the NTPase Cycle from Molecular Simulations.[1]

## 2.1 Overview

The C-terminus domain of non-structural 3 (NS3) protein of the *Flaviviridae* viruses (e.g. HCV, dengue, West Nile, Zika) is a nucleoside triphosphatase (NTPase) -dependent superfamily 2 (SF2) helicase that unwinds double-stranded RNA while translocating along the nucleic polymer. Due to these functions, NS3 is an important target for antiviral development yet the biophysics of this enzyme are poorly understood. Microsecond-long molecular dynamic simulations of the dengue NS3 helicase domain are reported from which allosteric effects of RNA and NTPase substrates are observed. The presence of a bound single-stranded RNA catalytically enhances the phosphate hydrolysis reaction by affecting the dynamics and positioning of waters within the hydrolysis active site. Coupled with results from the simulations, electronic structure calculations of the reaction are used to quantify this enhancement to be a 150-fold increase, in qualitative agreement with the experimental enhancement factor of 10-100. Additionally, protein-RNA interactions exhibit NTPase substrate-induced allostery, where the presence of a nucleoside (e.g. ATP or ADP) structurally perturbs residues in direct contact with the phosphodiester backbone of the RNA. Residue-residue network analyses highlight pathways of short ranged interactions that connect the two active sites. These analyses identify motif V as a highly connected region of protein structure through which energy released from either active site is hypothesized to move, thereby inducing the observed allosteric effects. These results lay the foundation for the design of novel allosteric inhibitors of NS3.

[1] Russell B. Davidson[a], Josie Hendrix[a], Brian J. Geiss[b,c], Martin McCullagh[a]; [a] Department of Chemistry, Colorado State University, Fort Collins, CO, USA, [b] Department of Microbiology, Immunology, and Pathology, Colorado State University, Fort Collins, CO, USA, [c] School of Biomedical Engineering, Colorado State University, Fort Collins, CO, USA

## 2.2 Author Summary

Non-structural protein 3 (NS3) is a *Flaviviridae* (e.g. Hepatitis C, dengue, and Zika viruses) helicase that unwinds double stranded RNA while translocating along the nucleic polymer during viral genome replication. As a member of superfamily 2 (SF2) helicases, NS3 utilizes the free energy of nucleoside triphosphate (NTP) binding, hydrolysis, and product unbinding to perform its functions. While much is known about SF2 helicases, the pathways and mechanisms through which free energy is transduced between the NTP hydrolysis active site and RNA binding cleft remains elusive. Here we present a multiscale computational study to characterize the allosteric effects induced by the RNA and NTPase substrates (ATP, ADP, and $P_i$) as well as the pathways of short-range, residue-residue interactions that connect the two active sites. Results from this body of molecular dynamics simulations and electronic structure calculations are highlighted in context to the NTPase enzymatic cycle, allowing for development of testable hypotheses for validation of these simulations. Our insights, therefore, provide novel details about the biophysics of NS3 and guide the next generation of experimental studies.

## 2.3 Introduction

Flaviviruses (family *Flaviviridae*) are small ($\sim$11 kilobases) positive-sense, single-stranded RNA (ssRNA) viruses that include members such as dengue (serotypes 1-4), Zika, West Nile, yellow fever, and Japanese Encephalitis viruses. The dengue virus (DENV) is a public health threat that causes serious morbidity and mortality globally[16,17]. Infection with DENV can result in “break-bone" fever, an extraordinarily painful disease with symptoms ranging from a mild fever to a fatal hemorrhagic syndrome[18]. There are approximately 50 million serious infections and 20,000 deaths each year, and dengue infections are a leading cause of mortality in children in a number of Latin and Asian countries[16]. Dengue viruses have re-emerged in the United States, and a growing number of locally acquired infections in Florida, Texas, and Hawaii have been reported over the last decade. Despite a reinvigorated effort due to the recent Zika epidemic[19], there are currently no approved small molecule antivirals to treat Flavivirus-induced diseases.

One of the primary antiviral targets in *Flaviviridae* is the nonstructural protein 3 (NS3), which plays a critical role in the viral replication cycle[20–30]. NS3 is a multifunctional protein found in all *Flaviviridae*, possessing an N-terminal serine protease domain responsible for proteolytically cleaving the viral polyprotein during translation[31] and a C-terminal helicase/nucleoside triphosphatase (NTPase)/RNA triphosphatase domain[32–37]. In a nucleoside triphosphate (NTP) hydrolysis-dependent mechanism, the NS3 helicase domain (NS3h) unwinds double-stranded RNA (dsRNA) while translocating along the nucleic polymer. These functions are required to resolve the dsRNA replication intermediate into fully-mature positive strand RNAs (see Ref. 38 for a recent review). Mutations in the NS3 helicase and NTPase active sites are seen to abrogate NS3 function as well as decrease viral survival[39–41], demonstrating the importance of these enzymatic functions to the flavivirus life cycle. Drugs identified to inhibit DENV NS3h suffer from specificity issues because they are either NTPase inhibitors[42] or RNA/DNA mimics such as ivermectin[28], suramin[29] or aurintricarboxylic acid[30]. Therefore, it is of interest to further elucidate the mechanism of DENV NS3h with molecular resolution to help identify new and specific target regions for antiviral therapeutics.

The *Flaviviridae* NS3h have been classified as a superfamily 2 (SF2) helicase (NS3/NPH-II subfamily; a DEx/H helicase) where the NTPase cycle (Figure 2.1) provides the free energy needed to unwind dsRNA and translocate along the nucleic substrate in a 3′ to 5′ direction[43]. Structurally, NS3h are monomeric helicases composed of three subdomains; subdomains 1 and 2 (red and orange in the inset of Figure 2.1) are RecA-like folds that are structurally conserved across all SF1 and SF2 helicases, whereas subdomain 3 (green) is unique to the NS3/NPH-II subfamily and contains some of the least conserved portions of the protein. In Figure 2.1, an adenosine triphosphate (ATP; purple) molecule is bound within the NTPase active site between subdomains 1 and 2. Also, an RNA substrate (blue) is bound within the RNA-binding cleft, separating subdomains 1 and 2 from subdomain 3. The 5′ terminus of the RNA is positioned at the top of the protein in Figure 2.1 and the ds/ss RNA junction is hypothesized to be just above this region of the protein.

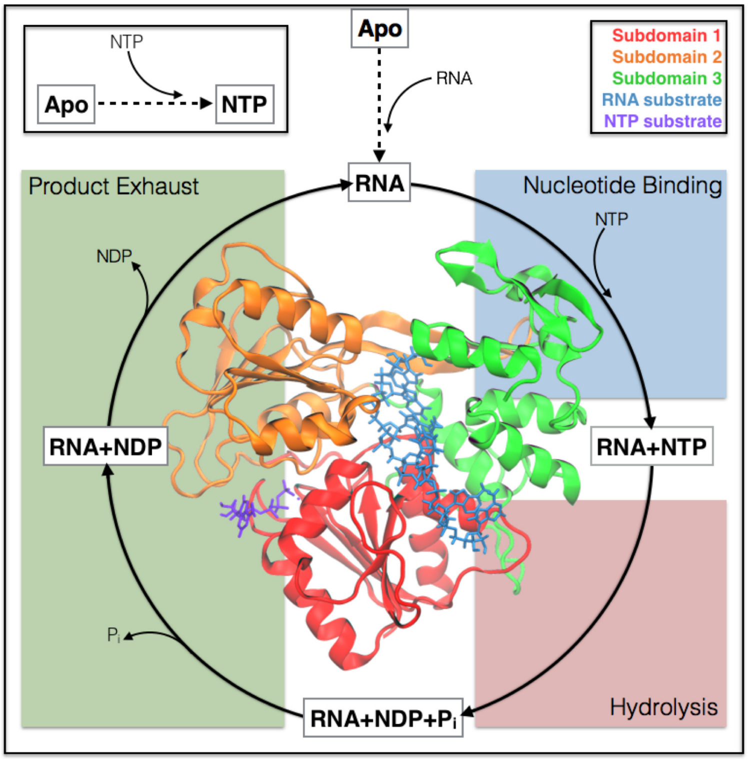

**Figure 2.1: The NTPase cycle of NS3h.** A schematic depicting the hypothesized substrate cycle that NS3h moves through during the NTPase function. Free energy released from this cycle powers the unwinding of dsRNA and unidirectional translocation along the nucleic polymer. The protein structure (inset) demonstrates the tertiary structure of NS3h as well as the positions of the RNA-binding cleft (ssRNA substrate colored blue) and the NTPase active site (ATP molecule colored purple).

The NS3/NPH-II subfamily of SF2 helicases exhibit both RNA-stimulated NTPase activity and NTPase-dependent helicase activity[32–37]. These experimentally observed phenomena suggest that (1) the presence of RNA affects the NTPase active site, thereby activating the NTPase cycle and (2) this cycle is the source of free energy needed to perform work on the RNA (translocation and unwinding). In Figure 2.1, the enzymatic cycle for the NTPase function is depicted by four dynamic events: RNA is bound within the RNA-binding cleft and activates the NTPase cycle, NTP binds, NTP is hydrolyzed, and finally products (nucleoside diphosphate – NDP – and inorganic phosphate – $H_2PO_4^-$, $P_i$) are released. To date, it is unclear which stage(s) of the cycle are responsible for the translocation and unwinding functions of NS3h. Furthermore, the biophysical couplings between NTPase and helicase active sites are still poorly understood[43].

One of the better studied *Flaviviridae* NS3h is that of the Hepatitis C virus (HCV; family: *Flaviviridae hepacivirus*)[44–55]. Utilizing both ensemble[44–50] and single molecule[51–53,56,57] techniques, studies have provided insights into the kinetic steps of the HCV NS3h translocation function. These studies, alongside crystallography studies of various *Flaviviridae* NS3h, suggest that the NS3 enzyme tracks along the phosphodiester backbone of the nucleic oligomer, unwinding one base-pair per hydrolysis event[50–52]. To explain these experimental results, various models describing the translocation mechanism have been reported, depicting NS3h as a Brownian[48–50] or backbone stepping motor[51,54–56] protein. These models envision the coupling between NTPase and helicase functions through different biophysical mechanisms, yet the models are not mutually exclusive and are limited in temporal and spatial resolution[58,59].

Luo *et al.* reported a set of crystal structures of the DENV NS3h in important protein-substrate complexes of the NTPase cycle (bolded text in Figure 2.1)[60]. From these structures, major allosteric influences of RNA-binding were seen in the NTPase active site. For example, Luo and coworkers noted that the presence of an RNA substrate shifts the carboxylate group of Glu285 (motif II) into a more catalytically relevant structure for the hydrolysis reaction. Mutation of the Glu285 residue abrogates NTPase and helicase activities[40]. These static structures have provided novel insights into RNA-induced protein structural changes yet provide limited insight into the NTPase cycle or translocation and unwinding functions of NS3h.

Previous theoretical studies of helicases have focused on a broad range of enzymes such as PcrA (SF1)[61–64], transcription terminator Rho (SF5)[65], SV40 (SF3)[66], and various NS3h enzymes[67–71]. Of the theoretical studies on NS3h, Perez-Villa *et al.* reported microsecond-long molecular dynamics (MD) simulations of the HCV NS3h-ssRNA systems in the presence and absence of ATP and ADP. The reported simulations were used to interrogate the thermodynamics of these substrate states with various conformations of the NTPase active site[67]. While the reported results are of interest for NS3h, the authors provide limited insight into the molecular mechanisms at play during the NTPase cycle. Other theoretical studies of the NS3h enzyme are limited in timescales (tens to hundreds of ns of simulation), substrate states modeled, or spatial resolution (e.g. coarse grained elastic network model)[68–71]. Therefore, theoretical modeling of the NS3h enzyme has yet to elucidate further details about the structural and dynamic couplings within NS3h in light of the NTPase cycle.

We report here a multiscale theoretical study of the DENV NS3h enzyme at each substrate state along the NTPase cycle. RNA-induced allostery on the NTPase active site is reported wherein the presence of an RNA substrate alters the positioning and dynamics of waters within the hydrolysis active site. Inspired by this observation, minimum energy electronic structure calculations are performed to investigate the energy landscape of the hydrolysis reaction. Additionally, investigations into NTPase substrate-induced allostery on the RNA-binding cleft suggest that NS3h interacts with RNA in a NTPase substrate-dependent manner. Umbrella sampling (US) simulations are performed to enhance the sampling of a proposed elementary step of the translocation mechanism observed during the unbiased simulations. Finally, analyses of the correlated motions between residues are used to identify allosteric pathways that connect the two active sites. It is through these pathways that we hypothesize that free energy released during the NTPase cycle is transduced to the RNA-binding cleft and utilized to perform work on the RNA. This study of the substrate states of DENV NS3h lays the foundation for further study of the NTPase cycle and marks the most complete picture of the molecular mechanism of the NS3 NTPase/helicase to date.

## 2.4 Methods and Models

### 2.4.1 Starting Structures and System Preparation

A subset of the crystal structures reported by Luo *et al.*[60] of the Dengue NS3h (serotype 4) are used as the initial structures for all-atom, explicit solvent MD simulations. Specifically, the binary complex of NS3h with a seven-residue ssRNA substrate (PDB ID: 2JLU) is used to model the ssRNA substrate state, while the ternary structures of ssRNA+ATP (2JLV), ssRNA+ADP+P$_i$ (2JLY), and ssRNA+ADP (2JLZ) model the pre-hydrolysis, post-hydrolysis, and product release states of the NTPase cycle, respectively. The Apo (2JLQ) and ATP (2JLR) substrate states are also simulated and used as experimental controls for our investigation into allostery.

The RNA-bound structures of DENV NS3h were crystalized as dimers of the protein[60]. For these systems, chain A of the structure is used as the starting conformation. Furthermore, the A conformers are chosen for residues with multiple side chain conformations. In all crystal structures with ATP substrates, the crystalized $Mn^{2+}$ divalent cation is converted into a $Mg^{2+}$. For the ATP crystal structure (2JLR), residues of the protease linker region were poorly resolved and so are transferred from the Apo (2JLQ) structure after aligning the neighboring amino acid backbones in both systems.

### 2.4.2 Molecular Dynamics Simulations

All-atom, explicit solvent MD simulations are performed for the six substrate states of DENV NS3 and presented in Figure 2.1 (denoted Apo, ATP, ssRNA, ssRNA+ATP, ssRNA+ADP+$P_i$, and ssRNA+ADP). The simulations are performed using the GPU-enabled AMBER14 software[72], ff14SB[15] parameters for proteins, and ff99bsc0$_{\chi \text{ OL3}}$[73,74] parameters for RNA. Parameters for ATP[75], ADP[75], $P_i$ (provided in Supporting Information (SI)), and $Mg^{2+}$[76] are also used. For each system, the crystal structures are solvated in TIP3P water boxes with at least a 12 Å buffer between the protein and periodic images. Crystallographic waters are maintained. Sodium and chloride ions are added to neutralize charge and maintain a 0.10 M ionic concentration. The Langevin dynamics thermostat and Monte Carlo barostat are used to maintain the systems at 300 K and 1 bar. Direct nonbonding interactions are calculated up to a 12 Å distance cutoff. The SHAKE algorithm is used to constrain covalent bonds that include hydrogen[77]. The particle-mesh Ewald method[78] is used to account for long-ranged electrostatic interactions. A 2 fs integration time step is used, with energies and positions written every 2 ps. The minimum amount of simulation performed for each system is one trajectory of 1.5 $\mu$s, with the first 200 ns of simulation sacrificed to equilibration of the starting structures. Simulation of the ssRNA system is performed to 2 $\mu$s. For both the ATP and ssRNA+ATP systems, two 1.5 $\mu$s simulations are performed. The total amount of unbiased simulation reported here on the described structures is 12.5 $\mu$s.

### 2.4.3 Umbrella Sampling Simulations

US simulations are performed to enhance sampling of a hypothesized elementary translocation event wherein the biased collective variable is the distance between the central carbon of the guanidinium group of Arg387 to the phosphorous atom of phosphate 4 in the RNA. These simulations are run for the ssRNA, ssRNA+ATP, ssRNA+ADP+P$_i$, and ssRNA+ADP systems, using the same protocol as the unbiased simulations with the addition of a bias. For each substrate state, a minimum of 22 sampling windows are simulated for 50 ns each with harmonic wells positioned every 0.5 Å and ranging from 3.50 to 14.00 Å. Harmonic force constants are 20 kcal mol$^{-1}$ Å$^{-2}$. Further simulation and additional windows are run in regions of collective variable space with poor sampling. The weighted histogram analysis method (WHAM)[79] is used to analyze the results of these simulations, with bin sizes of 0.1 Å. Bootstrapping is used to approximate error bars for the probability density and free energy plots shown. The total amount of biased simulations reported here is 5.12 $\mu$s.

### 2.4.4 Electronic Structure Calculations

Electronic structure calculations are performed at the $\omega$B97X-D/6-31+G* level of theory[80] using the Guassian 09 version B.01 program[81]. The $\omega$B97X-D functional is chosen due to its broad applicability[82,83] and a recent study demonstrating its energetic accuracy for a variety of phosphate hydrolysis reactions[84]. The QM system is composed of a truncated ATP molecule (truncated to methyl triphosphate, MTP), functional groups of nine surrounding protein residues (Pro195, Gly196, Lys199, Glu285, Ala316, Gly414, Gln456, Arg460, and Arg463), a Mg$^{2+}$ ion, and seven water molecules. The amino acids are truncated at various positions (more detail in SI) using hydrogen atoms. For each residue, the position of the terminal heavy atom is frozen to maintain the active site geometry. This yielded a total of 138 atoms in the QM calculations.

These calculations are performed on active site conformations pulled from the unbiased MD simulations of the ssRNA+ATP and ATP substrate states, thereby investigating the influence of observed RNA structural allostery on the hydrolysis reaction mechanism and energy landscape.

Frames used for the initial reactant state structures were selected by visualizing MD frames in which a lytic water is present. Through visual and RMSD analyses of such frames, a single frame was chosen to represent the population of catalytically relevant structures. The hydrolysis reaction is then monitored by optimizing the reactants (MTP+lytic water), products (MDP+$HPO_4^{2-}$), and a single transition state (TS) in between. The initial TS and product state structures were created from the previous optimized structure. The minima are confirmed using a Hessian calculation. The TS is confirmed by examining the direction of the single imaginary frequency. Following geometry optimization, frequency calculations are performed to obtain gas-phase, zero-point energy corrected free energies for each active site conformation.

### 2.4.5 Data Analysis

Unless stated otherwise, analyses of MD trajectories are performed using Python 2.7 and the MDAnalysis module (version 0.15.0)[85]. Matplotlib is used for plotting data[86]. VMD is used for visualization of trajectories and production of structural figures[87–89]. For each substrate state, a single frame from the trajectories is used when presenting structural details of the respective substrate state. Further information on choosing these “exemplar” structures is given in the SI. Additionally, details of all analyses performed can be found in the SI. All scripts for the analyses are available on Github (https://github.com/mccullaghlab/DENV-NS3h).

## 2.5 Results and Discussion

For clarity, we present and discuss our results in three sections. The first and second sections independently report observed RNA-induced and NTPase substrate-induced structural allosteries, respectively. The focus of the RNA-induced allostery section is on the structural changes seen in the NTPase active site due to bound RNA. Similarly, the NTPase substrate-induced allostery section highlights changes seen in the structure and dynamics of the RNA-binding cleft due to the presence of different nucleoside substrates. In the final section, correlated motions between residues are used to highlight pathways through which these structural allosteric effects are induced.

### 2.5.1 RNA-Induced Allostery

To date, no biophysical explanation has been proposed for the 10 to 100-fold increase in NTPase turnover rate observed for DENV NS3h in the presence of RNA[37]. Crystallographic studies of the DENV NS3h structure have identified static structural allostery due to RNA binding[60], yet a dynamic picture and interpretation of these influences are still missing. In this section, comparisons of the simulations of the Apo, ATP, ssRNA, and ssRNA+ATP substrate states are used to depict structural rearrangements induced by RNA. These RNA-induced allosteries are observed to affect the positioning and dynamics of waters within the NTPase active site. These novel insights gained from the comparisons of the MD simulations inspire the reported electronic structure calculations of the reactant, transition, and product states of the hydrolysis reaction. In combination, these results demonstrate that the observed enhancement of NTPase activity originates from the RNA-induced destabilization of the lytic water.

**The RNA-binding Loop and $\alpha 2$.**

The most marked structural difference between DENV NS3h with or without ssRNA is the change in conformation of the RNA-binding loop (L$\beta 3 \beta 4$; Thr244 to Glu255). The crystal structures of DENV NS3h from Luo *et al.*[60] with no RNA present (Apo, 2JLQ; ATP, 2JLR) resolve this loop in a “closed” conformation while the crystal structures with bound RNA all have this loop in an “open” conformation. Figure 2.2(A) depicts both conformations and the relative position of the loop with respect to the RNA-binding cleft and NTPase active site. In the “closed” conformation, the RNA-binding loop is covering part of the RNA-binding cleft while, in the “open” conformation, this loop contacts the phosphodiester backbone of the RNA as well as amino acids of $\alpha$-helix 2 ($\alpha 2$). Transitions from “closed” to “open” conformations are not sampled during our MD simulations of the Apo and ATP systems demonstrating that the crystal structure conformations are minima in the solution phase free energy surfaces.

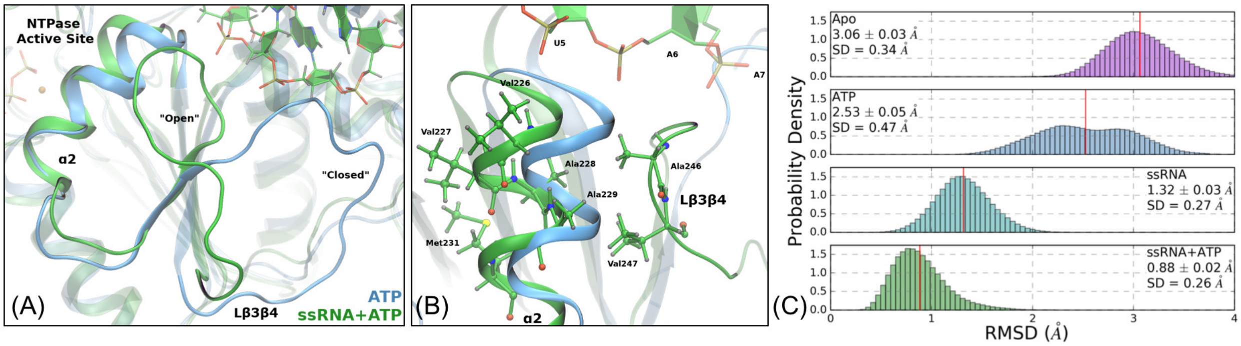

**Figure 2.2: RNA-induced displacement of L$\beta 3\beta 4$ and $\alpha 2$.** (A) Depiction of the "open" and "closed" structural states of L$\beta 3\beta 4$ for exemplar structures of ATP (blue) and ssRNA+ATP (green) simulations. (B) Hydrophobic interactions between L$\beta 3\beta 4$ and $\alpha 2$ stabilize the "open" conformation. Furthermore, Val227 and Met231 ($\alpha 2$) are pushed in towards the NTPase active site when L$\beta 3\beta 4$ is in the "open" conformation. (C) RMSD of $\alpha 2$ (residues 224 to 235) backbone atoms referenced against the ssRNA+ATP crystal structure (PDB ID: 2JLV).

The RNA-induced structural change of L$\beta 3\beta 4$ affects the position of $\alpha 2$ as highlighted in Figure 2.2(B), where the top of $\alpha 2$ is displaced in towards the NTPase active site when L$\beta 3\beta 4$ is in the "open" conformation. This conformation is stabilized by hydrophobic contacts between Ala246 and Val247 (L$\beta 3\beta 4$) and Val226, Ala228, Ala229 ($\alpha 2$). When in the "closed" conformation, this hydrophobic pocket is not formed and leaves the top of $\alpha 2$ exposed to solvent.

The structural deviation of $\alpha 2$ is quantified by computing the root mean square deviation (RMSD) of the backbone atoms of $\alpha 2$ (residues 224 to 235) relative to the ssRNA+ATP crystal structure (2JLV). The distributions of this metric are presented in Figure 2.2(C) with the largest structural deviations seen in the simulations of the Apo and ATP substrate states. Bound RNA decreases the RMSD values while an ATP substrate shows minimal influence. Therefore, these RNA-induced hydrophobic interactions between L$\beta 3\beta 4$ and $\alpha 2$ stabilize the structural conformation of $\alpha 2$ where the top of the helix is pushed in towards the NTPase active site. Interestingly, Val227 and Met231 are the residues in $\alpha 2$ that have prominent positions in the NTPase active site. While these hydrophobic side chains likely have minimal influence on the hydrolysis reaction mechanism, their structural shift into the hydrolysis active site reduces the volume of the pocket.

**Motif II.**

Motif II (Walker B) is a set of highly conserved amino acid residues within NTPase enzymes and is known to play an important role in the catalysis of the hydrolysis reaction[40]. In DENV and other *Flaviviridae*, motif II is the DEAH sequence (residues 284 to 287) where Asp284 and Glu285 are positioned in the rear of the NTPase active site. Luo *et al.* noted that the presence of RNA shifts the carboxylate group of Glu285 from a magnesium-bound position to a position more conducive to coordinating the lytic water[60]. In this RNA-induced position, Glu285 is ideally located to act as a base where it can accept a proton from the lytic water, thereby increasing the nucleophilicity of the attacking group during the hydrolysis reaction.

Our MD simulations maintain these starting conformations and support the deduced importance of Glu285. Snapshots of the Glu285 positions in the ATP and ssRNA+ATP simulations are shown in Figure 2.3 (A) and (B), respectively. The highlighted water demonstrates the position of a lytic water in the NTPase active site, relative to the $\gamma$-phosphorus atom. Structurally, with RNA bound, the carboxylate side chain of Glu285 is pulled away from the coordination sphere of the $Mg^{2+}$ cation and is moved into plane with the terminal phosphoanhydride bond. In either position, Glu285 is observed to hydrogen bond with the lytic water yet, in the RNA-induced position, the lytic water is positioned in a more ideal environment for nucleophilic attack (quantified in the next section).

Both Asp284 and Glu285 are major structural landmarks within the NTPase active site and have no direct interactions with the RNA substrate. Rather, the origin of the RNA-induced structural rearrangement of motif II residues is attributed to RNA-induced displacement of residues down the linear amino acid sequence, specifically Phe288 and Asp290. Figure 2.3 (C) shows the structural alignment of the ATP and ssRNA+ATP structures (same frames as in panels (A) and (B)), focusing on residues Glu285 to Asp290. The structural deviations of the residues highlighted in Figure 2.3 (C) are quantified with an RMSD analysis of the backbone atoms of residues 284 to 290, referenced against the ssRNA+ATP crystal structure (Figure 2.3 (D)). There is a shift of $\sim 1.3$ Å in these atoms when comparing RNA-bound systems (ssRNA, ssRNA+ATP) and no RNA systems (Apo, ATP).

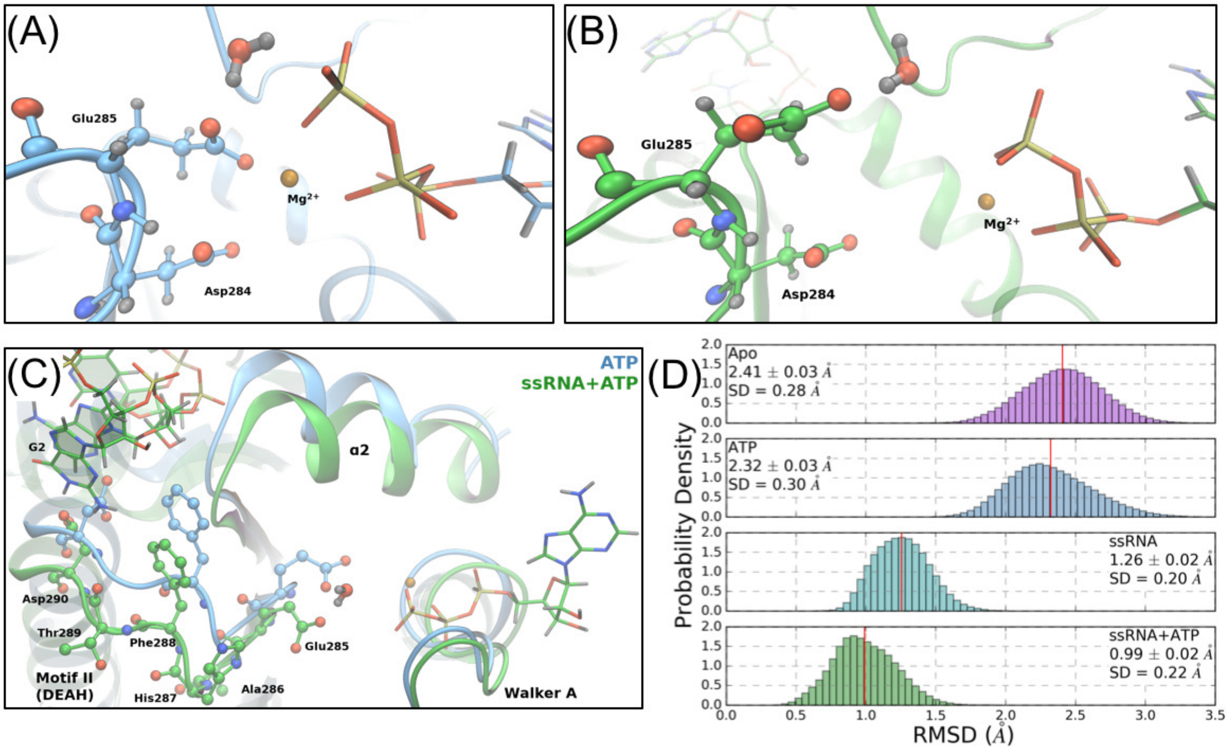

**Figure 2.3: RNA-induced allostery on motif II.** The Asp284 and Glu285 positioning relative to the $\gamma$-phosphate of the ATP molecule, for the ATP (A) and ssRNA+ATP (B) systems. In each panel, the highlighted water molecule is identified as the most lytic-like water within the active site. (C) Structural alignment of the same frames shown in (A) and (B), highlighting the RNA-induced backbone shift of residues Glu285 to Asp290. Phe288 and Asp290 are highlighted in both systems due to their prominence in the RNA-binding cleft. (D) RMSD of the backbone atoms of residues 284 to 290 referenced against the ssRNA+ATP crystal structure (PDB ID: 2JLV).

Therefore, bound RNA causes a backbone shift of the post-motif II residues (e.g. Phe288, Asp290) that propagates to the residues within the NTPase active site.

### Water positioning and dynamics within the NTPase active site.

RNA allosterically affects the positions of amino acids within the NTPase active site, yet it is unclear how these structural rearrangements influence the hydrolysis cycle. When comparing the ATP and ssRNA+ATP simulations, the positions and dynamics of the ATP molecule and $Mg^{2+}$ cation are minimally affected by the presence of RNA. Alternatively, waters within the NTPase active site are observed to be greatly influenced by the presence of bound RNA. For example, the average number of water molecules found within the NTPase active site decreases from $30.0 \pm 0.7$

molecules for the Apo substrate state to $21.72 \pm 0.08$ molecules in the ssRNA state. A similar but reduced trend is observed when comparing the ATP ($15.0 \pm 0.2$ water molecules) and ssRNA+ATP ($12.8 \pm 0.4$) simulations.

The translational and rotational dynamics of water molecules within the NTPase active site are also influenced by bound RNA, as shown graphically in Figure 2.4(A) and (B). The mean squared displacement (MSD, panel (A)) is a metric describing the average squared distance traveled by water molecules within the NTPase active site over a time interval, where large slopes indicate fast diffusion of water. The MSD metric for the Apo substrate state (purple) has a large slope relative to the ssRNA substrate state, demonstrating that waters in the NTPase active site diffuse more slowly when an RNA is bound within the binding cleft. Although much less dramatic, a similar trend is seen in the ATP and ssRNA+ATP states. The O-H bond autocorrelation metric (panel (B)) describes the rotational motions of water molecules within the active site, thereby looking at water reorientation; slower decay of this metric indicates slower reorientation times. Similar to the MSD results, the ssRNA-bound systems have extended O-H bond correlation times relative to the control states (Apo and ATP), indicating that rotational motions of water molecules within the NTPase active site are slowed by the RNA.

Considering the hypothesized SN2 mechanism of the hydrolysis reaction[90,91], ideal nucleophilic attack by a lytic water on the $\gamma$-phosphorous atom ($P_\gamma$) of ATP is described by an attack angle of 180° with respect to the terminal phosphoanhydride ($P_\gamma$-$O_{\beta,\gamma}$) bond. The distance between the lytic water oxygen ($O_{\text{wat}}$) and $P_\gamma$ will decrease to a bonded distance of $\sim$ 1.7 Å over the course of this reaction. Therefore, the $P_\gamma$-$O_{\text{wat}}$ distance and $O_{\beta,\gamma}$-$P_\gamma$-$O_{\text{wat}}$ angle are used as geometric collective variables that describe the nucleophilic attack of a lytic water. Projecting the positions of waters within the NTPase active site onto these two coordinates allows for comparisons of the positioning of catalytically relevant water in the ATP and ssRNA+ATP substrate simulations. The two-dimensional heat maps of this projection are shown in the SI (Figures Figure 2.16 and Figure 2.17) for both of the substrate states.

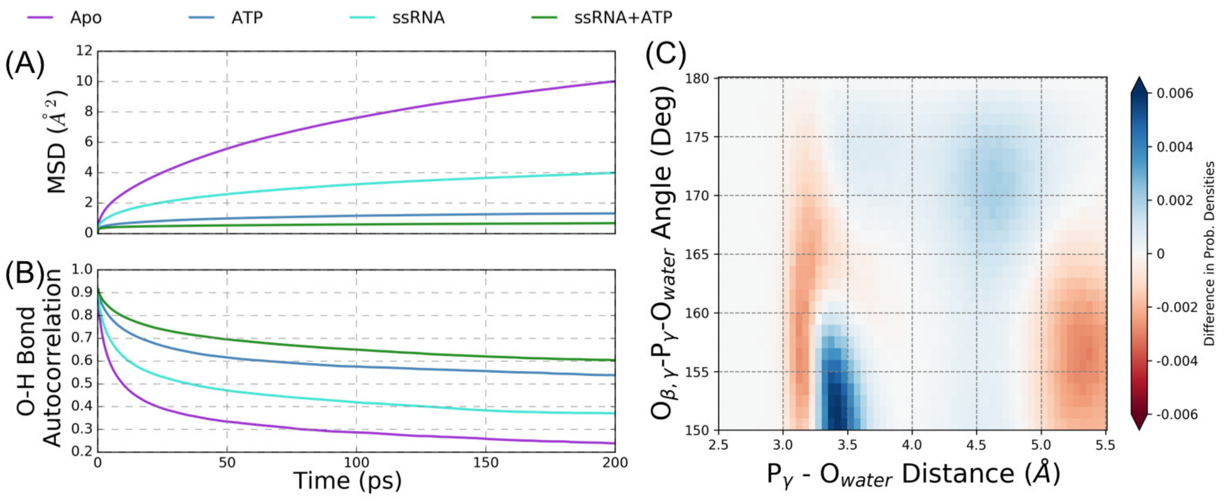

**Figure 2.4: Water dynamics and positioning within the NTPase active site.** (A) Mean square displacement (MSD) and (B) O-H bond autocorrelation metrics for the Apo, ATP, ssRNA, and ssRNA+ATP simulations that describe the translational and rotational motions of waters within the active site. (C) The difference between the ssRNA+ATP and ATP probability densities of water positions within the NTPase active site, projected onto the $O_{\beta,\gamma}$-$P_\gamma$-$O_{\text{wat}}$ angle and $P_\gamma$-$O_{\text{wat}}$ distance. These axes are used to project water positions into catalytically relevant space relative to the ideal position of a lytic water in the hydrolysis reaction.

The difference between the probability densities for the ssRNA+ATP and ATP simulations is shown in Figure 2.4(C), where positive values (blue) correspond to increased probability density in ssRNA+ATP versus ATP states. Therefore, the presence of RNA causes water molecules in lytic positions of the NTPase active site to shift into more ideal (larger) nucleophilic angles while pushing competing waters at short distances to lower angles. Motivated by the electronic structure calculations reported in the next section, geometric cutoffs are used to quantify these observations by defining a conical volume of the NTPase active site within which waters are identified as lytic: waters with a $P_\gamma$-$O_{\text{wat}}$ distance less than 5.0 Å and an $O_{\beta,\gamma}$-$P_\gamma$-$O_{\text{wat}}$ angle greater than 155° are defined as lytic. The probability of observing a frame with water in a lytic position is 72.93% ± 0.07% for the ATP system and 79.08% ± 0.09% for the ssRNA+ATP system.

In total, these results demonstrate that RNA affects the dynamics and positioning of waters within the NTPase active site. These effects are propagated from the RNA binding cleft to the NTPase active site through structural rearrangements of $L\beta3\beta4$, $\alpha2$, and motif II. Although it is difficult to fully deconvolute the specific influences of these structural allosteries on the water molecules in the active site, we propose that the observed influence of RNA on number and dynamics of water molecules originates from the structural rearrangement of $\alpha 2$, where Val227 and Met231 become more prominent in the hydrolysis active site when RNA is bound. These hydrophobic residues not only exclude water molecules from the active site but also slow the translational and rotational motions of water molecules. This RNA-induced effect can be thought of as a entropic destabilization of the NTPase active site, where the RNA decreases the phase space that the water molecules can populate. Furthermore, the RNA-induced structural rearrangement of the Glu285 carboxylate group leads to the observed increased in probability of lytic water molecules. Through the backbone displacement of motif II residues, the Glu285 side chain is pulled away from the $Mg^{2+}$ cation and into plane of the $\gamma$-phosphate group, thereby creating a local protein environment that stabilizes water molecules into more ideal positions for nucleophilic attack. This effect is interpreted as a direct destabilization of the lytic water in the hydrolysis reaction.

**Electronic Study of the NTP Hydrolysis Reaction.**

The impact of the RNA-induced repositioning of the lytic water on the hydrolysis reaction is investigated using density functional theory (DFT) calculations of an abbreviated NTPase active site where conformations are pulled from the unbiased MD simulations. Active site geometry optimizations are performed on the ATP and ssRNA+ATP substrate states where the hydrolysis reaction is modeled as a concerted SN2 mechanism using a reactant state (ATP*), a transition state (TS), and a product state ($HPO_4^{2-}$). Geometry optimized potential energies and gas phase free energy corrections are used to compute the free energy landscape of the hydrolysis reaction for the respective substrate state, as presented in Figure 2.5. Figure S1 highlights the full selection of the NTPase active site (amounting to 138 atoms) that are included in the DFT calculations. For clarity, the geometries presented in Figure 2.5 only include the triphosphate, lytic water, $Mg^{2+}$, and Glu285 atoms.

The free energy landscape of the ATP substrate state is presented in Figure 2.5(A), where the reactant structure has the lytic water 3.41 Å away from the gamma phosphorus and at an angle of 157° between the water oxygen and the $O_{\beta,\gamma}$-$P_\gamma$ bond. The TS structure is found to be 30.5

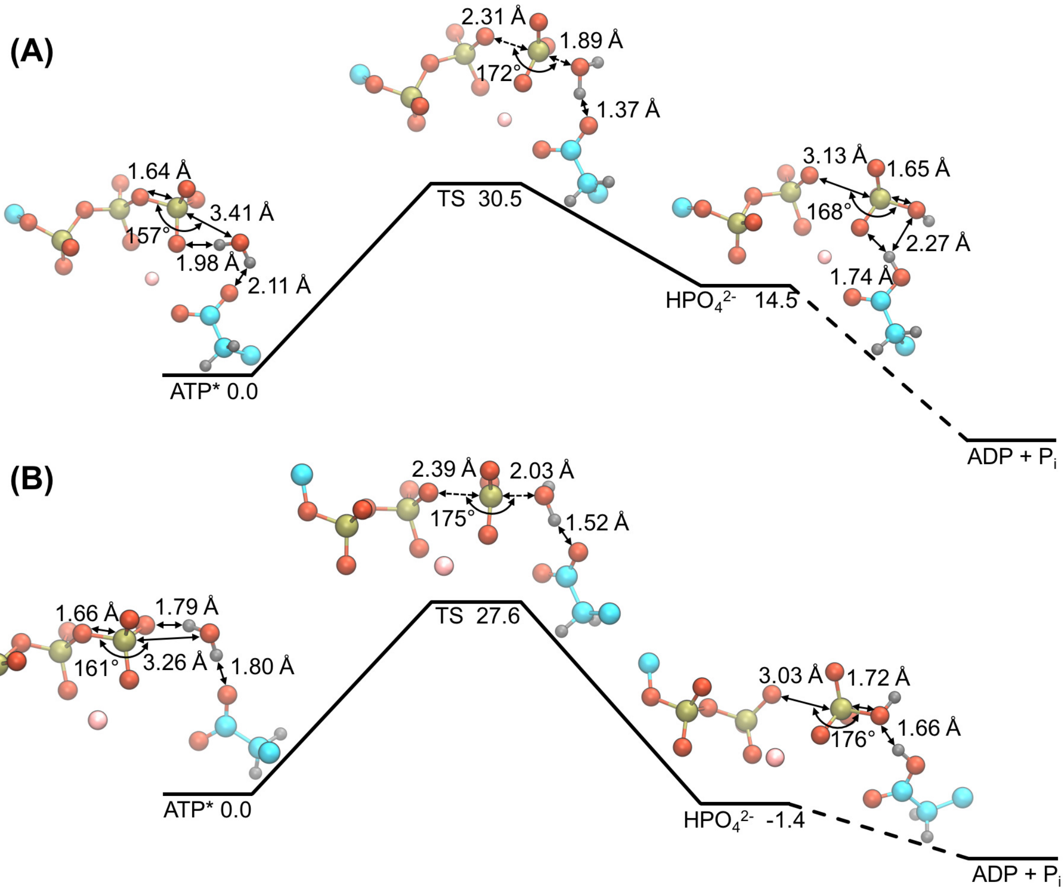

**Figure 2.5: Energy landscape and structures of the NTP hydrolysis reaction in the active site of DENV NS3h for the ATP (A) and ssRNA+ATP (B) substrate states.** DFT calculations were performed using the $\omega$B97X-D/6-31+G* level of theory. A total of 138 atoms were included in the quantum mechanical calculations (see supporting information for full structures) but only the triphosphate, lytic water, $Mg^{2+}$, and Glu285 side chain atoms are shown here for clarity. Important distances and angles are included in the structural representations of each state. All energies are reported in units of kcal mol$^{-1}$.

kcal mol$^{-1}$ above the reactant state, with a $P_\gamma$-$O_{wat}$ distance of 1.89 Å and $O_{\beta,\gamma}$-$P_\gamma$-$O_{wat}$ angle of 172°. Additionally, the $O_{\beta,\gamma}$-$P_\gamma$ bond distance has increased from 1.64 Å to 2.31 Å. Over the transition, the lytic water molecule reorients relative to the $\gamma$-phosphate and Glu285 atoms. Via this reorientation, a proton from the lytic water is partially transferred to the carboxylate group of Glu285, as seen in the difference in the water O-H bond distance between TS and reactant states ($\Delta O_{wat}$-$H_{wat}$ = d($O_{wat}$-$H_{wat}$)$_{TS}$ - d($O_{wat}$-$H_{wat}$)$_{ATP*}$ = 0.12 Å). The product state was found following the TS in which the proton has completely transferred to Glu285, forming the $HPO_4^{2-}$ molecule.

Panel (B) of Figure 2.5 depicts the hydrolysis reaction landscape and structures for the ssRNA+ATP substrate state. The reactant structure has the lytic water 3.26 Å away from the gamma phosphorus and at an $O_{\beta,\gamma}$-$P_\gamma$-$O_{\text{wat}}$ angle of 161°. In the TS structure, the $P_\gamma$-$O_{\text{wat}}$ and $O_{\beta,\gamma}$-$P_\gamma$ distances become 2.03 Å and 2.39 Å, respectively, while the $O_{\beta,\gamma}$-$P_\gamma$-$O_{\text{wat}}$ angle increases to 175°. Unlike in the ATP substrate state, the O-H bond distance of the lytic water is minimally perturbed when comparing the reactant and TS structures ($\Delta O_{\text{wat}}$-$H_{\text{wat}}$ = 0.04 Å). Rather, the proton transfer step is completed during the transition from the TS to the product state. Overall, the calculated activation barrier height for the ssRNA+ATP substrate state is 27.6 kcal mol$^{-1}$, corresponding to a 2.8 kcal mol$^{-1}$ decrease in barrier height relative to the ATP substrate state landscape.

The overall reaction, ATP$(aq)$ $\xrightarrow{\text{NS3h}}$ ADP$(aq)$ + P$_i(aq)$, is expected to be exergonic for both substrate states due to NS3h being hydrolysis active in the presence and absence of RNA[37]. While the energy landscape for the ATP substrate state (panel (A)) does not demonstrate an exergonic reaction, we hypothesize that neither product states presented in Figure 2.5 adequately model the final product state of the hydrolysis reaction. This hypothesized thermodynamic product state requires a proton transfer from Glu285 to the $HPO_4^{2-}$ molecule as well as an unbinding event of the $HPO_4^{2-}$ molecule from the $Mg^{2+}$ coordination sphere. Optimization of such a product state is infeasible due to the limited description of the protein environment in these DFT calculations. Additionally, it is assumed that the energy barriers of these subsequent events are much smaller than the hydrolysis reaction barrier and so, are disregarded in the current study.

The calculated activation barrier heights of both substrate states are in good agreement with previous DFT studies of NTP hydrolysis in a protein environment[91–97] and in aqueous solution[98–100]. The differences observed between the ATP and ssRNA+ATP free energy landscapes of the hydrolysis reaction are mainly attributed to the different positions of the Glu285 carboxylate group. For either substrate state, this functional group acts as a base that increases the nucleophilicity of the lytic water. When unbound from the $Mg^{2+}$ coordination sphere, Glu285 performs this function more effectively, stabilizing the lytic water at a shorter $P_\gamma$-$O_{\text{wat}}$ distance and a larger $O_{\beta,\gamma}$-$P_\gamma$-$O_{\text{wat}}$ angle. Additionally, comparisons of the associative ($P_\gamma$-$O_{\text{wat}}$ distance) and dissociative ($O_{\beta,\gamma}$-$P_\gamma$

distance) reaction coordinates suggest that the RNA-induced structural rearrangement of the active site leads to slight changes in the hydrolysis mechanism. For both substrate states, the change in $P_\gamma$-$O_{wat}$ distance from reactant to TS states (ATP: $\Delta P_\gamma$-$O_{wat}$ = -1.52 Å; ssRNA+ATP: $\Delta P_\gamma$-$O_{wat}$ = -1.23 Å) is larger in magnitude than the respective change in $O_{\beta,\gamma}$-$P_\gamma$ distance (ATP: $\Delta O_{\beta,\gamma}$-$P_\gamma$ = 0.67 Å; ssRNA+ATP: $\Delta O_{\beta,\gamma}$-$P_\gamma$ = 0.73 Å). These values suggest that the hydrolysis reaction proceeds through an asynchronous SN2 hydrolysis mechanism where the nucleophilic attack by the lytic water is enhanced by the local protein environment[101,102]. Further comparison of these values demonstrates that the hydrolysis reaction for the ATP substrate state proceeds through a more "associative"[102] transition state than does the reaction for the ssRNA+ATP substrate state. The changes in the reaction coordinates discussed above as well as the changes in the $O_{wat}$-$H_{wat}$ distance (ATP: $\Delta O_{wat}$-$H_{wat}$ = 0.12 Å; ssRNA+ATP: $\Delta O_{wat}$-$H_{wat}$ = 0.04 Å) and the $O_{Glu285}$-$H_{wat}$ distance (ATP: $\Delta O_{Glu285}$-$H_{wat}$ = -0.74 Å; ssRNA+ATP: $\Delta O_{Glu285}$-$H_{wat}$ = -0.28 Å) collectively suggest that the ssRNA+ATP TS structure is more reactant-like than the TS structure of the ATP substrate state.

**Theoretical Enhancement Factor of the RNA-Stimulated NTPase Activity.**

The results from the last two subsections are consistent with the hypothesis that the biophysical origin of the experimentally observed RNA-stimulated NTPase activity[37] is two-fold: (1) the RNA-induced structural changes of L$\beta3\beta4$, $\alpha2$, and motif II affect the probability of water molecules to be located in lytic positions, and (2) the same RNA-induced structural changes alter the activation barrier of the hydrolysis reaction. To account for both effects, a reaction scheme is proposed to describe the NTPase function within NS3h, where a fast equilibrium exists between active site conformations with and without the presence of a lytic water. This equilibrium is followed by the slow, irreversible hydrolysis reaction. Using this scheme, the theoretical observed rate constant for the NTP hydrolysis reaction is $k_{obs} = K_{eq}k_{hydrol}$. For both ATP and ssRNA+ATP substrate states, the $K_{eq}$ is defined as the respective ratio of probabilities of observing a MD frame with and without a lytic water. The hydrolysis rate constant is quantified using an Arrhenius rate equation where the Boltzmann factor accounts for the activation energy barrier observed in the electronic structure calculations. The ratio of observed rate constants is defined as the theoretical enhancement factor of the RNA-stimulated NTPase activity,

$$\text{Enhancement Factor} = \frac{k_{obs}^{ssRNA+ATP}}{k_{obs}^{ATP}} = \frac{K_{eq}^{ssRNA+ATP}}{K_{eq}^{ATP}} \exp\left(-\Delta E_a/RT\right)$$

where $\Delta E_a$ is the difference in activation energies between the ssRNA+ATP and ATP substrate states ($E_a^{ssRNA+ATP} - E_a^{ATP}$). The Arrhenius prefactor is assumed to be constant for both ATP and ssRNA+ATP substrate states and thus does not contribute to the enhancement factor.

Taking the ratio of the ssRNA+ATP and ATP rates results in a theoretical enhancement factor of 150, which is consistent with the experimentally observed enhancement factor (10 to 100)[37]. Error analyses for this calculation are at least three fold: (1) statistical error from the sampling in MD simulations, (2) error in the force field, and (3) error in the DFT calculations. Propagation of the statistical uncertainties of the probability of observing a lytic water lead to an error in the enhancement factor of 0.002. The errors in the force field and the DFT energies[103] are difficult if not impossible to estimate. Uncertainty of the enhancement factor is comprised mainly of errors in the activation energies obtained from the DFT calculations since they are present in the exponential term in the Arrhenius equation. Therefore, an error of 1 kcal mol$^{-1}$ in the activation barriers is used as a conservative estimate for the DFT method. Using these approximated uncertainties, the minimum and maximum values for the calculated enhancement factor are observed to bound the best estimate (150) by an order of magnitude on both sides. Even with this large range, the calculated enhancement factor maintains the qualitative narrative that RNA-induced structural allostery of L$\beta3\beta4$, $\alpha2$, and motif II leads to the repositioning of the lytic water within the NTPase active site as well as affects the energetics of the hydrolysis reaction.

### 2.5.2 NTPase Substrate-Induced Allostery

Experimental studies have shown that the NS3h helicase functions (translocation and unwinding) are NTPase dependent, yet it is unclear which equilibrium states and/or dynamic events of the NTPase cycle are the source of the necessary free energy for these functions[35,36]. All previously developed models describing these functions have deduced that the NTPase cycle drives conformational changes in the RNA-binding cleft, thereby cycling the protein-RNA interactions leading to unidirectional translocation and melting of the duplex/single stranded nucleic junction[48–51,54–56]. Yet, limited structural allostery attributed to the NTPase substrates (e.g. ATP, ADP, and $P_i$) is observed in the crystal structures of DENV NS3h[60]. Therefore, a subset of the MD simulations reported here (ssRNA, ssRNA+ATP, ssRNA+ADP+$P_i$, and ssRNA+ADP) is used to interrogate protein-RNA interactions as well as identify protein structural changes that have NTPase substrate-dependent behaviors.

### Protein-RNA Contacts.

For the RNA-bound substrate states modeled here, the RNA oligomer is seven residues long with the first five residues (5′ end) strongly bound within the RNA-binding cleft of the protein. The remaining two nucleic residues are poorly resolved in the crystal structures[60] and are highly fluctional in the MD simulations. For the NS3/NPH-II helicase subfamily, amino acids in motifs Ia, Ib, IV, IVa, and V are observed in crystal structures to have contact with the phosphodiester backbone of the nucleic oligomer[54,60], yet little is known about the dynamic role that these residues play during translocation and unwinding of the nucleic substrate[43]. As observed in our MD simulations, protein-RNA contacts are dominated by electrostatic interactions between highly conserved residues in these motifs and the first four phosphate groups of the ssRNA, as highlighted in Figure 2.6. Specifically, arginines (225 and 387), threonines (224, 244, and 408), backbone amides (Arg225, Ile365, Arg387), and $\alpha$-helix dipole moments ($\alpha 2$ as well as subdomain 2 $\alpha$-helices 1, $\alpha 1'$, and 2, $\alpha 2'$)[104] are observed to stabilize RNA through interactions with the phosphate groups.

### Asymmetry of the Protein-RNA Interactions.

Nonbonding interaction energies are used to provide a quantitative description of the relative strength of the protein-RNA interactions. Comparisons of these data between substrate states of the NTPase cycle provide insight into the hypothesized NTPase dependence of protein-RNA interactions. The pair-wise sum of Lennard-Jones and short-range, unscreened electrostatic energies are

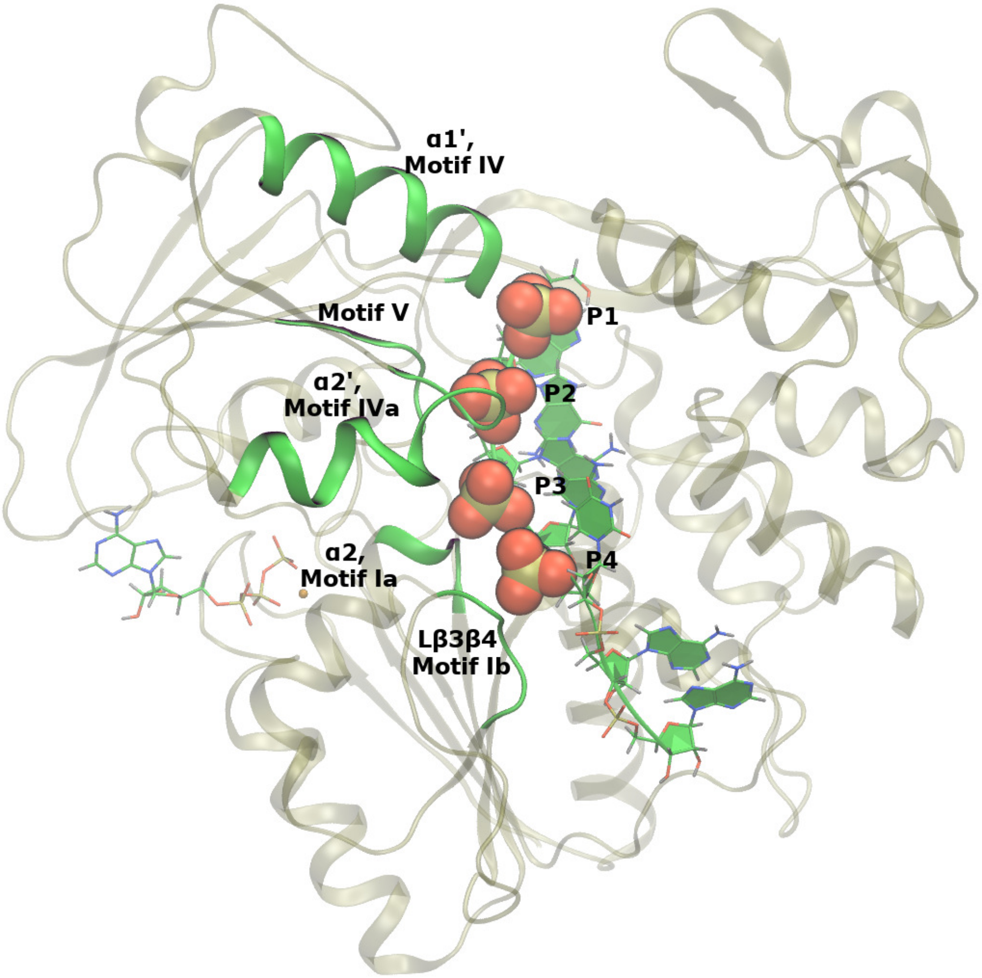

**Figure 2.6: Protein-RNA contacts.** Motifs Ia, Ib, IV, IVa, V of NS3h make strong contact with the phosphodiester backbone of the RNA. These contacts are dominated by electrostatic interactions between the phosphate groups of RNA and highly conserved amino acid residues. The four strongly bound phosphate groups (labeled P1 through P4) are highlighted with space filling representations.

calculated using the *lie* analysis function in AMBER14 cpptraj[105]. Table 2.1 shows the nonbonding interaction energies between RNA phosphates 1 through 4 (P1-4) (in total and individually) and all protein residues. An interaction cutoff of 12 Å was used for these calculations.

The totals of interaction energies between the protein and P1-4 demonstrate that the ssRNA+ATP (-625 ± 3 kcal mol$^{-1}$) and ssRNA+ADP+P$_i$ (-649 ± 5 kcal mol$^{-1}$) structures have more stable protein-RNA interactions than the ssRNA (-582 ± 6 kcal mol$^{-1}$) and ssRNA+ADP (-579 ± 4 kcal mol$^{-1}$) structures. These results suggest that the presence of the $\gamma$-phosphate group (or P$_i$) in the NTPase active site has a stabilizing effect on the strongly bound phosphate groups of the RNA.

**Table 2.1: Nonbonding interaction energies between RNA phosphate groups (named P1 through P4) and all protein residues.** Units for all values shown are kcal mol$^{-1}$. An interaction cutoff of 12 Å is used. Short-range, electrostatic energies were calculated with a dielectric of 1.

| | **ssRNA** | **ssRNA+ATP** | **ssRNA+ADP+P$_i$** | **ssRNA+ADP** |
|---|---|---|---|---|
| **P1-4** | -582 ± 6 | -625 ± 3 | -649 ± 5 | -579 ± 4 |
| **P1** | -62 ± 2 | -106 ± 2 | -126 ± 4 | -118 ± 3 |
| **P2** | -157 ± 2 | -192 ± 1 | -193 ± 1.3 | -158 ± 1.9 |
| **P3** | -206 ± 3 | -188.4 ± 0.4 | -190.2 ± 0.9 | -176.9 ± 0.5 |
| **P4** | -155 ± 3 | -139.3 ± 0.6 | -139.7 ± 0.9 | -126 ± 1.3 |

This result is in direct disagreement with experimental results of HCV NS3 where it is observed that the protein has a decrease in affinity for nucleic oligomers in the presence of ATP[47,49]. One possible reason for this discrepancy is the lack of a realistic nucleic polymer that extends above and below the RNA-binding cleft in the simulations reported here. Modeling more complex RNA structures is left for further study.

The interaction energies between protein atoms and individual phosphate groups presented in Table 2.1 allow for comparisons of local protein-RNA interactions in the various substrate states. Protein-phosphate 2 (P2) energies show similar trends as the total energies, where the presence of the $\gamma$-phosphate group stabilizes P2 interactions by $\sim$ 35 kcal mol$^{-1}$. Additionally, while similar in total energies, the ssRNA and ssRNA+ADP substrate states have very different local energies, suggesting different protein-RNA contacts. Specifically, the ssRNA substrate state has stronger interactions with P3 (-206 ± 3 kcal mol$^{-1}$) and P4 (-155 ± 3 kcal mol$^{-1}$) than the other three systems (e.g. ssRNA+ATP, -188.4 ± 0.4 kcal mol$^{-1}$, -139.3 ± 0.6 kcal mol$^{-1}$) and weaker interactions with P1 (ssRNA: -62 ± 2 kcal mol$^{-1}$; ssRNA+ATP: -106 ± 2 kcal mol$^{-1}$). This demonstrates that a bound nucleoside (ATP or ADP) causes a shift in protein-RNA interactions from 3′ (P3 and P4) to 5′ (P1) RNA residues.

### Hypothesized Elementary Step in the Translocation Mechanism.

Visual analysis of the protein-P3 and -P4 interactions in the ssRNA simulation highlighted a rare event where the guanidinium side chain of Arg387 (motif IVa) transitions from coordinat-

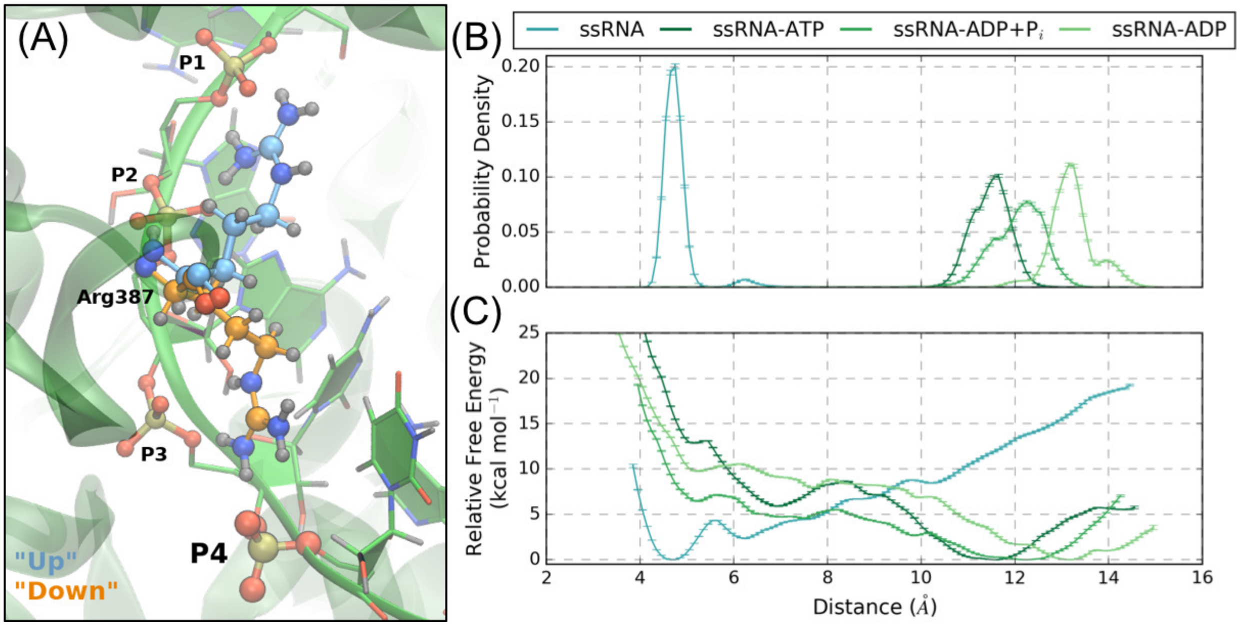

**Figure 2.7: NTPase substrate-dependent interactions between Arg387 of motif IVa and RNA phosphate groups.** (A) The guanidinium group of Arg387 is observed to transition from the “up" conformation to the “down" conformation, respectively colored blue and orange. (B) Probability densities and (C) free energy surfaces from the US simulations performed to model the “up" to “down" transition of the Arg387 side chain. Short collective variable distances correspond to the “down" conformation. As emphasized by the line colors, the trend of these results show that ssRNA favors the “down" conformation while the other substrate states favor the “up" conformation, suggesting a NTPase substrate-dependence of the Arg387 conformational states.

ing P1 and P2 to P3 and P4. The two conformations of this event are depicted in Figure 2.7(A): the “up" conformation (colored blue) has the guanidinium group coordinating P1 and P2 while, in the “down" conformation (colored orange), the side chain coordinates P3 and P4. During the ssRNA+ATP, ssRNA+ADP+P$_i$, and ssRNA+ADP simulations, Arg387 is stable in the “up" conformation. In the ssRNA simulation, the “up" to “down" transition occurs once, after which no reverse transitions occur. Furthermore, after Arg387 transitions to the “down" conformation, two concerted events are observed to occur: (1) Lys388 (motif IVa) coordinates P1 and P2, taking up Arg387’s previous position, and (2) P4 partially unbinds from the RNA-binding cleft.

Arg387 and Lys388 are highly conserved residues in the NS3/NPH-II subfamily of SF2 helicases and have been classified as motif IVa residues (positioned at the N-terminus of $\alpha2'$). Lys388 is more solvent exposed than Arg387 and fluctuates around the phosphate groups of the RNA. The $\alpha2'$ secondary structure is generally solvent exposed with its dipole-axis pointing towards P2. Historically, an arginine residue observed to coordinate adjacent phosphate groups of RNA has been termed an arginine fork[106,107]. Arg387 is one such arginine fork that has been observed to have functional importance to the helicase functions of NS3h. In previous studies of HCV NS3h, the analogous arginine residue (Arg393) was mutated to an alanine, resulting in abrogation of nucleic binding, translocation, and unwinding functions of the mutant HCV NS3h[108]. A recent crystal structure of the Zika virus NS3h in complex with an RNA (PDB ID: 5GJB) has been reported where the analogous motif IVa arginine (Arg388) is positioned in the “down" conformation[109]. This crystal demonstrates that the Arg387 transition observed during DENV NS3h simulation corresponds to a realistic conformation of NS3h:RNA complexes. Furthermore, the high sequence conservation of Arg387 and the mutational results from HCV NS3h suggest that this motif IVa arginine fork binds the RNA and plays an important role in the helicase functions of NS3h.

Due to the concerted nature of the Arg387 transition and the unbinding of P4, we propose that the observed “up" to “down" transition is potentially an elementary step in the translocation mechanism of NS3h, where the guanidinium side chain conformations are NTPase substrate-dependent. Therefore, we investigate the thermodynamics of the side chain conformational states using one-dimensional US simulations for the ssRNA, ssRNA+ATP, ssRNA+ADP+P$_i$, and ssRNA+ADP systems. The biased collective variable is the distance between the central carbon of the Arg387 guanidinium group to the phosphorous atom of P4, where the “up" and “down" conformations correspond to long and short distances, respectively. The resulting probability densities and relative free energy surfaces from the US simulations are shown in Figure 2.7 (B) and (C), respectively. The relative changes in free energy between the “up" and “down" side chain conformations ($\Delta$ G$_{US}$) are -7.52$\pm$0.05 kcal mol$^{-1}$, 13.885$\pm$0.05 kcal mol$^{-1}$, 7.45$\pm$0.06 kcal mol$^{-1}$, and 10.91$\pm$0.05 kcal mol$^{-1}$ for the ssRNA, ssRNA+ATP, ssRNA+ADP+P$_i$, and ssRNA+ADP systems, respectively. Therefore, the Arg387 side chain states are observed to have a NTPase substrate-dependence where the ssRNA system energetically favors the “down" conformation while the ssRNA+ATP, ssRNA+ADP+P$_i$, and ssRNA+ADP substrate states favor the “up" conformation.

Unbinding of P4 is not observed for US windows corresponding to the “down" conformation in the ssRNA+ATP, ssRNA+ADP+$P_i$, and ssRNA+ADP US simulations.

These results support the hypothesis that the Arg387 side chain conformational states are NTPase substrate-dependent and exemplify a large shift in protein-RNA interactions. Considering the full NTPase cycle (Figure 2.1), Arg387 thermodynamically favors the “down" conformation in the ssRNA substrate state where the guanidinium group coordinates P3 and P4. Subsequently moving through the NTPase substrate states, Arg387 is expected to transition to the “up" conformation and coordinate P1 and P2. Therefore, transitions between the Arg387 conformational states are predicted to impart a $3'$ to $5'$ direction to interactions between NS3h and RNA. Furthermore, the concerted events that were observed to followed the Arg387 conformational change (Lys388 coordinating P1 and P2; partial unbinding of P4) are novel events that have potentially important implications for the translocation of NS3h along the phosphodiester backbone of ssRNA. These results support the hypothesis that the Arg387 conformational states represent states of a NTPase substrate-dependent, elementary step in the unidirectional translocation mechanism of NS3h.

### 2.5.3 Allosteric Pathways

The current view of allosteric regulation focuses on signal transduction through complex, 3-dimensional networks, brought about by intrinsic structural and/or dynamic changes along pathways connecting two distal, non-overlapping active sites[110–112]. These allosteric pathways are described by coupled short-range, residue-residue interactions that lead to long-range correlations. In the previous two sections, RNA-induced and NTPase substrate-induced structural rearrangements have been presented. In this section, these allosteric structural changes are absorbed into a unified description of the allosteric pathways connecting the RNA-binding cleft with the NTPase active site.

Dynamic network analyses, such as residue-residue correlations, have been used to identify allosteric pathways within proteins from simulation[110–114]. A growing body of literature has highlighted the functional importance of such pathways as well as the fundamental residue-residue interactions leading to their emergence[110–116]. We report here residue-residue distance correlation analyses that are used to identify the allosteric pathways present within the DENV NS3h protein. Focus is given to the motifs discussed in the previous sections ($\alpha2$, motifs II and IVa) due to the observed structural rearrangements. Additionally, the correlation heat maps are used to identify segments of the protein that experience strong correlations with numerous other regions of the protein, such as motif V. While motif V does not experience substrate-induced structural rearrangements, the strong correlations between motif V and motifs in both the NTPase active site and RNA binding cleft are hypothesized to have functional importance in the signal transduction mechanism of allosteric regulation. Unlike the previous two sections, comparisons between substrate states (RNA-bound and NTPase substrate-bound) are not considered here. Instead, focus is given to the discussion of the residue-residue distance correlation analysis of the ssRNA+ATP substrate state.

**Correlations Between Motifs.**

Figure 2.8(A) shows the average residue-residue distance correlation heat map for the ssRNA+ATP substrate state, where residue pairs with strong correlated and anti-correlated motions are colored red or blue, respectively. The correlation heat map is abridged by setting correlation values to zero if the average COM-COM distance of residue pairs is greater than 15 Å. This simplification limits the correlation analysis to residue pairs within a close proximity, thereby identifying the short-range interactions that build up to the pathways connecting the RNA-binding cleft and the NTPase active site. Correlation heat maps for the other substrate states are presented in SI (Figures Figure 2.18 to Figure 2.22).

As expected, there is strong correlation along the linear sequence of residues, as seen in Figure 2.8(A) along the diagonal. Secondary structures experience more extended linear sequence correlations than non-structured regions (thickness of diagonal sections). The RecA-like $\beta$-sheets in subdomains 1 (residues 188 to 326) and 2 (residues 327 to 481) produce the honeycomb patterns observed in the heat map, due to the extremely stable tertiary structure observed in NS3h. Lines drawn on Figure 2.8(A) highlight a range of 20 residues centered on $\alpha2$, motifs II and IVa.

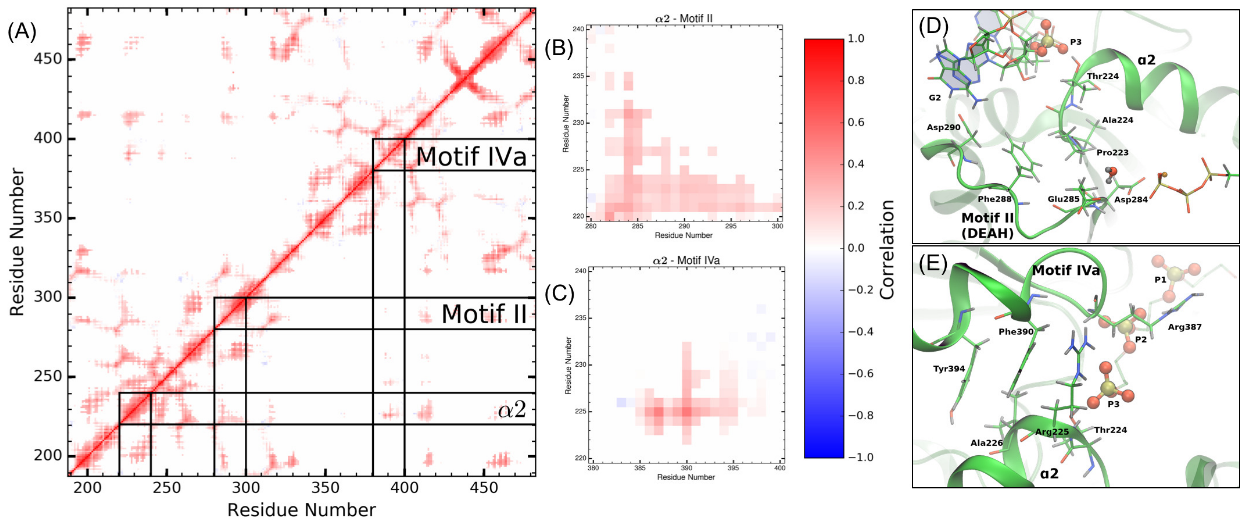

**Figure 2.8: Correlated motions of protein motifs observed to experience RNA- or NTPase substrate-induced allostery.** (A) The average COM-COM residue pair correlation heat map for the ssRNA+ATP system abridged with a distance cutoff of 15 Å. Lines drawn highlight the structural motifs discussed in the previous two sections ($\alpha2$, motifs II and IVa). Panels (B) and (C) are magnifications of off-diagonal regions in (A) that correspond to the correlations between $\alpha2$ and motif II or motif IVa, respectively. Hotspots within these regions identify the short-range residue-residue interactions that couple the structures. Panels (D) and (E) provide structural depiction of these residue-residue interactions.

Panels (B) and (D) of Figure 2.8 highlight the $\alpha2$/motif II off-diagonal region of the heat map and the structural features leading to the observed correlations between these two segments. Specifically, Asp284 and Glu285 of motif II experience correlated motions with most of $\alpha2$ (residues 220 to 231) and especially strong correlations with Ala222, Pro223, and Thr224. Thr224 directly coordinates phosphate 3 of the RNA substrate. Therefore, panels (B) and (D) highlight a pathway through Glu285, Ala222, Pro223, and Thr224 that connects residues of paramount importance in the hydrolysis reaction (Glu285) with residues that directly coordinate the phosphodiester backbone of the RNA (Thr224).

In a similar fashion, the motif IVa structure is observed to have strong correlations with residues in $\alpha2$, as shown in panels (C) and (E) of Figure 2.8. Phe390 and, to a lesser extent, Tyr394 of motif IVa are observed to have strong correlated motions with residues 222 to 232 of $\alpha2$. Alternatively, from the perspective of $\alpha2$, the guanidinium group of Arg225 has strong and maintained electrostatic interaction with the backbone carbonyl group of Arg387. Through this short-range interaction, Thr224, Arg225, and Ala226 experience correlated motions with residues 385 to 395 of motif IVa. As observed in Figure 2.8(E), the side chains of Arg225, Phe390, and Tyr394 are positioned in a $\pi$ stacking-like structure, bookended by RNA phosphate 3 as well as residues in the NTPase active site (not shown).

Direct coupling between motifs II and IVa is minimal due to the large distances between the residues in both motifs, as shown in the respective off-diagonal region of Figure 2.8(A). Rather, short-range interactions between residues in $\alpha2$ (Ala222, Pro223, Thr224, Arg225) with residues of motifs II (Asp284, Glu285) and IVa (Arg387, Phe390, Tyr394) act as channels through which free energy released from the NTPase cycle or helicase functions can be transferred from the NTPase active site to the phosphodiester backbone of the RNA (and vice versa). Structural or dynamic perturbation of these pathways, via mutation or small molecule binding, are hypothesized to affect the efficiency of energy transduction.

**Motif V.**

The residue-residue correlation analyses are used to identify structural regions of NS3h that exhibited strong correlations with other regions of the protein. Particularly, this analysis highlighted residues 407 to 420 of motif V. Figure 2.9 (A) shows the motif V correlation heat map section for the ssRNA+ATP system, where lines are drawn to highlight the strong coupling between these residues and residues of motifs I, II, III, IV, IVa, VI as well as $\alpha2$. Structurally, the highly correlated nature of motif V is explained by the position of these residues in relation to the NTPase active site, RNA-binding cleft, and protein residues important in either active site, as shown in Figure 2.9(B). Motif V consists of residues in subdomain 2 and has a complex secondary structure (loop into short $\alpha$-helix into loop). The backbone amide of Gly414 actively coordinates the lytic water of the hydrolysis reaction. The alcohol group of Thr408 coordinates phosphate 2 of the RNA. Therefore, the linear sequence of motif V is a direct pathway connecting the NTPase active site with the RNA-binding cleft. Limited structural changes are observed for this motif in the presence or absence of RNA or NTPase substrates. Rather, motif V is observed to have strongly correlated motions with the previously mentioned structural regions for all substrate states. Similar to the

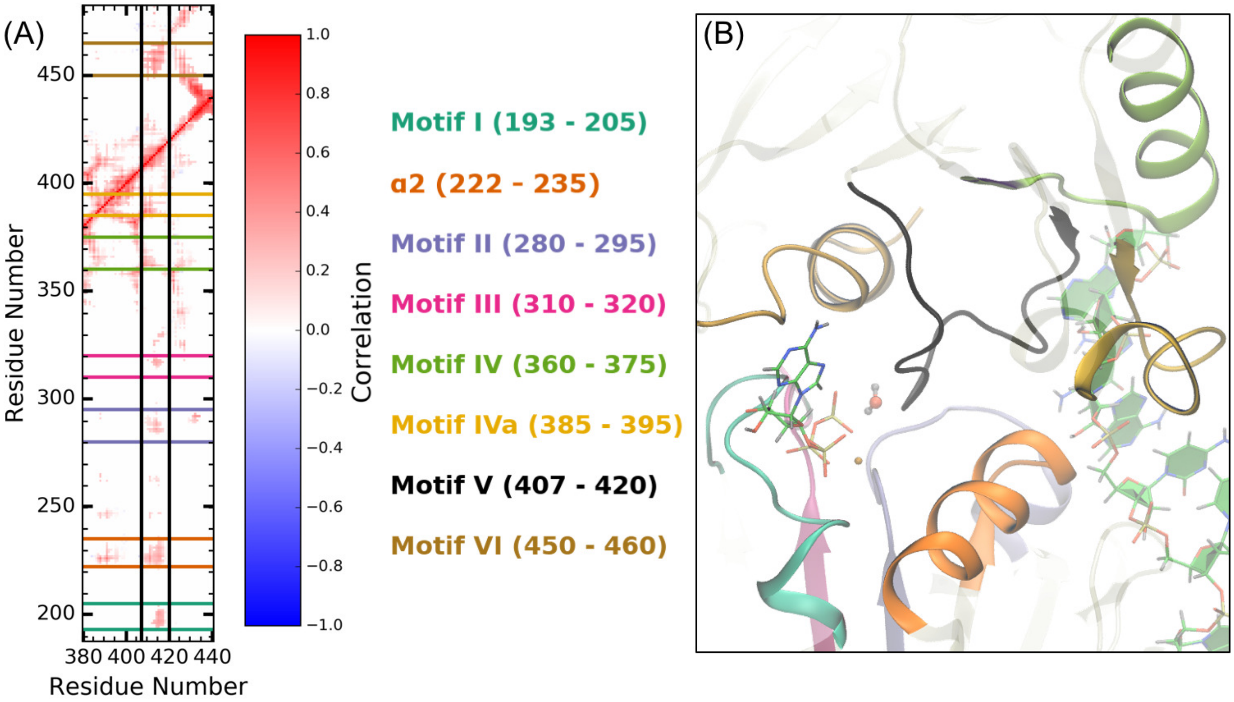

**Figure 2.9: Motif V is a highly correlated and centralized structure within subdomains 1 and 2.** (A) Vertical segment of the ssRNA+ATP correlation heat map focusing on motif V (residues 407 to 420). Conserved motifs that have strong correlations with motif V are highlighted by horizontal lines on the heat map, colored as shown in the legend. (B) ssRNA+ATP exemplar structure depicting the central position of motif V in relation to the NTPase active site and the conserved motifs highlighted in panel (A). The ATP and lytic water molecules are shown to highlight the proximal location of motif V residues with respect to the NTPase active site.

previously discussed pathways, motif V is hypothesized to be another pathway for free energy transduction from one active site to another.

## 2.6 Conclusions

Through analyses of the reported simulations, molecular observables of RNA- and NTPase substrate-induced allostery were identified. Specifically, an RNA bound within the RNA-binding cleft affects the dynamics and positioning of water molecules within the NTPase active site. This allosteric influence is conferred from the RNA-binding cleft to the hydrolysis active site through structural rearrangements of L$\beta3\beta4$, $\alpha2$, and motif II. These RNA-induced structural changes lead to an entropic destabilization of the NTPase active site as well as a direct destabilization of the lytic water. Inspired from these results, electronic structure calculations were used to investigate the energetics of NTP hydrolysis reaction. The energetic landscapes obtained from the DFT calculations demonstrate that RNA decreases the activation barrier as well as affects the mechanism of the hydrolysis reaction. Combining these results into a kinetic model allowed for the calculation of a theoretical RNA-stimulated NTPase activity enhancement factor of 150, which qualitatively matches the experimentally observed enhancement factor. Therefore, results from MD and DFT calculations provide novel, multiscale insight into the RNA-induced allosteric effects that stimulate the catalysis of the NTP hydrolysis reaction in DENV NS3h.

Unlike RNA, the NTPase substrates are smaller perturbations to the NS3h structure and dynamics. Protein-RNA interaction energies were used to investigate the NTPase substrate-dependence of protein-RNA contacts in the unbiased MD simulations. From these analyses, the protein-RNA phosphodiester backbone interactions were observed to be NTPase substrate-dependent. The presence of the $\gamma$-phosphate (or $\mathrm{P}_i$) of the NTPase substrate was observed to strengthen the protein-RNA contacts. Furthermore, the localized nonbonding interaction energies demonstrate a large shift in protein-RNA contacts, originating in part from the side chain conformational states of Arg387. Results from US simulations demonstrate that the Arg387 side chain conformational states exemplify NTPase substrate-dependent protein-RNA interactions. With the purview of the NTPase cycle, transitions between conformational states leads to $3'$ to $5'$ translocation. Therefore, we hypothesize that the transition between Arg387 side chain conformations is an elementary step in the unidirectional translocation mechanism of NS3h along the phosphodiester backbone of RNA.

Finally, consideration of these allosteric effects independent of one another provides an incomplete picture of the biophysics of the NS3h protein. Residue-residue correlation analyses were used to identify structural regions of the protein that experienced correlated motions with other regions. These analyses were used to describe the allosteric pathways that connect $\alpha 2$ with motifs II and IVa. The short-range, residue-residue interactions were presented that connect the RNA-binding cleft to the NTPase active site. Furthermore, the correlation heat maps allow for identification of regions of the protein that experience strong correlated motions with numerous other regions. Motif V is one such example, where the segment of 13 residues has strong coupled motions with seven other motifs in subdomains 1 and 2. This highly correlated nature suggests that motif V functions as a centralized communication hub that connects distal portions of the protein structure.

Complete modeling of a revolution of the NTPase cycle in the DENV NS3h presents a significant challenge for current computational methodologies. Rather, we have divided the cycle into equilibrium substrate states and dynamic events where the protein transitions from one substrate state to another. The simulations reported have modeled the important NTPase cycle substrate states, leading to novel insights into the function and underlying biophysics of the DENV NS3h enzyme with focus given to the allosteric connections between the RNA-binding cleft and NTPase active site. We hypothesize that the observed allosteric effects and pathways have important roles in the transduction of energy from one active site to another during the dynamic events of the NTPase cycle. Therefore, the results reported have laid an initial foundation for theoretical investigations into the dynamic events of the NTPase cycle.

Beyond further theoretical modeling of NS3h, transverse relaxation-optimized spectroscopy (TROSY) NMR, mutational biochemical studies, and targeted small molecule binding experiments can be envisioned to test the hypothesis and results presented. Nuclear magnetic resonance has been previously used to study dynamics within isolated HCV NS3 helicase subdomains[117–119], but the size of the full dengue NS3h domain is too large for traditional NMR approaches due to line width increasing with increased molecular mass. However, TROSY NMR has been developed that may allow for experimental monitoring of fast NS3h dynamics[120]. We anticipate that perturbation of the wild-type structure or dynamics of the allosteric pathways in NS3h will lead to abrogation of NTPase and/or helicase functions, and are currently developing and testing assays to test each step with dengue NS3h. Residues active in the allosteric pathways are viable targets for mutational studies where varying a specific amino acid residue is expected to alter the short-range, residue-residue interactions and lead to a destabilization of the pathway connecting the two active sites. This is hypothesized to result in reductions in enzymatic activity and would be observable in our biochemical assays. Additionally, these pathways are viable targets for theoretical and experimental small molecule drug docking experiments with focus given to molecules that disrupt the residue-residue interactions along the pathway. Co-crystallization of specific conformation-binding molecules may lock NS3h into transition-state conformations that can help verify our computational studies. Molecular candidates also have the potential for inhibiting NS3h function during replication and being specific to the NS3/NPH-II subfamily of SF2 helicases.

## 2.7 Funding

This work was supported by the National Science Foundation-supported (grant No. ACI-1548562) Extreme Science and Engineering Discovery Environment (XSEDE) (allocation No. CHE160008), Colorado State University to MM for start-up funds, and the National Science Foundation Research Experience for Undergraduates (CHE-1461040) and Fixman Summer Theory program to JH. We would also like to acknowledge the support of NVIDIA Corporation with the donation of one of the Titan X Maxwell GPUs used for this research.

## 2.8 Supporting Information

### 2.8.1 System Preparation

**Molecular Dynamics Simulations** Energy minimization of MD starting structures was performed in two steps: 1) 10,000 steps of minimization with the protein, substrates, and crystallographic waters restrained and 2) another 10,000 steps with all atoms unconstrained. Initial heating of each simulation was performed where minimized structures were heated from 0 K to 300 K over 0.5 ns. The heating simulation was performed with 75 kcal mol$^{-1}$ harmonic restraints on all non-solvent atoms. Equilibration of the protein structure at 300 K was performed by slowly releasing the applied restraints over the course of 2 ns. This minimization and heating protocol was used for all reported simulations.

**QM structures** As mentioned in the article, reactant starting structures for the density functional theory (DFT) calculations were grabbed from the body of MD frames that posses a water molecule in a lytic position. To demonstrate that the initial structures used in the DFT calculations adequately represent the population of frames with lytic waters, the RMSD of protein side chain heavy atoms included in the QM region was calculated for all MD frames, referenced against the DFT starting structures. For the ATP substrate state, the average RMSD for this selection is 1.2 ± 0.4 Å. For the ssRNA+ATP substrate state, the average RMSD for this selection is 1.2 ± 0.3 Å. The high relative standard deviation of these values is associated with fluctuations of Ala316 atoms during the MD simulations.

The QM region consists of 138 atoms corresponding to the methyl triphosphate, $Mg^{2+}$ cation, six water molecules, and ten amino acid residues closest to the triphosphate tail of the ATP molecule (shown in Figure 2.10). The amino acid residues were truncated at various positions with hydrogen atoms. For example, a hydrogen atom was added to the backbone amide of Pro195 as well as the $C_\alpha$ atom of Gly196. Lys199 was truncated at $C_\epsilon$ atom. For Thr200, the alcohol group is truncated with a hydrogen, thereby bringing the number of water atoms to seven and number of amino acid residues down to nine (as reported in the article). Glu285 is truncated at the $C_\beta$ atom. Ala316 is truncated on the backbone amide group and $C_\alpha$ atoms. Similarly, Gly414 is truncated at the backbone amide and carbonyl atoms. Gln456 is truncated at the $C_\gamma$ atom. Both Arg460 and Arg463 are truncated at the $C_\gamma$ atoms. These truncations maintain the coordination sphere of $Mg^{2+}$ as well as the important contacts between the triphosphate tail and amino acid functional groups. Waters included in the QM calculations were chosen due to their proximity to the terminal phosphoanhydride bond as well as the $Mg^{2+}$ cation.

### 2.8.2 Data Analysis

**Equilibration and Convergence of Simulations** Global and local protein structural metrics (e.g. radius of gyration, RMSD referenced against the respective crystal structures; Figure 2.11 to Figure 2.13) are used to determine that the protein structure in each simulation had adequately con-

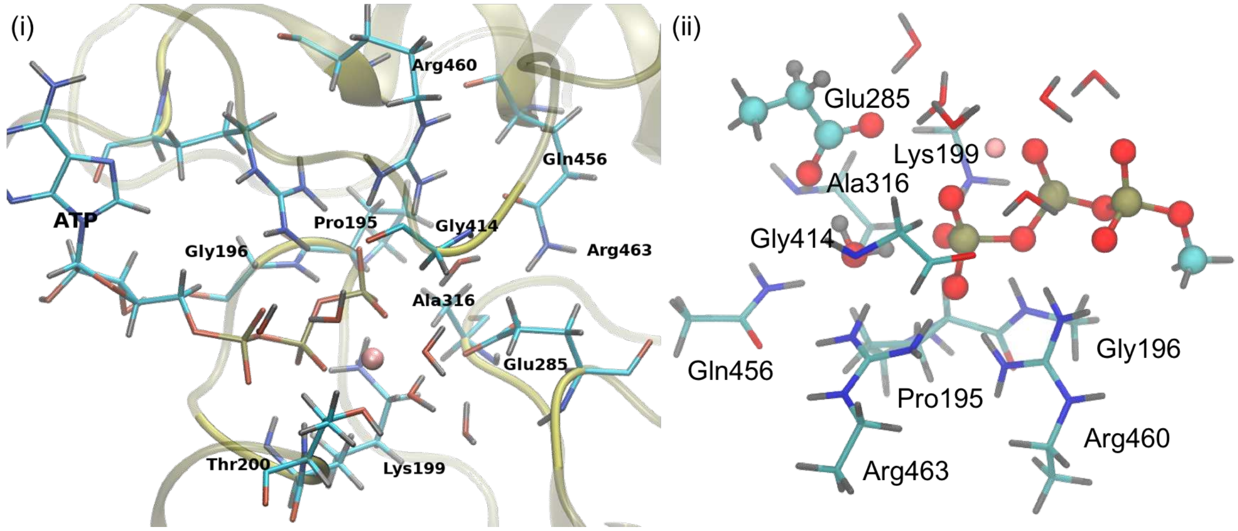

**Figure 2.10:** Structural representation of the QM region. (A) The QM region within the broader protein structure. Residue labels included in the QM region are provided. Thr200 is included because the alcohol group of the side chain was replaced with a water molecule in the electronic structure calculations. (B) Depiction of the QM region with full truncation used in the electronic structure calculations. Atoms highlighted with the CPK representation are shown in the energy landscape (Fig 5) in the article.

verged away from the crystal structure after 200 ns of simulation. Further convergence of the simulations is measured by separating the full trajectories into 50 ns windows that are treated as independent trajectories for the purpose of convergence and error analyses. Comparisons between these windows are made to identify long-timescale changes occurring during the microsecond-long trajectories. Additionally, for analysis metrics discussed in the paper, averages of these windows are used to determine the standard error of the mean for the full dataset. For the substrate states with a pair of microsecond-long trajectories (ATP and ssRNA+ATP), each replica is given 200 ns to equilibrate from the starting structures, after which analysis results from both replicas are combined into the reported ensemble averages. Error analysis of these averages include the 50 ns windows from both replicas.

**Alignment Landmarks** The $\beta$-sheets in the RecA-like subdomains of DENV NS3h (subdomains 1 and 2) are observed to be extremely stable during the reported MD simulations as well as across all substrate states. Therefore, this collection of secondary structures acts as a good reference point for analyzing structural and dynamic changes in the rest of the protein. Specifically, the $C_\alpha$ atoms of the residues making up these secondary structures (listed in Table 2.2) are used as the landmark for all analyses requiring structural alignment of the protein. Figure 2.14 shows the aligned RMSD of this landmark for all trajectories as referenced against the ssRNA+ATP crystal structure (PDB ID: 2JLV). Similarly, Figure 2.15 shows the RMSD of all heavy-atoms for residues comprising these $\beta$-sheets, referenced against the ssRNA+ATP crystal structure and using the alignment landmark. Structurally, these $\beta$-sheets are conserved across the SF1 and SF2 helicase families and so we hypothesize that this landmark can be efficiently used for broad structural comparisons within these families.

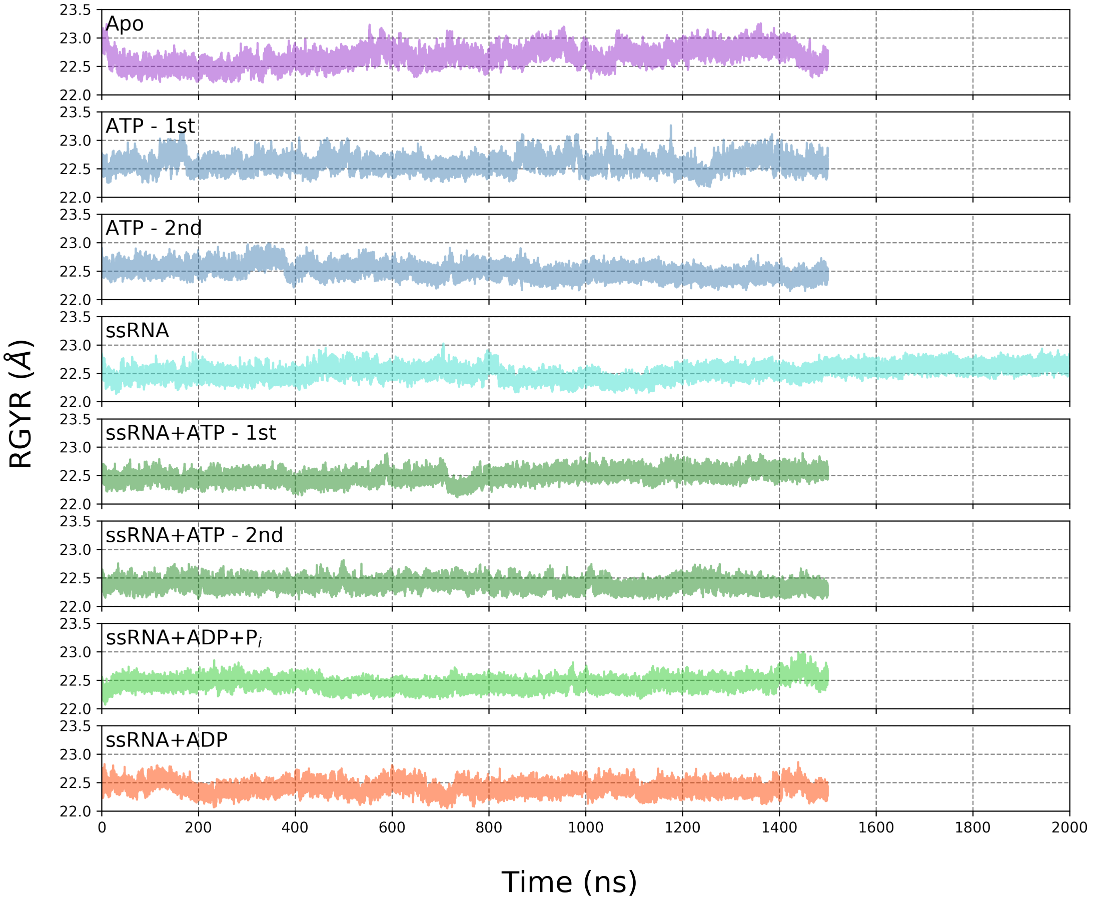

**Figure 2.11:** RGYR of the protein and substrates. As seen through this metric (e.g. Apo, ssRNA+ADP+P$_i$, and ssRNA+ADP), the RGYR metric deviates from the starting values during the first 200 ns of simulation.

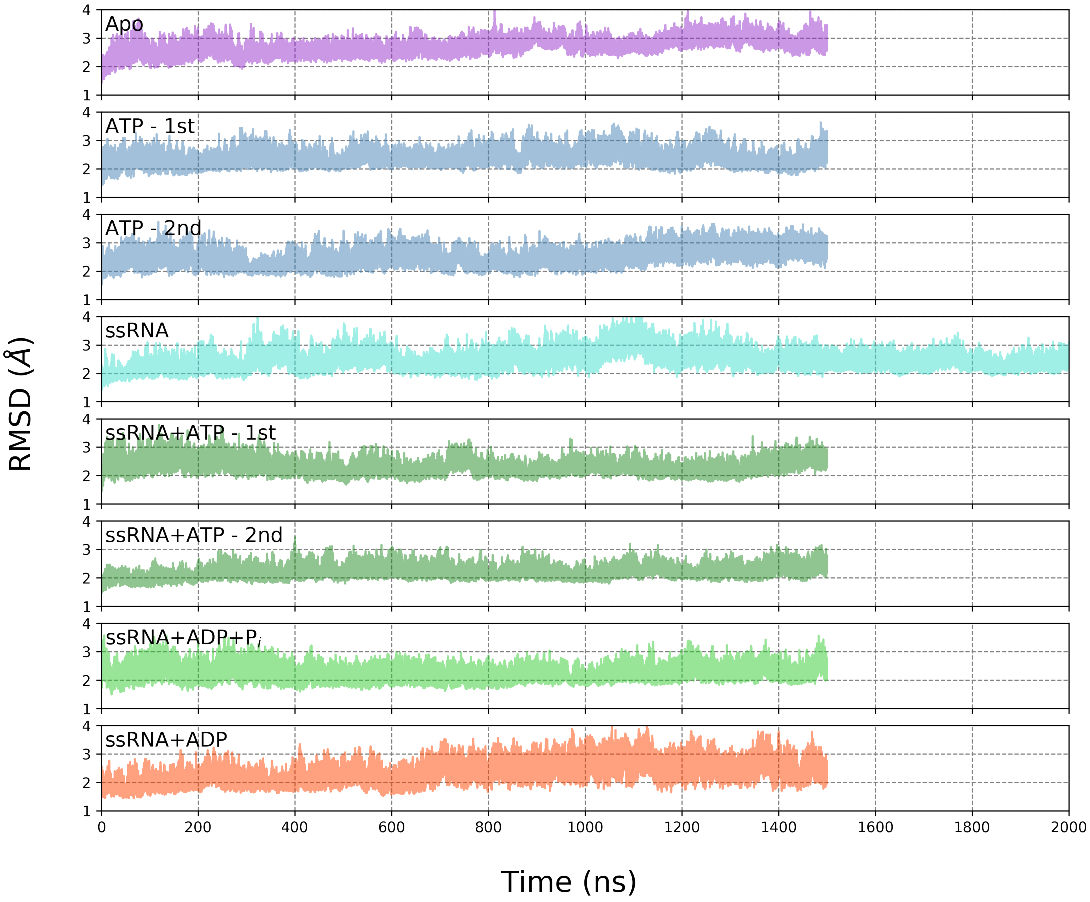

**Figure 2.12:** RMSD of all heavy atoms of the protein, referenced against the respective crystal structures. Before the calculation, every frame is aligned to the $\beta$-sheets discussed in the Alignment Landmarks section.

**Exemplar Structures** All structural figures presented in this manuscript were created in a consistent fashion: for each substrate state, a single frame from the total number of frames was chosen to be the "exemplar" structure for the substrate state. The exemplar structure corresponds to the frame with the minimum RMSD value of all protein, heavy atoms referenced against the average structure of the equilibrated portion of the microsecond-long trajectories.

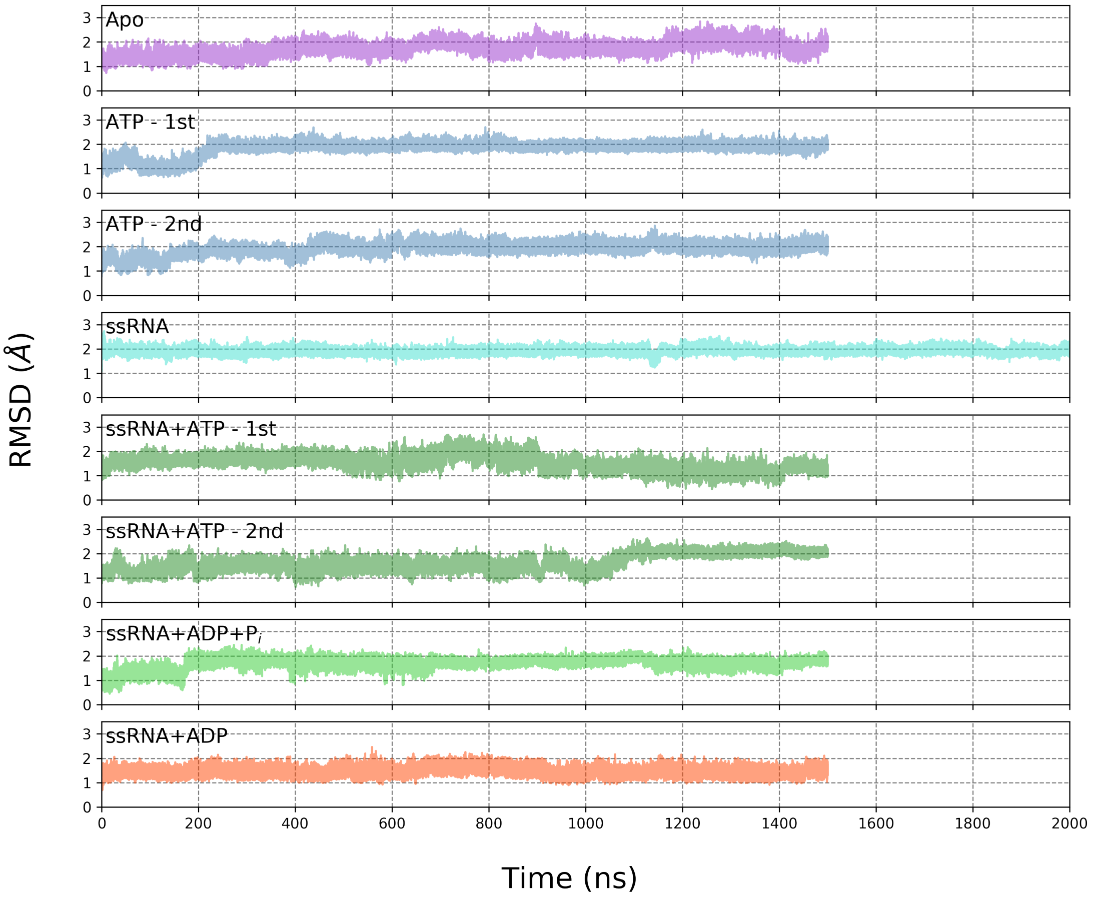

**Figure 2.13:** RMSD of all heavy atoms of motif II residues (residue IDs: 284 to 291), referenced against the respective crystal structures. Before the calculation, every frame is aligned to the $\beta$-sheets discussed in the Alignment Landmarks section.

**Definition of the Center of the NTPase Active Site** The center of the NTPase active site was determined by measuring the center of mass of a subset of protein residues that are prominent within the active site. These residues (numbers: 194 to 202, 227, 230, 231, 284, 285, 314, 316, 326, 412 to 416, 455, 456, 459, 460, and 463) were identified using protein-ATP(ADP) distance analyses and visual analysis of trajectories. Waters within an 8 Å radius of the COM coordinate are considered to be within the active site and are analyzed further.

**Table 2.2:** Residue numbers for each $\beta$-sheet in the alignment landmark.

| Secondary Structure | Residue Numbers |
|---|---|
| $\beta 1$ | 187-192 |
| $\beta 2$ | 217-222 |
| $\beta 3$ | 240-242 |
| $\beta 4$ | 257-261 |
| $\beta 5$ | 279-283 |
| $\beta 6$ | 309-314 |
| $\beta 1'$ | 332-336 |
| $\beta 2'$ | 357-361 |
| $\beta 3'$ | 381-385 |
| $\beta 4'$ | 403-407 |
| $\beta 5'$ | 420-425 |
| $\beta 6'$ | 470-474 |

**Definition of Lytic Waters** Of the waters within the NTPase active site, a lytic water was defined as a water with an $P_\gamma$-$O_{\text{wat}}$ distance less than 5 Å and $O_{\beta,\gamma}$-$P_\gamma$-$O_{\text{wat}}$ angle greater than 155°. The definition of the angle cutoff is supported by the electronic structure results, where the lytic water in the ssRNA+ATP and ATP substrate state reactant structures have an $O_{\beta,\gamma}$-$P_\gamma$-$O_{\text{wat}}$ angle of 161° and 157°, respectively. These geometric cutoffs are used to identify frames in the ATP and ssRNA+ATP MD simulations that have water(s) near ideal lytic positions. The reported probabilities in section 2.5.1 were calculated by counting the number of frames with a lytic water and dividing by the total number of frames. The error for this measurement was approximated by assuming a Poisson distribution for counting experiments.

The two-dimensional heat maps of NTPase active site water positions projected on the $P_\gamma$-$O_{\text{wat}}$ distance and $O_{\beta,\gamma}$-$P_\gamma$-$O_{\text{wat}}$ angle are presented in Figure 2.16 and SFigure 2.17.

**Water Mean Squared Displacement and O-H Bond Autocorrelation** The mean squared displacement (MSD) and O-H bond autocorrelation analyses were measured for waters within the NTPase active site. To obtain the necessary time resolution, 1 ns trajectories were initiated from restart files every 50 ns. Frames were written every 0.2 ps, providing 5000 frames per trajectory to calculate the MSD and O-H bond autocorrelation metrics as a function of time. Rather than use the

$\beta$-sheet alignment landmark, the NTPase active site residues were used as the alignment landmark for this analysis thus providing a more localized alignment. Both metrics were averaged over the post-equilibration trajectories (26 trajectories for each system except for the ssRNA system which had 36 trajectories).

**Nonbonding Interaction Energies** The linear interaction energy (*lie*) function of AMBER14 cpptraj was used to calculate the pairwise sum of Lennard-Jones (LJ) and short-range, unshielded electrostatic energies between the protein and the RNA substrate.[105] A 12 Å distance cutoff was applied for both LJ and electrostatic energies. Reported errors in the interaction energies were calculated by measuring the standard error of the averages of the 50 ns windows (discussed above).

**Correlation Analyses** The calculation and utility of residue-residue distance correlation analyses have been thoroughly discussed in the literature.[111–114] Nodes in the correlation matrix were defined as the center of mass of each residue. The correlation heat map is abridged by applying a distance cutoff of 15 Å where correlation values are set to zero if the average COM-COM distance is greater than the cutoff.

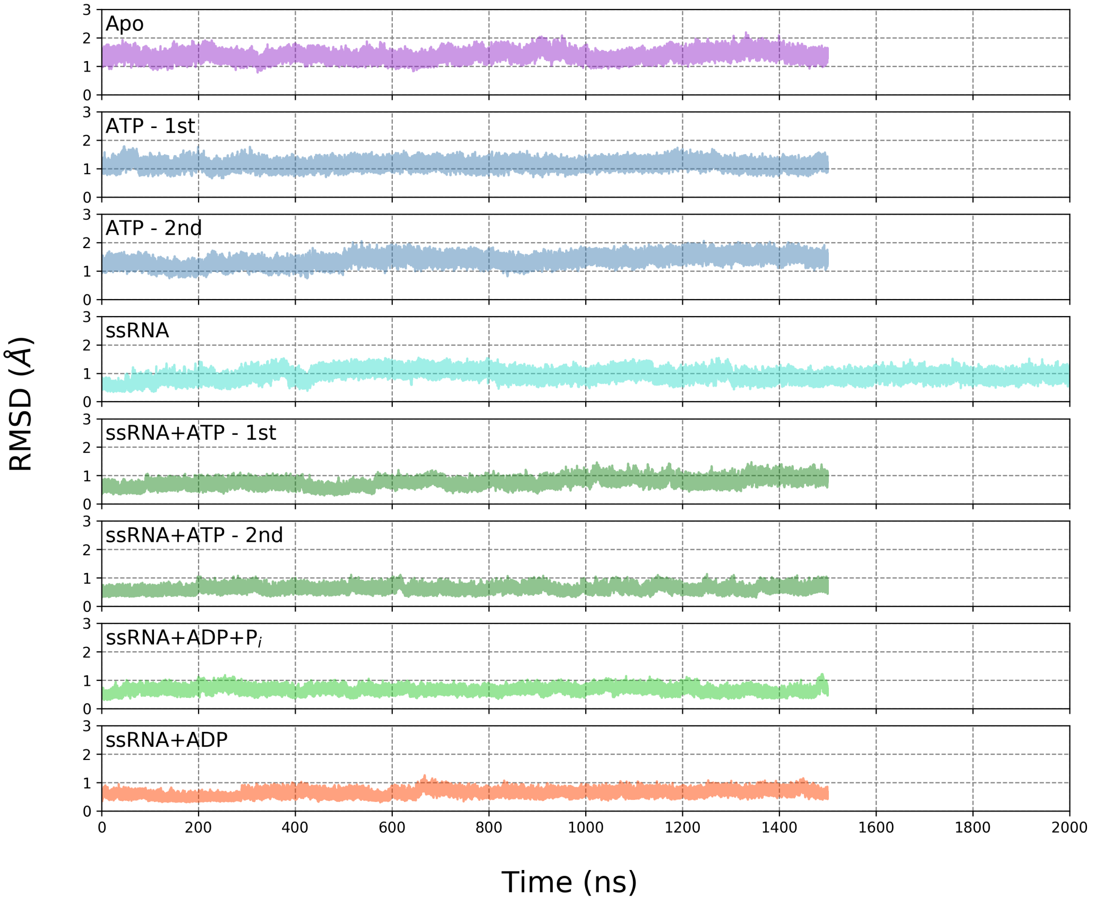

**Figure 2.14:** RMSD of the C$\alpha$ atoms of the RecA-like $\beta$-sheets referenced against the ssRNA+ATP crystal structure (PDB ID: 2JLV) with the alignment landmark applied. Residue numbers for this atom selection are provided in Table 2.2. Generally, RMSD values are small and uniform for all simulations and thus support the use of this alignment landmark.

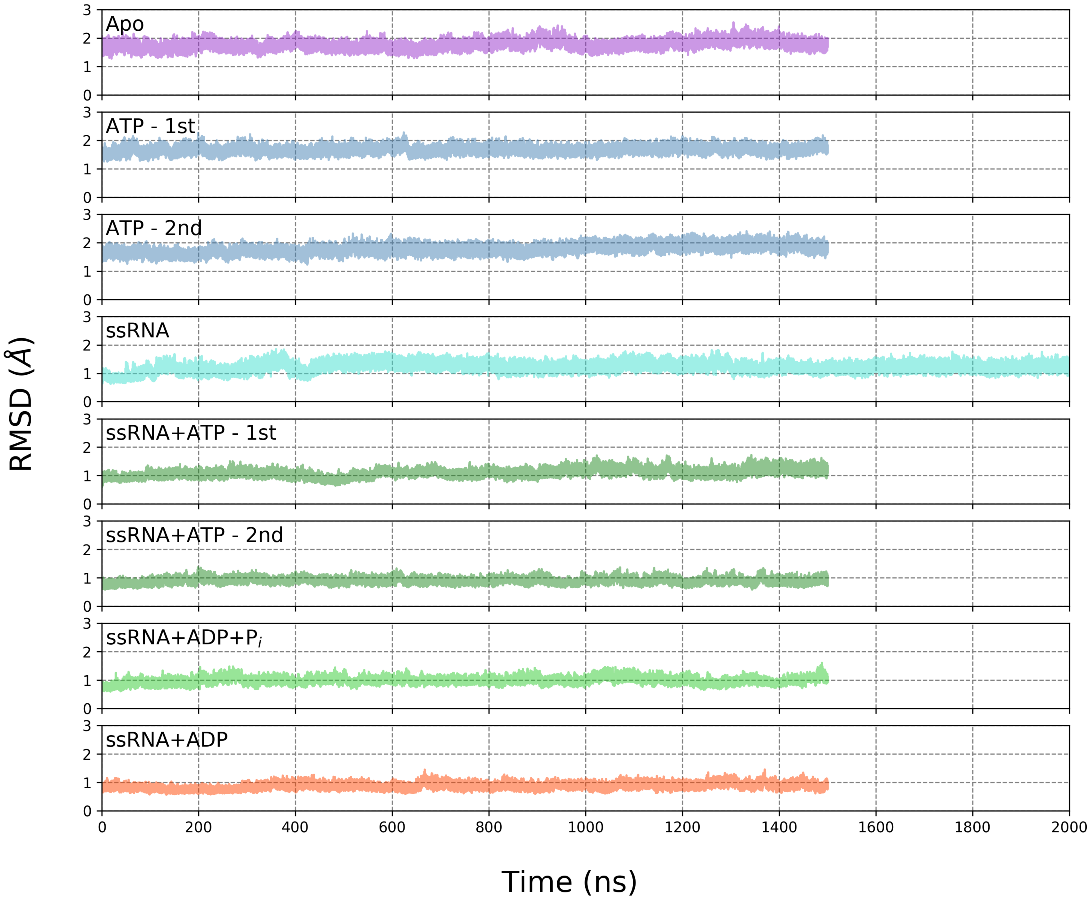

**Figure 2.15:** RMSD of the RecA-like $\beta$-sheets referenced against the ssRNA+ATP crystal structure (PDB ID: 2JLV) with the alignment landmark applied. Residue numbers for this atom selection are provided in Table 2.2.

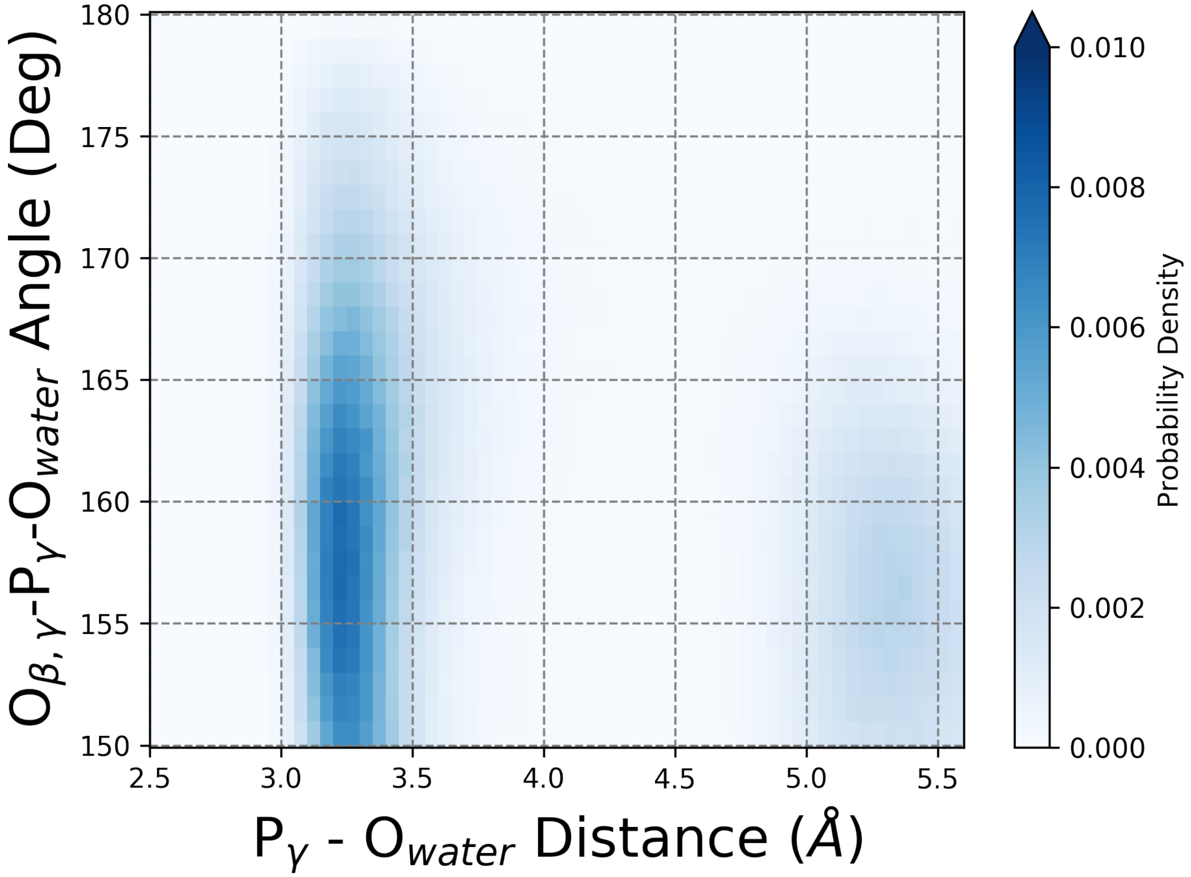

**Figure 2.16:** Probability density heat map of water molecule positions within the NTPase active site for the ATP substrate state. The water positions are projected onto the $P_\gamma$-$O_{wat}$ distance and $O_{\beta,\gamma}$-$P_\gamma$-$O_{wat}$ angle. Waters within 5 Å and above an angle of 155° are defined as lytic waters.

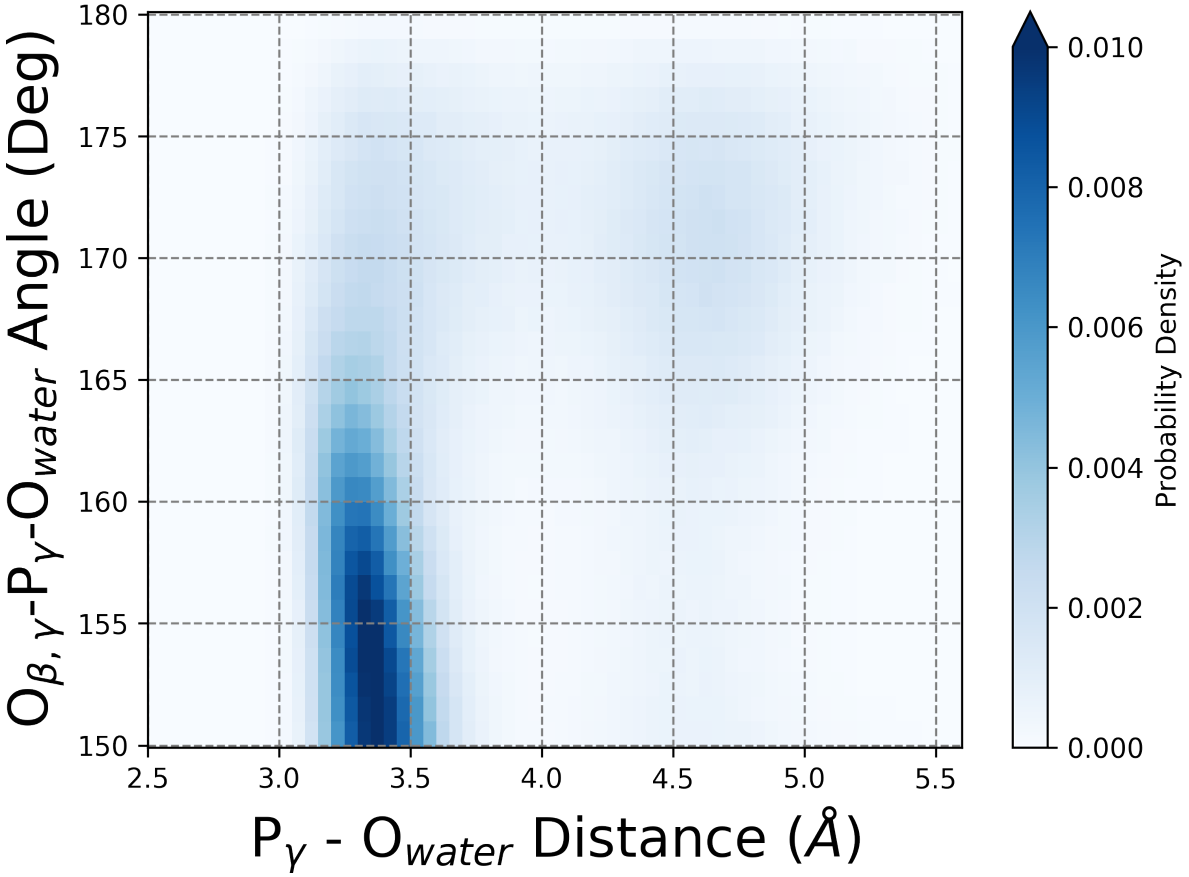

**Figure 2.17:** Probability density heat map of water molecule positions within the NTPase active site for the ssRNA+ATP substrate state. The water positions are projected onto the $P_\gamma$-$O_{wat}$ distance and $O_{\beta,\gamma}$-$P_\gamma$-$O_{wat}$ angle. Waters within 5 Å and above an angle of 155° are defined as lytic waters.

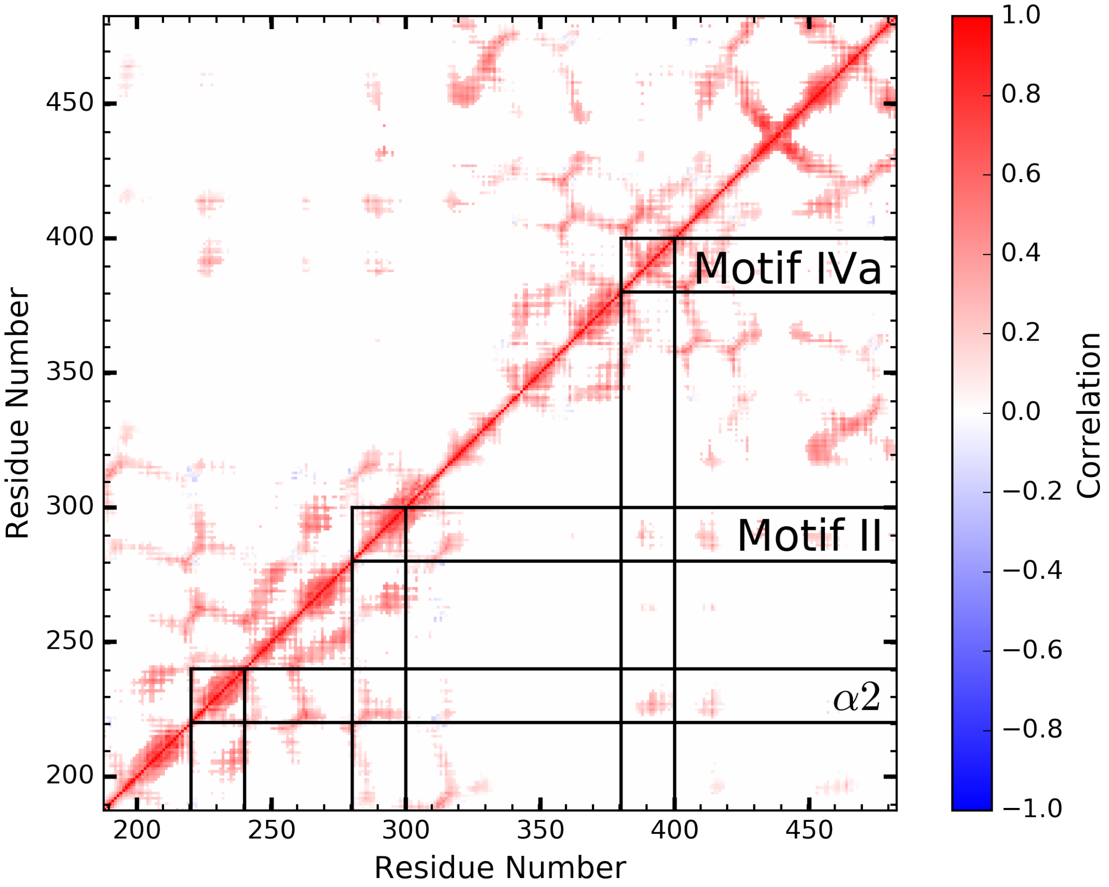

**Figure 2.18:** Residue-residue correlation heat map for the Apo substrate state. Residues of $\alpha 2$, motif II, and motif IVa are highlighted by the drawn lines.

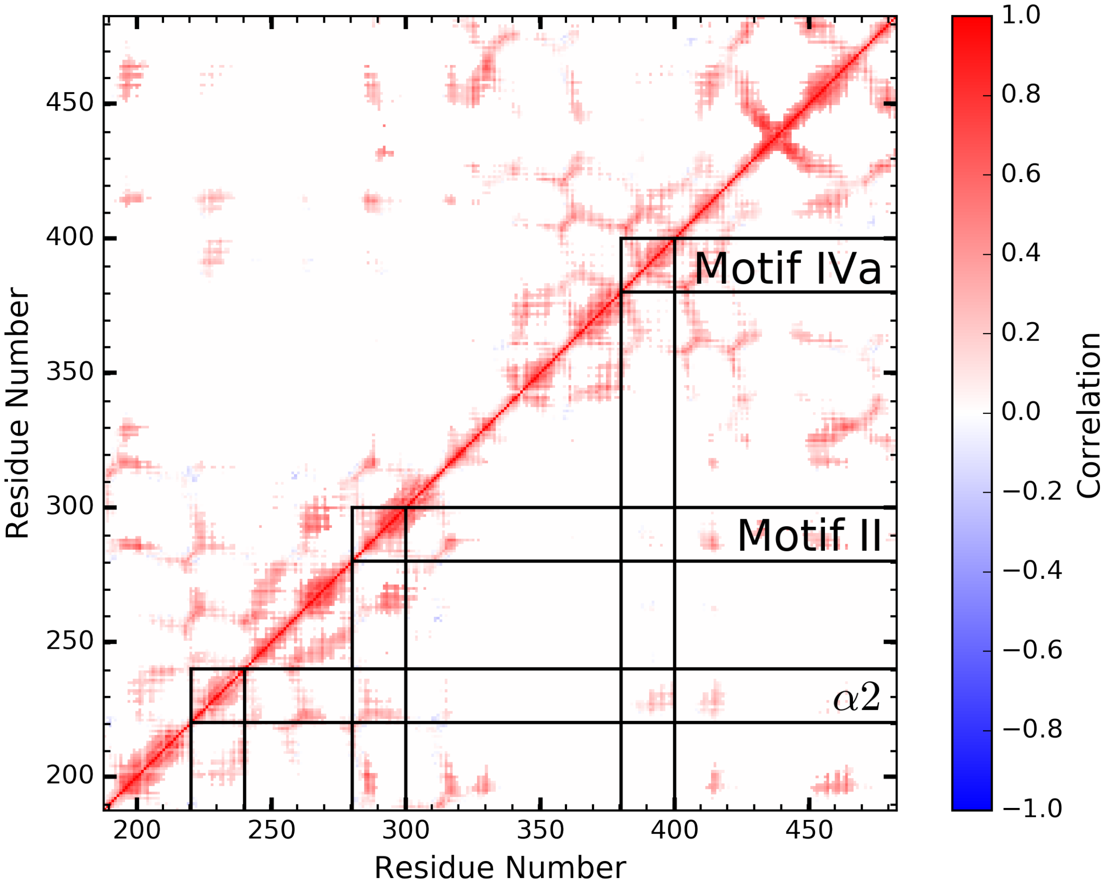

**Figure 2.19:** Residue-residue correlation heat map for the ATP substrate state. Residues of $\alpha 2$, motif II, and motif IVa are highlighted by the drawn lines.

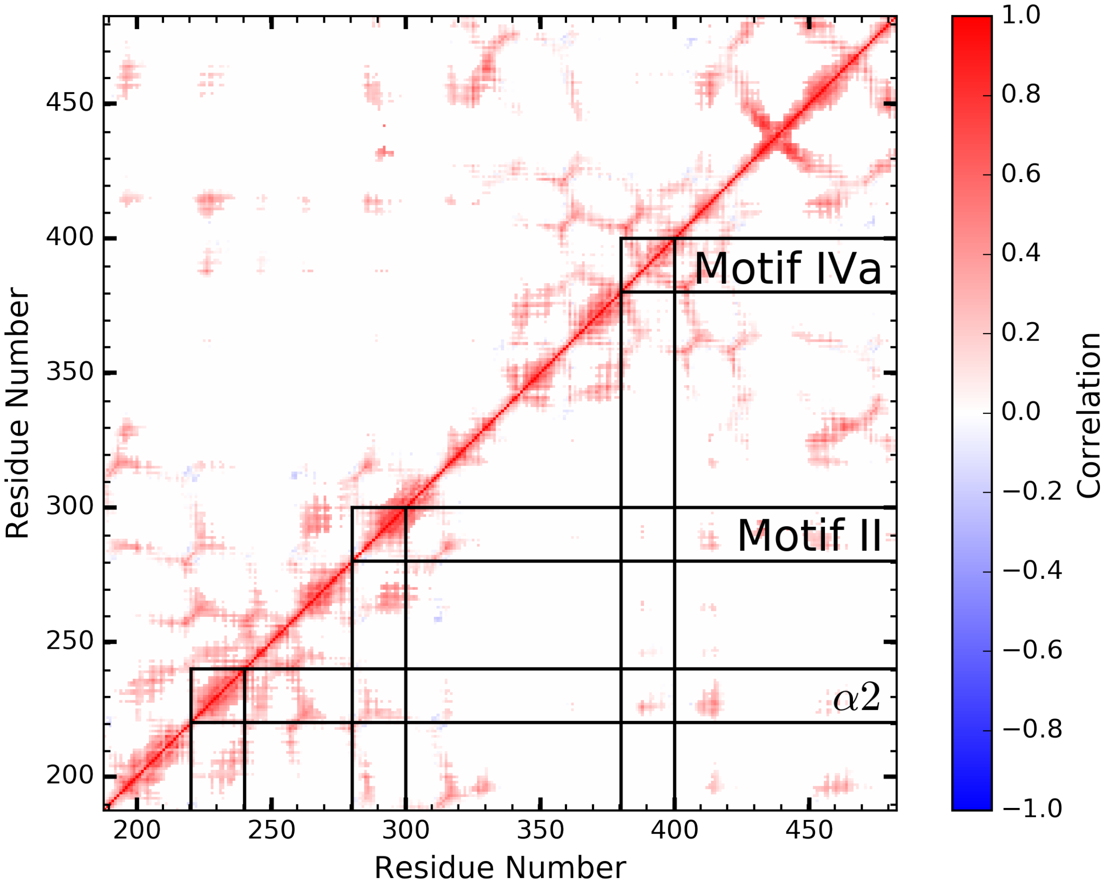

**Figure 2.20:** Residue-residue correlation heat map for the ssRNA substrate state. Residues of $\alpha2$, motif II, and motif IVa are highlighted by the drawn lines.

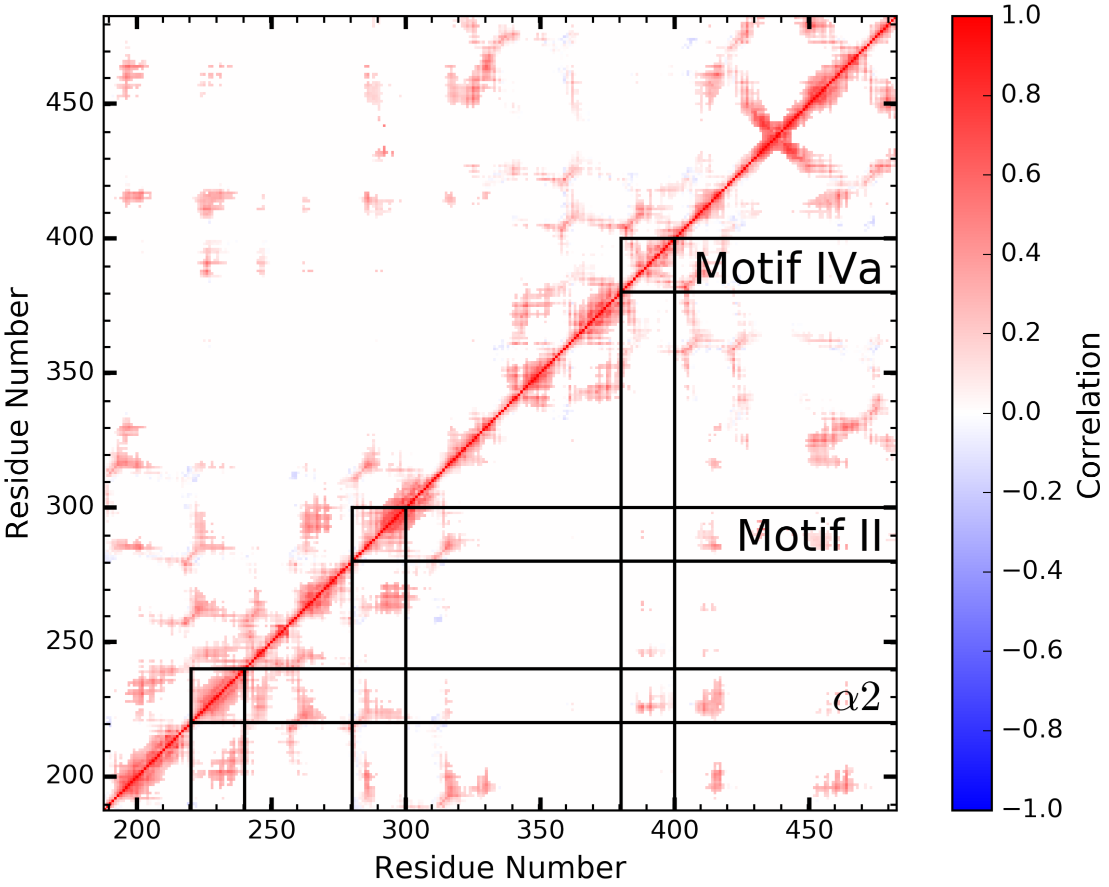

**Figure 2.21:** Residue-residue correlation heat map for the ssRNA+ADP+P$_i$ substrate state. Residues of $\alpha2$, motif II, and motif IVa are highlighted by the drawn lines.

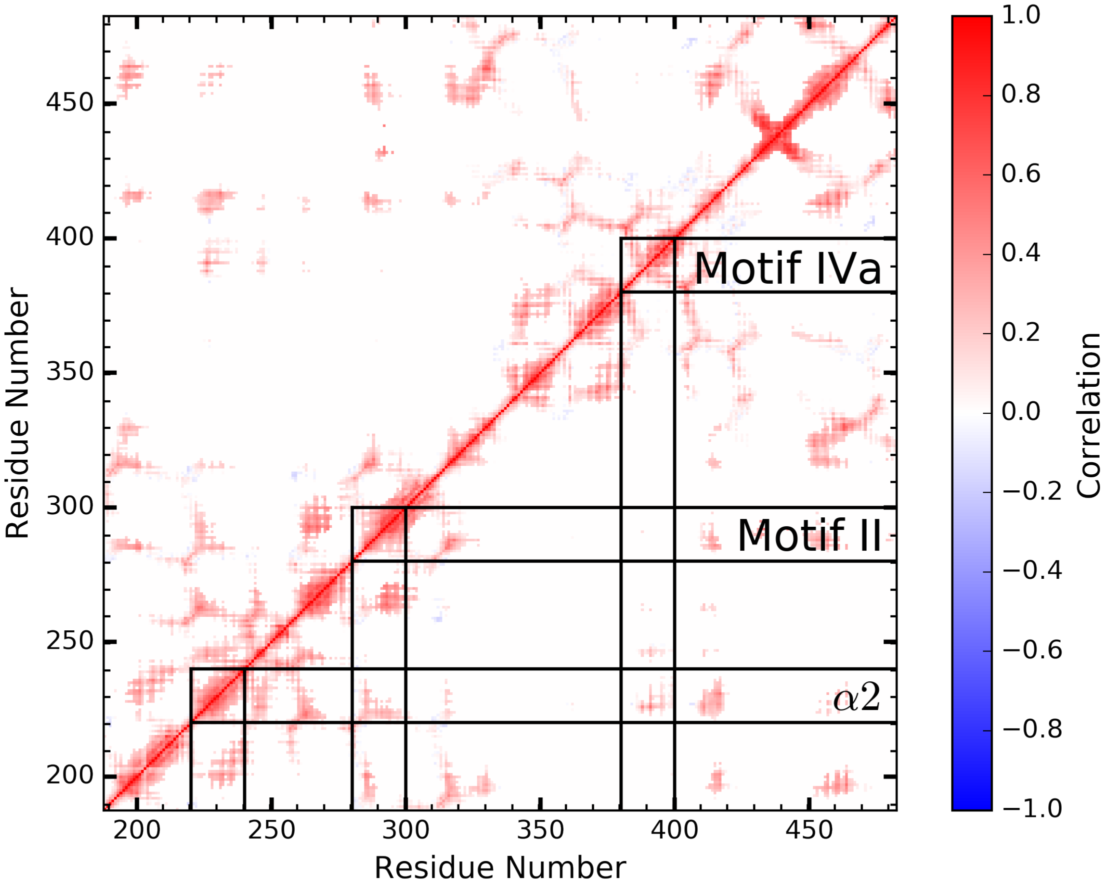

**Figure 2.22:** Residue-residue correlation heat map for the ssRNA+ADP substrate state. Residues of $\alpha 2$, motif II, and motif IVa are highlighted by the drawn lines.

# Chapter 3

# RNA-dependent Structures of the RNA-binding Loop in the Flavivirus NS3 Helicase.[2]

## 3.1 Overview

The flavivirus NS3 protein is a helicase that has pivotal functions during the viral genome replication process, where it unwinds double-stranded RNA and translocates along the nucleic acid polymer in a nucleoside triphosphate hydrolysis-dependent mechanism. Crystallographic and computational studies of the flavivirus NS3 helicase have identified the RNA-binding loop as an interesting structural element, which may function as an origin for the RNA-enhanced NTPase activity observed for this family of helicases. Microsecond-long unbiased molecular dynamics as well as extensive replica exchange umbrella sampling simulations of the Zika NS3 helicase have been performed to investigate the RNA-dependence of this loop's structural conformations. Specifically, the effect of the bound single-stranded RNA (ssRNA) oligomer on the putative "open" and "closed" conformations of this loop are studied. In the Apo substrate state, the two structures are nearly isoergonic ($\Delta G_{O\rightarrow C} = -0.22$ kcal mol$^{-1}$), explaining the structural ambiguity observed in Apo NS3h crystal structures. The bound ssRNA is seen to stabilize the "open" conformation ($\Delta G_{O\rightarrow C} = 1.97$ kcal mol$^{-1}$) through direct protein-RNA interactions at the top of the loop. Interestingly, a small ssRNA oligomer bound over 13 Å away from the loop is seen to affect the free energy surface to favor the "open" structure while minimizing barriers between the two states. The mechanism of the transition between "open" and "closed" states is characterized as are residues of importance for the RNA-binding loop structures. From these results, the loop is hypothesized to be a viable region in the protein for targeted small-molecule inhibition and mutagenesis studies,

[2] Russell B. Davidson[a], Josie Hendrix[a], Brian J. Geiss[b,c], Martin McCullagh[d]; [a] Department of Chemistry, Colorado State University, Fort Collins, CO, USA, [b] Department of Microbiology, Immunology, and Pathology, Colorado State University, Fort Collins, CO, USA, [c] School of Biomedical Engineering, Colorado State University, Fort Collins, CO, USA, [d] Department of Chemistry, Oklahoma State University, Stillwater, OK, USA

where stabilization of the “closed” RNA-binding loop will negatively impact RNA-binding and the RNA-enhanced NTPase activity.

## 3.2 Introduction

The flaviviruses (family *Flaviviridae*) present major health threats to the tropical and subtropical regions of the world. Over half of the world’s population lives in areas that are susceptible to infections from this viral family, which includes dengue, Zika, and West Nile viruses.[16,121] Outbreaks of Zika virus (ZIKV) in 2013 and 2014, in French Polynesia, as well as 2015, in Brazil, have been correlated with severe congenital malformations (microcephaly) as well as neurological complications such as Guillain-Barré syndrome.[19] The most recent ZIKV epidemic in the Americas sparked major research endeavors into identification of novel antiviral drugs against Zika and other flaviviruses.[19,122–125] Towards this goal, fundamental research on the viral enzymes’ structures and functions will aid the identification of antiviral targets.

The nonstructural protein 3 (NS3) of flaviviruses has been identified as one such target due to its pivotal role during the viral replication cycle.[24,26,125–134] As one of eight nonstructural proteins encoded by flaviviruses, NS3 is a multifunctional enzyme that possess an N-terminal serine protease domain and a C-terminal helicase/nucleoside triphosphatase (NTPase)/RNA triphosphatase domain.[32–37] The N-terminal protease is responsible for cleaving the viral polyprotein during translation. The C-terminal helicase domain (NS3h) unwinds the double-stranded RNA replication intermediate during the viral genome replication process.[38] In order to do so, NS3h binds and translocates along a single stranded RNA (ssRNA) polymer in a nucleoside triphosphate (NTP) hydrolysis-dependent mechanism. Therefore, the NS3h enzyme represents a complex molecular machine that has pivotal functions in the replication of the viral genome.

NS3h is categorized as a superfamily 2 (SF2) helicase (NS3/NPH-II subfamily) that hydrolyzes NTP to unwind double stranded RNA and translocate along the nucleic acid oligomer in a $3'$ to $5'$ direction.[43] A structural representation of ZIKV NS3h is shown in Figure 3.1A. The enzyme has three subdomains; subdomains 1 and 2 are RecA-like structures with large $\beta$-sheets forming the core of the subdomain while subdomain 3 is unique to the NS3 subfamily and less conserved structurally across the *Flaviviridae* viruses. The RNA-binding cleft, highlighted in orange, is a large channel that separates subdomain 3 from subdomains 1 and 2, with the 5′ terminus of the ssRNA oligomer positioned at the top of the protein in Figure 3.1A. The NTP hydrolysis active site, highlighted in purple, is positioned between the beta sheets of subdomains 1 and 2.

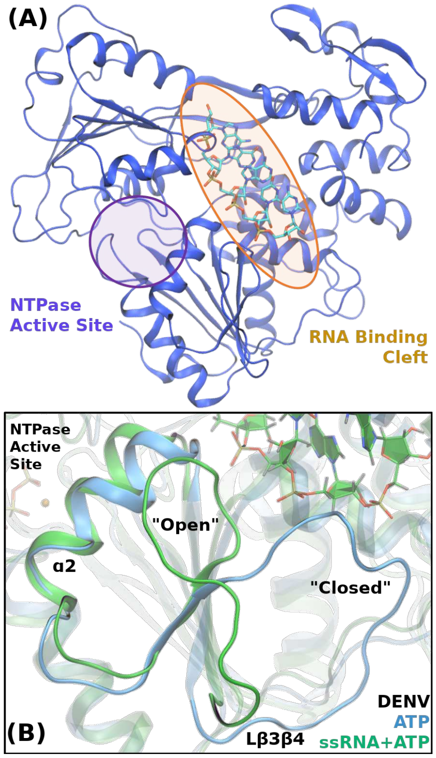

**Figure 3.1: Conformational states of the RNA-binding loop of the flavivirus NS3h.** (A) The ssRNA substrate state of the Zika NS3h (PDB 5GJB). The NTPase and RNA-binding clefts are highlighted in purple and orange, respectively. The 3′ terminus of the single-stranded RNA interacts with subdomain 1 and predominantly the RNA-binding loop. This loop region, also named L$\beta3\beta4$, is depicted in panel B, which has been adapted from Davidson *et al.*[7]. RNA-induced "open" and "closed" conformations of L$\beta3\beta4$ are depicted using the dengue NS3h ATP (blue) and ssRNA+ATP (green) crystal structures due to lack of the well-resolved "closed" conformations in Zika crystal structures.

Both active sites and respective functions are strongly coupled; NS3h is known to exhibit RNA-stimulated NTPase activity and NTPase-dependent helicase activity.[32–37,135] Coupling between these two active sites has been the subject of numerous recent studies.[7,71,135–138] The RNA-binding loop, shown in Figure 3.1B, is highlighted in this body of research as a site of allosteric structural change induced by the RNA oligomer, which we hypothesize is one of the origins of the RNA-enhanced NTPase activity.[7] This loop structure represents a large region of the RNA-binding cleft where the 3′ terminus of the bound RNA interacts with subdomain 1. Positioned between $\beta3$ and $\beta4$ of the subdomain 1 $\beta$-sheet, the RNA-binding loop (also termed L$\beta3\beta4$) is seen to have two major conformational states in flavivirus NS3h crystal structures. The dengue NS3h crystal structures presented in Figure 3.1B depict these two states: the "open" conformation has the loop closely interacting with $\alpha2$ (termed the spring helix by Gu and Rice[139]) while, in the "closed" conformation, the loop is positioned in close proximity to $\alpha3$ and is blocking the lower portion of the RNA-binding cleft.

During development of a dengue virus vaccine (DENVax), mutation of a NS3h residue in the RNA-binding loop (E250A) has been identified as one of three mutations producing a significantly attenuated phenotype.[133,134] It is unclear from these vaccine development studies why such a mutation results in the observed phenotype, specifically an increased temperature sensitivity and decreased viral replication. Yet, these results suggest that slight modification to the RNA-binding loop negatively affects the viral replication process.

The body of crystal structures available on the Protein Data Bank (PDB) of ZIKV and other flavivirus NS3h has left ambiguity as to the functional importance of the conformational states of the RNA-binding loop. Recently, Jain *et al.* highlighted the L$\beta3\beta4$ structure in their Apo ZIKV NS3h crystal structure (PDB ID: 5JRZ) relative to structures seen in other flavivirus NS3h.[136] In this crystal structure, the RNA-binding loop was resolved in the "open" conformation, which was insightful because the "open" conformation had previously been associated with RNA-bound structures.[60,140] In non-RNA-bound conformations, L$\beta3\beta4$ had always been resolved in the "closed" conformation or poorly resolved due to high flexibility of the loop region.

Additionally, a pair of molecular dynamics studies have recently provided insight into the L$\beta3\beta4$ structural states. Mottin *et al.* reported multiple 100 ns simulations of the ZIKV Apo and ssRNA-bound structures.[71] Large structural fluctuations of L$\beta3\beta4$ were observed in the Apo simulations with dampened conformational fluctuations seen in the ssRNA simulations. Minimal quantification of the observed L$\beta3\beta4$ structural states were performed for these simulations. In Davidson *et al.*, we reported micro-second long simulations of dengue NS3h structures where the "open" conformation maintained direct interactions with residues in $\alpha2$, which is seen to push the helix in towards the NTPase active site and, thereby, affect the behaviors of active site water molecules.[7] These results led us to hypothesize that the RNA-binding loop is an allosteric site where RNA affects the NTPase activity through the loop-$\alpha2$ interactions observed in the "open" conformation.

We report here a directed study of the L$\beta3\beta4$ conformational states in an effort to understand the effect RNA has on the loop. Micro-second long MD simulations of the ZIKV Apo, ssRNA, and an artificially-altered ssRNA:NS3h systems were used to sample the loop conformations in different RNA-bound substrate states. In the simulation of the Apo system, a transition from the "open" to "closed" conformation is observed and quantified using numerous methods. Finally, replica exchange umbrella sampling (REUS) simulations were used to enhance the sampling of the RNA-binding loop's conformations in the three substrate states, allowing us to quantify the RNA-induced effect on the loop's structural free energy landscape. This body of simulations demonstrates that an RNA oligomer perturbs the L$\beta3\beta4$ free energy surface to favor the "open" conformation over the "closed" conformation, even when the RNA is far removed from the loop.

## 3.3 Methods

### 3.3.1 Starting Structures and System Preparation

Initial structures for the reported all-atom, explicit solvent MD simulations originated from the ZIKV Apo NS3h (PDB: 5JRZ) and binary NS3h:ssRNA (PDB: 5GJB) crystal structures.[136,140] Additionally, a ssRNA-bound system was artificially created from the 5GJB conformation, where three of the five nucleotides were removed from the $3'$ end of the RNA oligomer. The remaining RNA in this structure is positioned at the top of the RNA-binding cleft, $\sim$ 13.5 Å away from the closest L$\beta3\beta4$ residue. This structure and respective simulations will be referred to as ssRNA$_{1-2}$.

### 3.3.2 Molecular Dynamics Simulations

All-atom, explicit solvent MD simulations were performed for the three conformations discussed above (denoted Apo, ssRNA, and ssRNA$_{1-2}$). The simulations were performed using the GPU-enabled AMBER16 and AMBER18 software[141,142], ff14SB[15] parameters for proteins, and ff99bsc0$_{\chi\,\mathrm{OL3}}$[73,74] parameters for RNA. For each system, the starting structures were solvated in TIP3P water boxes with at least a 12 Å buffer between the protein and periodic images. Crystallographic waters were maintained. Sodium and chloride ions were added to neutralize charge and maintain a 0.10 M ionic concentration. The Langevin dynamics thermostat and Monte Carlo barostat were used to maintain the systems at 300 K and 1 bar. Direct nonbonding interactions were calculated up to a 12 Å distance cutoff. The SHAKE algorithm was used to constrain covalent bonds that include hydrogen.[77] The particle-mesh Ewald method was used to account for long-ranged electrostatic interactions.[78] A 2 fs integration time step was used with energies and positions written every 5 ps. The minimum amount of simulation performed for each system was one trajectory of 1.0 $\mu$s, with the first 200 ns of simulation sacrificed to equilibration of the starting structures. Simulation of the Apo system was performed to 1.3 $\mu$s in order to thoroughly sample the "closed" L$\beta3\beta4$ conformation.

### 3.3.3 Adaptive Sampling of L$\beta3\beta4$ Transition

Neither the ssRNA nor ssRNA$_{1-2}$ systems sampled the L$\beta3\beta4$ conformation transition during the respective microsecond MD simulations. To obtain structures of the "closed" conformation, three independent, 100 ns steered molecular dynamics (SMD) simulations were used to slowly pull L$\beta3\beta4$ into poorly sampled structural space.[143] For these SMD trajectories, the pulling collective variable was the distance between the Val250 (L$\beta3\beta4$) and Arg269 ($\alpha3$) C$\alpha$ atoms. A pulling force constant of 20 kcal mol$^{-1}$ Å$^{-2}$ was used; all other simulation parameters were maintained as above. The rate of pulling was approximately 0.5 Å ns$^{-1}$.

Frames from this set of SMD trajectories were subsequently used to initiate 40 independent, unbiased trajectories, each 50 ns long. Selection of starting frames for these trajectories was performed using a two dimensional projection of the unbiased and SMD trajectories onto essential dynamics (discussed in the Supporting information, SI) eigenvectors. Regions of poor sampling in this projection space were identified and frames associated with these regions were used in these independent, unbiased trajectories. This body of simulation represents an extra 2 $\mu$s of simulation for both the ssRNA and ssRNA$_{1-2}$ systems.

### 3.3.4 Replica Exchange Umbrella Sampling (REUS) Simulations

The REUS method was used to enhance the sampling of the L$\beta3\beta4$ conformational space for the Apo, ssRNA, and ssRNA$_{1-2}$.[144] The distance between Ala230 ($\alpha2$) and Ala247 (L$\beta3\beta4$) C$\alpha$ atoms was used as the biased collective variable. This single atomic displacement distance was chosen because it has the largest free energy barrier from the Apo unbiased simulation's sampling of the "open" to "closed" conformational change. Additionally, the Ala230 C$\alpha$ atom has small positional variance due to being in $\alpha2$; the large changes in this distance collective variable occurring during the transition are strongly correlated with the L$\beta3\beta4$ transition.

For each of the three systems described above, 40 windows were used in the REUS simulations. Equilibrium wells of these windows range from 3.625 Å to 18.250 Å with these equilibrium wells separated by 0.375 Å. Harmonic, biasing force constants were 20 kcal mol$^{-1}$ Å$^{-2}$. Window exchanges were attempted every 25 ps, with an average accepted rate of 0.37 attempts. The biased CV values are written every 200 fs, while frames and energies are written every 2 ps. For each window, a total of 46 ns of REUS simulations was performed, amounting to a total of 1.84 $\mu$s of enhanced sampling of the L$\beta3\beta4$ structures for each of the systems.

Due to the protocol used to obtain structures of the "closed" conformation in the ssRNA and ssRNA$_{1-2}$ systems, the first 10 ns of the REUS simulations were not included in free energy analyses. We use the eigenvector method for umbrella sampling (EMUS) analysis package to calculate the stitched free energy surfaces for these REUS simulations.[145]

### 3.3.5 Data Analysis

Analyses of MD trajectories were performed using Python 3.7.4 and the MDAnalysis module (version 0.19.2).[85] Matplotlib was used for plotting data.[86] VMD was used for visualization of trajectories and production of structural figures.[87–89] All analysis scripts are available on Github (https://github.com/mccullaghlab/ZIKV-Lb3b4). Additional analysis details are provided in the SI.

## 3.4 Results and Discussion

The structure-function relationship of the RNA-binding loop is poorly understood, especially when considering the RNA-dependence of the loop's structure. 56 monomeric flavivirus NS3h structures have been reported on the Protein Data Bank (PDB), of which a large proportion have poorly resolved electron densities for L$\beta3\beta4$ atoms. Additionally, a multiple sequence alignment of the flavivirus NS3h sequences indicates that the loop is one of the least conserved regions of the protein, which generally suggests low functional importance of sequence positions. See the SI for further discussion of the structural and bioinformatic analyses of flavivirus NS3h. Yet, as presented in Figure 3.1B, large structural changes are seen in the loop when comparing RNA-bound and Apo structures, a common trend across all flaviviruses. This ambiguity around the RNA-binding loop's functional importance has motivated our molecular dynamics study of the loop and, more specifically, its RNA-dependent conformational states.

Results of this study are presented in two sections. The first section highlights the L$\beta3\beta4$ structural states observed during the unbiased, microsecond-long MD simulations of Apo, ssRNA, and ssRNA$_{1-2}$ systems. Focus is given to the thorough quantification of a transition between the "open" and "closed" loop conformational states, observed during the Apo NS3h simulation, as well as specific residues hypothesized to be functionally important during this transition. The second section reports free energy surfaces of the structural transition, as obtained from REUS enhanced sampling simulations, to highlight the effect that the bound RNA oligomer has on L$\beta3\beta4$'s structure.

### 3.4.1 L$\beta3\beta4$ Conformations in MD Simulations.

Microsecond-long simulations of Zika NS3h Apo (5JRZ), ssRNA (5GJB), and an artificially-altered ssRNA system (ssRNA$_{1-2}$) have been performed to sample the RNA-binding loop's conformations in the presence or absence of an RNA oligomer. Starting structures for all three systems have L$\beta3\beta4$ in the "open" conformation, directly interacting with residues of $\alpha2$. Shown in Figure 3.2A, the root mean square deviation (RMSD) of the loop backbone atoms for these simulations, referenced against the 5GJB structure, indicate that large structural deviations occur during the Apo simulation while RNA-bound systems maintain the "open" loop conformation. The small structural deviations seen during the ssRNA simulation are expected due to strong interactions between the 3′ RNA nucleotides and protein residues at the top of the loop. Surprisingly, artificial removal of three 3′ nucleotides has a limited effect on the L$\beta3\beta4$ conformational sampling, as seen in Figure 3.2. This suggests that an RNA-binding loop conformational change is a rare event that may be affected via indirect RNA-loop interactions.

Large structural deviations of the loop have been observed by Mottin *et al.* in 100 ns MD simulations of the 5JRZ and 5JMT crystal structures.[71] A large transition between the "open" and "closed" loop conformations is observed in our Apo simulation, quantified by RMSD and structurally depicted in Figure 3.2. During the initial 900 ns of this simulation, L$\beta3\beta4$ fluctuates about the "open" crystal structure conformation (small RMSD values, red to light green colors). The transition to the "closed" structure begins to occur at 900 ns (light green to turquoise), seen respectively by a large jump in the loop backbone atom RMSD. The loop begins to sample the "closed" conformation by 1 $\mu$s (turquoise to blue), although it is unclear if the transition has completed due to the poor resolution of the "closed" conformation in ZIKV crystal structures.

**Residues of interest for L$\beta3\beta4$ conformations.**

Certain residues in the local region of L$\beta3\beta4$ are hypothesized to have important functions in relation to the loop's structural states as well as protein-RNA interactions. Such residues are highlighted in Figure 3.3 and will be discussed in further detail here to present their hypothesized or observed importance.

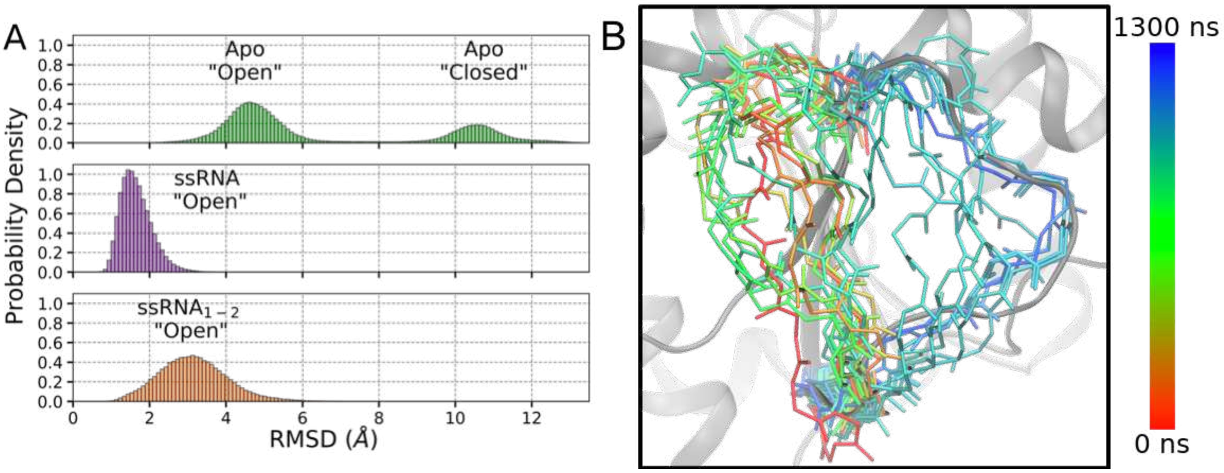

**Figure 3.2: Structural fluctuations of L$\beta$3$\beta$4 during the MD simulations.** (A) Root mean square deviation (RMSD) analysis of residues 246 to 254, relative to the 5GJB crystal structure. Large RMSD values, seen in the Apo results, indicate a large structural shift away from the "open" loop conformation occurring during that simulation. (B) Structural representation of L$\beta$3$\beta$4 backbone atoms over many time steps of the Apo simulation. At $\sim$ 900 ns (light green to turquoise), the loop structure begins to transition from the "open" to "closed" conformation.

The three arginine residues highlighted in Figure 3.3A are ideally positioned within the RNA-binding cleft to function as arginine forks, residues that can strongly coordinate two adjacent phosphate groups of nucleic substrates.[106,107] Arg226 sits at the top of $\alpha$2, directly interacts with co-crystallized RNA in the 5GJB structure, and is a highly conserved residue of Motif Ia (0.13% sequence variance in flavivirus NS3h sequences).[43] Numerous crystal structures have resolved the guanidinium functional group of Arg226 in coordination with the RNA phosphate backbone of 3′ terminal nucleotides. Additional arginine residues that are highlighted in panel A, Arg242 (0.67% sequence variance) and Arg269 (1.88% sequence variance), are positioned in $\beta$3 and $\alpha$3, respectively. These residues have not been observed in crystal structures to coordinate RNA due to the limited resolution of the co-crystallized RNA. Yet, their sequence conservation and ideal positioning within the RNA-binding cleft support our hypothesis that these arginine residues play important roles in binding RNA.

Residues Thr245 and Thr246, shown in Figure 3.3B, and are positioned at the structural change between $\beta$3 and L$\beta$3$\beta$4. Although less conserved than the residues discussed above, both Thr245 (7.80%) and Thr246 (58.0% sequence variance) are seen to have large structural changes that occur

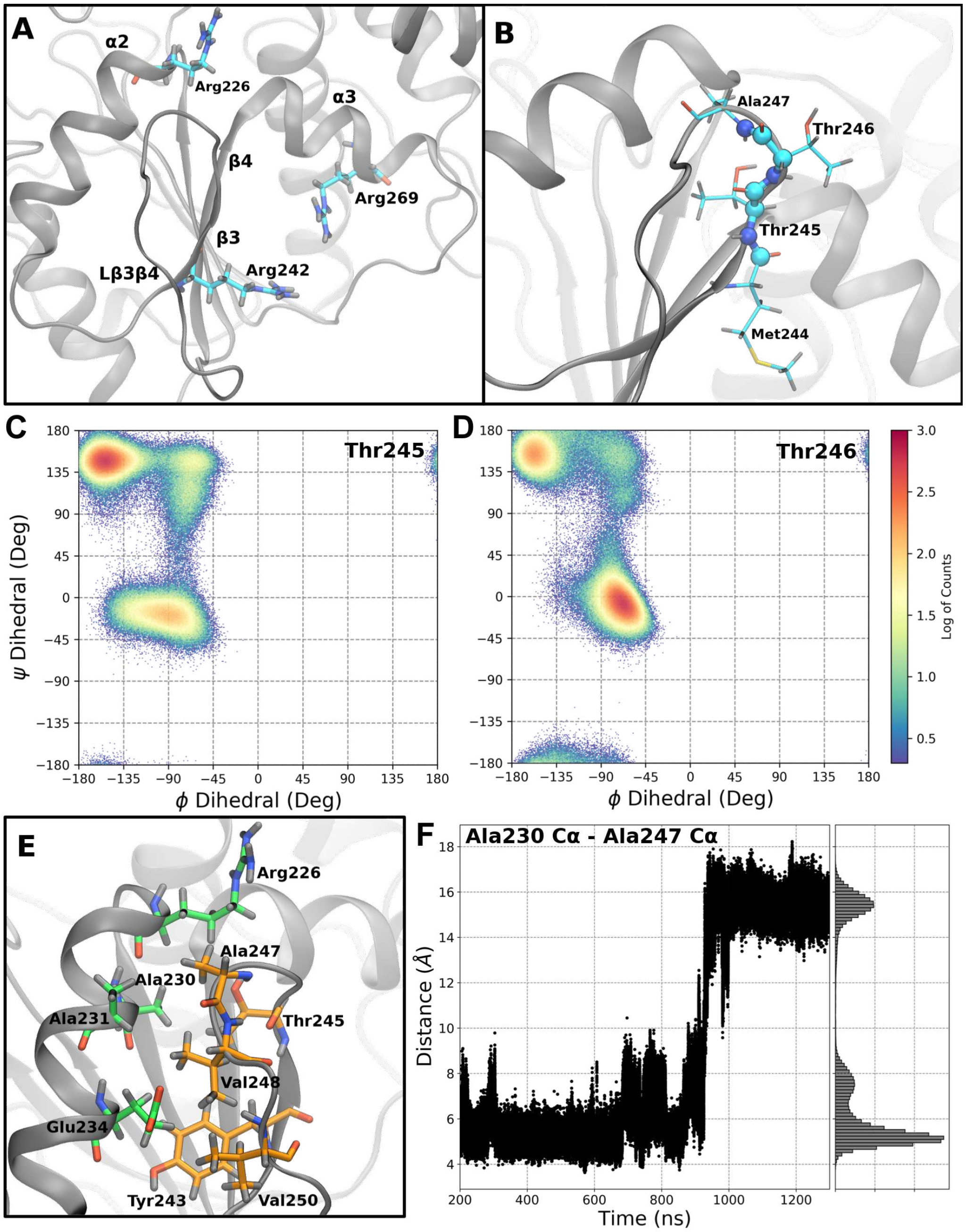

**Figure 3.3: Residues and collective variables that are good descriptors of the "open" to "closed" transition and conformational states.** (A) Arginine residues in the RNA-binding cleft, local to L$\beta3\beta4$. (B) Thr245-Thr246 are residues positioned at the transition between $\beta3$ and L$\beta3\beta4$. Ramachandran plots of Thr245 (C) and Thr246 (D) highlight a dihedral switch in these two residues occurring during the loop structural transition. (E) Residues in $\alpha2$ (green) and L$\beta3\beta4$ (orange) that form a small, stable hydrophobic pocket between the two secondary structures. (F) The time evolution of the distance between Ala230 ($\alpha2$) and Ala247 (L$\beta3\beta4$) C$\alpha$ atoms describes the breakup of the hydrophobic pocket and subsequent loop structural transition.

during the “open” to “closed” transition observed in the Apo simulation. Specifically, the loop’s “open” to “closed” conformational change is well described by the simultaneous switching of the backbone dihedrals of these two residues, presented in the Ramachandran plots shown in Figure 3.3C and D. This observation is supported by these residues’ dihedral values in the flavivirus NS3h crystal structures (see SI). In the “open” conformation of 5JRZ, Thr245 and Thr246 have $\phi$ and $\psi$ dihedral values of (-155°, 141°) and (-79°, -1°), respectively. Conversely, in crystal structures with L$\beta3\beta4$ in the poorly resolved, “closed” conformation (e.g. 5K8L), these residues have $\phi$ and $\psi$ dihedral values of (-76°, -12°) and (-141°, 168°), respectively. The Ramachandran plots of Thr245 and Thr246 (Figure 3.3C and D, respectively) have strong sampling in both dihedral spaces associated with the “open” and “closed” loop conformations. The time evolution of these four dihedrals are shown in Supporting Figure 3.9-3.12. While the loop shifts from the “open” to the “closed” conformation, large concerted flips in the Thr244 and Thr245 $\psi$ dihedrals are observed, followed by correlated fluctuations of these residues’ $\phi$ dihedrals. The dihedral states of these two residues describe the backbone structural conformation of the top of L$\beta3\beta4$ and so, are hypothesized to be strong decisors of the loop’s conformational state. We hypothesize that these dihedrals will switch during transitions between the “open” and “closed” conformations of the loop for all flavivirus NS3h.

A set of hydrophobic residues in $\alpha2$ and L$\beta3\beta4$ form a small, hydrophobic pocket when the RNA-binding loop is in the “open” conformation.[7] This hydrophobic pocket is structurally depicted in Figure 3.3E, where Ala230 and Ala231 ($\alpha2$) and Ala247, Val248, and Val250 (L$\beta3\beta4$) are in direct interaction with each other. Additionally, Tyr243 ($\beta3$) and aliphatic portions of Arg226, Glu234, and Thr245 are neighboring this small hydrophobic region. These stabilizing, hydrophobic interactions are lost in the “closed” conformation of L$\beta3\beta4$, as quantified by the C$\alpha$-C$\alpha$ distance between Ala230 and Ala247 shown in Figure 3.3F. This distance metric describes the breakup of the depicted hydrophobic pocket and is one of the atomic, pairwise distances with the largest change during the “open” to “closed” transition observed in the Apo simulation.

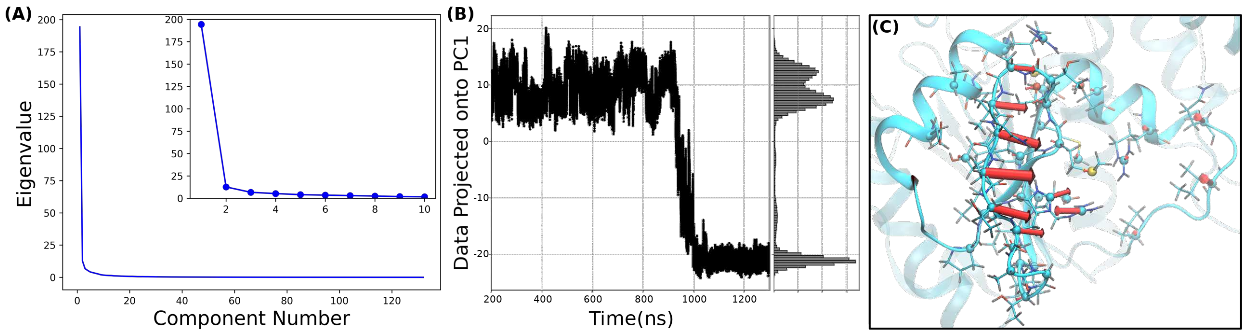

**Figure 3.4: Zika NS3h Apo's essential dynamics in the L$\beta$3$\beta$4 region.** (A) The scree plot for the essential dynamics analysis indicates that PC1 is the only major eigenvector to be considered. (B) Projection of the trajectory's data onto PC1 clearly separates the "open" (positive values) and "closed" (negative values) L$\beta3\beta4$ conformations. (C) The porcupine plot of the PC1 highlights the correlated fluctuations of the loop residues' cartesian coordinates during the transition between the two conformational states.

### Essential Dynamics of the L$\beta$3$\beta$4 Transition.

The presented RMSD, dihedral, and distance collective variables clearly delineate the "open" and "closed" conformations of L$\beta3\beta4$, but there are additional important coordinates necessary to describe the observed transition. To quantify these large-scale motions, the essential dynamics (ED) of L$\beta3\beta4$ were analyzed. Principal component analysis of the L$\beta3\beta4$ cartesian coordinates will appropriately quantify the two structural states because the transition between "open" and "closed" conformations represents the largest covariance motions in the coordinate space. Choice of coarse-grained sites for the ED analysis were guided by the multiple sequence analysis results (see SI); residues near the loop region with high sequence conservation were prescribed two sites, describing the backbone and side chain fluctuations independently. Backbone atoms were used to describe the residue-level fluctuations for sequence positions with low conservation or small side chains (e.g. Gly, Ala, and Val). The protocol for determining these atomic coarse-grained sites, detailed in the SI, resulted in 44 atoms that were used to describe the L$\beta3\beta4$ transition.

The L$\beta3\beta4$ transition dominates the first eigenvector of the ED analysis, as demonstrated in the scree plot shown in Figure 3.4A. This eigenvector accounts for 75% of the total variance of the atom selection. The remaining eigenvectors have sufficiently small magnitude eigenvalues and, so, are disregarded for the remainder of this study. The cartesian coordinate data projected onto eigenvector 1 is presented in Figure 3.4B, where the transition is clearly observed. The "open" conformation is represented by positive projected data values. Additionally, the transition seen at $\sim$900 ns in Figure 3.2 is also observed in the projected data. The "closed" conformation is described by large magnitude negative values, with smaller magnitude values representing intermediate loop conformations during the transition.

Additionally, the first eigenvector highlights the structural motions that occur during the "open" to "closed" transition. Figure 3.4C shows the porcupine plot of this eigenvector. The residues with large magnitude vectors in this figure have the largest covariance during the "open" to "closed" conformational change described by PC1. Specifically, C$\alpha$ atoms of the loop have the largest magnitude cartesian vectors, indicating strong representation of these residues' fluctuations in the most dominant eigenvector of the ED analysis. As expected, the direction of these vectors are associated with the "open" to "closed" transition observed in the MD simulation. Interestingly, Arg242 has a large magnitude vector as well pointing in the opposite direction of the loop residues' large magnitude vectors.

### 3.4.2 Free Energy Landscape of the RNA-binding Loop Conformations.

We have performed extensive replica exchange umbrella sampling (REUS) simulations to enhance the sampling of the conformational space of L$\beta3\beta4$ in the Apo, ssRNA, and ssRNA$_{1-2}$ systems. Unfortunately, PC1 cannot be used as the biased collective variable in these REUS simulations due to the non-trivial task of applying a bias on the linear combination of atoms' coordinates. Instead, the distance between Ala230 ($\alpha2$) and Ala247 (L$\beta3\beta4$) C$\alpha$ atoms was used as proxy for the PC1 reaction coordinate. As previously discussed, this distance represents a large collective variable space and describes the breakup of the hydrophobic interactions between $\alpha2$ and L$\beta3\beta4$ (see Figure 3.3). Finally, the dividing surface between the "open" and "closed" conformational states was chosen as the respective collective variable value associated with the highest free energy barrier between the two equilibrium wells. With this dividing surface defined for each system in the Ala230-Ala247 C$\alpha$ distance collective variable space, the relative free energy of the

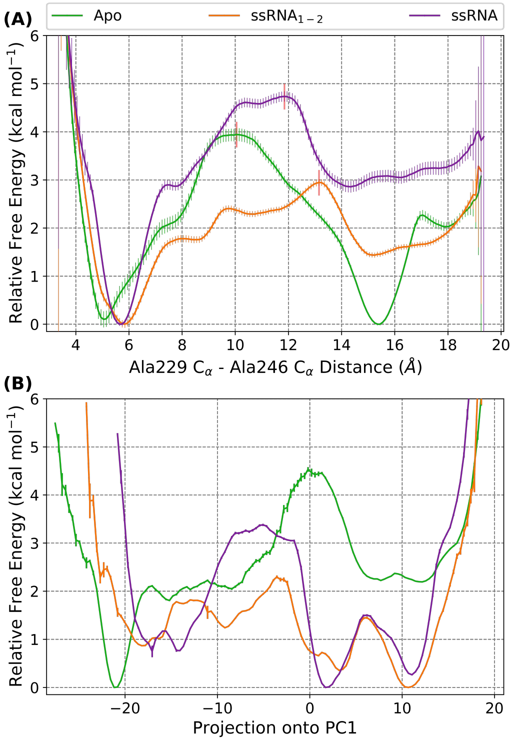

**Figure 3.5: Free energy surfaces for the Apo, ssRNA, and ssRNA$_{1-2}$ systems as projected on the biased CV and PC1.** (A) Small and large distance values represents the "open" and "closed" conformations, respectively. The transition barrier, shown as a vertical red line, is used to separate the two conformational states of L$\beta3\beta4$. Error bars were calculated by EMUS, which accounts for the decorrelation time of the collective variable. (B) Positive and negative values correspond to the "closed" and "open" conformations, respectively. Error bars were measured using bootstrapping and so likely under approximate the error in the free energy surfaces.

two L$\beta3\beta4$ conformations was approximated by integrating the free energy surface on either side of the divide.

### Relative free energy differences of Apo, ssRNA, and ssRNA$_{1-2}$ systems.

The free energy surfaces for the biased CV are shown in Figure 3.5A, where small ($\sim$4 to 7 Å) and large ($\sim$14 to 18 Å) values correspond to the "open" and "closed" conformations, respectively. In this projection, the Apo substrate state has nearly isoergonic "open" and "closed" conformations ($\Delta G_{O\to C} = -0.22 \pm 0.04$ kcal mol$^{-1}$), while the two RNA-bound states favor the "open" conformation ($\Delta G_{O\to C} = 0.95 \pm 0.02$ for ssRNA$_{1-2}$ and $1.97 \pm 0.03$ kcal mol$^{-1}$ for ssRNA). The small magnitude $\Delta G_{O\to C}$ for the Apo system corroborates the ensemble of observed

RNA-binding loop conformations in crystal structures, where both the “open” and poorly resolved “closed” conformations have been reported numerous times.

Both the $\Delta G_{O\to C}$ and barrier heights shown in Figure 3.5A highlight the destabilization of the “closed” conformation of L$\beta3\beta4$ in the presence of RNA. For the ssRNA system, five nucleotides were crystallized in the 5GJB crystal structure. The $3'$ end of this RNA oligomer is positioned just at the top of L$\beta3\beta4$, as seen in Supporting Figure 3.8B. The large $\Delta G_{O\to C}$ and barrier height for the ssRNA system indicate that the close proximity of the $3'$ nucleotides locks the RNA-binding loop into the “open” conformation.

Surprisingly, RNA perturbs the loop’s conformational free energy even when the RNA oligomer is $> 13$ Å away from the loop, as is the case for the ssRNA$_{1-2}$ system where three $3'$ nucleotides were artificially removed. The remaining, shortened RNA oligomer is positioned between subdomains 2 and 3, between the helical gate $\alpha$-helices. Yet, at this distance, the RNA decreases the free energy barrier between the “open” and “closed” loop conformations relative to both the Apo and ssRNA free energy surfaces. Additionally, this small RNA oligomer shifts the $\Delta G_{O\to C}$ from favoring the “closed” conformation (as the Apo system does) to favoring the “open” loop.

These results demonstrate that the RNA affects the L$\beta3\beta4$ free energy surface. In fact, the presence of even minimal RNA at the top of the RNA-binding cleft is seen to perturb the loop’s free energy surface in such a way as to enable the “closed” to “open” transition to occur more readily than in either of the other two substrate states. This may provide insight into the mechanisms of molecular recognition between NS3h and RNA. Specifically, incremental representations of an RNA oligomer bound within NS3h’s RNA-binding cleft are seen to modulate the free energy surface of L$\beta3\beta4$. Due to its approximately isoergonic $\Delta G_{O\to C}$ value, the Apo substrate state of NS3h is hypothesized to sample both the “closed” and “open” conformations of the RNA-binding loop; this is strongly supported by the crystal structures of the flavivirus NS3h. The free energy differences of the Apo and ssRNA$_{1-2}$ systems demonstrate that an RNA oligomer positioned between subdomains 2 and 3 ($> 13$ Å away from L$\beta3\beta4$) induces a population shift away from the “closed” conformation of the loop. This shift is aided by the diminished free energy barrier seen in the ssRNA$_{1-2}$ system's free energy surface, relative to Apo's. Further presence of RNA enhances this energetic shift in the free energy surface, as seen in the $\Delta G_{O \to C}$ for the ssRNA system.

Apo NS3h can sample both the "open" or the "closed" conformations, yet initial binding of RNA drives the free energy surface to favor "open" structures. This results in a population shift away from "closed" structures that is sustained by further binding of the RNA oligomer. Since we do not model the true binding event of RNA into NS3h, we cannot definitively state what mechanism best describes the molecular recognition of RNA by L$\beta 3 \beta 4$. Yet, the free energy surfaces reported above suggest that the "conformational selection" model of molecular recognition events may be at play for NS3h-RNA complexes, at least in regards to the RNA-binding loop.[146–148]

**Free energy surface of the L$\beta 3 \beta 4$ transition.**

As stated previously, the biased distance between Ala230 and Ala247 C$\alpha$ atoms was used as a proxy to describe the PC1 eigenvector from the Apo system's ED analysis. Therefore, the REUS simulations were projected onto PC1 to investigate the free energy surface of the loop conformational transition described by this eigenvector. The reweighted free energy surfaces associated with the projection onto PC1 are shown in Figure 3.5B. For these projected free energy surfaces, less emphasis is placed on the $\Delta G_{O \to C}$ values due to the larger collective variable space having lower quality sampling for all three systems. In conjunction with the sampling issues in this reweighted space, 1000 iterations of bootstrapping was used to estimate the error in the surfaces; this error is an approximate lower-bound for the true error in these results. Instead, focus is given to the features within these free energy surfaces and their mechanistic interpretations.

**Apo's L$\beta 3 \beta 4$ transition mechanism.** The Apo system's free energy surface in the projected PC1 space (green in Figure 3.5B) has two major free energy barriers, with the direction of the transition (i.e. "open" to "closed" or the reverse) greatly changing the energetics. From left to right, the first barrier is seen at PC1 $\approx$ -17 and separates the "closed" conformation from an intermediate "closed" conformation, where the barrier heights of 2.1 or 0.3 kcal mol$^{-1}$ are seen for the "closed" to intermediate and reverse transitions, respectively. Structural representations of these two states

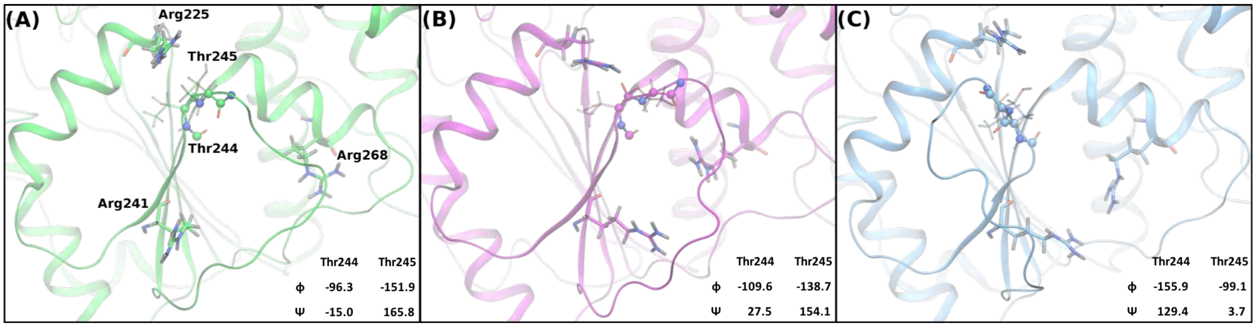

**Figure 3.6: Exemplar structures of the Apo system's L$\beta3\beta4$ conformations at local minima in the PC1 projected free energy surface.** (A) "Closed" conformation where Arg242 is solvent exposed and Thr245 and Thr246 dihedrals sample the expected "closed" values. (B) The intermediate "closed" conformation where Arg242 has transitioned to the opposite side of L$\beta3\beta4$, relative to its position in (A). (C) "Open" conformation where the Thr245 and Thr246 dihedrals have flipped into the putative "open" dihedral values and the Arg242 and Arg269 residues sit in the RNA-binding cleft.

are shown in Figure 3.6A and B, where the loop conformation remains closed, as supported by the dihedral values of Thr245 and Thr246 residues. The largest structural change that separates these two conformations is the Arg242 positioning. In the "closed" conformation, the side chain of this residue sits in a solvent exposed location while, in the intermediate state, the guanidinium group has repositioned into a protein-internal location on the opposite side of the L$\beta3\beta4$ backbone. This Arg242 repositioning is highlighted in the PC1 porcupine plot (Figure 3.4C) where the residue's large magnitude vector is aimed in the opposite direction of the loop's set of vectors. The Arg242 repositioning represents a large destabilization of the "closed" conformation with a very small barrier in the reverse direction, suggesting that the intermediate "closed" structure is rare and short lived.

The second barrier represents the transition between the intermediate "closed" and "open" conformations, seen at PC1 $\approx 0$. An exemplar structure of the "open" conformation is presented in Figure 3.6C, where the Thr245 and Thr246 dihedrals are seen in the "open" conformation dihedral space. As expected due to the L$\beta3\beta4$ structure seen in 5JRZ, arginines 226, 242, and 269 are all positioned in the RNA-binding cleft in the "open" conformation. For this transition, the two states are nearly isoergonic with a barrier heights of 2.5 and 2.3 kcal mol$^{-1}$ when transitioning from the intermediate to the "open" structure and the reverse, respectively. With respect to the

PC1 porcupine plot, this barrier represents the large structural change of the L$\beta3\beta4$ residues. The combination of the two barriers seen in the Apo system's PC1 free energy surface highlights the downhill energetics of the "open" to "closed" transition that was observed in the Apo unbiased simulations.

**ssRNA systems sample the PC1 transition poorly.** Both the ssRNA and ssRNA$_{1-2}$ systems' PC1-projected free energy surfaces have major differences from the Apo system's surface in Figure 3.5B. Important to note are regions on these surfaces with discontinuities or large changes in slope, especially in the transition barriers. The presence of such data indicates non-ergodic sampling at these positions in the PC1 space and suggests that the transition described by PC1 is not viable in the respective substrate state. Even with these caveats, the PC1-projected free energy surfaces of the ssRNA and ssRNA$_{1-2}$ systems are qualitatively similar to the results in the biased CV free energy surfaces. Both systems favor the "open" conformation of L$\beta3\beta4$ (positive values). The ssRNA$_{1-2}$ oligomer reduces the barrier heights seen in the PC1-projected space relative to those seen in the Apo or ssRNA free energy surfaces.

The ssRNA system's surface has one large barrier between the two conformational energy wells, approximated as 2.6 and 3.4 kcal mol$^{-1}$ for the "closed" to "open" transition and the reverse, respectively. The transition barrier for this system is the most prominent example of nonergodic sampling and, so, the barrier height is likely under approximated. Furthermore, no intermediate structures between the two wells are seen to be energetically stable and the sampling of negative-valued PC1 space is drastically narrowed. These results suggest that the PC1 eigenvector describes a transition mechanism between "open" and "closed" conformations that is not viable in the ssRNA system. Furthermore, this demonstrates that the energetics of the PC1 transition are strongly affected by the ssRNA oligomer as also observed in the biased CV energetics.

Similarly, the qualitative story observed in the biased CV free energy surface of the ssRNA$_{1-2}$ system is maintained in the PC1-projection surface. The small RNA oligomer stabilizes the "open" conformational state while also minimizing the barrier heights between "closed" and intermediate L$\beta3\beta4$ structures. The narrowing of PC1 sampling seen for the ssRNA system is not as drastic in the ssRNA$_{1-2}$ system, yet the "closed" conformational well is still displaced to more positive PC1 values. For this system, the structural origins of the RNA oligomer's effect on the L$\beta3\beta4$ free energy surface are difficult to identify from the body of simulations reported here. Further study of the numerous RNA-bound structural states of the flavivirus NS3h will need to be performed to fully deconvolute these results.

## 3.5 Conclusions

The L$\beta3\beta4$ conformations observed in the flavivirus NS3h crystal structures were investigated using all-atom, explicit solvent MD simulations of the Apo, ssRNA, and ssRNA$_{1-2}$ substrate states. A single observation of the "open" to "closed" conformational transition, seen in the unbiased Apo simulation, was studied to identify collective variables that efficiently quantified the structural changes observed. The backbone dihedrals of the Thr245 and Thr246 residues of L$\beta3\beta4$ are one such set of collective variables, where crystal structures and MD simulations highlight the large quantitative change in these dihedrals in the "open" and "closed" conformations. Additionally, an essential dynamics analysis of the observed transition produced a single, dominant eigenvector that described 75$\%$ of the positional covariance of the RNA-binding loop region. PC1 strongly separated the "open" and "closed" conformations and accounted for the large scale fluctuations of the full loop.

The free energy surfaces of the L$\beta3\beta4$ structural states were quantified using REUS simulations. Reweighting these free energy results into the PC1 collective variable space also allowed for us to study the energetics of the "open" to "closed" transition originally observed in the unbiased Apo simulation. In either the biased collective variable or PC1 spaces, the quantitative free energy results highlight the RNA-dependence of the L$\beta3\beta4$ structures. For Apo, a relatively large barrier separates the two structural states that are nearly isoergonic in the Ala230-Ala247 C$\alpha$ distance space. In the projected space, the transition from "open" to "closed" is much more energetically favorable than the reverse process. In this transition, two energetic steps are observed to occur: the backbone of the RNA-binding loop moves into the "open" conformation, as described by the

Thr245 and Thr246 dihedrals, followed by the side chain of Arg242 fluctuating to the opposite side of the loop to a position in the RNA-binding cleft.

For the ssRNA and ssRNA$_{1-2}$ systems, the "open" loop conformation is energetically favored in both free energy surfaces. Direct interactions between RNA and loop residues lead to an even larger barrier seen for the ssRNA system's biased collective variable free energy surface. Additionally, the transition described by PC1 is not viable in the ssRNA system as indicated by the nonergodic regions in the respective surface. Interestingly, the ssRNA$_{1-2}$ system has lower free energy barriers in both the biased collective variable and PC1 free energy surfaces. These results suggest that even a small RNA oligomer, bound in the RNA-binding cleft between subdomains 2 and 3, increases the rate at which L$\beta3\beta4$ transitions between the "closed" conformation (favored by Apo) and the "open" conformation. These results also suggest a possible conformational selection mechanism for the RNA-dependent L$\beta3\beta4$ structural states, where RNA positioned between subdomains 2 and 3 causes an energetic shift in the loop's structures to favor the RNA-bound, "open" conformation.

In light of our previous study of the dengue NS3h NTPase cycle, the L$\beta3\beta4$'s RNA-dependent free energy surface provides interesting insight into the hypothesis that RNA-induced enhancement of the NTP hydrolysis reaction originates, to some degree, from the "open" L$\beta3\beta4$ conformation. We previously reported that the number of water molecules within the hydrolysis active site is decreased as are these waters' translational and rotational motions when RNA is present.[7] Additionally, the positioning of water molecules within the active site was also observed to be affected by RNA. These effects are proposed to originate from RNA-induced structural changes in the hydrolysis active site, an example of which is a structural shift in $\alpha2$ brought on by the L$\beta3\beta4$ interactions while in the "open" conformation. Therefore, our free energy results suggest that initial binding states of the NS3h:RNA complex (modeled by ssRNA$_{1-2}$) enhance the L$\beta3\beta4$ transition from "closed" to "open", leading to structural shifts in $\alpha2$ that prime the NTPase active site for the hydrolysis reaction. Continued binding of RNA (modeled by ssRNA) locks in the $\alpha2$-L$\beta3\beta4$ interactions even further.

The structural origins of the ssRNA$_{1-2}$ oligomer's effect on L$\beta3\beta4$ have not been identified, yet our results highlight the RNA-binding loop as a potential target for small-molecule inhibition or mutagenesis studies. We hypothesize that destabilization or prevention of the "open" L$\beta3\beta4$ structure, while in the Apo or ssRNA$_{1-2}$ substrate states, would diminish the RNA-induced enhancement of the NTPase activity as well as lead to weakened protein-RNA interactions. The highly conserved Arg226, Arg242, and Arg269 are all hypothesized to function as arginine fork residues that strongly coordinate the phosphate groups of 3$'$ RNA nucleotides. Arg242 is also observed to play an important role in the "open" to "closed" transition described by PC1; stabilization of the solvent exposed conformation of Arg242 (seen in Figure 3.6) is proposed as a potential method of inhibiting the transition. Therefore, mutations or small molecules that perturb the wild-type behavior of L$\beta3\beta4$ would be detectable via NTPase, RNA binding, and helicase activity assays. Additionally, further computational studies of the NS3h:RNA structural states could provide insights into the short-range, residue-residue interactions through which the observed allosteric effect is propagated.

## 3.6 Funding

This work was supported by the National Science Foundation-supported (grant No. ACI-1548562) Extreme Science and Engineering Discovery Environment (XSEDE) (allocation No. CHE160008) and the National Science Foundation Research Experience for Undergraduates (CHE-1461040) and Fixman Summer Theory program to JH.

## 3.7 Supporting Information

### 3.7.1 Analyses

**Multiple Sequence Alignment**

The NIAID Virus Pathogen Database and Analysis Resource (ViPR)[149] (accessed on May 30th, 2018 through the web site at http://www.viprbrc.org/) was used to collect the available *Flaviviridae* NS3 sequences, totaling 51,816 independent sequences from all four genuses of the virus family.

The Biopython module[150] (version 1.74) was used to preprocess this population of NS3 sequences. Specifically, this preprocessing analysis removed uninteresting sequences (not an NS3h sequence, too short to be full NS3h, or poorly resolved). Subsequently, the CD-HIT software package was used to perform a three-step clustering analysis.[151,152] The first clustering step used a clustering threshold of 1.00, which was used to aggregate non-unique sequences into a single, representative sequence. This was done to normalize the weight that each unique sequence has in the population of sequences. The second step in the sequence clustering protocol used a clustering threshold of 0.4, which successfully clustered sequences by the genera of the *Flaviviridae* family (flavivirus, hepacivirus, pestivirus, and pegivirus). Finally, the third step of clustering was performed on the flavivirus sequences with a threshold of 0.8. This last step was used to remove poorly sampled clusters of sequences, which were either poor quality sequences or associated with flaviviruses (e.g. Apoi virus) that had few representative sequences. After the preprocessing analysis and clustering protocol, 1,488 flaviviral sequences remained and were subsequently analyzed in a multiple sequence alignment (MSA).

The MAFFT software[153] was used to perform three independent MSA analyses: (1) a progressive alignment with iterative refinement, (2) local alignment with iterative refinement methods using WSP and consistency scores, and (3) global alignment with iterative refinement methods using WSP and consistency scores. Output from these separate MSA analyses were then analyzed in the trimAl software[154] to quantitatively chose the highest quality MSA as well as remove sequence positions with gaps in at least 20% of the analyzed sequences. Finally, post-analysis of the MSA results measured the sequence position variance away from the consensus sequence. Specifically, the position frequency matrix was calculated from the ensemble of aligned sequences. The variance away from the most-probable residue at each sequence position was calculated from this data; these values are reported in the paper for specific residues of interest. Additioanlly, Figure 3.7 depicts the sequence logo of the MSA results, using dengue NS3h residue numbering.[155,156] The exact pre-processing, clustering, MSA, and post-processing protocols are available on Github (https://github.com/mccullaghlab/ZIKV-Lb3b4/bioinformatics).

Figure 3.8A highlights the sequence variance of sequence positions in the local region of L$\beta3\beta4$. Sequence positions with high conservation ($< 1\%$ variance) across all flavivirus sequences are colored blue, while positions with larger variance away from the consensus are colored from white to red, representing increasing variance. This structural representation of the MSA results highlight the poor conservation of the loop residues while sequence positions in $\alpha2$, $\beta3$, $\beta4$, and $\alpha3$ secondary structures are more strongly conserved.

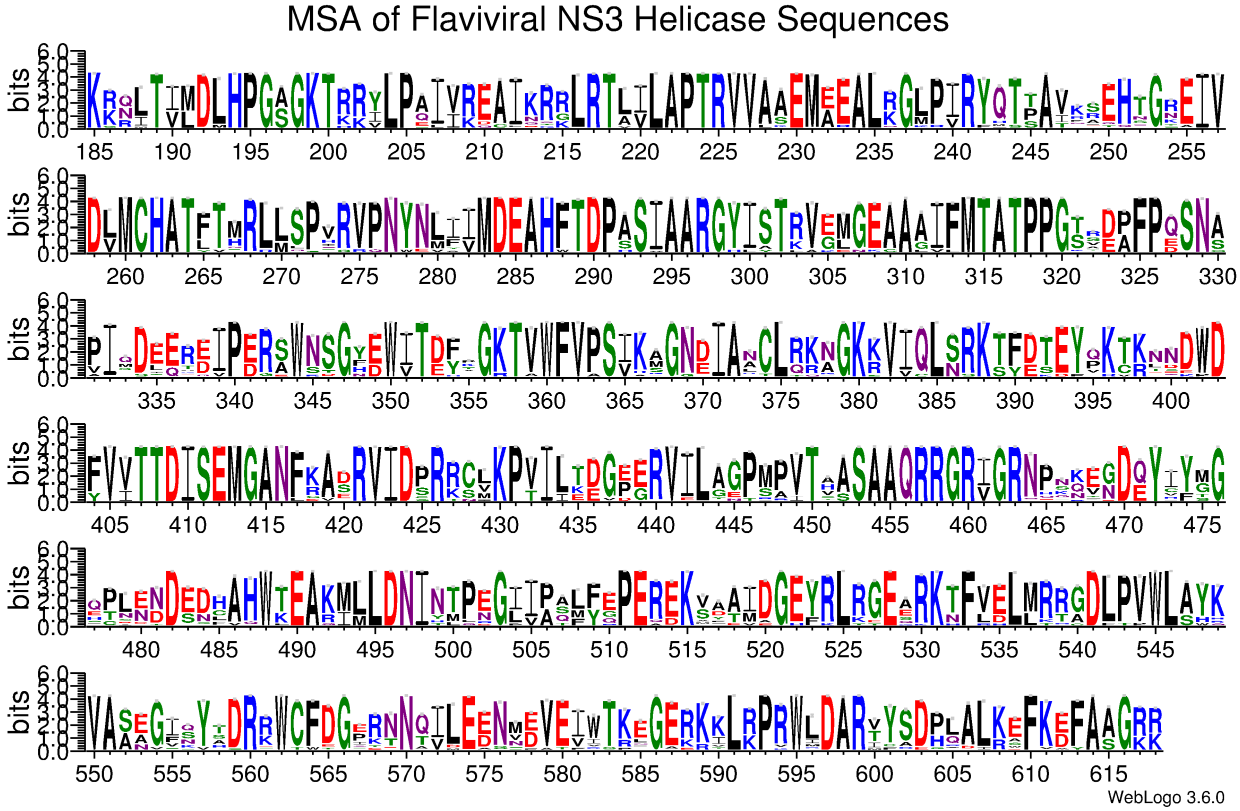

**Figure 3.7:** Sequence logo[155,156] of the flaviviral NS3h multiple sequence alignment results, using residue numbering from the DENV NS3h 2JLV structure. An equiprobable background composition of amino acid usage was assumed. Amino acids are colored based on their side chain chemistry: polar residues (green), neutral (purple), basic (blue), acidic (red), and hydrophobic (black). The relative height of each residue letter describes the relative frequency of observing the respective residue at that sequence position. The overall height of the column describes how conserved the sequence position is.

### Crystal Structure Alignment

In conjunction with the MSA analysis of flavivirus NS3h sequences, a structural alignment of all available crystal structures of the flavivirus NS3h was performed to highlight the structural heterogeneity of the RNA-binding loop and surrounding areas. In total, 56 monomer structures of flavivirus NS3h are available on the Protein Data Bank (PDB). The dengue NS3h:ssRNA:ATP crystal structure (2JLV) was used as the alignment reference structure. The alignment landmark reported in Davidson *et al.*[7] was used. Specifically, the C$\alpha$ atoms of the core $\beta$-sheets of subdomains 1 and 2 were used for alignment due to the strongly maintained tertiary structure of these subdomains.

Panel B of Figure 3.8 depicts a representative body of crystal conformations of ZIKV NS3h, focusing on the L$\beta3\beta4$ structural region. Generally, the "closed" conformation of the loop is poorly resolved except for the initial residues of the loop. Conversely, structures with the loop in the "open" conformation are well resolved and have similar structures to the RNA-bound (5GJB) crystal. Panels C and D of this figure highlight the arginine fork residues found in the L$\beta3\beta4$ local region of the RNA-binding cleft as well as the Thr245 and Thr246 residues, which are seen in structural analyses to be strong decisors of the two conformational states of the loop.

### Essential Dynamics of the L$\beta3\beta4$ Structure

The largest covariance motions of the RNA-binding loop were analyzed in the unbiased Apo simulation. Specifically, residues in $\alpha2$, $\beta3$, L$\beta3\beta4$, $\beta4$, and $\alpha3$ were of interest when quantifying the "open" to "closed" conformational change observed in the unbiased simulation. Of these secondary structures, residues that face away from the RNA-binding cleft were not considered. The MSA results guided the choice of coarse-graining used in this essential dynamics (ED) analysis. If a sequence position is strongly conserved (low sequence position variance), then the residue fluctuations are described with a two site coarse-graining: the C$\alpha$ atom used to describe backbone fluctuations and a side chain atom to describe the side chain fluctuations. An exception to this are residues with small side chains (e.g. Gly, Ala, and Val) where the fluctuation of the side chain is highly correlated with the backbone fluctuations. Residues with low conservation are coarse-

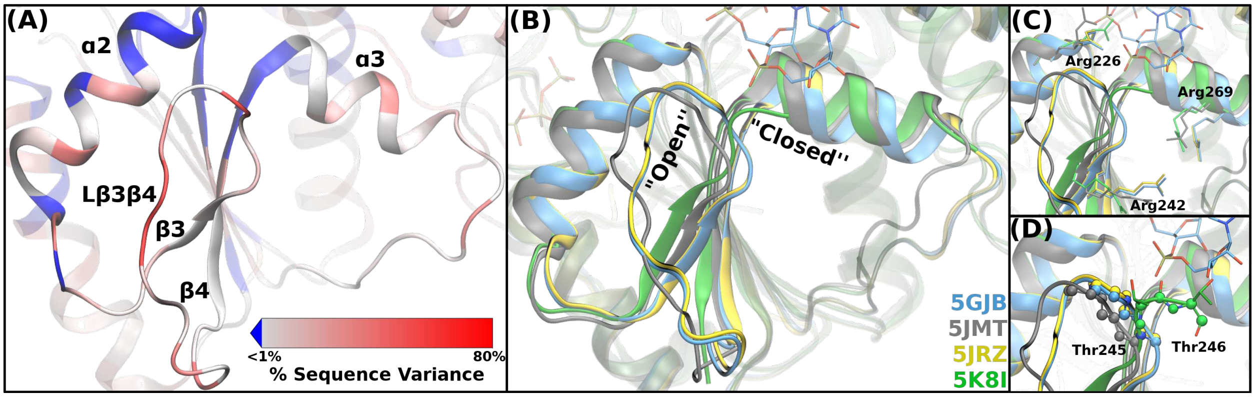

**Figure 3.8:** Alignment of flaviviral NS3h sequences and crystal structures. (A) Structural representation of the MSA results. Sequence positions are colored based on percent variance away from the consensus sequence. Highly conserved positions are colored blue, while less conserved residues are colored from white to red with increasing variance. (B) Crystal structure alignment of a subset of ZIKV NS3h, focusing on the L$\beta3\beta4$ region of subdomain 1. 5K8I is one of the few crystal structures with the L$\beta3\beta4$ structure in a "closed"-like position, albeit largely unresolved. (C) The strongly conserved Arg226, Arg242, and Arg269 residues hypothesized to function as arginine forks. (D) The large L$\beta3\beta4$ structural change can be seen in the large dihedral shifts of the Thr245-Thr246 residue pair.

grained as a single site at the C$\alpha$ atom. This bioinformatics-guided coarse-graining resulted in 44 atoms, incorporating all five secondary structures in the region of interest. A principal component analysis (PCA) of these atoms' cartesian coordinates result in eigenvectors describing the essential dynamics of the L$\beta3\beta4$ conformational transition observed in the Apo simulation.

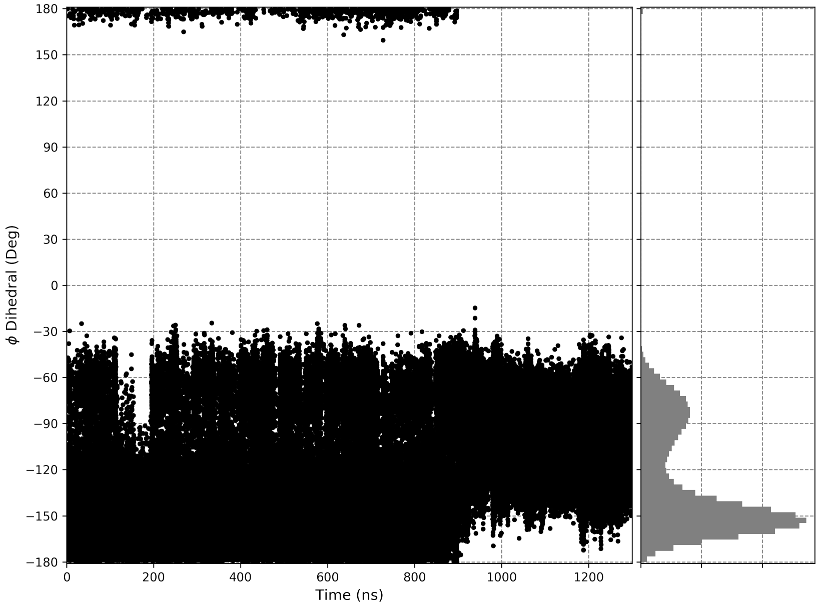

**Figure 3.9:** $\phi$ dihedral of Thr245 in the unbiased Apo simulation. The "open" to "closed" conformational change in L$\beta3\beta4$ occurs at approximately 900 ns.

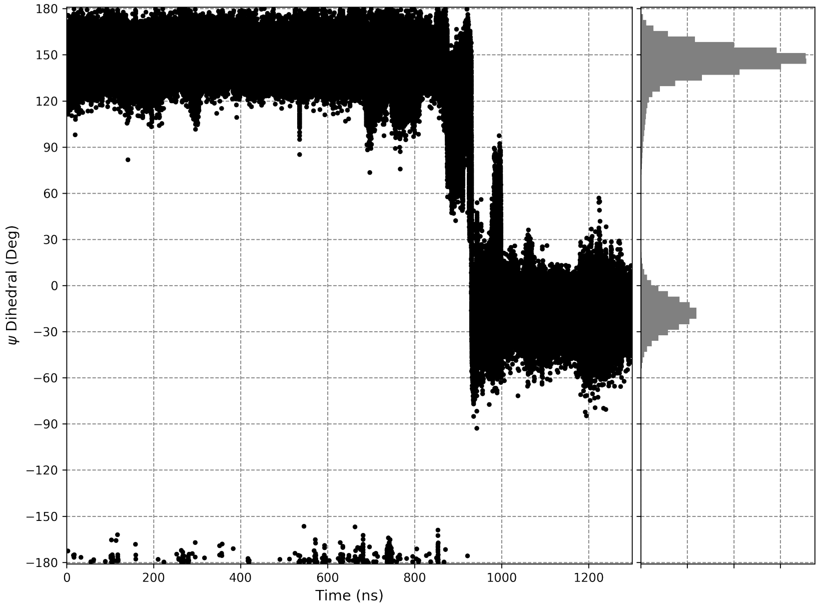

**Figure 3.10:** $\psi$ dihedral of Thr245 in the unbiased Apo simulation. The "open" to "closed" conformational change in L$\beta3\beta4$ occurs at approximately 900 ns.

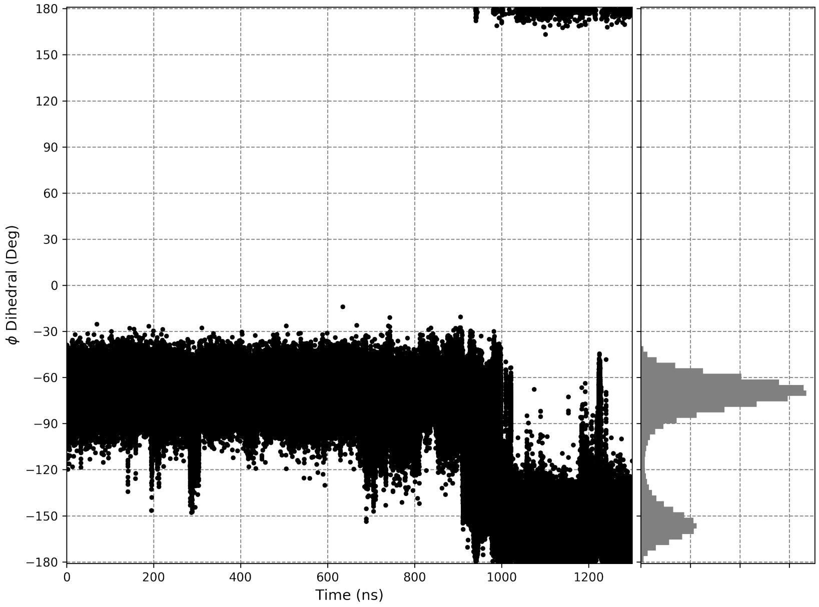

**Figure 3.11:** $\phi$ dihedral of Thr246 in the unbiased Apo simulation. The “open” to “closed” conformational change in L$\beta3\beta4$ occurs at approximately 900 ns.

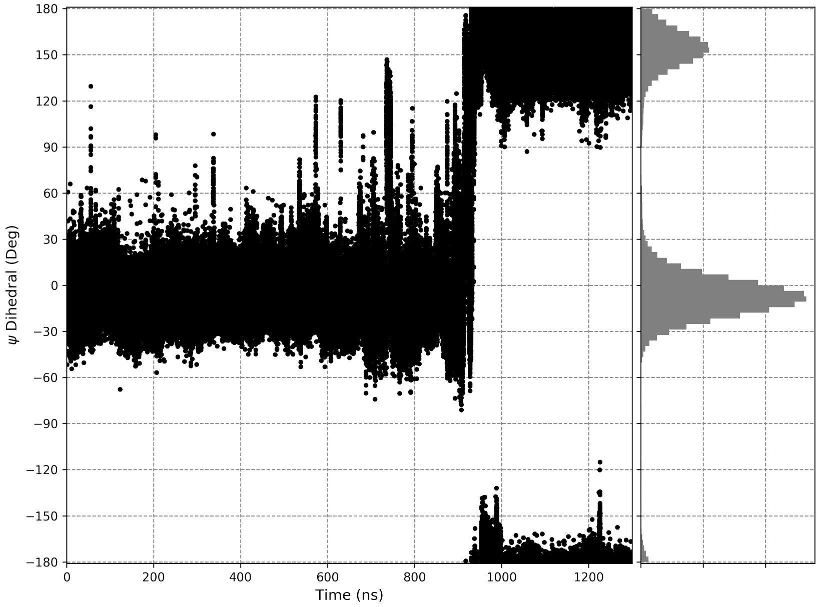

**Figure 3.12:** $\psi$ dihedral of Thr246 in the unbiased Apo simulation. The “open” to “closed” conformational change in L$\beta3\beta4$ occurs at approximately 900 ns.

# Chapter 4

# Residue-Level Allostery Propagates Through the Effective Coarse-Grained Hessian.[3]

## 4.1 Overview

The long-ranged coupling between residues that gives rise to allostery in a protein is built up from short-ranged physical interactions. Computational tools used to predict this coupling and its functional relevance have relied heavily on the application of graph theoretical metrics to residue-level correlations measured from all-atom molecular dynamics (aaMD) simulations. The short-ranged interactions that yield these long-ranged residue-level correlations are quantified by the effective coarse-grained Hessian. Here we compute an effective harmonic coarse-grained Hessian from aaMD simulations of a benchmark allosteric protein, IGPS, and demonstrate the improved locality of this graph Laplacian over two other connectivity matrices. Additionally, two centrality metrics are developed that indicate the direct and indirect importance of each residue at producing the covariance between the effector binding pocket and the active site. The residue importance indicated by these two metrics is corroborated by previous mutagenesis experiments and leads to unique functional insights. In contrast to previous computational analyses, our results suggest that fP76-hK181 is the most important contact for conveying direct allosteric paths across the HisF-HisH interface. The connectivity around fD98 is found to be important at affecting allostery through indirect means.

[3] Peter T. Lake[a,†], Russell B. Davidson[a,†], Glen M. Hocky [b], Martin McCullagh[a]; [a] Department of Chemistry, Colorado State University, Fort Collins, CO, USA, [b] Department of Chemistry, New York University, New York, NY, USA, † Contributed equally to this work.

## 4.2 Introduction

Allostery refers to the long-range functional coupling of sites in a macromolecule through networks of short-ranged interactions. This phenomenon can be a pivotal component of a protein's function[157] as demonstrated in GPCR signaling,[158] coupling of ATP binding and hydrolysis to mechanical work in motor proteins[7,159–161] and activation of oxygen binding in hemoglobin.[13,116,162–165] Many of these processes are initiated by the binding of an effector molecule that modulates the activity at a distal active site. This behavior has become an increasingly important target in the field of drug development due to possible improvements in selectivity over orthosteric sites.[1,166–169] Intriguingly, allostery can be incorporated into an enzyme's function even on the short time scales of directed evolution studies to produce a desired increase in activity.[170] Therefore, it is highly desirable to be able to identify and predict allostery, as well as the interactions ("allosteric pathways") involved in these processes.

There are many well established experimental techniques for characterizing allostery, including activity based assays to investigate non-Michaelis-Menten kinetic behavior[9], H/D mass spectrometry,[171,172] as well as structural based approaches such as X-ray crystallography,[173] cryoEM,[174] and NMR.[175–183] Structural techniques can be used to identify residues that interact directly with or are structurally perturbed by the effector molecule by comparing *apo* and effector bound states of a protein. Identifying allosteric pathways is significantly more challenging. While pathways have been identified using experimental techniques in well-studied proteins such as hemoglobin,[13,184–186] the combination of NMR spectroscopy and computational techniques have allowed for the most robust description of allosteric pathways.[178,181,182,187,188]

Computational techniques used to investigate allostery rely on graph theoretical approaches to identify important residues or connections that convey information from the effector binding pocket to the active site.[180,189–191] A weighted graph is constructed by defining nodes (residues) and edges (bonds) that connect nodes. These edge weights populate a pairwise adjacency matrix, $\boldsymbol{A}$, that has finite values between connected residues. The closely related graph Laplacian, $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A}$ where $\boldsymbol{D}$ is a diagonal matrix with elements $D_{ii} = \sum_j A_{ij}$, can be readily constructed the adjacency matrix. A variety of edge and node centrality metrics have been developed using either the adjacency matrix of graph Laplacian to identify important amino acids contributing to the allostery in a protein.[180,188,192]

Two computational techniques have been used to generate these graphs: bioinformatics and molecular dynamics. Sequence-based bioinformatic methods, working under the assumption that allosteric pathways are functionally important and thus evolutionarily conserved, use residue pairwise sequence co-evolution to define a graph.[180,186,193] These methods do not explicitly take into account the 3D structure of a given protein as this information should be implicitly captured in sequence covariance. Additionally, structural databases can be used to build pairwise, knowledge-based potentials that describe the interactions observed in e.g. a protein's crystal structure.[194] None of these models, however, directly account for the ensemble nature of protein structure and the associated allosteric behavior.[164]

All-atom molecular dynamics (aaMD) is a well-established method to sample the configurational ensemble of a protein.[195–200] Analyses of such simulations have been used to identify allosteric importance based on residue interaction networks[180,201–203] or positional covariance-based metrics.[113,187,188,203,204] Both of these models are used to describe the average residue-residue couplings from the ensemble of protein configurations observed. Interestingly, these methods lead to dramatically different graphs: the residue interaction networks will be localized in space while the positional covariance will be delocalized. Thus, while aaMD provides an appealing measure of protein configurational space, the appropriate graph Laplacian to describe residue-level correlations is not well understood.

An alternative to using some form of the positional covariance as the adjacency matrix is to use the Hessian, which is a graph Laplacian constructed from the second spatial derivative of the Hamiltonian. Elastic Network Models (ENM) have been used to define a residue-level coarse-grained Hessian and study the contributions of the normal modes to allostery for a variety of systems.[205] While standard ENM models qualitatively capture low frequency motions of proteins, the lack of residue specificity in the model yields low fidelity with all-atom models in the mid to high frequency range.[206] Bottom-up coarse-graining approaches, such as REACH[207,208] and hENM,[209] have been developed to yield anisotropic ENMs with increased residue specificity. These methods are designed to capture higher fidelity with all-atom normal modes across the spectrum.[210] We propose to connect these two fields of study, namely protein allostery and coarse-grained potentials, and use the resulting effective coarse-grained Hessian as the graph Laplacian to study protein allostery.

In the next section, we demonstrate that the effective coarse-grained Hessian is the appropriate graph Laplacian to quantify residue-level allostery from aaMD simulations. Motivated by the physical interpretation of the Hessian, we define two centrality metrics that indicate a Hessian element's importance at conveying covariance between a selected set of sources and sink residues. Finally, we apply this method to simulations of the imidazole glycerol phosphate synthase (IGPS) heterodimer, which is a well established benchmark allosteric system. *Our definition and resulting weights of allosteric pathways provide new insight into the physics underlying allostery, a key functional component of most proteins.*

## 4.3 Theoretical Framework

Structural allostery can be described as the positional change at one site due to the application of an external force at a different site. A linear response of a system to an external force, $\boldsymbol{f_{ex}}$, can be written as

$$\boldsymbol{\Delta x} = \frac{1}{T}\boldsymbol{C}\boldsymbol{f_{ex}}, \tag{4.1}$$

where $\boldsymbol{\Delta x}$ is the change in the equilibrium position due to the perturbation, $\boldsymbol{C}$ is the matrix of covariance of particle positions, and $T$ is temperature in units of energy.[211] (4.1) demonstrates the importance of positional covariance to predicting allostery, but does not indicate anything about the short-ranged physical interactions leading to this behavior.

Within a graph theoretical framework, the short-ranged interactions that lead to long-ranged correlations are what should populate the graph Laplacian of the system. The Hessian is the appropriate graph Laplacian for molecular systems because (1) Hessian elements are only finite for short-ranged physical interactions, (2) the Hessian is, by construction, a symmetric positive semi-definite matrix as all graph Laplacians are, and (3) the covariance is related to the generalized inverse of the Hessian. The final point is most well-known in elastic network models (ENMs) for which the covariance and Hessian are related by the Moore-Penrose pseudoinverse ($\boldsymbol{H}^{+} \equiv \frac{1}{T}\boldsymbol{C}$). This relationship is consistent with the Gaussian Markov random field literature in which the graph Laplacian and the covariance are related by pseudoinverses.[212] In the context of allostery in proteins, the covariance is typically measured at a residue-level even if the underlying simulations have atomic resolution. Thus, the Hessian of interest is the second derivative of an effective coarse-grained Hamiltonian. Determining the effective Hessian from a measured covariance is a non-trivial problem due to the difficulties in converging a measured covariance.[213] Here we use a previously described coarse-graining procedure to generate an effective harmonic Hessian based on a measured covariance.

### 4.3.1 The Effective Harmonic Hessian from All-Atom Molecular Dynamics Simulations

Explicit solvent aaMD simulations can provide a detailed and accurate structural ensemble picture of proteins under physiological conditions.[195–200] These simulations have been used to generate $3N \times 3N$ covariance matrices that have been used to assess $N-$residue level structural allostery. Motivated by the idea that the structural ensemble in a single free energy well from aaMD is well represented by a harmonic system,[214–218] effective harmonic Hamiltonians have been fit to mapped all-atom data.[207–209] In this work, we construct an effective harmonic Hessian using a slightly modified heterogeneous-ENM (hENM) procedure[209] as described in the Supporting Information (SI) Computational Methods section.

The result of the hENM procedure is a $N \times N$ force constant matrix, $\boldsymbol{k}$, that is optimized to reproduce pairwise particle variances. The $3 \times 3$ tensor element of the $3N \times 3N$ Hessian is defined as

$$\boldsymbol{H}_{ij} = -k_{ij}\hat{\boldsymbol{R}}_{ij} \otimes \hat{\boldsymbol{R}}_{ij}, \tag{4.2}$$

where $\hat{\boldsymbol{R}}_{ij} = \frac{\langle \boldsymbol{r}_i \rangle - \langle \boldsymbol{r}_j \rangle}{|\langle \boldsymbol{r}_i \rangle - \langle \boldsymbol{r}_j \rangle|}$ and $\langle \boldsymbol{r}_i \rangle$ is the average position of node $i$. The $3N \times 3N$ covariance can then be reconstructed as $\boldsymbol{C} = T\boldsymbol{H}^{+}$. The corresponding covariance calculated from the hENM Hessian demonstrates good correlation with the measured aaMD covariance (Figure 4.5).

### 4.3.2 Allosteric Paths in the Hessian

To investigate allosteric pathways, we start by dictating that the sum over these pathways yields the covariance between a selected pair of source and sink residues. This is motivated by the idea that the covariance is the physical observable that dictates the linear response between two residues. Employing a property from graph theory that, given a weighted adjacency matrix, $\boldsymbol{A}$, $(\boldsymbol{A}^{\ell})_{ij}$ is the weighted sum of all walks of length $\ell$ between nodes $i$ and $j$, we can express the covariance in the following way,

$$\boldsymbol{C} = \sum_{\ell=0}^{\infty} \boldsymbol{A}^{\ell} = (\boldsymbol{I} - \boldsymbol{A})^{-1}\,, \tag{4.3}$$

where $I$ is the identity matrix and the last equality arises from identifying the infinite sum as an example of the Neumann series. At this juncture, it is natural to write that $\boldsymbol{A} = \boldsymbol{I} - \boldsymbol{C}^{-1}$ but, in the context of molecular simulations, $\boldsymbol{C}$ is a singular matrix and, thus, not invertible due to the removal of center-of-mass translation and rotation.

If we consider $\boldsymbol{C}$ to be strictly invertible and map the infinite number of walks to a finite set of paths, $P_{ij}$, it can be shown that

$$\frac{1}{T}\boldsymbol{C}_{ij} = \sum_{n \in P_{ij}} \boldsymbol{H}_{ii}^{-1} \prod_{\langle \alpha, \beta \rangle \in n} -\boldsymbol{H}_{\alpha\beta} \times \left\{ \left[ \boldsymbol{H}_{\{\alpha\}\mathrm{prev}} \right]^{-1} \right\}_{\beta\beta}, \tag{4.4}$$

where $\langle \alpha, \beta \rangle$ is an edge in the path and subscript $\{\alpha\}\mathrm{prev}$ denotes the principle submatrix of $\boldsymbol{H}$ obtained by removing all nodes previously visited in the path. The inverse terms in (4.4) arise from mapping walks to paths; traversing all loops from $\alpha$ to itself results in terms related to the conditional variance of $\alpha$. The only terms that couple two nodes together in (4.4) are $\boldsymbol{H}_{\alpha\beta}$ which are the terms in the Hessian. We note that these terms in the Hessian are $3 \times 3$ tensors for a 3D system.

Paths are usually referenced in terms of a length $\ell$ as opposed to a weight. This distinction can be made by considering a path length as being the negative of a sum of the adjacency matrix on a logarithmic scale as opposed to the weight which is a product of elements in the adjacency matrix. In this way, the paths with the largest weights have the shortest lengths. We study these paths by sampling them in a Monte Carlo scheme; even though the paths observed are not necessarily all the shortest paths within some number, the algorithm is much more efficient than the typically used Floyd-Warshall algorithm.[219,220] This method of sampling paths also leads us to consider these paths as part of an ensemble, all of which can contribute. The Monte Carlo sampling is also readily extended to account for path lengths which are not strictly a sum of pairwise interactions. The procedure is detailed in the SI Theory section.

A similar but more complicated procedure can be followed to determine paths that yield the covariance in a singular covariance matrix. The most crucial idea obtained from performing the full derivation is that fully connected paths, those with finite Hessian values connecting a given source and sink, are not the only contributions to the covariance. Additional terms, that can be described as broken paths, contribute to the covariance due to a coupling through the null-space of the singular matrix. The relevant outcome of this for the current work is that residues that contribute to direct paths are not the only residues that affect the covariance between a source and sink. This motivates the need for additional importance metrics to study allosteric interactions in a given graph.

### 4.3.3 Hessian Derivative as a Centrality Metric

The effective harmonic Hessian lends itself to another centrality metric. One can consider how changing a single spring constant alters the covariance between a given set of sources and sinks. This leads to an edge-based centrality metric we call the derivative metric, defined by

$$\delta_{edge}^{(ij)} = \frac{d||\boldsymbol{C}_{mn}||^2}{dk_{ij}}, \tag{4.5}$$

where $||\boldsymbol{C}_{mn}||^2$ is the squared Frobenius norm of the covariance tensor between source $m$ and sink $n$. We map this metric down to a node-based centrality metric by

$$\delta_{node}^{(i)} = \sum_j k_{ij} \delta_{edge}^{(ij)}. \tag{4.6}$$

This mapping of the edge metric to a node metric is motivated by considering a fractional change of all the spring constants connecting a given node. These values can be positive and negative and vary by orders of magnitude. Simulations for this study yielded values of $\delta_{node}$ ranging from $-5.57 \times 10^{-3}$ Å$^4$ to $2.20 \times 10^{-4}$ Å$^4$.

## 4.4 Results and Discussion

### 4.4.1 Model protein - IGPS

Imidazole glycerol phosphate synthase (IGPS) is an enzyme that functions in both the purine and histidine biosynthesis pathways of plants, fungi, archaea, and bacteria. In *Thermatoga maritima*, IGPS is a heterodimeric protein complex of HisH and HisF proteins (depicted in white and orange respectively in Figure 4.1). HisH catalyzes the hydrolysis of glutamine into ammonia and glutamate. Nascent ammonia is then shuttled across the HisF–HisH interface and through the $(\beta/\alpha)_8$ barrel of HisF (an approximate distance of 25 Å). At the HisF cyclase active site, the ammonia reacts with N'-[(5'-phosphoribulosyl)formimino]-5-aminoimidazole-4-carboxamide ribonucleotide (PRFAR) to form imidazole glycerol phosphate (IGP) and 5-aminoimidazole-4-carboxamide (AICAR). These two reactions are strongly coupled with a 1:1 stoichiometry, despite the large distance that separates the two active sites.[221,222] In addition to the concerted mechanism between HisF and HisH, IGPS is classified as a V-type allosteric enzyme, such that the rate of the glutaminase reaction is critically dependent on the presence of the PRFAR ligand. Experimental assays have quantified this strong allosteric activation to be an approximate 4,900-fold increase in activity relative to basal levels.[223] Allosteric activation is also observed in the presence of the cyclization products, IGP and AICAR, but at reduced magnitudes.[223,224]

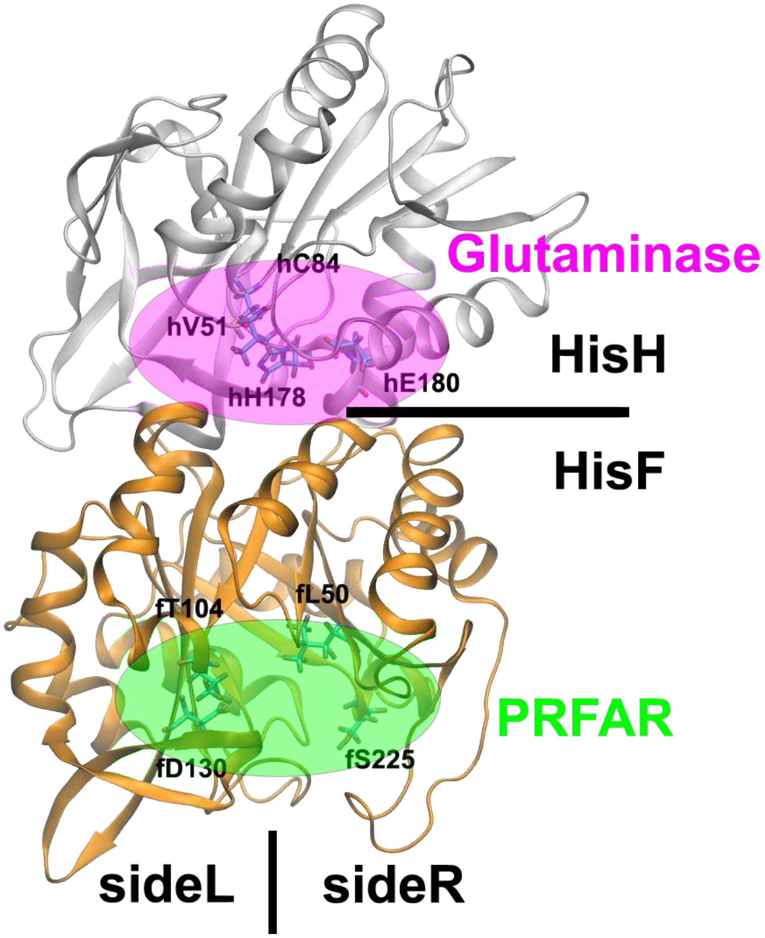

**Figure 4.1: Structural depiction of the IGPS heterodimer.** The protein is composed of HisH and HisF monomers. The PRFAR ligand binds in the pocket indicated by the green oval and source residues fL50, fT104, fD130 and fS225. The glutaminase active site, labeled in pink, is located in the HisH monomer near the interface. We chose four sink residues hV51, hC84, hH178 and hE180 to identify this pocket.

The allostery of the IGPS enzyme has been studied for nearly two decades using a broad range of experimental and theoretical methodologies. Mutational studies, coupled with biochemical assays and NMR experiments, have highlighted residues that are pivotal for the functionality of the protein complex as well as dynamic effects induced by the HisF ligand.[181,182,223–231] Additionally, MD simulations and graph-theoretic analyses have been applied to the IGPS system to study the allosteric effect of PRFAR and the allosteric paths that couple the two distal active sites.[111,113,181,187,188,203,226,232–234] For this work, we have performed aaMD simulations of *apo* IGPS as well as one mutant variant, totaling 1$\mu$s of trajectory for each system. Subsequently, an ef-

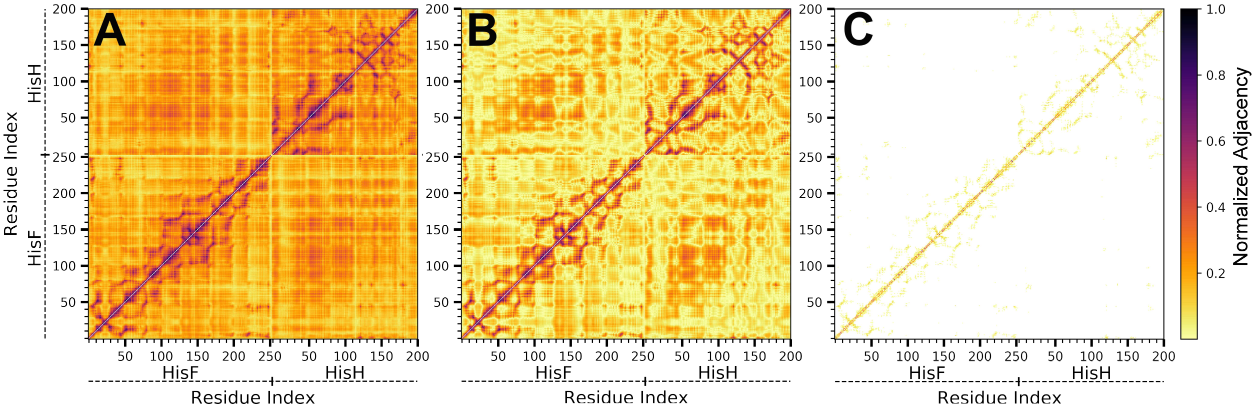

**Figure 4.2: Normalized adjacency matrices computed from $4 \times 250$ ns all-atom molecular dynamics simulations of IGPS protein in its *apo* form.** A) The linear mutual information adjacency, rMI,[235] B) the Pearson correlation adjacency and C) the effective harmonic Hessian.

fective harmonic Hessian has been constructed and analyzed for both systems (SI Computational Methods).

### 4.4.2 Adjacency Matrix Comparison

Adjacency matrices based on the covariance and mutual information (MI) have been used to study allostery in IGPS.[113,203,234] The theoretical framework outlined above demonstrates the importance of a third adjacency matrix, one based on the Hessian. In order to compare and contrast between these three, we consider absolute values of normalized adjacency matrices. These matrices are limited to values between 0 and 1, with values near one indicating strong connectivity and values near zero indicating weak connections. Residue-level normalized adjacency matrices based on our $1\mu$s MD simulation of *apo* IGPS are shown in Figure 4.2.

Mutual information (MI) is a measure of correlation between two distributions based on the difference between marginal and joint Shannon entropies. The numeric values of this quantity range from zero to infinity, but can be "normalized" by computing $r_{MI} = \sqrt{1 - \exp(-\frac{2}{3}MI)}$ where this manipulation leads to pairs with large MI having $r_{MI} \approx 1$ and uncoupled pairs having $r_{MI} \approx 0$. We note, however, that neither the MI nor the rMI matrices are positive semi-definite, which differentiates it from the two other matrices discussed here. The motivation for using this "adjacency" matrix over, for example, the normalized 1D covariance matrix, is that it accounts for correlation along perpendicular degrees of freedom.[235] The residue-level $r_{MI}$ matrix, computed from a linear MI, for our *apo* IGPS simulations is depicted in Figure 4.2A. This plot indicates strong correlations along the primary sequence as indicated by dark colors in the near off-diagonal elements of the matrix. Further off-diagonal dark coloring indicates secondary structural elements that also convey strong correlation. The break between the HisF and HisH proteins is also evident as there are clear dividing lines at residue 253. Despite stronger near diagonal correlation, the $r_{MI}$ adjacency matrix demonstrates long-range contacts even in the HisF–HisH coupling sub-matrix in the bottom-right or top-left of the matrix.

An alternative to $r_{MI}$ is the Pearson correlation matrix. This adjacency matrix is derived from the covariance matrix, which is positive semi-definite. In a 3D system of $N$ particles, the covariance matrix is a $3N \times 3N$ matrix or, equivalently, an $N \times N$ matrix of $3 \times 3$ tensor elements. A typical manipulation is to reduce the $3N \times 3N$ matrix to an $N \times N$ matrix by taking the trace of each tensor element. While this manipulation still yields a positive semi-definite matrix, we do note that it ignores couplings in orthogonal degrees of freedom upon mapping the system to a 1D system.[235] The resulting $N \times N$ matrix can be further manipulated to the Pearson correlation matrix with values between -1 and 1 by dividing by the square root of the diagonal elements of each corresponding row and column. Taking the absolute value of the Pearson correlation for the *apo* IGPS system yields the normalized adjacency matrix depicted in Figure 4.2B. While there are quantitative differences observed between $r_{MI}$ and the Pearson correlation, the qualitative behavior remains the same; there are strong contacts along the primary sequence but there are finite contacts for almost all elements of the matrix.

The third adjacency matrix we consider is based on the effective harmonic Hessian. We compute this from the $3N \times 3N$ covariance matrix and the average structure from a simulation following the modified hENM procedure as discussed in the Theory section and further elaborated in the SI. The Hessian is a symmetric positive semi-definite matrix by construction, and the normalized adjacency matrix can be constructed in a manner analogous to what is done for the Pearson correlation. The Hessian-based normalized adjacency matrix (Figure 4.2C) shows strong primary

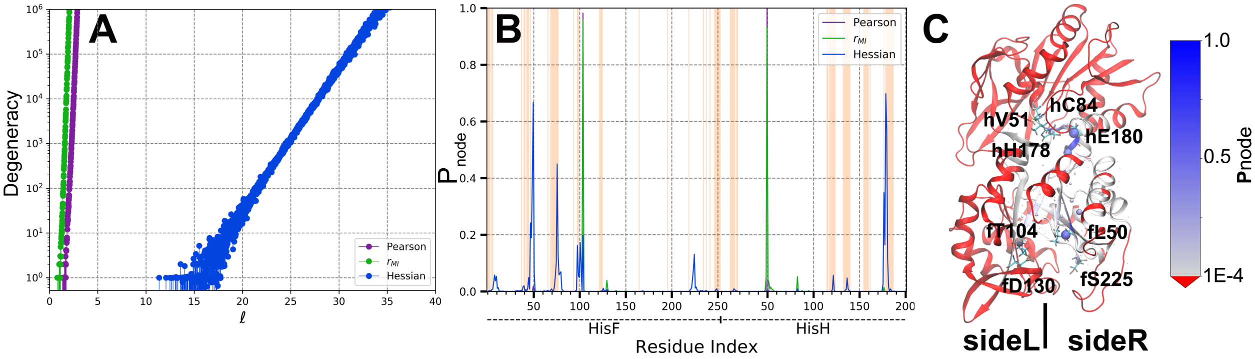

**Figure 4.3: Comparison of paths and resulting centralities for different adjacency matrices for the IGPS protein.** A) The path length degeneracy as a function of path length, $\ell$. B) The probability of observing a given node in a sampled path, $P_{node}$. C) The probability of a given node in a sampled path for the Hessian-based adjacency matrix coloring residues in a structural representation of IGPS.

sequence connectivity in agreement with the correlation and $r_{MI}$ adjacency matrices. Secondary structure connectivity is observed in the near off-diagonal of Figure 4.2C in agreement with the strongly correlated off-diagonal regions of Figure 4.2A and B. These connections are distinctly weaker in the Hessian-based adjacency, however, and the farther off-diagonal elements are zero indicating little long-range connectivity in the graph. This adjacency matrix is consistent with the idea of short-ranged physical interactions that yield long-range correlation.

### 4.4.3 Direct Paths Comparison

Paths that convey the covariance between a set of sources and sinks are attractive physical interpretations of allostery. Direct paths, ones with complete connectivity in the adjacency matrix, are an incomplete picture of the covariance between any given source and sink but are readily sampled. We use direct paths here to compare and contrast adjacency matrices. In this work, we sample paths between four source residues (fLeu50, fThr104, fAsp130, fSer225) that span the effector-binding pocket (green oval in Figure 4.1) and four sink residues (hVal51, hCys84, hHis178, hGlu180) in the glutaminase active site (pink oval in Figure 4.1 of IGPS that have been previously identified as important.[187,236,237] Path sampling is performed in a Monte-Carlo scheme as described in the Theoretical Framework section and SI.

Paths between sources and sinks in the Hessian-based adjacency matrix are longer and less degenerate than those observed in the Pearson or $r_{MI}$ matrices. Figure 4.3A depicts the degeneracy of paths as a function of path length, $\ell$, for the Pearson correlation, $r_{MI}$ and Hessian paths. The paths in the Pearson correlation and $r_{MI}$ adjacency matrices behave similarly; paths of extremely short length are found and the degeneracy grows rapidly with path length. This can be understood by observing finite values in all elements of these two matrices. The paths on the Hessian are longer and the degeneracy of paths grows much less rapidly than either the Pearson or r$_{MI}$ paths. Again, this can be understood by the smaller number of connections in the adjacency matrix leading to both longer paths as well as more unique, short path lengths. These results demonstrate that the paths on the Hessian use a small number of short-ranged interactions to produce long-range correlation.

Paths in the Hessian sample different nodes than paths in the other two adjacency matrices. A residue centrality metric, $P_{node}$, can be defined as the probability of observing each residue in the sampled paths. This metric is plotted as a function of residue number in Figure 4.3B for each adjacency matrix. The $r_{MI}$ and Pearson adjacency matrices both have high probability of observing the source residue fThr104 in the paths ($P_{node} \sim 0.8$) but little probability of observing the other three source residues (fLeu50, fAsp130, fSer225). Similarly, the sink residue probabilities are dominated by hVal51 ($P_{node} \sim 0.8$) with some contribution from hCys84 ($P_{node} \sim 0.4$). This is due to the highly correlated nature of residues fThr104 and hVal51 yielding a direct path between the two that is much shorter than all other paths. The Hessian-based adjacency matrix, on the other hand, yields significant probability of observing source residue fLeu50 and sink residue hGlu180. There are, however, finite probabilities of observing the other source and sink residues in the Hessian paths.

The Hessian-based paths go through known important regions of IGPS. Figure 4.3C depicts these P$_{node}$ results on the structure of the heterodimer, highlighting the localization of residues sampled in paths to sideR of HisF as indicated by the white to blue coloring. These paths mainly propagate from fLeu50 through a hydrophobic network that spans the interface of the beta barrel and surrounding alpha helices in HisF. This includes residue fVal48 ($P_{node} = 0.19$; to be discussed later). These results are consistent with previous results indicating the importance of sideR.[187,188]

Interestingly, the Hessian-based paths show significant probability of observing residues fPro76 ($P_{node} = 0.45$) and hLys181 ($P_{node} = 0.58$) as the connection between HisF and HisH. This is in contrast to some previous results indicating the importance of a salt-bridge at the interface between fAsp98 ($P_{node} = 0.16$) and hLys181. We leave further discussion of this to a subsequent section in which we compare to experimental results.

### 4.4.4 Derivative Centrality Metric of Hessian

Paths are an appealing approach to identify important residues between binding sites, but direct paths are not the only contributors to covariance between source and sink residues. How a mutation will ultimately affect the covariance between nodes via paths is not obvious. Using the Hessian as the adjacency matrix suggests another analysis to compare to mutagenesis experiments. Given the physical interpretation of the finite Hessian elements, we can consider how the covariance between a source and sink is affected by changes in the Hessian elements. The details of this method, termed the derivative centrality metric, are provided in the Theory section. This type of thinking parallels work from Rocks *et al.* but has not been applied directly to allostery in a protein.[238,239]

After applying a 12 Ådistance cut-off to the derivative edge metric, contacts at the HisF–HisH interface are highlighted. A structural representation of the edges with large magnitude derivative values from (4.5) is provided in Figure 4.4A. The highest density of large magnitude (green) edges span the heterodimer interface, specifically at a region identified previously by Amaro *et al.* and Rivalta *et al.* to be modulated by PRFAR-binding.[187,234] These past results are interpreted as a strengthening of the spring constants that span the interface in this region, leading to increased frequency of the reported "breathing motion". Results from the derivative centrality metric then suggest that the strengthening of these spring constants impacts the covariance between source and sink residues, thus indicating an indirect allosteric effect on the glutaminase active site due to PRFAR binding.

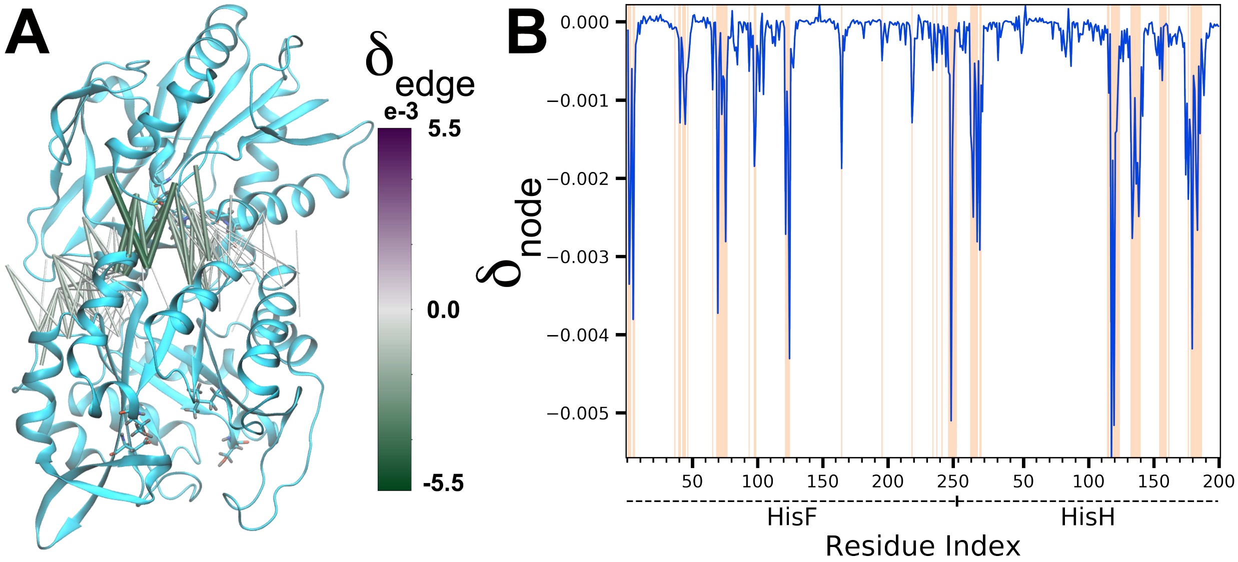

**Figure 4.4: Hessian-based derivative centrality metric $\delta_{node}$ of IGPS apo.** A) A structural representation edges with large derivative metric values B) Derivative node metric values as a function of residues. Residues that have connections across the HisF–HisH interface are colored by vertical orange lines.

The derivative node metric ((4.6)) on *apo* IGPS highlights the importance of a second position at the interface between HisF and HisH. These values are plotted as a function of residue number in Figure 4.4B with interfacial residues highlighted in orange. A cluster of three residues with large magnitude $\delta_{node}$ values are hP119 ($\delta_{node} = -5.6 \times 10^{-3}$ Å$^{4}$), hM121 ($\delta_{node} = -5.2 \times 10^{-3}$ Å$^{4}$) and fV125 ($\delta_{node} = -4.3 \times 10^{-3}$ Å$^{4}$). These three residues sit at the interface between HisF and HisH, on SideL of the region discussed above. The interface is closer together at this position and the interfacial contacts in the Hessian matrix are observed to be stronger. The large magnitude $\delta_{node}$ values of these three residues suggest that the perturbation of their respective contacts will have a large impact on the covariance between sources and sinks. Due to the limited focus on this region of IGPS, we proffer this cluster of residues as potential targets for mutagenesis or inhibitory binding studies.

The path-based and derivative centrality metrics highlight different residues. This can be observed by noting that bridge residues, highlighted in orange, tend to be underrepresented in Figure 4.3B and highly represented in Figure 4.4B. The structural comparison also indicates that $P_{node}$ highlights residues on SideR of the protein. On the other hand, the derivative centrality met-

**Table 4.1: Residue importance for glutaminase related allostery in IGPS.** A comprehensive list of published kinetic assay results for glutaminase activity of single point mutants of IGPS is included as well as our node centrality metrics, $P_{node}$ and $\delta_{node}$, for each of the mutated residues.

| Residue | Mutation | Ratio of Activated/Basal $k_{cat}/K_M$ | $P_{node}$ | $\delta_{node}$ ($10^{-3}$ Å$^4$) |
|---|---|---|---|---|
| fR5 | fR5A$^{a,b}$<br>fR5H$^{a,b}$<br>fR5K$^{a,b}$ | 870<br>71<br>440 | 0.0068 | -3.8049 |
| fV12 | fV12A$^{c}$ | 3095 | 0.0131 | -0.0201 |
| fK19 | fK19A$^{c}$<br>fK19A$^{a,b}$<br>fK19R$^{a,b}$ | 92<br>45<br>110 | 0.0000 | 0.0225 |
| fV48 | fV48A$^{c}$ | 139 | 0.1891 | -0.3954 |
| fD98 | fD98A$^{a,d}$<br>fD98A$^{c}$ | 2<br>54 | 0.1620 | -1.8466 |
| fK99 | fK99A$^{a,b}$<br>fK99R$^{a,b}$ | 9100<br>700 | 0.0547 | -1.4076 |
| fT104 | fT104A$^{a,e}$ | 40 | 0.1973 | 0.1068 |
| fQ123 | fQ123A$^{a,d}$ | 2000 | 0.0000 | -0.8889 |
| hN12 | hN12A$^{a,d}$ | 100 | 0.0004 | -1.6660 |
| hK181 | hK181A$^{a,d}$ | 2000 | 0.5797 | -4.1806 |
| WT$^{b,d}$ | | $4500-4900$ | – | – |

Mutant residues are reported using *T. maritima* IGPS notation. $^a$ denotes mutations performed in *S. cerevisiae* IGPS. $^b$ is from Ref. 223, $^c$ is from Ref. 182, $^d$ is from Ref. 227 and $^e$ is from Ref. 234.

ric highlights only the interfacial residues, which span both sides of the protein. The correlation between these metrics is plotted in Figure 4.6. As presented in that figure, the majority of residues highlighted in the derivative metric are not observed in the direct paths between sources and sinks. The few exceptions to this are interfacial residues such as hK181 that are both at the interface, as well as in the direct paths. The frequency of such residues is relatively rare, thus we conclude that these two analyses provide different and complementary information.

### 4.4.5 Comparison to Experimental Results for IGPS

The results from our simulations and centrality metrics on the effective harmonic coarse-grained Hessian compare favorably to kinetic experiments. To assess the mechanistic role of specific residues, we will focus on comparing our results to kinetic assay experiments of wild-type (WT) and mutant IGPS proteins. These assays compute the Michaelis-Menten enzymatic efficiency metric, k$_{cat}$/K$_{M}$, in the activated (effector bound) and basal states of the protein. The ratio of the enzymatic efficiency in these two states indicates the allosteric enhancement factor of the enzyme. The WT enzyme exhibits an allosteric enhancement factor of 4,500 to 4,900[182,223] while most mutant proteins have a smaller enhancement factor relative to WT, indicating reduced allosteric activation of glutaminase activity by PRFAR (see Table 4.1).

The mutation of a residue found in an allosteric pathway is hypothesized to have an effect on the allosteric activation of the enzyme. Of the residues studied by mutagenesis (Table 4.1), we find that only residues hK181, fT104, fV48 and fD98 have a significant (above $\sim 6\%$) probability of being in the direct paths between the PRFAR binding pocket and glutaminase active site. Of these, single point mutations of fT104, fV48 and fD98 to alanines diminish allosteric enhancement factor by over an order of magnitude compared to WT. This is consistent with the picture that altering residues in the allosteric paths disrupts the ability of the enzyme to properly convey covariance between pockets.

Interestingly, a single point alanine mutation to hK181 only causes a factor of two decrease in allosteric enhancement factor when compared to WT. This result calls into question the importance of the salt-bridge between hK181 and fD98 that has been previously implicated as of extreme importance in IGPS allosteric paths. As was mentioned in the previous paths sections, we find that the residues fP76 and hK181 are more sampled in the paths than fD98. Additionally, in the IGPS *apo* Hessian, the fP76-hK181 and fD98-hK181 spring constants are 7.322 and 0.411 kcal mol$^{-1}$ Å$^{-2}$, respectively. The interaction between fP76 and the aliphatic portion of hK181's side chain is stronger than the salt bridge observed between fD98-hK181. We hypothesized that these trends in interactions would be well maintained in an hK181A mutant. To test this, we performed a simulation of the mutant and found a comparable force constant for the fP76-hA181 edge (2.844 kcal mol$^{-1}$ Å$^{-2}$) and no force constant between fD98-hA181. Therefore, the alanine mutation of hK181 constitutes a small perturbation to the direct allosteric paths, which explains the small decrease in observed allosteric enhancement factor.

Residues that are at the HisF–HisH interface but not in allosteric paths can also have a large effect on the covariance between sources and sinks as demonstrated by our derivative metric. In Table 4.1, interfacial residues consist of fR5, fK99, fQ123 and hN12. The $\delta_{node}$ values for these residues are $-3.8049 \times 10^{-4}$ Å$^4$, $-1.4076 \times 10^{-4}$ Å$^4$, $-0.8889 \times 10^{-4}$ Å$^4$ and $-1.6660 \times 10^{-4}$ Å$^4$ respectively. Experimentally, mutations to these residues are all shown to decrease allosteric enhancement relative to WT. This suggests that the mutations have changed the interaction network around the mutated residue, which, in turn, caused a change to the covariance between PRFAR binding pocket residues and glutaminase active site residues. Interestingly, mutation to fK99 can either have a decrease (fK99R) or an increase (fK99A) in activity relative to WT. The derivative metric does not indicate how the change in covariance affects allosteric enhancement just that it will change. It will thus be of interest to study fK99 further to investigate how these two mutants affect the covariance.

If the path and derivative centrality metrics provide a complete picture of allostery in IGPS, then residues not in paths and not at the HisF–HisH interface will have little effect on the experimentally measured allosteric enhancement factor. Residue fV12 is observed in only 1% of paths and has a $\delta_{node}$ of only $-0.0201 \times 10^{-4}$ Å$^4$. A valine to alanine mutation at this position has little affect on the allosteric enhancement factor relatively to WT, suggesting that the $P_{node}$ and $\delta_{node}$ results do provide a complete picture for this residue. In contrast, fK19 has $P_{node} = 0$ and $\delta_{node} = 0.0225 \times 10^{-4}$ Å$^4$ yet all three mutations listed in Table 4.1 have a large effect on allosteric enhancement factor. Amaro *et al.* performed aaMD simulations to model the unbinding of the PRFAR ligand from which fK19 was observed to play an important role in the recognition and binding of the effector molecule.[234] Thus, a mutation to the fK19 residue is likely to decrease the allosteric enhancement factor by simply reducing the binding affinity of PRFAR and this behavior might not perturb the allostery between binding pockets.

## 4.5 Conclusions

In this work, we provided evidence that the effective coarse-grained Hessian is the appropriate graph Laplacian to consider in the context of allostery. The Hessian only contains finite values for short-ranged physical interactions and can be rigorously tied to the covariance for harmonic systems. We use a previously developed coarse-graining protocol, hENM, to compute the best effective harmonic Hessian that captures the residue-level covariance of all-atom molecular dynamics systems.

With the Hessian as the graph Laplacian, we develop two centrality metrics to highlight important residues that contribute to allostery. Both of these metrics are applied to the IGPS protein to investigate interactions between the PRFAR binding pocket and the glutaminase active site. The first of these metrics is based on direct path sampling and recapitulates the known importance of sideR in the allostery of IGPS. Paths in the Hessian-based adjacency matrix are found to be significantly longer and to be less diluted than paths found on two previously used adjacency matrices.

The second centrality metric we develop is based on the derivative of the covariance as a function of a given Hessian element. This metric is motivated by mutagenesis experiments in which one perturbs the interactions around a mutation site as compared to wild type. This metric identifies residues at the interface between HisF and HisH as being important for the allosteric network between the PRFAR binding pocket and the glutaminase active site. The correlation between the derivative and path centrality metrics is found to be minimal, suggesting that these two metrics provide different yet important information about the covariance between the two pockets.

Results from the path and derivative centrality metrics on the effective Hessian corroborate and functionally explain previous mutagenesis experiments. Interestingly, the combination of experimental and simulation results suggest that fP76-hK181 is a more important HisF–HisH interfacial connection for allosteric paths than the previously implicated fD98-hK181. Additionally, we propose a novel sight for targeted mutational or inhibitory binding studies based on the results obtained from our two centrality metrics.

## 4.6 Funding

This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562. We would specifically like to acknowledge the San Diego Supercomputer Center Comet and Pittsburgh Super Computing Bridges resources used under XSEDE allocation CHE160008 awarded to MM.

## 4.7 Supporting Information

### 4.7.1 Computational Methods

**Heterogeneous elastic network models**

Here we will closely follow the procedure known as heterogenous ENM (hENM) to generate an effective harmonic Hessian that best reproduces the aaMD measured covariance.[209] Our adapted hENM algorithm is as follows:

1. Initiate a trial $N \times N$ force constant matrix, $\boldsymbol{k}$.

2. Create the $3N \times 3N$ Hessian matrix, $\boldsymbol{H}$ using the tensor matrix elements $\boldsymbol{H}_{ij} = -k_{ij} \cdot \hat{r}_{ij} \otimes \hat{r}_{ij}$ for $i \neq j$ and $\boldsymbol{H}_{ii} = \sum_{j \neq i} \boldsymbol{H}_{ij}$

3. Get the predicted $3N \times 3N$ covariance, $\boldsymbol{C}$, by performing the Moore-Penrose psuedo inverse of $\boldsymbol{H}$, $\boldsymbol{C} = T\boldsymbol{H}^{+}$.

4. Update the force constants based on

$$k_{ij}^{n+1} = k_{ij}^{n} - \alpha \left( \frac{1}{\sigma_{ij}^{n}} - \frac{1}{\sigma_{ij}^{target}} \right) \tag{4.7}$$

where $\sigma_{ij}^n = \hat{r}_{ij} \left[ \boldsymbol{C}_{ii}^n + \boldsymbol{C}_{jj}^n - \boldsymbol{C}_{ij}^n - \boldsymbol{C}_{ji}^n \right] \hat{r}_{ij}$ is the variance of node $i$ and $j$ projected along the separation vector and $\alpha$ is a minimization step size parameter (we found $\alpha = 1 \times 10^{-2}$ to yield stable convergence).

5. Repeat steps 2-4 until a converged Hessian matrix has been produced.

### 4.7.2 All-Atom MD Simulations

All-atom, explicit solvent MD simulations are performed for the *apo* substrate state of *T. maritima* IGPS (PDB: 1GPW).[222] Specifically, chains C and D are used from this crystal structure. Crystallographic waters are maintained. The active site mutation, fN11D, present in this crystal structure was converted back to wild type.

The simulations are performed using the GPU-enabled AMBER18 software[142] and the ff14SB[15] parameters for proteins atoms. Using tleap, the starting structure is solvated in a TIP3P water box with at least a 12 Å buffer between the protein and periodic images. Sodium and chloride ions are added to neutralize charge and maintain a 0.10 M ionic concentration. Direct nonbonding interactions are calculated up to a 12 Å distance cutoff. The SHAKE algorithm is used to constrain covalent bonds that include hydrogen[77]. The particle-mesh Ewald method[78] is used to account for long-ranged electrostatic interactions. Before simulation began, a two stage minimization was performed: (1) 10,000 steps of conjugate gradient optimization were performed to minimize water positioning (substrate atoms restrained with a 75 kcal mol$^{-1}$ Å$^{-2}$ force constant) and (2) an additional 10,000 minimization steps with no restraints applied. The system was slowly heated from 25 K to 303 K over 1 ns. Additionally, 4 ns of NVT simulation was performed to equilibrate the cubic box volume. Finally, the simulations were run in the NTP ensemble, using the Langevin dynamics thermostat and Monte Carlo barostat to maintain the systems at 303 K and 1 bar. A 2 fs integration time step is used, with energies and positions written every 5 ps. An initial 500 ns simulation was run, from which the structure at 250 ns was used to initialize three more independent trajectories. These new trajectories were given different random number seeds and were run for an additional 250 ns. We have a total of 1 $\mu$s of trajectory to analyze the *apo* state of IGPS.

### 4.7.3 Theory

**Linear Response Solution to Allosteric Pathways**

**Path Sampling**

The emphasis of the ensemble of paths makes sampling the paths an apt way of studying them. For this purpose, we present a Markov chain simulation to sample the paths. This work is in contrast to previous studies, namely WISP,[113] that focus on finding the exact shortest paths up to some number. Instead, we statistically sample a distribution of paths. This algorithm is shown to be more efficient than using WISP to accomplish the same result.

The set of objects that are studied using the Markov chain simulation is the set of paths from a predefined source and sink nodes. Each path is then assigned an effective length. For the purpose of this work we consider some adjacency matrix $\boldsymbol{A}$ and functionalize it into a cost matrix $\boldsymbol{\chi}$ in an analogous way as the covariance is related to the Pearson correlation. Namely,

$$\chi_{ij} = -\log\left(\frac{|A_{ij}|}{\sqrt{A_{ii}A_{jj}}}\right) \tag{4.8}$$

In this study, we consider the paths of three adjacency matrices: the covariance, $r_{MI}$, and the Hessian. The length $\ell$ of a path is defined as a sum over values in the cost matrix corresponding to the edges in the path. The distribution of the paths that are studied is of the form

$$P_{path} \propto \exp\left[-\ell/\tau\right], \tag{4.9}$$

where the free parameter $\tau$ is an effective temperature of the simulation. The Markov chain simulation we present is not limited to these definitions of path lengths and path probabilities; more complex functions of the weights and the path distributions can readily be used.

A trial move in the Monte Carlo simulation consists of randomly choosing a node in the graph that is neither the source nor sink node from a uniform distribution, and add/remove the node to/from the path if it is currently/not in the present path. Any trial move that destroys the path, as will happen when an edge with a zero in the adjacency matrix is used, the move is rejected. The following values associated with the probability of adding the selected node $n$ between each pair of connected nodes $i$ and $j$ in the path not containing $n$ is then calculated

$$p_{ij} = \exp\left[-\frac{\chi_{in} + \chi_{nj} - \chi_{ij}}{\tau}\right]. \tag{4.10}$$

If the node is not in the current path, the trial move is a path with node $n$ randomly inserted between two connected nodes with the probability proportional to $p_{ij}$. Detailed balance is achieved by using a Metropolis algorithm whereby the trial path is accepted with a probability $P_{Met}$ of

$$P_{Met} = \min\left[1, \left(\sum_{\langle ij\rangle} p_{ij}\right)^{\pm 1}\right], \tag{4.11}$$

where the upper sign is for when the trial move removes a node from the path and the lower sign is for node addition.

The difference between using the Hessian as the adjacency matrix and the equation we derived relating the Hessian to the covariance in terms of paths can be mocked in the sampling on a few levels. As a first pass, the Hessian can be treated as a $3N \times 3N$ object where an edge in a graph is described by a $3 \times 3$ submatrix. The nature of the Hamiltonian we are using to find spring constants of the system produces a $3 \times 3$ submatrix describing the interaction of node $i$ and $j$ of the form $k_{ij}\hat{\boldsymbol{r}}_{ij}\hat{\boldsymbol{r}}_{ij}^{\intercal}$. The weight $w$ of a path can then be considered to be the product of these spring constants, treated in the same fashion as above, as

$$\begin{aligned} w &= \prod_{\langle i,j\rangle} \frac{k_{ij}}{\sqrt{k_{ii}k_{jj}}}\hat{\boldsymbol{r}}_{ij}\hat{\boldsymbol{r}}_{ij}^{\intercal} \\ &= \hat{\boldsymbol{r}}_{i_1 i_2}\hat{\boldsymbol{r}}_{i_{M-1}i_M}^{\intercal} \prod_{\langle i,j,k\rangle} \cos\theta_{ijk} \prod_{\langle i,j\rangle} \frac{k_{ij}}{\sqrt{k_{ii}k_{jj}}}, \end{aligned} \tag{4.12}$$

where $\langle i, j, k\rangle$ is the set of all three consecutive nodes in the path. The two products in this equation can be interpreted as the scalar weight of the path and the tensor part is treated separately. Taking the length of the path to be equal to $\ell = -\log w$, these paths can similarly sampled with the following change to the $p_{ij}$ defined above,

$$p_{ij} = \left( \frac{\cos\theta_{i-1,i,n} \cos\theta_{i,n,j} \cos\theta_{n,j,j+1}}{\cos\theta_{i-1,i,j} \cos\theta_{i,j,j-1}} \right)^{1/\tau} \exp\left[ -\frac{\chi_{in} + \chi_{nj} - \chi_{ij}}{\tau} \right]. \tag{4.13}$$

Notice that this definition of path length is not trivially studied by exhaustive search algorithms such as that proposed by WISP, but it readily studied using this Markov chain simulation framework.

A simulation of the paths consists of $N$ trial moves between writing the path to file, where $N$ is the number of nodes in the graph, and then $10^6$ paths are generated to sample the ensemble. The choice of $\tau$ is dependent on the desired ensemble of paths to sample. For the present study, $\tau$ is chosen such that the first thousand shortest paths are well sampled so that our results can be compared to that of WISP.

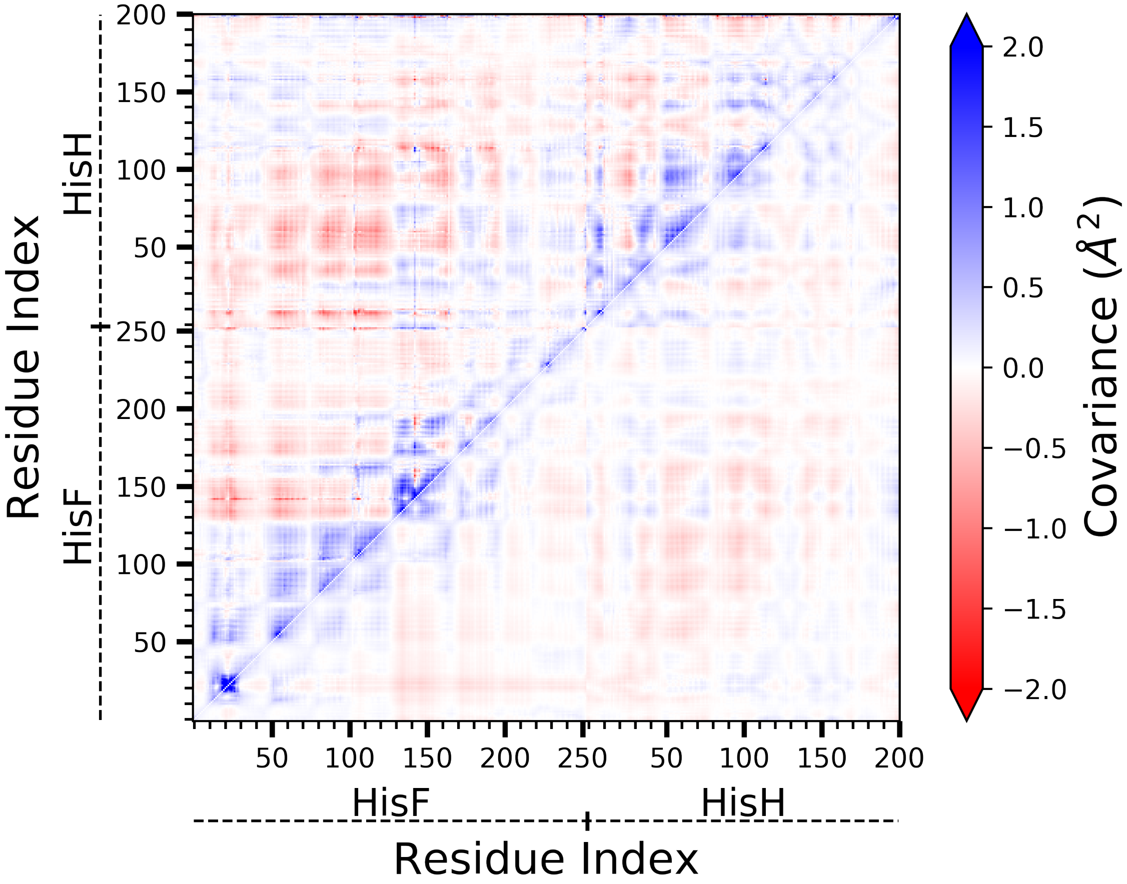

**Figure 4.5:** Covariance matrices computed through the hENM procedure (bottom right) and raw simulation (top left) from 1 $\mu$s all-atom molecular dynamics of *apo* IGPS.

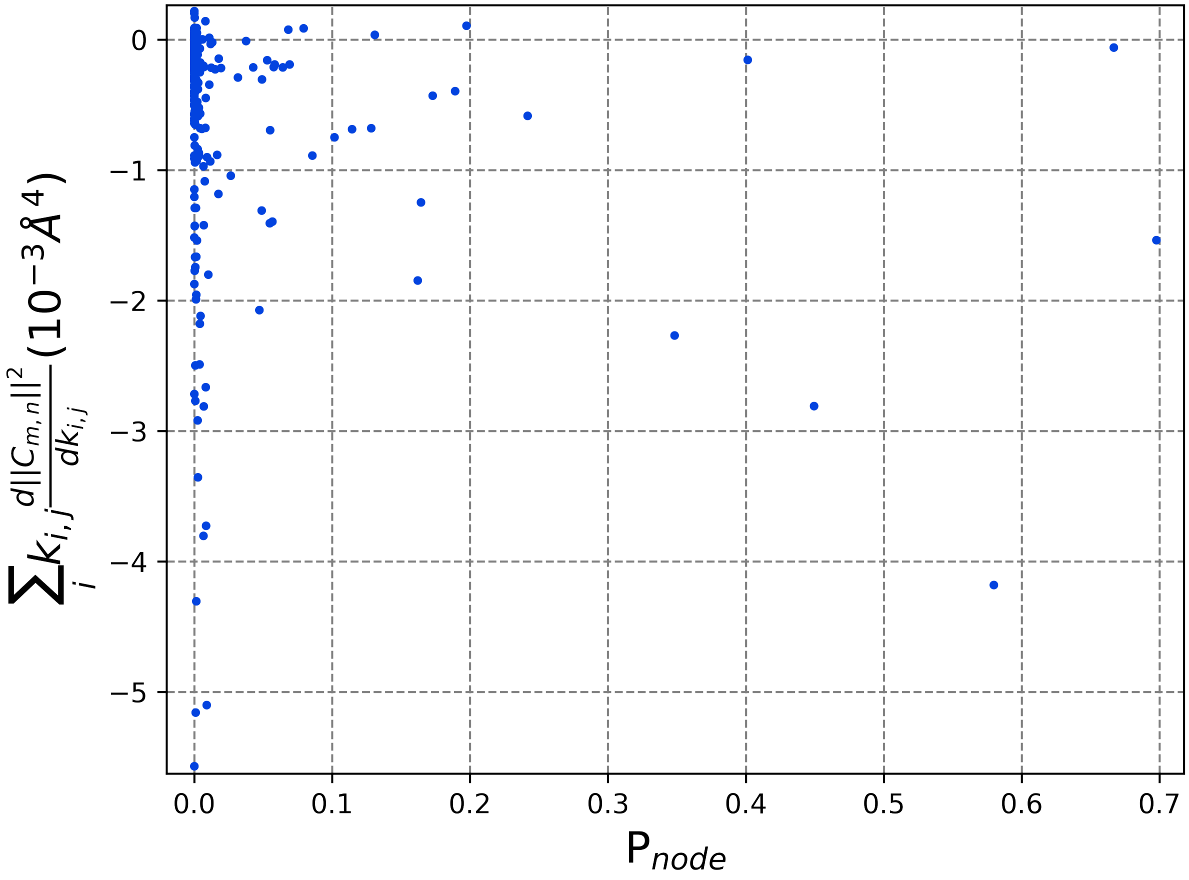

**Figure 4.6:** Correlation of Hessian derivative metric and $P_{\text{node}}$.

# Chapter 5

# Conclusions

The previous three chapters have focused on the study of allostery in the flavivirus NS3h (Chapters 2 and 3) and bacterial IGPS (Chapter 4) proteins. Allostery is the biochemically-observed phenomenon where the presence of a ligand, termed the effector molecule, enhances or inhibits or regulates, in some fashion, the activity of a protein's functions. From a molecular or atomic perspective, this phenomenon can generally be described by direct interactions between the effector molecule and the protein at the allosteric active site that lead to some structural or dynamic perturbations at the protein's orthosteric site. The mechanisms of allostery is system-specific and poorly understood due to the complexity of the phenomenon.

The study of allostery has been broken down into three scientific questions that focus on various aspects of the phenomenon:

1. What is the allosteric effect? How does the effector molecule perturb the orthosteric site?

2. What are the interactions between the enzyme and effector that lead to this allosteric effect?

3. How are the orthosteric and allosteric sites coupled?

Specific aims of study to answer these questions can utilize molecular dynamics (MD) simulations to glean novel, atomic-resolution insights about allosteric proteins, specifically focusing on the interactions between the protein and its ligands. This dissertation is a presentation of my research, which has begun to answer these three questions for the flaviviral NS3h and bacterial IGPS proteins. Both of these enzymes have wild-type functions that are regulated using allosteric mechanisms, as observed in experimental studies, yet these mechanisms are incompletely understood at a fundamental, atomistic resolution. Therefore, the modeling of these protein systems that I report here has forwarded our understanding of the allosteric mechanisms at play utilized by these proteins.

Specifically, in Chapter 2, comparative analysis of a set of all-atom MD simulations of the dengue NS3h was used to highlight allosteric effects for both the NTP hydrolysis active site and the RNA-binding cleft. This body of simulations was used to identify that RNA decreases the number of waters, slows their dynamics, and affects their positioning within the hydrolysis active site as well as affects the hydrolysis reaction's energetics. Similarly, the hydrolysis active site's substrates (nucleoside triphosphates) are shown to perturb the protein-RNA interactions within the RNA-binding cleft. This work directly highlights aspects of the allosteric effects (question 1) that are observed in the dengue NS3h system and are hypothesized to be in play for other flavivirus NS3h as well.

In Chapter 3, the structural states of a large region of the RNA-binding cleft of Zika NS3h are studied to identify the direct protein-RNA interactions that function as origins of the observed RNA-induced allosteric effects on the NTP hydrolysis active site. Results from the presented set of unbiased and biased MD simulations highlight how RNA perturbs the structural free energy landscape of the RNA-binding loop (known as L$\beta3\beta4$). Residues of this local structural region are seen to have functional importance to the RNA-induced structural changes in L$\beta3\beta4$. Additionally, the free energy surfaces of the various RNA-bound states allow us to hypothesize a mechanism of NS3h "recognizing" the binding of an RNA oligomer and subsequently undergoing large structural changes to enable further binding. With the aid of structural and sequential alignments, I hypothesize that these observations are consistent for all of the flavivirus NS3h's.

Chapter 4 moves away from the study of flavivirus' NS3h and instead presents a novel methodology to study the short-range, residue-residue interactions that connect the orthosteric and allosteric active sites within a protein, using MD simulations to thoroughly sample residue-residue interactions. The past literature has attempted to study these interaction networks using, e.g., the Pearson correlation or linear mutual information matrices to quantify the strength of direct residue-residue interactions, which results in nonphysical descriptions of the protein network. Instead, the effective harmonic Hessian matrix and respective centrality metrics, presented in Chapter 4, quantify a physically-relevant representation of the protein network that highlights residues of importance to allosteric coupling between two active sites. The bacterial IGPS protein complex is used as a benchmark system to test the veracity of this new methodology, to considerable success.

## 5.1 Continued Work

Further study of both the flavivirus NS3h and bacterial IGPS systems are needed to further elucidate the atomic-resolution mechanisms behind allostery. While the studies presented in this dissertation have provided novel insights into the respective systems' allostery, much work can be done to verify these results as well as answer the questions presented above in more detail. Ideas of continued research on both the flavivirus NS3h and effective harmonic Hessian projects are presented here.

### 5.1.1 Allostery of Flavivirus NS3h

The *Flaviviridae flavivirus* NS3h system is an exceptional benchmark system for the study of the mechanisms of allosteric regulation. The coupling between the NTP hydrolysis active site and the RNA-binding cleft are pivotal to the replication of the viral genome. Insights gained from continued research on these aspects of NS3h may provide the key to treatment of the respective flavivirus diseases.

As a first step to studying the couplings between the NTPase and helicase active sites, Du Pont *et al.* has utilized MD simulations as well as mutagenesis and virological studies of NS3h Motif V residues, which are seen in Chapter 1 to have strong correlated fluctuations with residues in both active sites.[7,138] Concurrent application of computational and experimental studies of NS3h highlighted specific residue mutations that resulted in an attenuated virus as well as the atomic-resolution explanation of this phenotype. Additionally, this study provides a strong research protocol for similar studies attempting to directly correlation experimental results with computational simulations (and vice versa) for the flavivirus NS3h system.

For example, a similar methodology to that used in Du Pont *et al.*[138] could be envisioned to examine the importance of $\alpha 2$ (Motif Ia) and L$\beta 3 \beta 4$ residues in relation to the RNA-induced allosteric effect on the NTPase active site as well as RNA-binding. Perturbations to the certain residues in this region are hypothesized to affect the L$\beta3\beta4$ conformational states, which is hypothesized to be observable in NTPase activity and RNA-binding biochemical assays. The experimentally performed mutations could be modeled in MD simulations to provide atomic resolution understanding of the *in situ* results.

Additionally, the $\alpha2$ and L$\beta3\beta4$ secondary structures represent the region of NS3h with the shortest distance separating the RNA-binding cleft and the NTPase active site. Therefore, I hypothesize that it is through this region's network of short-ranged, residue-residue interactions that RNA perturbs the hydrolysis active site most. This hypothesis can be tested by the application of the effective harmonic Hessian analysis to study how the two active sites in the protein are coupled. The application of this new methodology to the NS3h system will provide a novel, physically-relevant description of the protein that can further explain and validate biochemical and virological results. Additionally, the application of this description of the protein network can guide further computational and experimental studies in structural regions of the NS3h system that have yet to be identified by either computational or experimental studies to date.

These proposed research aims would continue to answer the three scientific questions associated with allostery.

### 5.1.2 NS3h as a Molecular Machine

In addition to the virological importance of the flavivirus NS3h, the protein is an exemplar molecular machine that utilizes the hydrolysis of nucleoside triphosphates (NTPs) to power the helicase functions. In this way, the conversion of chemical energy into mechanical work by NS3h is a fascinating phenomenon that is incompletely understood. Research on the conversion, transfer, and utilization of free energy within a protein matrix will provide novel understanding of the biophysics of energy transduction within molecular machines. These insights may begin to allow scientists to design new or utilize already-available molecular machines to perform novel functions at the molecular level.

In this regard, continued modeling of the NS3h NTP hydrolysis enzymatic cycle (see Figure 2.1) will provide new details on energy release from the hydrolysis cycle and, subsequently, how this energy is transduced to the RNA-binding cleft. Chapter 2 represents a foundational study of the equilibrium substrate states of this cycle, where we are hypothesizing that each substrate state is long-lived relative to the life time of the transitions between substrate states. With this body of simulations, we have sampled, to some degree, the ensemble of NTPase active site and RNA-binding cleft conformational states. There is further work to be done on the quantification of the various substrate states to identify the biophysically relevant structures of both the NTPase active site and RNA-binding cleft in regards to the energy conversion and transduction mechanisms. For example, the definition of the active and inactive conformations of the NTPase active site, as quantified by the lytic water analysis in Chapter 2, does not account for conformations of the protein residues in this region. Applying a conformational study, similar to the PCA analysis reported in Chapter 3, could independently quantify the structural states of the hydrolysis active site, allowing for us to correlate lytic water states with structural states. A second generation of quantum mechanical calculations or quantum mechanics molecular dynamics simulations could be performed to study the hydrolysis reaction in the identified structural states.

A novel research endeavor in this specific aim is to model the dynamic events of the hydrolysis active site, specifically the binding of NTP and unbinding of the hydrolysis products (nucleoside diphosphate and inorganic phosphate). These events are equally as likely to be sources of free energy release as the hydrolysis reaction and, therefore, deserve to be modeled so as to calculate the free energy surfaces of the respective events. Utilization of novel or cutting-edge simulation protocols will likely be necessary to obtain adequate sampling of the events due to the complexity of the collective variable space as well as conformational states of the NS3h system.

Finally, the identification of paths through which energy is transduced from the NTPase active site to the RNA-binding cleft is a major goal for this specific aim. The effective harmonic Hessian methodology, developed in Chapter 4, might provide physically-relevant insight towards this research goal. Yet, further development and verification of the methodology is needed to have confidence in the interpretation of the method's results in regards to energy transduction.

### 5.1.3 Further Development of the Effective Harmonic Hessian Method

The effective harmonic Hessian methodology is a physically-motivated metric to describe the short-ranged, residue-residue interactions through which allosteric signals are hypothesized to propagate. The methodology can be broken down into two steps: (1) quantification of the Hessian matrix that describes the interaction network and (2) centrality and path analyses to identify residues that are important in the allosteric mechanisms of the protein. There are improvements to be had in both steps of the current methodology.

Generally, the Hessian matrix is the pseudo-inverse of the residue-residue covariance matrix and, so, assumes that a residue-level coarse-graining of the protein system is acceptable when describing the direct, short-ranged interactions of interest. Additionally, the methodology has not implemented a state-space quantification. I hypothesize that conformational states of the protein system, as defined by a fine-grained description of the protein, will have quantifiable differences in the Hessian matrix, which may be important in regards to the functions of the protein (see the discussion of the NS3h hydrolysis active site above). Finally, the algorithm used to converge the Hessian matrix has not been optimized for the desired application nor has error been quantified for this calculation. The current convergence method does not efficiently reach a minimum when evaluating the model versus the training data. All of these aspects of the Hessian matrix calculation are potential sources for bias and error to enter into the results produced from this methodology. Research endeavors should be undertaken by users of this method so as to minimize the effect that these biases might have on their results.

Additionally, the development of physically-relevant centrality metrics is a continued research endeavor in the McCullagh group. The $P_{node}$ metric currently assumes that only the direct paths connecting the two active sites are important to the allosteric mechanisms within a protein system. This metric assumes that paths that do not connect the two active sites but rather connect the active sites to other regions of the protein network (termed "broken paths") are unimportant. There is some doubt to the validity of this assumption and so metrics are being developed to account for the ensemble of broken paths within the protein network. Additionally, there is ambiguity in the interpretation of the derivative metric. What does the sign of the $\delta_{node}$ value signify? How do the magnitudes of this metric correlate with experimental observations? Can this metric be verified by an experimental observable? These questions highlight research endeavors underway in the McCullagh group in regards to the development of centrality metrics that have direct, physical interpretations.

## 5.2 A Broad Perspective

The research presented in this dissertation represents foundational work for the specific aims described above. Beyond those proposed research endeavors, I believe my work has the potential to lead to broader impacts in the fields of virology and biophysics. Utilizing all of the computational techniques described in this dissertation, one can begin to provide the degree of understanding of a protein system, at an atomic resolution, to confidently propose residues to experimentally perturb and predict the resulting phenotype. Currently, the strong connection between the McCullagh and Geiss groups allows for direct testing of hypotheses developed from the computational study of the flaviviral NS3h. Continued development of the collaborative workflow between these two groups has the potential to produce an efficient research pipeline that can intake MD simulations of NS3h, develop experimentally-testable hypotheses as to perturbing the wild-type protein, and subsequently test these hypotheses using biochemical and virological methods. The development of such a pipeline is nontrivial due to the huge differences in time- and length-scales that computational and experimental techniques report on. Yet, results reported in this dissertation as well as in Du Pont *et al.*[138] indicate that the development of such a pipeline is underway.

Prediction of biochemical or virological phenotypes from MD simulations is exceptionally exciting when considering the development of antiviral therapeutics and vaccines. Intelligent design of these antiviral treatments is currently infeasible due to the hugely complex interactions at play during the life cycle of viruses. The current strategies to identify small-molecule inhibitors or vaccines to treat virus infections use high-throughput methods that are cost- and time-inefficient as well as are naive of the underlying, atomic-resolution interactions being perturbed. With the research pipeline discussed above, we can begin to identify specific residues in the protein that we hypothesize will attenuate the wild-type functioning of the protein when perturbed by mutation or small-molecule drug binding. This will greatly speed up the identification and development process of antiviral treatments. Using the flavivirus NS3h as a benchmark system, similar research pipelines can be envisioned for a broad range of protein systems. In this light, my research has provided the initial, detailed insights into the NS3h structure and function that sparked the development of this research pipeline.

# Bibliography

(1) Nussinov, R.; Tsai, C.-J. *Cell* **2013**, *153*, 293–305.

(2) Nussinov, R.; Tsai, C.-J.; Ma, B. *Annu. Rev. Biophys.* **2013**, *42*, 169–189.

(3) Nussinov, R.; Tsai, C. J.; Liu, J. **2014**, *136*, 17692–17701.

(4) Ahuja, L. G.; Aoto, P. C.; Kornev, A. P.; Veglia, G.; Taylor, S. S. *Proc. Natl. Acad. Sci. U.S.A.* **2019**, *116*, 15052–15061.

(5) Ahuja, L. G.; Taylor, S. S.; Kornev, A. P. **2019**, *71*, 685–696.

(6) Yuan, Y.; Tam, M. F.; Simplaceanu, V.; Ho, C. *Chem. Rev.* **2015**, *115*, 1702–1724.

(7) Davidson, R. B.; Hendrix, J.; Geiss, B. J.; McCullagh, M. *PLoS Comput. Biol.* **2018**, *14*, e1006103–28.

(8) Monod, J.; Wyman, J.; Changeux, J. P. *J. Mol. Biol.* **1965**, *12*, 88–118.

(9) Koshland, D. E.; Némethy, G; Filmer, D *Biochemistry* **1966**, *5*, 365–385.

(10) Einav, T.; Mazutis, L.; Phillips, R. *Journal of Physical Chemistry B* **2016**, *120*, 6021–6037.

(11) Shaanan, B. *J. Mol. Biol.* **1983**, *171*, 31–59.

(12) Fermi, G.; Perutz, M. F.; Shaanan, B.; Fourme, R. *J. Mol. Biol.* **1984**, *175*, 159–174.

(13) Motlagh, H. N.; Wrabl, J. O.; Li, J.; Hilser, V. J. *Nature* **2014**, *508*, 331–339.

(14) Wodak, S. J.; Paci, E.; Dokholyan, N. V.; Berezovsky, I. N.; Horovitz, A.; Li, J.; Hilser, V. J.; Bahar, I.; Karanicolas, J.; Stock, G.; Hamm, P.; Stote, R. H.; Eberhardt, J.; Chebaro, Y.; Dejaegere, A.; Cecchini, M.; Changeux, J. P.; Bolhuis, P. G.; Vreede, J.; Faccioli, P.; Orioli, S.; Ravasio, R.; Yan, L.; Brito, C.; Wyart, M.; Gkeka, P.; Rivalta, I.; Palermo, G.; McCammon, J. A.; Panecka-Hofman, J.; Wade, R. C.; Di Pizio, A.; Niv, M. Y.; Nussinov, R.; Tsai, C. J.; Jang, H.; Padhorny, D.; Kozakov, D.; McLeish, T. *Structure* **2019**, *27*, 566–578.

(15) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. *J. Chem. Theory Comput.* **2015**, *11*, 3696–3713.

(16) Bhatt, S.; Gething, P. W.; Brady, O. J.; Messina, J. P.; Farlow, A. W.; Moyes, C. L.; Drake, J. M.; Brownstein, J. S.; Hoen, A. G.; Sankoh, O.; Myers, M. F.; George, D. B.; Jaenisch, T.; Wint, G. R. W.; Simmons, C. P.; Scott, T. W.; Farrar, J. J.; Hay, S. I. *Nature* **2013**, *496*, 504–507.

(17) Saeedi, B. J.; Geiss, B. J. *Wiley Interdiscip. Rev. RNA* **2013**, *4*, 723–735.

(18) Henchal, E. A.; Putnak, J. R. *Clin. Microbiol. Rev.* **1990**, *3*, 376–396.

(19) Musso, D.; Gubler, D. J. *Clin. Microbiol. Rev.* **2016**, *29*, 487–524.

(20) Kadaré, G; Haenni, A. L. *J. Virol.* **1997**, *71*, 2583–2590.

(21) Borowski, P.; Niebuhr, A.; Schmitz, H.; Hosmane, R. S.; Bretner, M.; Siwecka, M. A.; Kulikowski, T. *Acta Biochim. Pol.* **2002**, *49*, 597–614.

(22) Raney, K. D.; Sharma, S. D.; Moustafa, I. M.; Cameron, C. E. *J. Biol. Chem.* **2010**, *285*, 22725–22731.

(23) Leung, D; Schroder, K; White, H; Fang, N. X.; Stoermer, M. J.; Abbenante, G; Martin, J. L.; Young, P. R.; Fairlie, D. P. *J. Biol. Chem.* **2001**, *276*, 45762–45771.

(24) 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. *Antimicrob. Agents Chemother.* **2013**, *57*, 1902–1912.

(25) 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. *ACS Chem. Biol.* **2015**, *10*, 1887–1896.

(26) 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. *ACS Infect. Dis.* **2015**, *1*, 140–148.

(27) 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. *Antiviral Res.* **2017**, *139*, 49–58.

(28) 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 *J. Antimicrob. Chemother.* **2012**, *67*, 1884–1894.

(29) Basavannacharya, C.; Vasudevan, S. G. *Biochem. Biophys. Res. Commun.* **2014**, *453*, 539–544.

(30) Shadrick, W. R.; Mukherjee, S.; Hanson, A. M.; Sweeney, N. L.; Frick, D. N. *Biochemistry* **2013**, *52*, 6151–6159.

(31) Chambers, T. J.; Weir, R. C.; Grakoui, A; McCourt, D. W.; Bazan, J. F.; Fletterick, R. J.; Rice, C. M. *Proc. Natl. Acad. Sci. U.S.A.* **1990**, *87*, 8898–8902.

(32) Wengler, G; Wengler, G *Virology* **1991**, *184*, 707–715.

(33) Warrener, P; Tamura, J. K.; Collett, M. S. *J. Virol.* **1993**, *67*, 989–996.

(34) Kuo, M. D.; Chin, C; Hsu, S. L.; Shiao, J. Y.; Wang, T. M.; Lin, J. H. *J. Gen. Virol.* **1996**, *77*, 2077–2084.

(35) Utama, A.; Shimizu, H.; Morikawa, S.; Hasebe, F.; Morita, K.; Igarashi, A.; Hatsu, M.; Takamizawa, K.; Miyamura, T. *FEBS Lett.* **2000**, *465*, 74–78.

(36) Borowski, P; Niebuhr, A; Mueller, O; Bretner, M; Felczak, K; Kulikowski, T; Schmitz, H *J. Virol.* **2001**, *75*, 3220–3229.

(37) Wang, C.-C.; Huang, Z.-S.; Chiang, P.-L.; Chen, C.-T.; Wu, H.-N. *FEBS Lett.* **2009**, *583*, 691–696.

(38) Klema, V. J.; Padmanabhan, R.; Choi, K. H. *Viruses* **2015**, *7*, 4640–4656.

(39) Matusan, A. E.; Pryor, M. J.; Davidson, A. D.; Wright, P. J. *J. Virol.* **2001**, *75*, 9633–9643.

(40) Benarroch, D.; Selisko, B.; Locatelli, G. A.; Maga, G.; Romette, J.-L.; Canard, B. *Virology* **2004**, *328*, 208–218.

(41) Sampath, A.; Xu, T.; Chao, A.; Luo, D.; Lescar, J.; Vasudevan, S. G. *J. Virol.* **2006**, *80*, 6686–6690.

(42) Lim, S. P.; Wang, Q.-Y.; Noble, C. G.; Chen, Y.-L.; Dong, H.; Zou, B.; Yokokawa, F.; Nilar, S.; Smith, P.; Beer, D.; Lescar, J.; Shi, P.-Y. *Antiviral Res.* **2013**, *100*, 500–519.

(43) Fairman-Williams, M. E.; Guenther, U.-P.; Jankowsky, E. *Curr. Opin. Struct. Biol.* **2010**, *20*, 313–324.

(44) Serebrov, V.; Pyle, A. M. *Nature* **2004**, *430*, 476–480.

(45) Beran, R. K. F.; Bruno, M. M.; Bowers, H. A.; Jankowsky, E.; Pyle, A. M. *J. Mol. Biol.* **2006**, *358*, 974–982.

(46) Zhang, C; Cai, Z; Kim, Y. C.; Kumar, R; Yuan, F; Shi, P. Y.; Kao, C; Luo, G *J. Virol.* **2005**, *79*, 8687–8697.

(47) Preugschat, F; Averett, D. R.; Clarke, B. E.; Porter, D. J. *J. Biol. Chem.* **1996**, *271*, 24449–24457.

(48) Levin, M. K.; Patel, S. S. *J. Biol. Chem.* **2002**, *277*, 29377–29385.

(49) Levin, M. K.; Gurjar, M. M.; Patel, S. S. *J. Biol. Chem.* **2003**, *278*, 23311–23316.

(50) Levin, M. K.; Gurjar, M.; Patel, S. S. *Nat. Struct. Mol. Biol.* **2005**, *12*, 429–435.

(51) Myong, S.; Bruno, M. M.; Pyle, A. M.; Ha, T. *Science* **2007**, *317*, 513–516.

(52) Dumont, S.; Cheng, W.; Serebrov, V.; Beran, R. K.; Tinoco, I.; Pyle, A. M.; Bustamante, C. *Nature* **2006**, *439*, 105–108.

(53) Cheng, W.; Arunajadai, S. G.; Moffitt, J. R.; Tinoco, I.; Bustamante, C. *Science* **2011**, *333*, 1746–1749.

(54) Gu, M.; Rice, C. M. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 521–528.

(55) Büttner, K.; Nehring, S.; Hopfner, K.-P. *Nat. Struct. Mol. Biol.* **2007**, *14*, 647–652.

(56) Hopfner, K.-P.; Michaelis, J. *Curr. Opin. Struct. Biol.* **2007**, *17*, 87–95.

(57) Yodh, J. G.; Schlierf, M.; Ha, T. *Q. Rev. Biophys.* **2010**, *43*, 185–217.

(58) Pyle, A. M. *Annu. Rev. Biophys.* **2008**, *37*, 317–336.

(59) Appleby, T. C.; Anderson, R.; Fedorova, O.; Pyle, A. M.; Wang, R.; Liu, X.; Brendza, K. M.; Somoza, J. R. *J. Mol. Biol.* **2011**, *405*, 1139–1153.

(60) Luo, D.; Xu, T.; Watson, R. P.; Scherer-Becker, D.; Sampath, A.; Jahnke, W.; Yeong, S. S.; Wang, C. H.; Lim, S. P.; Strongin, A.; Vasudevan, S. G.; Lescar, J. *EMBO J.* **2008**, *27*, 3209–3219.

(61) Dittrich, M.; Schulten, K. *Structure* **2006**, *14*, 1345–1353.

(62) Yu, J.; Ha, T.; Schulten, K. *Biophys. J.* **2006**, *91*, 2097–2114.

(63) Yu, J.; Ha, T.; Schulten, K. *Biophys. J.* **2007**, *93*, 3783–3797.

(64) Dittrich, M; Yu, J; Schulten, K. *Top. Curr. Chem.* **2007**, *268*, 319–347.

(65) Ma, W.; Schulten, K. *J. Am. Chem. Soc.* **2015**, *137*, 3031–3040.

(66) Yoshimoto, K.; Arora, K.; Brooks, C. L. *Biophys. J.* **2010**, *98*, 1449–1457.

(67) Pérez-Villa, A.; Darvas, M.; Bussi, G. *Nucleic Acids Res.* **2015**, *43*, 8725–8734.

(68) Zheng, W.; Liao, J.-C.; Brooks, B. R.; Doniach, S. *Proteins* **2007**, *67*, 886–896.

(69) Flechsig, H.; Mikhailov, A. S. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 20875–20880.

(70) Mastrangelo, E.; Bolognesi, M.; Milani, M. *Biochem. Biophys. Res. Commun.* **2012**, *417*, 84–87.

(71) Mottin, M.; Braga, R. C.; da Silva, R. A.; da Silva, J. H.; Perryman, A. L.; Ekins, S.; Andrade, C. H. *Biochemical and Biophysical Research Communications* **2017**, *492*, 643–651.

(72) Case, D. A.; Babin, V; Berryman, J. T.; Betz, R.; Cai, Q.; Cerutti, D.; Cheatham, T.E., I.; Darden, T.; Duke, R.; Gohlke, H.; Goetz, A.; Gusarov, S.; Homeyer, N.; Janowski, R.; Kaus, J.; Kolossváry, I; Kovalenko, A.; Lee, T.; LeGrand, S.; Luchko, T.; Luo, R.; Madej, B.; Merz, K.; Paesani, F.; Roe, D.; Roitberg, A.; Sagui, C.; Salomon-Ferrer, R.; Seabra, G.; Simmerling, C.; Smith, W.; Swails, J.; Walker, R.; Wang, J.; Wolf, R.; Wu, X.; Kollman, P. AMBER 14., University of California, San Francisco, 2014.

(73) Banáš, P.; Hollas, D.; Zgarbová, M.; Jurečka, P.; Orozco, M.; Cheatham, T. E.; Šponer, J.; Otyepka, M. *J. Chem. Theory Comput.* **2010**, *6*, 3836–3849.

(74) Zgarbová, M.; Otyepka, M.; Šponer, J.; Mládek, A.; Banáš, P.; Cheatham, T. E.; Jurečka, P. *J. Chem. Theory Comput.* **2011**, *7*, 2886–2902.

(75) Meagher, K. L.; Redman, L. T.; Carlson, H. A. *J. Comp. Chem.* **2003**, *24*, 1016–1025.

(76) Li, P.; Roberts, B. P.; Chakravorty, D. K.; Merz, K. M. *J. Chem. Theory Comput.* **2013**, *9*, 2733–2748.

(77) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. *J. Comp. Phys.* **1977**, *23*, 327–341.

(78) Cheatham, T. E. I.; Miller, J. L.; Fox, T.; Darden, T. A.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 4193–4194.

(79) Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. *J. Comput. Chem.* **1992**, *13*, 1011–1021.

(80) Chai, J.-D.; Head-Gordon, M. *Phys. Chem. Chem. Phys.* **2008**, *10*, 6615–6620.

(81) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi,

M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. Gaussian 09, Revision B.01., Wallingford CT, 2009.

(82) Schenker, S.; Schneider, C.; Tsogoeva, S. B.; Clark, T. *J. Chem. Theory Comput.* **2011**, *7*, 3586–3595.

(83) Smith, S. A.; Hand, K. E.; Love, M. L.; Hill, G.; Magers, D. H. *J. Comput. Chem.* **2013**, *34*, 558–565.

(84) Duarte, F.; Barrozo, A.; Åqvist, J.; Williams, N. H.; Kamerlin, S. C. L. *J. Am. Chem. Soc.* **2016**, *138*, 10664–10673.

(85) Michaud-Agrawal, N.; Denning, E. J.; Woolf, T. B.; Beckstein, O. *J. Comput. Chem.* **2011**, *32*, 2319–2327.

(86) Hunter, J. D. *Comput. Sci. Eng.* **2007**, *9*, 90–95.

(87) Humphrey, W; Dalke, A; Schulten, K *J. Mol. Graph.* **1996**, *14*, 33–38.

(88) Frishman, D; Argos, P *Proteins* **1995**, *23*, 566–579.

(89) Stone, J. *An Efficient Library for Parallel Ray Tracing and Animation.*, MA thesis, Computer Science Department, University of Missouri-Rolla, 1998.

(90) Grigorenko, B. L.; Rogov, A. V.; Topol, I. A.; Burt, S. K.; Martinez, H. M.; Nemukhin, A. V. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 7057–7061.

(91) McCullagh, M.; Saunders, M. G.; Voth, G. A. *J. Am. Chem. Soc.* **2014**, *136*, 13053–13058.

(92) Akola, J; Jones, R. O. *J. Phys. Chem. B* **2006**, *110*, 8121–8129.

(93) Freedman, H.; Laino, T.; Curioni, A. *J. Chem. Theory Comput.* **2012**, *8*, 3373–3383.

(94) McGrath, M. J.; Kuo, I. F. W.; Hayashi, S.; Takada, S. *J. Am. Chem. Soc.* **2013**, *135*, 8908–8919.

(95) Hsu, W.-L.; Furuta, T.; Sakurai, M. *J. Phys. Chem. B* **2016**, *120*, 11102–11112.

(96) Okimoto, N; Yamanaka, K; Ueno, J; Hata, M; Hoshino, T *Biophys. J.* **2001**, 2786–2794.

(97) Sun, R.; Sode, O.; Dama, J. F.; Voth, G. A. *J. Chem. Theory Comput.* **2017**, *13*, 2332–2341.

(98) Akola, J; Jones, R. O. *J. Phys. Chem. B* **2003**, *107*, 11774–11783.

(99) Harrison, C. B.; Schulten, K. *J. Chem. Theory Comput.* **2012**, *8*, 2328–2335.

(100) Glaves, R.; Mathias, G.; Marx, D. *J. Am. Chem. Soc.* **2012**, *134*, 6995–7000.

(101) Zalatan, J. G.; Herschlag, D. *J. Am. Chem. Soc.* **2006**, *128*, 1293–1303.

(102) Rosta, E.; Kamerlin, S. C.; Warshel, A. *Biochemistry* **2008**, *47*, 3725–3735.

(103) Kim, M.-C.; Sim, E.; Burke, K. *Phys. Rev. Lett.* **2013**, *111*, 073003.

(104) Sengupta, D.; Behera, R. N.; Smith, J. C.; Ullmann, G. M. *Structure* **2005**, *13*, 849–855.

(105) Roe, D. R.; Cheatham, T. E. *J. Chem. Theory Comput.* **2013**, *9*, 3084–3095.

(106) Calnan, B. J.; Tidor, B.; Biancalana, S.; Hudson, D.; Frankel, A. D. *Science* **1991**, *252*, 1167–1171.

(107) Tao, J; Frankel, A. D. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 2723–2726.

(108) Lam, A. M. I.; Keeney, D.; Frick, D. N. *J. Biol. Chem.* **2003**, *278*, 44514–44524.

(109) Tian, H.; Ji, X.; Yang, X.; Zhang, Z.; Lu, Z.; Yang, K.; Chen, C.; Zhao, Q.; Chi, H.; Mu, Z.; Xie, W.; Wang, Z.; Lou, H.; Yang, H.; Rao, Z. *Protein Cell* **2016**, *7*, 562–570.

(110) Sethi, A.; Eargle, J.; Black, A. A.; Luthey-Schulten, Z. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 6620–6625.

(111) Van Wart, A. T.; Eargle, J.; Luthey-Schulten, Z.; Amaro, R. E. *J. Chem. Theory Comput.* **2012**, *8*, 2949–2961.

(112) Doshi, U.; Holliday, M. J.; Eisenmesser, E. Z.; Hamelberg, D. *Proc. Natl. Acad. Sci. U.S.A.* **2016**, *113*, 4735–4740.

(113) Van Wart, A. T.; Durrant, J.; Votapka, L.; Amaro, R. E. *J. Chem. Theory Comput.* **2014**, *10*, 511–517.

(114) Seo, M.-H.; Park, J.; Kim, E.; Hohng, S.; Kim, H.-S. *Nat. Commun.* **2014**, *5*, 3724.

(115) Cecchini, M.; Houdusse, A.; Karplus, M. *PLoS Computational Biology* **2008**, *4*, ed. by Jacobson, M. P., e1000129.

(116) Cui, Q.; Karplus, M. *Protein Sci.* **2008**, *17*, 1295–1307.

(117) Liu, D.; Wang, Y.-S.; Gesell, J. J.; Wyss, D. F. *J. Mol. Biol.* **2001**, *314*, 543–561.

(118) Liu, D.; Wyss, D. F. **2001**, *19*, 283–284.

(119) Gesell, J. J.; Liu, D.; Madison, V. S.; Hesson, T.; Wang, Y. S.; Weber, P. C.; Wyss, D. F. *Protein Eng. Des. Sel.* **2001**, *14*, 573–582.

(120) Pervushin, K. V.; Wider, G; Riek, R; Wuthrich, K. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 9607–9612.

(121) Messina, J. P.; Kraemer, M. U.; Brady, O. J.; Pigott, D. M.; Shearer, F. M.; Weiss, D. J.; Golding, N.; Ruktanonchai, C. W.; Gething, P. W.; Cohn, E.; Brownstein, J. S.; Khan, K.; Tatem, A. J.; Jaenisch, T.; Murray, C. J.; Marinho, F.; Scott, T. W.; Hay, S. I. *eLife* **2016**, *5*, e15272.

(122) Ekins, S.; Perryman, A. L.; Horta Andrade, C. *PLoS Neg. Trop. Dis.* **2016**, *10*, ed. by Diemert, D. J., e0005023.

(123) Ekins, S.; Mietchen, D.; Coffee, M.; Stratton, T. P.; Freundlich, J. S.; Freitas-Junior, L.; Muratov, E.; Siqueira-Neto, J.; Williams, A. J.; Andrade, C. *F1000Res* **2016**, *5*, 150.

(124) Gupta, A. K.; Kaur, K.; Rajput, A.; Dhanda, S. K.; Sehgal, M.; Khan, M. S.; Monga, I.; Dar, S. A.; Singh, S.; Nagpal, G.; Usmani, S. S.; Thakur, A.; Kaur, G.; Sharma, S.; Bhardwaj, A.; Qureshi, A.; Raghava, G. P. S.; Kumar, M. *Sci. Rep.* **2016**, *6*, 32713.

(125) Cox, B. D.; Stanton, R. A.; Schinazi, R. F. *Antivir. Chem. Chemother.* **2015**, *24*, 118–126.

(126) Leung, D.; Schroder, K.; White, H.; Fang, N. X.; Stoermer, M. J.; Abbenante, G.; Martin, J. L.; Young, P. R.; Fairlie, D. P. *J. Biol. Chem.* **2001**, *276*, 45762–45771.

(127) Borowski, P.; Niebuhr, A.; Schmitz, H.; Hosmane, R. S.; Bretner, M.; Siwecka, M. A.; Kulikowski, T. *Acta Biochim. Pol.* **2002**, *49*, 597–614.

(128) Raney, K. D.; Sharma, S. D.; Moustafa, I. M.; Cameron, C. E. *J. Biol. Chem.* **2010**, *285*, 22725–31.

(129) Basavannacharya, C.; Vasudevan, S. G. *Biochem. Biophys. Res. Commun.* **2014**, *453*, 539–44.

(130) 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. *ACS Chem. Biol.* **2015**, *10*, 1887–1896.

(131) 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. *Antiviral Res* **2017**, *139*, 49–58.

(132) Munawar, A.; Beelen, S.; Munawar, A.; Lescrinier, E.; Strelkov, S.; Munawar, A.; Beelen, S.; Munawar, A.; Lescrinier, E.; Strelkov, S. V. *Int. J. Mol. Sci.* **2018**, *19*, 3664.

(133) Kinney, R. M.; Butrapet, S.; Chang, G. J. J.; Tsuchiya, K. R.; Roehrig, J. T.; Bhamarapravati, N.; Gubler, D. J. *Virology* **1997**, *230*, 300–308.

(134) Osorio, J. E.; Huang, C. Y. H.; Kinney, R. M.; Stinchcomb, D. T. *Vaccine* **2011**, *29*, 7251–7260.

(135) Swarbrick, C. M.; Basavannacharya, C.; Chan, K. W.; Chan, S. A.; Singh, D.; Wei, N.; Phoo, W. W.; Luo, D.; Lescar, J.; Vasudevan, S. G. *Nucleic Acids Res.* **2017**, *45*, 12904–12920.

(136) Jain, R.; Coloma, J.; Garcia-Sastre, A.; Aggarwal, A. K. *Nat. Struct. Mol. Biol.* **2016**, *23*, 752–754.

(137) Cao, X.; Li, Y.; Jin, X.; Li, Y.; Guo, F.; Jin, T. *Nucleic Acids Res.* **2016**, *44*, 10505–10514.

(138) Du Pont, K. E.; Davidson, R. B.; McCullagh, M.; Geiss, B. J. *J. Biol. Chem.* **2019**, jbc.RA119.011922.

(139) Gu, M.; Rice, C. M. *J. Biol. Chem.* **2016**, *291*, 14499–509.

(140) Tian, H.; Ji, X.; Yang, X.; Zhang, Z.; Lu, Z.; Yang, K.; Chen, C.; Zhao, Q.; Chi, H.; Mu, Z.; Xie, W.; Wang, Z.; Lou, H.; Yang, H.; Rao, Z. *Protein & Cell* **2016**, *7*, 562–570.

(141) Case, D. A.; Betz, R.; Cerutti, D.; Cheatham, T. I.; 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. AMBER16., San Francisco., 2016.

(142) Case, D. A.; Ben-Shalom, I.; Brozell, S.; Cerutti, D.; Cheatham, T. I.; Cruzeiro, V.; Darden, T.; Duke, R.; Ghoreishi, D.; Gilson, M.; Gohlke, H.; Goetz, A.; Greene, D.; Harris, R.; Homeyer, N.; Izadi, S.; Kovalenko, A.; Kurtzman, T.; Lee, T.; LeGrand, S.; Li, P.; Lin, C.; Liu, J.; Luchko, T.; Luo, R.; Mermelstein, D.; Merz, K.; Miao, Y.; Monard, G.; Nguyen, C.; Nguyen, H.; Omelyan, I.; Onufriev, A.; Pan, F.; Qi, R.; Roe, D.; Roitberg, A.; Sagui, C.; Schott-Verdugo, S.; Shen, J.; Simmerling, C.; Smith, J.; Salomon-Ferrer, R.; Swails, J.; Walker, R.; Wang, J.; Wei, H.; Wolf, R.; Wu, X.; Xiao, L.; York, D.; Kollman, P. AMBER18., San Francisco., 2018.

(143) Isralewitz, B.; Gao, M.; Schulten, K. *Curr. Opin. Struct. Biol.* **2001**, *11*, 224–230.

(144) Fukunishi, H.; Watanabe, O.; Takada, S. *J. Chem. Phys.* **2002**, *116*, 9058–9067.

(145) Thiede, E. H.; Van Koten, B.; Weare, J.; Dinner, A. R. *J. Chem. Phys.* **2016**, *145*, 084115.

(146) Tsai, C. J.; Ma, B; Nussinov, R *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 9970–9972.

(147) Boehr, D. D.; Nussinov, R.; Wright, P. E. *Nat. Chem. Biol.* **2009**, *5*, 789–796.

(148) Kumar, S.; Ma, B.; Tsai, C.-J.; Sinha, N.; Nussinov, R. *Protein Sci.* **2008**, *9*, 10–19.

(149) Pickett, B. E.; Sadat, E. L.; Zhang, Y.; Noronha, J. M.; Squires, R. B.; Hunt, V.; Liu, M.; Kumar, S.; Zaremba, S.; Gu, Z.; Zhou, L.; Larson, C. N.; Dietrich, J.; Klem, E. B.; Scheuermann, R. H. *Nucleic Acids Res.* **2012**, *40*, D593–D598.

(150) Cock, P. J.; Antao, T.; Chang, J. T.; Chapman, B. A.; Cox, C. J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B.; De Hoon, M. J. *Bioinformatics* **2009**, *25*, 1422–1423.

(151) Fu, L.; Niu, B.; Zhu, Z.; Wu, S.; Li, W. *Bioinformatics* **2012**, *28*, 3150–3152.

(152) Li, W.; Godzik, A. *Bioinformatics* **2006**, *22*, 1658–1659.

(153) Katoh, K.; Misawa, K.; Kuma, K.; Miyata, T. *Nucleic Acids Res.* **2002**, *30*, 3059–3066.

(154) Capella-Gutiérrez, S.; Silla-Martínez, J. M.; Gabaldón, T. *Bioinformatics* **2009**, *25*, 1972–1973.

(155) Crooks, G. E.; Hon, G.; Chandonia, J.-M.; Brenner, S. E. *Genome Res.* **2004**, *14*, 1188–1190.

(156) Schneider, T. D.; Stephens, R. M. *Nucleic Acids Res.* **1990**, *18*, 6097–6100.

(157) Gunasekaran, K; Ma, B.; Nussinov, R. *Proteins: Struct., Funct., Bioinf.* **2004**, *57*, 433–443.

(158) Canals, M.; Sexton, P. M.; Christopoulos, A. *Trends Biochem. Sci.* **2011**, *36*, 663–672.

(159) Kelley, J. A.; Knight, K. L. *J. Biol. Chem.* **1997**, *272*, 25778–25782.

(160) Vale, R. D.; Milligan, R. A. *Science* **2000**, *288*, 88–95.

(161) Hersch, G. L.; Burton, R. E.; Bolon, D. N.; Baker, T. A.; Sauer, R. T. *Cell* **2005**, *121*, 1017–1027.

(162) Bohr, C.; Hasselbalch, K; Krogh, A. *Skandinavisches Archiv Für Physiologie* **1904**, *16*, 402–412.

(163) Pauling, L. *Proc. Natl. Acad. Sci. U.S.A.* **1935**, *21*, 186–191.

(164) Nussinov, R. *Chem. Rev.* **2016**, *116*, 6263–6266.

(165) Liu, J.; Nussinov, R. *PLoS Comput. Biol.* **2016**, *12*, e1004966.

(166) Lu, S.; Li, S.; Zhang, J. *Med. Res. Rev.* **2014**, *34*, 1242–1285.

(167) Cortina, G. A.; Kasson, P. M. *Curr. Opin. Struct. Biol.* **2018**, *52*, 80–86.

(168) Tautermann, C. S.; Binder, F.; Büttner, F. H.; Eickmeier, C.; Fiegen, D.; Gross, U.; Grundl, M. A.; Heilker, R.; Hobson, S.; Hoerer, S.; Luippold, A.; Mack, V.; Montel, F.; Peters, S.; Bhattacharya, S.; Vaidehi, N.; Schnapp, G.; Thamm, S.; Zeeb, M. *J. Med. Chem.* **2019**, *62*, 306–316.

(169) An, X.; Lu, S.; Song, K.; Shen, Q.; Huang, M.; Yao, X.; Liu, H.; Zhang, J. *J. Chem. Inf. Model.* **2019**, *59*, 597–604.

(170) Buller, A. R.; van Roye, P.; Cahn, J. K. B.; Scheele, R. A.; Herger, M.; Arnold, F. H. *J. Am. Chem. Soc.* **2018**, *140*, 7256–7266.

(171) Konermann, L. *Nat. Struct. Mol. Biol.* **2016**, *23*, 511–512.

(172) Berry, L.; Poudel, S.; Tokmina-Lukaszewska, M.; Colman, D. R.; Nguyen, D. M. N.; Schut, G. J.; Adams, M. W. W.; Peters, J. W.; Boyd, E. S.; Bothner, B. *Biochim. Biophys. Acta Gen. Subj.* **2018**, *1862*, 9–17.

(173) Daily, M. D.; Gray, J. J. *Proteins: Struct., Funct., Bioinf.* **2007**, *67*, 385–399.

(174) Fan, G.; Baker, M. R.; Wang, Z.; Seryshev, A. B.; Ludtke, S. J.; Baker, M. L.; Serysheva, I. I. *Cell Res.* **2018**, *28*, 1158–1170.

(175) Walma, T.; Spronk, C. A. E. M.; Tessari, M.; Aelen, J.; Schepens, J.; Hendriks, W.; Vuister, G. W. *J. Mol. Biol.* **2002**, *316*, 1101–1110.

(176) Fuentes, E. J.; Gilmore, S. A.; Mauldin, R. V.; Lee, A. L. *J. Mol. Biol.* **2006**, *364*, 337–351.

(177) Gianni, S.; Walma, T.; Arcovito, A.; Calosci, N.; Bellelli, A.; Engström, A.; Travaglini-Allocatelli, C.; Brunori, M.; Jemth, P.; Vuister, G. W. *Structure* **2006**, *14*, 1801–1809.

(178) Manley, G.; Rivalta, I.; Loria, J. P. *J. Phys. Chem. B* **2013**, *117*, 3063–3073.

(179) Grutsch, S.; Brüschweiler, S.; Tollinger, M. *PLoS Comput. Biol.* **2016**, *12*, e1004620.

(180) O'Rourke, K. F.; Gorman, S. D.; Boehr, D. D. *Comput. Struct. Biotechnol. J.* **2016**, *14*, 245–251.

(181) Rivalta, I.; Lisi, G. P.; Snoeberger, N.-S.; Manley, G.; Loria, J. P.; Batista, V. S. *Biochemistry* **2016**, *55*, 6484–6494.

(182) Lisi, G. P.; East, K. W.; Batista, V. S.; Loria, J. P. *Proc. Natl. Acad. Sci. U.S.A.* **2017**, *114*, E3414–E3423.

(183) Xu, Y.; Zhang, D.; Rogawski, R.; Nimigean, C. M.; McDermott, A. E. *Proc. Natl. Acad. Sci. U.S.A.* **2019**, *116*, 2078–2085.

(184) Perutz, M. F. *Nature* **1970**, *228*, 726–739.

(185) Perutz, M. F.; Wilkinson, A. J.; Paoli, M; Dodson, G. G. *Annu. Rev. Biophys. Biomol. Struct.* **1998**, *27*, 1–34.

(186) Süel, G. M.; Lockless, S. W.; Wall, M. A.; Ranganathan, R. *Nat. Struct. Biol.* **2003**, *10*, 59–69.

(187) Rivalta, I.; Sultan, M. M.; Lee, N.-S.; Manley, G. A.; Loria, J. P.; Batista, V. S. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, E1428–36.

(188) Negre, C.; Morzan, U. N.; Hendrickson, H. P.; Pal, R; Lisi, G. P.; Loria, J. P.; Rivalta, I; Ho, J; Batista, V. S. *Proc. Natl. Acad. Sci. U.S.A.* **2018**, *115*, E11201–E11208.

(189) Dokholyan, N. V. *Chem. Rev.* **2016**, *116*, 6463–6487.

(190) Schueler-Furman, O.; Wodak, S. J. *Curr. Opin. Struct. Biol.* **2016**, *41*, 159–171.

(191) Stolzenberg, S.; Michino, M.; LeVine, M. V.; Weinstein, H.; Shi, L. *Biochim. Biophys. Acta* **2016**, *1858*, 1652–1662.

(192) Del Sol, A.; Fujihashi, H.; Amoros, D.; Nussinov, R. *Protein Sci.* **2006**, *15*, 2120–2128.

(193) Lockless, S. W.; Ranganathan, R *Science* **1999**, *286*, 295–299.

(194) Kaya, C.; Armutlulu, A.; Ekesan, S.; Haliloglu, T. *Nucleic Acids Res.* **2013**, *41*, W249–55.

(195) Karplus, M; Petsko, G. A. *Nature* **1990**, *347*, 631–639.

(196) Van Gunsteren, W. F. *Curr. Opin. Struct. Biol.* **1993**, *3*, 277–281.

(197) Karplus, M.; McCammon, J. A. *Nat. Struct. Biol.* **2002**, *9*, 646–652.

(198) Karplus, M. *Biopolymers* **2003**, *68*, 350–358.

(199) Karplus, M; Kuriyan, J *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6679–6685.

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

(201) Kong, Y.; Karplus, M. *Proteins: Struct., Funct., Bioinf.* **2009**, *74*, 145–154.

(202) Vijayabaskar, M. S.; Vishveshwara, S. *Biophys. J.* **2010**, *99*, 3704–3715.

(203) Ribeiro, A. A.; Ortiz, V. *J. Chem. Theory Comput.* **2014**, *10*, 1762–1769.

(204) Proctor, E. A.; Kota, P.; Aleksandrov, A. A.; He, L.; Riordan, J. R.; Dokholyan, N. V. *Chem. Sci.* **2015**, *6*, 1237–1246.

(205) Haliloglu, T.; Bahar, I. *Curr. Opin. Struct. Biol.* **2015**, *35*, 17–23.

(206) Ming, D.; Wall, M. E. *Phys. Rev. Lett.* **2005**, *95*, 337–4.

(207) Moritsugu, K.; Smith, J. C. *Biophys. J.* **2007**, *93*, 3460–3469.

(208) Moritsugu, K; Smith, J. C. *Comput. Phys. Commun.* **2009**.

(209) Lyman, E.; Pfaendtner, J.; Voth, G. A. *Biophys. J.* **2008**, *95*, 4183–4192.

(210) Li, M.; Zhang, J. Z. H.; Xia, F. *Chem. Phys. Lett.* **2015**, *618*, 102–107.

(211) Ikeguchi, M.; Ueno, J.; Sato, M.; Kidera, A. *Phys. Rev. Lett.* **2005**, *94*, 478–4.

(212) Besag, J. *J. R. Stat. Soc. Series B Stat. Methodol.* **1974**, *36*, 192–225.

(213) Friedman, J; Hastie, T; Tibshirani, R *Biostatistics* **2008**, *9*, 432–441.

(214) Brooks, B.; Karplus, M. *Proc. Natl. Acad. Sci. U.S.A.* **1983**, *80*, 6571–6575.

(215) Atilgan, A. R.; Durell, S. R.; Jernigan, R. L.; Demirel, M. C.; Keskin, O; Bahar, I *Biophys. J.* **2001**, *80*, 505–515.

(216) Zheng, W.; Doniach, S. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 13253–13258.

(217) Leioatts, N.; Romo, T. D.; Grossfield, A. *J. Chem. Theory Comput.* **2012**, *8*, 2424–2434.

(218) Bastolla, U. *WIREs Comput Mol Sci* **2014**, *4*, 488–503.

(219) Dijkstra, E. W. *Numer. Math.* **1959**, *1*, 269–271.

(220) Floyd, R. W. *Commun. ACM* **1962**, *5*, 345.

(221) Chaudhuri, B. N.; Lange, S. C.; Myers, R. S.; Chittur, S. V.; Davisson, V.; Smith, J. L. *Structure* **2001**, *9*, 987–997.

(222) Douangamath, A.; Walker, M.; Beismann-Driemeyer, S.; Vega-Fernandez, M. C.; Sterner, R.; Wilmanns, M. *Structure* **2002**, *10*, 185–193.

(223) Myers, R. S.; Jensen, J. R.; Deras, I. L.; Smith, J. L.; Davisson, V. J. *Biochemistry* **2003**, *42*, 7013–7022.

(224) Lisi, G. P. P.; Manley, G. A. A.; Hendrickson, H.; Rivalta, I.; Batista, V. S. S.; Loria, J. P. *Structure* **2016**, *24*, 1155–1166.

(225) Klem, T. J.; Chen, Y; Davisson, V. J. *J. Bacteriol.* **2001**, *183*, 989–996.

(226) Amaro, R. E.; Myers, R. S.; Davisson, V. J.; Luthey-Schulten, Z. A. *Biophysical Journal* **2005**, *89*, 475–487.

(227) Myers, R. S.; Amaro, R. E.; Luthey-Schulten, Z. A.; Davisson, V. J. *Biochemistry* **2005**, *44*, 11974–11985.

(228) Lipchock, J. M.; Loria, J. P. *Biomol. NMR Assign.* **2008**, *2*, 219–221.

(229) Lipchock, J.; Loria, J. P. *J. Biomol. NMR* **2009**, *45*, 73–84.

(230) Lipchock, J. M.; Loria, J. P. *Structure* **2010**, *18*, 1596–1607.

(231) Lisi, G. P.; Currier, A. A.; Loria, J. P. *Front. Mol. Biosci.* **2018**, *5*, 4.

(232) Amaro, R.; Tajkhorshid, E.; Luthey-Schulten, Z. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 7599–7604.

(233) Amaro, R.; Luthey-Schulten, Z. *Chem. Phys.* **2004**, *307*, 147–155.

(234) Amaro, R. E.; Sethi, A.; Myers, R. S.; Davisson, V. J.; Luthey-Schulten, Z. A. *Biochemistry* **2007**, *46*, 2156–2173.

(235) Lange, O. F.; Grubmüller, H. *Proteins: Struct., Funct., Bioinf.* **2006**, *62*, 1053–1061.

(236) Tesmer, J. J. G.; Klem, T. J.; Deras, M. L.; Davisson, V. J.; Smith, J. L. *Nat. Struct. Biol.* **1996**, *3*, 74–86.

(237) Zalkin, H.; Smith, J. L. *Adv. Enzymol. Relat. Areas Mol. Biol.* **1998**, *72*, 87–144.

(238) Rocks, J. W.; Pashine, N.; Bischofberger, I.; Goodrich, C. P.; Liu, A. J.; Nagel, S. R. *Proc. Natl. Acad. Sci. U.S.A.* **2017**, *114*, 2520–2525.

(239) Rocks, J. W.; Ronellenfitsch, H.; Liu, A. J.; Nagel, S. R.; Katifori, E. *Proc. Natl. Acad. Sci. U.S.A.* **2019**, *116*, 2506–2511.