UNIVERSITY OF CALIFORNIA, SAN DIEGO

Numerical and Computational Solutions for Biochemical Kinetics, Druggability, and Simulation

A dissertation submitted in partial satisfaction of the requirements for the degree of Doctor of Philosophy

in

Chemistry with a Specialization in Computational Science

by

Lane William Votapka

Committee in charge:

Professor Rommie E. Amaro, Chair
Professor Partho Ghosh
Professor J. Andrew McCammon
Professor Michael Norman
Professor Navtej Toor
Professor John Weare

2016

The Dissertation of Lane William Votapka is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

_____
Chair

University of California, San Diego

2016

DEDICATION

To my parents

EPIGRAPH

The fundamental laws necessary for the mathematical treatment of a large part of physics and the whole of chemistry are thus completely known, and the difficulty lies only in the fact that application of these laws leads to equations that are too complex to be solved.

Paul Dirac, 1929

Computers are famous for being able to do complicated things starting from simple programs.

Seth Lloyd, 2002

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

BD, Brownian dynamics

CADD, computer-aided drug discovery

CS, consensus site

DelEE, Delphi ensemble electrostatics

FLOP, floating point operation

GUI, graphical user interface

HA, hemagglutinin

MD, molecular dynamics

MFPT, mean first passage time

MSCS, multiple solvent crystal structures

MSM, Markov state model

NA, neuraminidase

NAMD, Nanoscale Molecular Dynamics

ODE, ordinary differential equation

PDB, protein data bank

PDE, partial differential equation

PKA, protein kinase A

POVME, Pocket Volume Measurer

QSAR, quantitiative structure-activity relationships

REL1, RNA editing ligase 1

RMSD, root mean squared deviation

SOD, superoxide dismutase

TnC, troponin C

VMD, Visual Molecular Dynamics

WISP, Weighted Implementation of Suboptimal Paths

LIST OF SUPPLEMENTAL FILES

Movie 1: Markov Models.mp4

Movie 2: Milestoning.mp4

LIST OF FIGURES

LIST OF TABLES

PREFACE

As Prof. Dirac said, the equations that govern all of chemistry are entirely known. Take for example the equation that bears his name. The Dirac equation provides one of the most complete descriptions of atomic/molecular realism available to us today. Except for situations of very close spatial or temporal proximity to a singularity or a deep gravity well, it accurately accounts for the quantum and relativistic phenomena in all but the most extreme physicochemical circumstances.

For all situations in our experience with any biological relevance such as the binding of two proteins or the diffusion of ions through a cell, the Dirac equation would yield a Herculean mathematical task: a challenge of mammoth proportions. A scientist would likely doom the prospect by pronouncing the statement: "this problem is hard".

Fortunately, for virtually all biomolecular applications, the Dirac equation is overkill. We may choose to disregard relativistic effects by resorting to the Schrödinger equation. We may further simplify the problem by approximating the quantum aspects of a chemical system using classical descriptions, implicit solvents, overdamped diffusion, ODEs, PDEs, Markov models, etc. Many levels of abstraction with countless applications of physical theories can be combined into a scientist's toolbox that we can use to examine, describe, model, simulate, approximate, emulate, and manipulate the universe around us.

During the Manhattan project, the mathematical genius John Von Neumann was responsible for predicting the hydrodynamics of the explosion immediately after

the initiation of a nuclear chain reaction. This calculation would allow them to predict whether a given bomb design would achieve the desired explosive yield, or would simply fizzle away; scattering the fissile material before it could participate in the reaction. Because of the difficulty of the hydrodynamic equations, Von Neumann employed countless ingenious approximations, linearizations, and assumptions in order to solve the problem. Later, when faced with the design of the larger, hotter, and more devastating H-bomb, Von Neumann and the others decided that the only tractable method to perform these computations was to take a numerical approach: timestep by miniscule timestep. Since computing machines had not yet been developed for routine scientific calculation, the scientists and mathematicians proceeded to compute the numerical solution by hand; spending many months and piles of paper to determine the outcome of the explosion. I would not be surprised if a computation of even superior accuracy could be accomplished in less than 24 hours with the laptop I'm using to type this document. In the year 1947, they finally invented the ENIAC supercomputer (it had sufficient speed to compute ~5000 simple additions/subtractions per second, or ~400 multiplications per second) to remove this agonizing burden from the scientists.

The computer is a wonderful (if sometimes infuriating) compatriot in scientific enterprise. Of course, a flippant suggestion to "use the computer" to solve our problems is more easily said than done. But a computational approach lowers the difficulty of a vast number of problems from practically impossible to attainable. Also, the continuous increase of computational power in recent decades, as well as advances in software and numerical techniques, is raising the computer's usefulness

all the time. What once required simplifications, menial hand calculations, or genius in order to approach the problem, can now often be done routinely by the computer, and can even be automated for use by the non-specialist.

So in order to accompany the countless theoretical tools, we also develop countless lines of code: software to remove the burden of meniality from the scientist; to unflinchingly and unquestioningly shovel through endless lists of arithmetic computations and through the branches of tortuous condition trees and algorithm workflows. But what makes the computer advantageous can also make it ruinous. It knows only what we tell it: it does *not* know "what we mean". Thus, if there is even a hint of obscurity or ambiguity in its instructions, it will either grind its program to a screeching halt, or worse, it will blunder off in some random direction, wasting time and resources, providing all the wrong answers. Worst of all, we may not even immediately realize that the answers are wrong!

Therefore, the proper use of a computer for science requires carefully designed, carefully coded software. Ideally, Each scientific program should be lovingly composed, complete with comments and unit-tests, all while keeping the satisfaction of the end-user in mind.

Sadly, such careful design is a pipe dream at this time. With the exception of a few well-written programs, much of our available scientific software is incomplete, thrown together, prone to errors, difficult to use, or all of the above. Most of the scientific software code that I have personally encountered is almost completely bereft of comments. Code accuracy, readability, and extendability are very important issues, and our scientific community is charged with a solemn task to improve this

situation. I must admit that I've not always expended the extra time and effort needed to craft my code to the proper level of perfection, so I'm really in no place to criticize. Perhaps our negligence is due to the fact that perfection is not our only focus: new techniques and theories are developing all the time, with huge numbers of programs that *need* to be written; and quickly. Therefore, a computational scientist is left with a dilemma of goals: quality versus quantity.

I was once told that science functioned much like a renaissance-era crafters' guild. Instead of a cabinetmaker who builds furniture or a goldsmith who fashions jewelry, we are the scientists who build theory and fashion experiments. Even the professor-postdoc-graduate student hierarchy resembles the master-journeyman-apprentice arrangement of antiquity. Examining some of the exquisite items made during the Baroque period in history indicates that many craftspeople of that age took pride in their work by building beautiful items intended to last and bring delight to the many generations that would follow.

But the industrial revolution and rising population has induced a high demand for large numbers of cheap, disposable items. This is both a blessing and a curse. We can buy usable furniture, tools, instruments, and other items cheaply. Unfortunately, our society has transformed into a factory-reliant throwaway culture that often disregards and devalues the creation of items made with skill and care. Most of our furniture and tools are not likely to last even the time span of our own lives. Perhaps similarly, our scientific code is made hastily, cheaply, and quickly, and is therefore largely useless beyond its immediate application. While this may be appropriate in some instances, there is no question that a bit of extra time and effort spent by a

primary developer can save later developers a proportionally larger amount of time if the former's code is being revisited. This extra effort could be as simple as code comments and intuitive variable names in a script, or as complex as coding language wrappers in a behemoth scientific software package like VMD or NAMD. I believe that the developers of both VMD and NAMD are reaping the rewards of widespread community use precisely because of the extensibility of their software. Their TCL interfaces allow for the potential of enhancements and other contributions by the community: several examples of which are presented in this dissertation. But there is always room for improvement. (For instance, at the time of this writing, the deepest levels of the NAMD code itself suffer a dreadful lack of comments and intuitive variable naming.)

Such is the situation in which we find ourselves. Perhaps I ought to start a good trend by applying these criticisms and recommendations to my own code and research: "Physician, heal thyself." I am optimistic that if these issues are widely addressed such that new approaches to solving problems are constantly innovated and produced, and such that the power of computation is brought rapidly and effectively to the entire biomedical community, we may see significant biomedical accomplishments in the near future, perhaps even going so far as the eradication of most human diseases in our lifetime.

Although we are not as centralized or focused as the Manhattan project was (or the Space Program, or the Human Genome project, or other great scientific initiatives of the past), they had no more daunting, and no more crucial of a task than that which faces the biomolecular research community today.

ACKNOWLEDGEMENTS

Receiving a Ph.D. in a field as interesting and challenging as our own is a great honor, and in my case, the fulfillment of a dream. I am deeply grateful and indebted to those who gave me the ability, by their various means, to realize this dream, and so I feel that I must defer the honor to them. This dissertation is the culmination of a glorious experience, and doubtlessly a gateway to many others.

First, I must thank my advisor, Rommie Amaro, who took a risk and gave me a chance to work in her lab at the very beginning; which has allowed us the opportunity to form a highly effective team, and a professional relationship of the highest caliber. I greatly admire Rommie's scientific and leadership abilities, and she has inspired me to reach for the loftiest goals. I am also grateful to my colleagues in the Amaro lab, as well as collaborators on the UCSD campus, who are perhaps too numerous to mention them all by name, nevertheless, my gratitude is no less abundant. I must also thank Prof. J. Andrew McCammon and others in the McCammon lab who, despite their own important projects and busy schedules, always came to my assistance when the need arose.

Chapter 2, in full, is a reprint of "Mechanism of 150-Cavity Formation in Influenza Neuraminidase", which was published in 2011 in Nature Communications, volume 2, page 388, by Rommie E. Amaro, Robert V. Swift, Lane W. Votapka, Wilfred W. Li and Robin M. Bush. The dissertation author was the third investigator and author of this paper.

Chapter 3, in full, is a reprint of "DelEnsembleElec: Computing Ensemble-Averaged Electrostatics Using DelPhi", which was published in 2013 in Communications in Computational Physics, volume 13, issue 1, pages 256-268, by Lane W. Votapka, Luke Czapla, Maxim Zhenirovskyy, and Rommie E. Amaro. The dissertation author was the primary investigator and author of this paper.

Chapter 4, in full, is a reprint of "Multistructural Hot Spot Characterization with FTProd", which was published in 2013 in Oxford Bioinformatics, volume 29, issue 3, pages 393-394, by Lane W. Votapka and Rommie E. Amaro. The dissertation author was the primary investigator and author of this paper.

Chapter 5, in full, is a reprint of "Variable Ligand- and Receptor-Binding Hot Spots in Key Strains of Influenza Neuraminidase", which was published in 2012 in the Journal of Molecular and Genetic Medicine, volume 6, page 293, by Lane W. Votapka, Özlem Demir, Robert V. Swift, Ross C. Walker, and Rommie E. Amaro. The dissertation author was the primary investigator and first author of this paper.

Chapter 6, in full, is a reprint of "Weighted Implementation of Suboptimal Paths (WISP): an Optimized Algorithm and Tool for Dynamical Network Analysis", which was published in 2014 in the Journal of Chemical Theory and Computation, volume 10, issue 2, pages 511-517, by Adam T. Van Wart, Jacob D. Durrant, Lane W. Votapka and Rommie E. Amaro. The dissertation author was the third investigator and author of this paper.

Chapter 7, in full, is a reprint of "POVME 2.0: An Enhanced Tool for Determining Pocket Shape and Volume Characteristics", which was published in 2014 in the Journal of Chemical Theory and Computation, volume 10, issue 11, pages

5047-5056, by Jacob D. Durrant, Lane W. Votapka, Jesper Sørensen, and Rommie E. Amaro. The dissertation author was the secondary investigator and author of this paper.

Chapter 8, in full, is a reprint of "Multiscale Estimation of Binding Kinetics Using Brownian Dynamics, Molecular Dynamics, and Milestoning", which was published in 2015 in PLOS Computational Biology, volume 11, issue 10, by Lane W. Votapka and Rommie E. Amaro. The dissertation author was the primary investigator and author of this paper.

Chapter 9, in full, is a reprint of "Bridging Scales Through Multiscale Modeling: A Case Study on Protein Kinase A", which was published in 2015 in Frontiers in Physiology, volume 6, by Britton W. Boras, Sophie P. Hirakis, Lane W. Votapka, Robert D. Malmstrom, and Rommie E. Amaro. The dissertation author was the third investigator and author of this paper.

Outside the academic sphere, I owe an enormous debt of love and gratitude to my supportive family, particularly to my parents: Gary and Susan Votapka. During the entire span of my life, they have been nurturing and supportive of me as any parent could ever be: constantly affirming, coaching, and instructing me in all matters. I also extend a great deal of gratitude to my sister Katie, my grandmother Betty, and other members of my family for all their love, help, and support. I am also grateful to the countless teachers and professors who have prepared me for this honor.

Finally, and most of all, I thank God, who through His power, authored nature, science, salvation and all that is beautiful in the universe. If I am anything good, it is because of Him.

VITA

2009 Bachelor of Science, **Point Loma Nazarene University**

2013 Master of Science **University of California, San Diego**

2016 Doctor of Philosophy **University of California, San Diego**

PUBLICATIONS

Votapka, LW and Amaro R.E., "Multiscale Estimation of Binding Kinetics Using Brownian Dynamics, Molecular Dynamics and Milestoning" *PLOS Computational Biology*, (2015)

Boras BW, Hirakis SP **Votapka LW**, Malmstrom RD, Amaro RE, McCulloch AD "Bridging Scales Through Multiscale Modeling: A Case Study on Protein Kinase A". *Frontiers in Physiology*, (2015)

Durrant JD, **Votapka LW**, Sørensen J, Amaro RE "POVME 2.0: An Enhanced Tool for Determining Pocket Shape and Volume Characteristics". *Journal of Chemical Theory and Computation*, (2014)

Van Wart AT, Durrant JD, **Votapka LW**, Amaro RE "Weighted Implementation of Suboptimal Paths (WISP): An optimized algorithm and tool for dynamical network analysis". *Journal of Chemical Theory and Computation*, (2014)

Votapka LW, Amaro RE "Multistructural Hot Spot Characterization with FTProd". *BMC Bioinformatics*, doi: 10.1093/bioinformatics/bts689. (2012)

Votapka LW, Demir Ö, Swift RV, Walker RC, Amaro RE "Variable ligand-and receptor-binding hot spots in key strains of influenza neuraminidase". *Journal of Molecular and Genetic Medicine* (2012)

Votapka LW; Czapla L; Zhenirovskyy M; Amaro RE "DelEnsembleElec: Computing ensemble-averaged electrostatics using DelPhi". *Communications in Computational Physics.* (2012)

Amaro RE; Swift RV; **Votapka LW**; Li WW; Walker RC; Bush R "Mechanism of 150-Cavity Formation in Influenza Neuraminidase". *Nature Communications*. 2: 388 doi 10.1038/ncomss1390 (2011)

FIELDS OF STUDY

Major Field: Chemistry and Biochemistry

The use and development of computer simulation, research programming, and scientific software for biomedical and biophysical applications.
Professor Rommie E. Amaro

ABSTRACT OF THE DISSERTATION

Numerical and Computational Solutions for Biochemical Kinetics, Druggability, and Simulation

by

Lane William Votapka

Doctor of Philosophy in Chemistry with a Specialization in Computational Science

University of California, San Diego, 2016

Professor Rommie E. Amaro

Computational tools provide the automation and power that enable detailed modeling and analysis of many biomolecular phenomena of interest. Open source programs and automated tools empower researchers and provide opportunities for improvement to existing software. In the past few years, I have developed several open-source scientific software packages for the purposes of automating difficult or menial tasks pertaining to computational biophysics. These software packages

involve the analysis of molecular dynamics simulations, Brownian dynamics simulations, electrostatics, pocket volume measurement, solvent fragment mapping, binding site characterization, milestoning theory, and allosteric network communications. In addition to allowing my research group and me to approach biomedical challenges that would otherwise be intractable, I hope and intend that these tools will be useful to the computational and theoretical biophysics research community.

Chapter 1: Introduction to the Dissertation - Overview of Methods and Chapter Content

This dissertation outlines the research I performed to produce and use scientific software and computational tools for the purposes of biomedical and biophysical inquiry. This software was designed with the intent to be reusable and extendable by other scientists, and is therefore an investment on behalf of the scientific community: intended to save time and also to extend research capabilities.

1.1 Overview of methods

Among many computational and theoretical resources I used for my projects, several major components merit brief description: molecular dynamics (MD), Brownian dynamics(BD), milestoning, electrostatics using the Poisson-Boltmann (PB) equation and Debye-Huckel theory, and computational solvent fragment mapping.

1.1.1 Molecular Dynamics

MD approximates the dynamics of a molecular system classically, and therefore can be defined as a numerical solution to Newton's equations of motion where the atoms are modeled as simple point particles. Given a carefully pre-

pared starting structure of atomic positions, atomic velocities, and position-dependent potential energy interaction functions between the atoms, the dynamics of a chemical system can be advanced in time by small increments. The longer-time trajectories of these simulations can then be examined to observe statistical mechanical quantities or interesting phenomena. We typically use MD to analyze the motions of proteins and other biomolecules. These trajectories are like a movie: showing the atoms of the biomolecule twisting and shifting. If a picture is worth a thousand words, then a movie is worth a thousand pictures. Unlike a static structure, the trajectory shows how the molecules would move: where they go, how they function, whether they deform or not. MD was used extensively in all my projects, and more formal, detailed information about MD exists in subsequent chapters 2, 3, 4, 5, 6, 7, 8 and 9.

1.1.2 Brownian Dynamics

In contrast, BD is a numerical method to simulate overdamped Langevin dynamics and can be described as a higher level of simplification and abstraction from MD. The BD simulation still includes all the atoms of the biomolecules and their substrates modeled as point particles, but they are rigidly constrained to drift and tumble with their constituent molecules: therefore, each molecule is a rigid body. Also, the water molecules and dissolved ions are replaced with a continuum: a field with charge, dielectric, and hydrodynamic properties intended to approximate those of an aqueous solution. If one is willing to accept the simplifying assumptions that BD introduces, as well as the potential inaccuracies, a relatively large range of

temporal and spatial scales are made available to the researcher for investigation into molecular activities. For instance, at the time of this writing, a MD trajectory that observes even a single binding event requires substantial time and cyberinfrastructure investment for a typical ligand-receptor system. In contrast, using BD, millions of binding events of a typical system can be observed within 24 hours on a multicore desktop computer. By counting the number of binding events versus the number of escape events, a probability of binding can be estimated. This can be used directly to compute a rate constant of binding, which can be compared to experiment or used to predict experimentally immeasurable rate constants of binding. I used BD in several of my major projects and more detailed information about BD can be found in chapters 8, 9 and 10.

1.1.3 Milestoning

Milestoning theory is very similar to Markov model theory: it computationally models kinetics and thermodynamics of processes, breaking a long process into multiple shorter ones, each of which is independent of the others, allowing for extensive parallelizability. Each of these independent trajectories generate statistics within a transition matrix and also an incubation time vector. These statistics are then propagated to compute the quantities of interest. There are a number of key differences between Markov models and milestoning. Many papers that provide extensive comparisons of the two methods[1-4]. In particular, if one has a system with a trajectory that traverses a space, the state that the system is in will be defined

differently for Markov models versus milestoning. In a Markov model, the state of the system is defined by the region or volume in which it can be found: a region within these surfaces. In milestoning, however, the state of the system is defined by the surface that the system last crossed. My project involved the use of milestoning to combine the trajectory results of MD and BD simulations to more accurately and efficiently compute rate constants of binding. Details of this project and of milestoning theory can be found in chapters 8 and 9.

1.1.4 The Poisson Boltzmann Equation and Debye-Huckel Theory

The Poisson-Boltzmann (PB) equation and the closely associated Debye-Huckel theory was discovered around the turn of the last century. As the name suggests, the equation is a combination of both Poisson's equation and a Boltzmann distribution. When solved for a particular physical system of a constellation of charges dissolved in a solution of electrolytes, the PB equation gives an approximation of the potential of mean force as a contribution by those electrolytes. The PB equation is essential for preparing a system BD simulations. I also generated a program to compute the ensemble-averaged electrostatics for a trajectory resulting from a MD simulation. This project, as well as additional detail about the PB equation and Debye-Huckel theory is outlined in chapter 3.

1.1.5 Computational Solvent Fragment Mapping

Computational solvent fragment mapping makes use of software to approximate true solvent fragment mapping: an experimental process designed to identify druggable "hot spots" on the surface of a biomolecule. Experimental solvent fragment mapping can be a difficult and expensive process: it involves dissolving the biomolecule of interest into a small-molecule solvent, and then crystallizing the biomolecule for X-ray structure determination. Computation solvent fragment mapping imitates this process by using a molecular mechanics forcefield such as CHARMM along with a docking protocol to predict the location of the hot spots by means of these small molecular fragments. The popular FTMAP server was developed to make computational solvent fragment mapping available to the wider scientific community. I developed a program to compute, combine, and analyze solvent fragment mapping across multiple structures to investigate dynamic information about the hot spots' evolving characteristics and transience. Details of the development and usage of the multistructural hot spot software is outlined in chapters 4 and 5.

1.2 Overview of Chapter Contents

Along with numerous individual scripts and programs intended to be useful for specific applications, I led or participated in the development of, either alone or in a group, five standalone programs, plugins, or script packages including: Delphi Ensemble Electrostatics (DelEE), FTProd, WISP, POVME2, and SEEKR. In the

subsequent chapters, I describe the details of the design, development, and practical use of each of these.

Chapter 2 describes the use of custom scripts to examine the variability of the volume and shape of the 150-pocket in influenza neuraminidase (NA). These scripts in turn used the pocket volume measuring program POVME 1.0, which was further improved in the work of Chapter 7 with the development of POVME 2.0.

Chapter 3 concerns the development of DelEE: a plugin for the molecular visualization program Visual Molecular Dynamics (VMD) that computes the ensemble-averaged electrostatics of a collection of structures. The electrostatics are computed using the DelPhi program: a Poisson-Boltzmann equation solver. The chapter includes a brief overview of Poisson-Boltzmann theory as well as justification for the usefulness of ensemble-averaged electrostatics. The plugin itself provides complete control over the calculation using a convenient graphical user interface (GUI).

Chapter 4 outlines the development of FTProd: another VMD plugin that receives as input, the output of FTMAP: a program that performs computational solvent fragment mapping. FTProd integrates the druggable hot spots and consensus sites that FTMAP determines across multiple structures, then clusters and characterized them according to user input, giving qualitative information concerning the changing identity, character, and transience of potentially important druggable sites on the surface of biomolecules.

A practical application of the FTProd tool has been included in chapter 5; where we examined the binding sites and druggable hot spots across several strains of influenza NA. In so doing, we determined potential interaction sites that may participate in substrate or inhibitor binding, the interface between monomers, or association with its counterpart glycoprotein hemagglutinin.

Chapter 6 describes the WISP project, a third VMD plugin designed to determine pathways of communication through the protein between allosteric sites by means of correlated motion.

Chapter 7 includes information about substantial improvements to the pocket volume measurement program: POVME. These improvements resulted in its subsequent version: POVME 2.0.

Chapter 8 is an account of research combining BD simulations with MD simulations using the theory of milestoning for the purposes of predicting rate constants of binding. As a pioneering study of the effectiveness of this method, benchmarking was performed on two spherically symmetric “toy” systems, as well as two biomedically relevant “real” systems with simple ligands: calcium ion binding to troponin C and superoxide anion binding to superoxide dismutase. Although yet unreleased at the time of this writing to the scientific community, this project made first use of the SEEKR software; a suite of scripts, tools, and programs to prepare, run, and analyze MD simulations, BD simulations, and milestoning calculations for the purposes of binding rate constant calculations.

Chapter 9 includes a review of the use of Markov models and milestoning for investigation into the dynamics of protein kinase A (PKA) using both MD and BD.

Chapter 2: Mechanism of 150-cavity Formation in Influenza Neuraminidase

The recently discovered 150-cavity in the active site of group-1 influenza A neuraminidase (NA) proteins provides a target for rational structure-based drug development to counter the increasing frequency of antiviral resistance in influenza. Surprisingly, the 2009 H1N1 pandemic virus (09N1) was crystalized without the 150-cavity characteristic of group-1 NAs. Here we demonstrate, through a total sum of 1.6 μs of biophysical simulations, that 09N1 NA exists in solution preferentially with an open 150-cavity. Comparison with simulations using avian N1, human N2, and 09N1 with a I149V mutation and an extensive bioinformatics analysis suggests that the conservation of a key salt bridge is a crucial mechanism in the stabilization of the 150-cavity across both subtypes. This result provides an atomic-level structural understanding of the recent finding that antiviral compounds designed to take advantage of contacts in the 150-cavity can inactivate both 2009 H1N1 pandemic and avian H5N1 viruses.

2.1 Introduction

Understanding the structural dynamics of the influenza glycoproteins has been a long-standing goal due to their direct impact on public health. The two major influenza glycoproteins, hemagglutinin (HA) and neuraminidase (NA), control entry and exit of the viral particles from the host cell, respectively. HA binds to sialic acid surface receptors on the host cell, whereas NA cleaves the terminal sialic acid

receptor linkage, facilitating viral shedding. The nine NA alleles have been dividedinto two groups based on phylogenetic analysis (group-1: N1, N4, N5, N8; group-2: N2, N3, N6, N7, N9)[5]. During the last century, influenza viruses carrying N1 (H1N1) or N2 (H2N2, H3N2) alleles have circulated in humans, first as pandemic strains and then, after subsequent adaptation to humans, as seasonal epidemic strains. Thus, a better understanding of the structural dynamics of N1 and N2 is particularly relevant for antiviral design.

Oseltamivir (Tamiflu) and zanamivir (Relenza), which target the NA, are currently the only antivirals approved by the FDA for the prophylaxis and treatment of influenza. These drugs, developed against available group-2 NA structures, represent some of the first successful rational structure-based drug development efforts[6]. The crystal structures of group-1 NAs revealed a never-before-seen 150-cavity adjacent to the sialic acid binding site[5]. It has been hypothesized, and very recently shown[7], that targeting the 150-cavity may allow the development of new antivirals with increased specificity and potency against group-1 enzymes. The increasing frequency of oseltamivir resistance in pre-2009 seasonal H1N1 viruses[8] and the occasional observation of oseltamivir-resistance among 2009 H1N1 pandemic viruses motivates new antiviral development. Having additional antivirals in our treatment arsenal would be advantageous, and potentially critical, if a highly virulent strain, *e.g.* H5N1, evolved the ability to undergo rapid transmission among humans, or if the already highly transmissible 2009 H1N1 pandemic virus was to evolve resistance to existing antiviral drugs.

Recently, it was revealed that the structure of the 2009 pandemic H1N1 NA lacked a 150-cavity, despite being a group-1 NA[9]. This surprising finding suggested that the 2009 pandemic N1 protein was structurally more similar to the group-2 NAs than to the group-1 NAs. Based on alignments of sequences representing all available NA crystal structures, highly conserved residues in the 150-loop and the 430-loop (residues 147-152, and 429-433, respectively, in N2 numbering) were hypothesized to functionally determine the structure of the 150-cavity[9]. In particular, I149 was found to be common between the 2009 pandemic N1 and group-2 NAs, whereas V149 was conserved among the other group-1 NAs. In the two solved N2 structures, which have somewhat atypical sequences, a salt bridge between D147 and H150 appeared to prevent the opening of the 150-loop, despite the presence of V149.

Here we test the hypothesis that position 149 is critical for determining the open or closed status of the 150-cavity. Our alternative hypothesis is that cavity status is plastic in the absence of a D147-H150 salt bridge, being dependent on loop conformations that are themselves flexible. Earlier computational studies of N1 from avian H5N1 showed that this isolate exhibited remarkable flexibility in the 150-loop[10]. The same avian N1 was also reported to contain a closed 150-loop under certain crystallization conditions[5] and additionally shown to be able to switch to a closed loop position during a MD simulation initiated from the co-crystallized oseltamivir-bound open-150-loop configuration[10]. The understanding that emerged was that the avian N1 was able to adopt a wide range of configurations in the 150-loop region, favoring an open conformation of the 150-cavity overall.

We examined the flexibility of the 150-cavity area in the 09N1 crystal structure through molecular dynamics simulations using 09N1 and other available structures of N1 and N2 alleles derived from human clinical isolates. In combination with the simulations, an extensive bioinformatics analysis for these alleles in the 150- and 430-loop regions offers new clues as to the controlling features of 150-cavity formation in these critical enzymes. Ultimately, we find that a key salt bridge appears to control the 150-cavity formation in both group-1 and group-2 enzymes, both of which are able to adopt flexible loop conformations in this critical region. We propose that this new structural understanding can be related to antiviral design for any of the influenza neuraminidase enzymes.

2.2 Results

2.2.1 Molecular Dynamics Simulations

To probe the effect of sequence on the atomic-level structure and dynamics of these critical enzymes, we performed four separate 100 ns molecular dynamics simulations for four tetrameric NA enzymes: 1) A/California/04/2009, an H1N1 virus isolated early in the 2009 pandemic (09N1, PDB: 3NSS)[11]. We note that the N1 allele in the pandemic strain had recently evolved from an Eurasian-lineage H1N1 swine virus[12]. 2) A mutant N1 that we engineered *in silico* from A/California/04/2009 by substituting Val for Ile at position 149 (09N1_I149V). 3) A/Vietnam/1203/04, an avian-derived H5N1 virus isolated from a human (VN04N1, PDB: 2HTY)[5]. 4) A/Tokyo/3/67, a seasonal human H2N2 virus (N2, PDB: 1NN2) [13]. We note that the

I149V mutation in A/Tokyo/3/67 is atypical for a human N2 allele (Table 2.1, Table 2.3).

Table 2.1: Variation in the 150- and 430-loops of N1 and N2 NA alleles of avian, swine and human influenza viruses

Consensus sequences contain amino acids at a frequency of at least 80%; major polymorphisms are indicated in white boxes, other boxes colored by residue. Sequences in bold font correspond to structures simulated in this paper.

host	N1 and N2 alleles	N	150-loop			430 loop		
			147	149	150	430	431	432
swine N1	N1 swine H1N1 classic lineage consensus	158	G	V	K	Q	P	K
	N1 swine H1N1 Eurasian lineage consensus	165	G	I/V	K	R	P	K
human N1	**N1 human H1N1 2009 pandemic consensus & A/California/04/2009**	**1806**	**G**	**I**	**K**	**R**	**P**	**K**
	N1 human H1N1 seasonal 2007-2009 consensus	1809	G	V	K	L	P	R
	N1 human H1N1 seasonal 1950-2007 consensus		G	V	K	R	P	R
	N1 human H1N1 seasonal 1930-40s consensus		G	V	K	R	P	K
	N1 human H1N1 1918 pandemic		G	V	K	Q	P	K
avian N1	**N1 avian consensus & A/Vietnam/1203/2004**	**2141**	**G**	**V**	**K**	**R**	**P**	**K**
human N2	N2 human H3N2 seasonal mid-2000s-present consensus	1727	N/D	V	R	R	K	E
	N2 human H3N2 seasonal 1990-mid 2000s consensus		D	V	H	R	K	Q/E
	N2 human H3N2 seasonal 1970-80s consensus		D	I	H	R	E	Q
	N2 human H3N2 1968 pandemic (NA of human H2N2 origin)		D	I	H	R	K	Q
	N2 human H2N2 seasonal A/Tokyo/3/1967 (atypical 149V)	**88**	**D**	**V**	**H**	**R**	**K**	**Q**
	N2 human H2N2 seasonal 1960s consensus		D	I	H	R	Q/K	Q
	N2 human H2N2 1957 pandemic (NA of avian origin)		G	I	H	R	P	Q
avian N2	N2 avian polymorphisms seen since the mid-1990s	1743	G	T/A/S	H	R	P	K/Q
	N2 avian consensus		G	I	H	R	P	Q

The homotetramer configuration of NA allows us to take advantage of multi-copy simulation sampling[14], amounting to the equivalent of nearly half a microsecond (400 ns) of sampling for each neuraminidase monomer, while accounting for realistic neighboring subunit effects within the structural dynamics. Alpha-carbon root mean square deviation (RMSD) plots for the tetramer systems and individual monomer chains exhibit stability over 100 ns, and there is good agreement between experimental and simulation-derived B-factors (Figures 2.4-2.7).

2.2.2 Pandemic 2009 H1N1 exhibits open 150-cavity

Our simulations reveal that the pandemic 09N1 NA is able to adopt open 150-cavity conformations in normal solution dynamics, and, in contrast to the crystal structure, it appears to favor an open 150-cavity conformation overall (Figure 2.1,

Table 2.1). In the simulations of 09N1, the 150-loop transitions to an open configuration by 50 ns in all chains of the tetramer (Figure 2.7). As a reference for open- and closed- loop structures, the PDBs 2HTY and 2HU4 were utilized, respectively. The open 150-cavity crystal structure (2HTY with hydrogen atoms added) exhibits a 150-cavity volume of 36 $Å^3$ as computed by POVME[15] (Table 2.5). Closed 150-cavity crystal structures (1NN2, 2HU4, 3NSS with hydrogen atoms added) were used as references and uniformly exhibit a volume of 0 $Å^3$. To quantify the extent to which structures within the dynamical ensemble adopt either a closed or open 150-cavity conformation, a time series of the pocket volume was computed over the course of the trajectory (Figure 2.2A, Figure 2.8). Structures were subsequently classified as open or closed based on 150-cavity volume, i.e. cavities with volumes greater than or equal to 18 $Å^3$, or at least half of the crystal structure open cavity volume, are considered "open." Through this method, we determined that the 09N1 system adopts an open 150-cavity during the majority of the simulation (60.8%, Table 2.2). We note that longer simulation times may further increase the percentage of 09N1 in the open conformation, overcoming the structural bias due to the simulation being initiated with a closed 150-cavity.

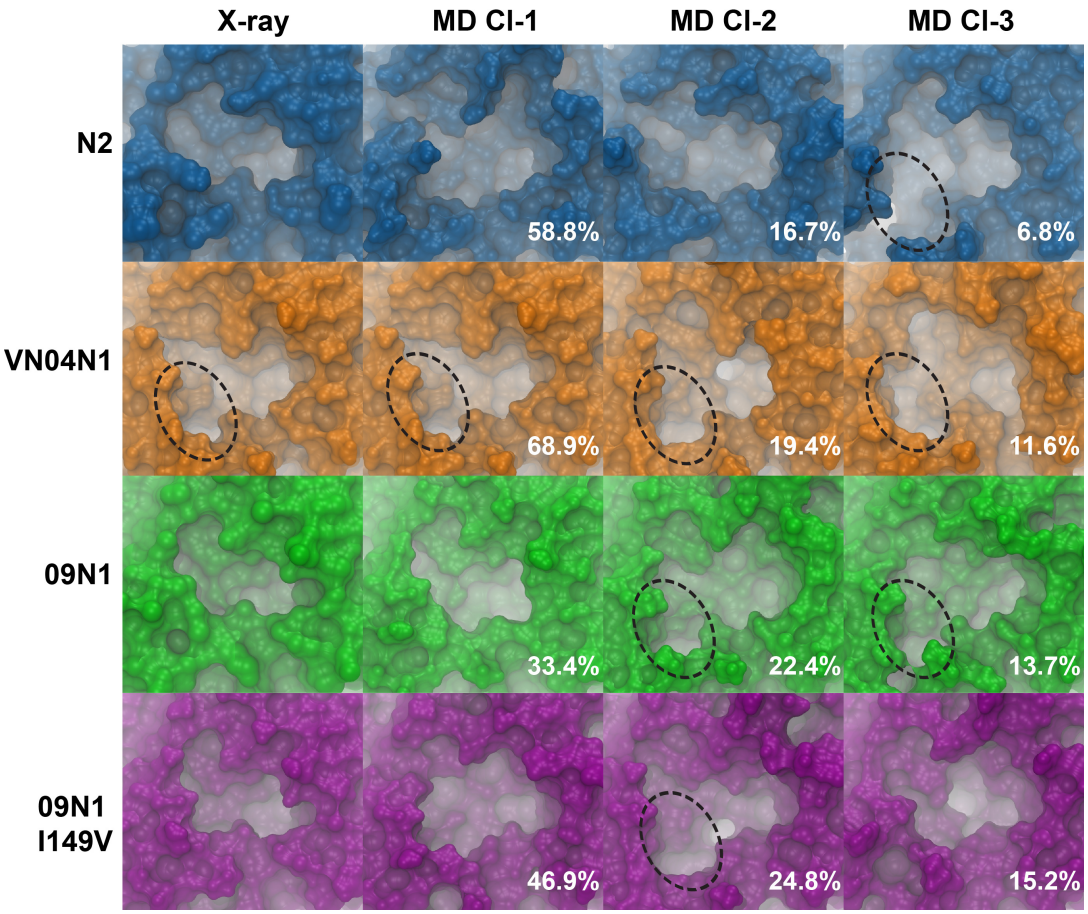

Figure 2.1: Solvent accessible surface area of NA binding site

The solvent accessible surface area of the NA binding site is shown, as computed by the MSMS [16] program, for the x-ray structure, and top three most dominant central member cluster structures (population percentages indicated in white text for each cluster), shown for A/Tokyo/3/67 (N2), A/Vietnam/1203/04 (VN04N1), A/California/04/2009 (09N1), and the 09N1_I149V mutant strain. The open 150-cavity, where present, is outlined with a dotted circle.

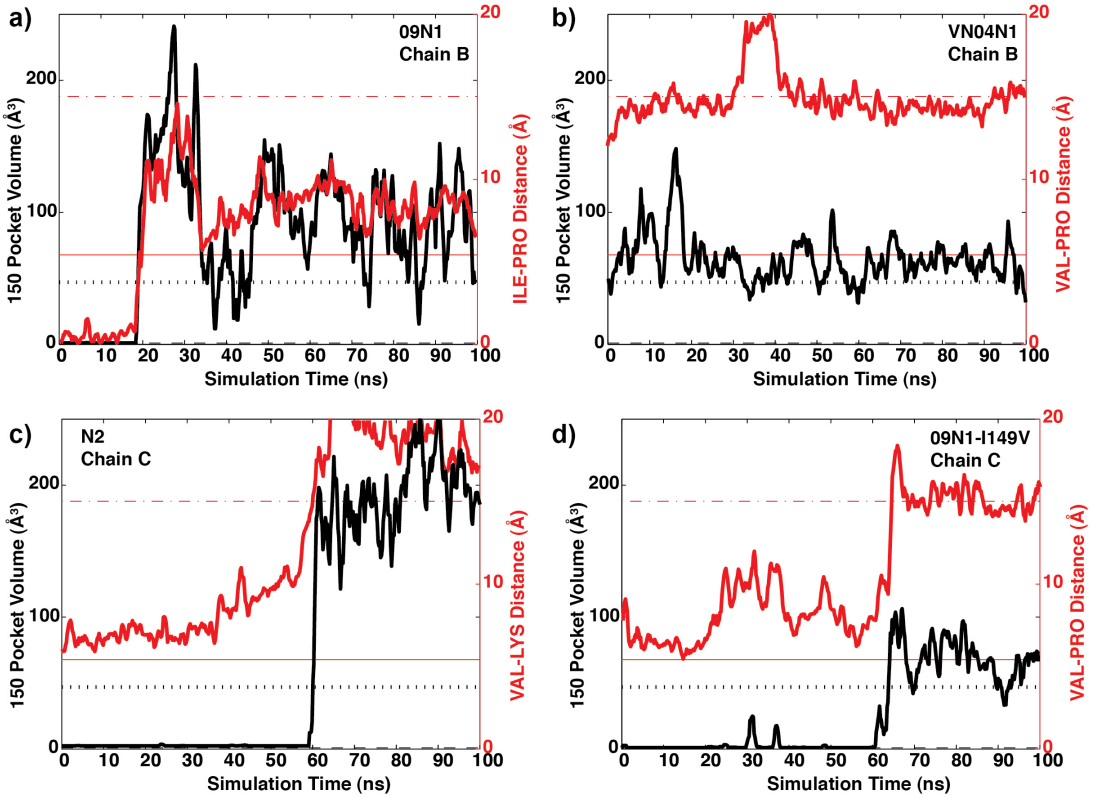

Figure 2.2: Time series analysis of 150-cavity volume and width for a particular monomer in each of the simulated systems

On the left side y-axis, the volume of the 150-cavity is computed over the course of simulation. The distance between alpha carbon of residue 431 (PRO in panels a, b, and d; LYS in c) and the closest sidechain carbon of residue 149 (Val149 panels b, c, d; ILE in a) is computed and shown in red and the right-side y-axis. The black and red dotted lines correspond to the open crystal structure (2HTY) volume and distance, respectively; whereas the black dashed and red solid lines correspond to the closed crystal structure (2HU4) volume and distance, respectively. The systems shown are A/Tokyo/3/67 (N2), A/Vietnam/1203/04 (VN04N1), A/California/04/2009 (09N1), and the 09N1_I149V mutant strain.

Table 2.2: Population Analysis based on open- or closed- 150 cavity

System (crystal structure)	Crystal structure state of 150-cavity	147.149,150…431	Total Open (%)	Total Closed (%)
N2 (1nn2)	Closed	D.VH…K	202 (10.1%)	1798 (89.9%)
VN04N1 (2hty)	Open	G.VK…P	1867 (93.4 %)	133 (6.6%)
09N1 (3nss)	Closed	G.IK…P	1215 (60.8 %)	785 (39.2%)
09N1_I149V (3nss*)	Closed	G.VK…P	742 (37.1 %)	1258 (62.9%)

RMSD-based clustering of the 150-loop residues enables an atomic-level population-based structural analysis. While the most populated cluster, i.e. the cluster that comprises at least 33% of the sampled ensemble, has a closed 150-loop configuration, structures within the next two most populated clusters adopt open 150-cavity configurations (Figure 2.1, Table 2.4). Figure 2.2 clearly shows that the second most dominant cluster from the 09N1 simulations has an open 150-loop, highly similar to the open-150-loop of VN04N1. By comparison, the VN04N1 150-cavity is consistently open throughout the simulations, being present for 93.4% of the trajectory (Figures 2.1 and 2.2B, Table 2.1, Figure 2.9). The formation of a stable and open 150-cavity in 09N1 indicates that the structural dynamics of the recent pandemic strain appear to be more similar to the classic group-1 isolates than to the group-2 isolates, in contrast to what the static crystal structure suggests. This finding provides an atomic-level structural understanding of how antiviral compounds designed to take advantage of contacts in the 150-cavity can be active against both the 2009 H1N1 and 2004 Vietnam H5N1 isolates, as very recently shown in ref. 3.

2.2.3 150-cavity formation controlled by a conserved salt bridge

The dynamics of the N2 strain reveal that a key salt bridge between conserved residues D147 and H150 controls the formation of the 150-cavity in N2. This ionic contact locks I149 in the space of the 150-cavity (Figure 2.3), as suggested in ref. 5. However, in each chain of the N2 tetramer simulation, this salt bridge intermittently breaks and then reforms; in chain C, at 60 ns the contact is lost again, after which the open 150-cavity forms, and contact to the 430-loop is lost (Figure 2.2C, Figure 2.10).

The loss of the D147-H150 salt bridge allows the 150-loop to move to the open position, even wider than the VN04N1 open 150-loop structure (Figure 2.2). RMSD-based clustering of the 150-loop indicates that while both the first and second most dominant configurations remain closed, the third most dominant configuration, representing 6.8% of the trajectory, exhibits an open 150-cavity (Figure 2.1). Volumetric calculations of the 150-cavity confirm that the open cavity conformation is present in 10% of the simulation and has a volume of 284 $Å^3$. For the remainder of the simulation, the salt bridge does not reform, and the wide-open 150-cavity therefore persists in one chain of the N2 tetramer.

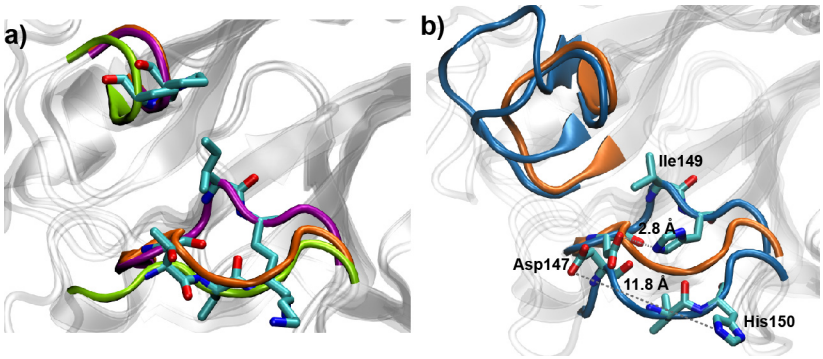

Figure 2.3: Structural variation in N1 and N2 clinical isolates
a) The 150- and 430-loop structures are shown for 09N1 crystal structure (purple), 09N1 second most dominant molecular dynamics (MD) cluster representative structure (green backbone), and VN04N1 crystal structure (orange), indicating that the pandemic N1 adopts an open 150-loop conformation. Gly147, Ile149, Lys150, and Pro431 are shown in stick representation**.** b) N2 150- and 430-loops from crystal and most dominant cluster representative structures are shown in blue, and open VN04N1 crystal structure are shown in orange. The D147-H150 salt bridge spontaneously ruptures in Chain C of N2, extending its initial contact from 2.8 Å in crystal structure to 11.8 Å in the most dominant MD-generated cluster structure, revealing a wide-open 150-cavity.

The spontaneous loss of this key contact under "physiologically relevant" simulation conditions provides a clear atomic-level model for 150-cavity formation in

the N2 clinical isolate. The loss of the salt bridge reduces the rigidity of the 150-loop, enabling the loop to sample more open conformations. Contacts with the neighboring 430-loop are simultaneously lost, and significant expansions of both the 150- and 430-cavities occur. (Figs. 2.1, 2.2C, and 2.3, Table 2.5 and Figure 2.8). Although the open 150-loop is energetically accessible in the N2 structures, its low population during the simulation makes it unlikely that this open 150-cavity would appear in x-ray crystallography experiments. Such a cavity would be able to accommodate compounds targeting the 150-cavity, albeit with a lower affinity, as very recently shown in ref. 2.3. In all the N1 proteins, D147 is replaced with an uncharged G147, and therefore no salt bridge is present to lock I/V149 in the 150-cavity space. This may explain why an open 150 cavity is characteristically observed in crystals, even in 09N1, which is able to adopt a stable open 150-cavity conformation. It also underscores the importance of considering solution-phase dynamics for these enzymes and not only crystallographic information, which is generally only able to provide one low-energy snapshot of the dynamic protein complex under crystalline conditions.

Among N1 alleles for which structures exist, the 2009 H1N1 pandemic isolate uniquely contains an I149. Thus, Li et al. [9] hypothesized that the additional extension of the I149 sidechain, compared to V149, may be a compensating factor in controlling the closed-loop structure, despite strict conservation of all other residues in this area. Structurally, the longer sidechain of I149 may facilitate van der Waals contacts to the neighboring 430-loop, and shift the population to a more closed 150-loop state; a V149 mutation would facilitate loss of contact between the 150- and

430-loop, shifting the population to a more open 150-cavity state. To test this hypothesis, we created the 09N1_I149V mutant strain *in silico* and performed an identical 100 ns simulation. Our results indicate that the effect of this mutation on 150-cavity status varies due to 150-loop flexibility. The time series data indicates that the I149V mutation caused Chain D to open almost immediately, Chain C to open after 60 ns, chain A to open intermittently, and had almost no effect on Chain B (Figure 2.12). Overall, the 09N1_I149V mutant is actually more closed, exhibiting the open 150-cavity less frequently, in only 37.1% of the simulation, compared to the normal 09N1 strain with the I149 present (Table 2.2, Figure 2.5D, Figure 2.12). Moreover, only one of the three most dominant structures, cluster 2, presents the open 150-cavity, and thus, the V149 by itself cannot explain the behavior of the 09N1_I149V.

2.2.4 Evolutionary analysis of sequence conservation in 150-loop region

To date, the evolutionary distribution of the 150-cavity among NA alleles has been inferred primarily from crystallographically resolved structures, which represent a limited subset of the genetic variation of NAs in nature. Based on those analyses, it seemed logical to attribute the occurrence of an open 150-cavity to having a V or I at position149, and by extension, to membership, in the group-1 and group-2 NAs respectively, as shown in Figure 3 of Li et al. Our dynamical analyses suggest that I/V149 is not as critical to 150-cavity status as the D147-H150 salt bridge, which warrants re-examination of the association of cavity status with NA group membership. We determined the distribution of genetic variation in the 150 and 450

loops among all avian, human and swine N1 and N2 sequences that had been deposited in GenBank and GISAID as of 12/8/2010 using phylogenetic analysis to construct consensus sequences for each major clade (see Supplementary Information (SI) section 2.5 for methodological details).

Our phylogenetic analyses (Table 2.1) show that no single amino acid position in the 150 or 430-loops clearly differentiates the N1 and N2 alleles, which are in group-1 and group-2, respectively. The D147-H150 salt bridge is not a defining characteristic of N2 alleles, as it is not present in avian viruses, which were the source of the N2 allele in the 1957 H2N2 human pandemic strain. Nor is the salt bridge found in human H3N2 viruses that have been circulating since 2008, due to fixation of a D147N mutation. Thus, neither the amino acid at position 149 nor the salt bridge are fixed characters that differentiate group 1 and group 2 NAs. Nor do they, at least by themselves, characterize viruses capable of infection of humans. Additional tests of our hypothesis require the acquisition of crystal structures of additional NA alleles, most critically, an N2 allele that contains both the D147-H150 salt bridge and I149.

2.3 Discussion

Our results highlight the importance of interpreting influenza neuraminidase sequence and structural data in light of the dynamical ensemble of conformations that are accessible to each NA protein. This work shows for the first time that both N1 and N2 clinical isolates exhibit flexibility in the 150-cavity neighboring the conserved sialic acid binding site. While it remains possible that the open and closed conformation observed in crystal structures may be due to differences in

crystallization conditions or procedures, our results indicate that the presence of the 150-cavity is not a strictly defining characteristic for group-1 or group-2 NA enzymes. Instead, it appears that both N1 and N2 enzymes are able to adopt an open 150-cavity within their solution phase structural ensemble, in various relative populations, which appear to be predominantly controlled by the presence of the D147-H/R150 salt bridge. This suggests a new paradigm for the understanding of the presence of the 150-cavity in both group-1 and group-2 neuraminidases. The inherent flexibility of the 150- and 430-loops may play a role in full glycan receptor recognition, and in particular, with facilitating recognition events with the distal sugar residues of different glycan receptors. It is likely that the opening and closing of the 150-cavity is required for natural sialoglycan substrates to fit into the active site, given the bulky nature of these glycans.

This study additionally underscores the need to consider dynamics in rationalizing the structure-function relationships of various antiviral-NA pairs. Ensemble-based drug discovery approaches[17] that account for full receptor flexibility towards neuraminidases that do not contain the D147-H150 salt bridge will likely present additional advances in the design of compounds that selectively target the 150-cavity, opening the possibility for receptor-specific inhibitors. In closing, we note that whether the flexibility of the NA binding site has an impact on receptor specificity, virus transmissibility, or pathogenicity remains to be seen and will likely require a better understanding of HA receptor binding domain dynamics for each of the NA/HA pairs found in humans[18].

2.4 Methods

2.4.1 Simulation Protocol

System setup was performed as follows for all simulated systems. Atomic coordinates were taken from 2HTY for A/Vietnam/1203/04 (VN04N1)[5], 3NSS for A/California/04/2009 (09N1)[9], and 1NN2 for A/Aichi/3/67 (N2)[13]. Protonation states for histidines and other titratable groups were determined at pH 6.5 by the PDB2PQR[19] web server using PROPKA[20] and manually verified. All crystallographically resolved water molecules and calcium ions were retained where possible and taken by homology from 2HTY if not present. The system was set up using the AMBER11[21] program xLeap using the AMBER99SB force field[22]. Disulfide bonds were properly enforced using the CYX notation in AMBER. A 10-12 Å pad of TIP3P waters was added to solvate to each system. Neutralizing counter ions were added to each system. In order to mimic experimental assay conditions, a 20 mM NaCl salt bath was introduced. System details and additional methodological information can be found in the Supplementary Information.

N1 and N2 tetramer simulations were performed with a version of the PMEMD module from AMBER 11 that was custom tuned for these specific simulations and the NICS Cray XT4 and SDSC Trestles supercomputers by SDSC under the NSF's TeraGrid Advanced User Support Program. The N1 and N2 apo tetramer complexes were minimized and equilibrated as follows. In order to alleciate any steric clashes prior to performing molecular dynamics the structures were minimized in a number of stages in which harmonic restraints of initially 5 kcal mol^{-1}

$Å^{-2}$ on all non-hydrogen protein atoms were slowly reduced over approximately 40,000 combined steepest descent and conjugate gradient minimization steps.

Following minimization, the system was linearly heated to 310 K in the NVT ensemble using a Langevin thermostat, with a collision frequency of 5.0 ps^{-1}, and harmonic restraints of 4 kcal mol^{-1} $Å^{-2}$ on the backbone atoms. Then, a further three 250 ps long runs at 310 K were conducted in the NPT ensemble with the restraint force constant being reduced by 1 kcal mol^{-1} $Å^{-2}$ each time and pressure controlled using a Berendsen barostat[23] with a coupling constant of 1 ps and a target pressure of 1 atm. A final 250 ps of NPT dynamics was run at 310 K without restraints and a Langevin collision frequency of $2ps^{-1}$. Production runs were then made for 100ns duration in the NVT ensemble at 310 K. As with the heating, the temperature was controlled with a Langevin thermostat (but with a 1.0 ps^{-1} collision frequency),. The time step used for all stages was 2 fs and all hydrogen atoms were constrained using the SHAKE algorithm[24]. Long range electrostatics were included on every step using the Particle Mesh Ewald algorithm [25] with a 4th order B-spline interpolation, a grid spacing of <1.0 Å, and a direct space cutoff of 8 Å. For all trajectories, the random number stream was seeded using the wallclock time in microseconds. The production trajectories for each monomer of the tetramer were extracted and concatenated to approximate 400 ns of monomer sampling.

2.4.2 RMSD Clustering

RMSD clustering was performed as implemented in the rmsdmat2 and cluster2 programs of the GROMOS++ analysis software [26]. 500 tetramer structures

were collected by sampling at 200 ps intervals. Monomer structures were then concatenated together, yielding a total of 2000 structures. Prior to clustering, external translational and rotational motions were removed by minimizing the RMSD distance of the alpha-carbon-atoms of the sampled structure to the equivalent atoms of the first frame of chain A. Using a 2.6 Å cutoff, clustering was then performed using the GROMOS++ clustering algorithm [27] in Gromacs [28] on the alpha-carbon atoms of the 6-residue subset, 146 to 152, which comprise the 150 loop. Each cluster contains a central structure, or "cluster representative member," called the "centroid," whose RMSD is equidistant to all other cluster members. The cluster representative's structural properties are considered characteristic of all cluster members. Cluster results are summarized in Table 2.2.

2.4.3 RMSD and B-factor calculations

B-factor calculations, as well as tetramer and monomer RMSD time series, were performed using the ptraj analysis tool in the AMBER 10 program suite [29]. Structures were sampled at 20ps intervals. Prior to performing each calculation, external translational and rotational motions were removed by minimizing the RMSD distance of the alpha-carbon atoms to the equivalent atoms of the first frame of the trajectory. RMSD and B-factor values were calculated for alpha-carbon atoms.

2.4.4 09N1 RMSD 150-loop measurements

RMSD values were measured using a custom, hand-written script in the VMD TCL-TK console [30]. Structures were sampled at 20ps intervals. Sampled structures

of each monomer were RMSD-aligned by alpha-carbon to the equivalent alpha-carbon atoms of the “reference” structure: chain A of PDB ID 2HTY, open reference; or chain A of PDB ID 2HU4, closed reference. Following alignment, the RMSD of the 150 loop of each monomer was measured with respect to the 150 loop of each reference structure. The 150 loops were defined as residues 146 to 152 for the 09N1 monomers, as well as for the open and closed reference structures.

2.4.5 Interatomic distance measurements

The distance separating the salt bridge pair ASP147 and HIS150 was measured using a custom, hand-written script in the VMD TCL-TK console. Structures were sampled at 20ps intervals. The distance between the two residues was defined as the distance separating centers of mass of the heavy atoms of the ASP147 carboxylate and the HIS 150 imidazole. The distance between residues 149 and 431 were measured for each step using a custom VMD script.

2.4.6 Neuraminidase Volume Population Analysis

The numbers of open or closed 150-cavity conformations out of a total of 2000 snapshots were computed. Any instantaneous volume equal to or greater than half the volume of the crystal structure of canonical group-1 serovar (2HTY exhibits a total 150-cavity volume of 36 $Å^3$) is considered to be "open". Otherwise the 150-cavity is considered "closed" (i.e., when it exhibits less than 18 $Å^3$). The volume of the 150-cavity was measured for each step by using POVME [15]; a pocket volume measuring algorithm. To measure the volume, we used a single inclusion sphere that

encompassed the 150-cavity. The POVME algorithm neglected the volume occupied by NA atoms and not spatially contiguous with a point specified within the 150-cavity. By rotating 90 degrees around the NA tetramer central axis, each of the other three 150-cavity sites were specified. The volume was thus measured for every snapshot of the simulation on all four chains of each NA.

2.4.7 Figures and Plots

Matlab was used to generate all plots and molecular images were created using VMD [30].

2.4.8 Consensus sequences

We downloaded all influenza A N1 and N2 gene sequences from humans, avians and swine that were greater than 600 base pairs in length from GenBank and GISAID on 12/8/2010. We aligned sequences using ClustalX 2.0[31] and constructed phylogenetic trees using MrBayes version 3.1.2[32] using the GTR+I+gamma model, as suggested by jmodeltest version 0.1.1[33] under the Akaike Information Criterion. All other MrBayes parameters were set to the default. We allowed MrBayes to run, sampling every 1,000 trees, until the Monte Carlo Markov chains converged as determined by Tracer software version 1.5[34]. We discarded the burn-in as determined by Tracer. Similar results were obtained using the neighbor-joining routine of PAUP* 4.0b10 [35](results not shown). Consensus sequences containing amino acids found at a frequency of at least 80% were constructed for each major evolutionary clade. Results

are shown, along with samples sizes, in Table 2.1; major polymorphisms are indicated in white boxes.

2.5 Supplementary Information

2.5.1 Supplementary Figures

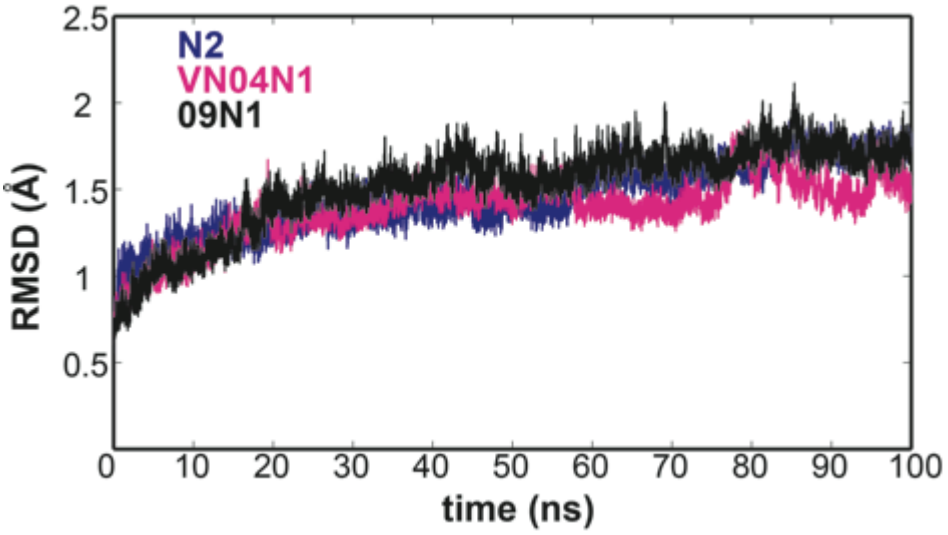

Figure 2.4: Time series for the RMSD over alpha-carbon atoms for tetramer systems
The time series for the RMSD over alpha-carbon atoms for tetramer systems is shown for each 100 ns simulation (for reference, N2 is shown in blue, VN04N1 in pink, and 09N1 in black). The plot indicates stability of the simulated systems.

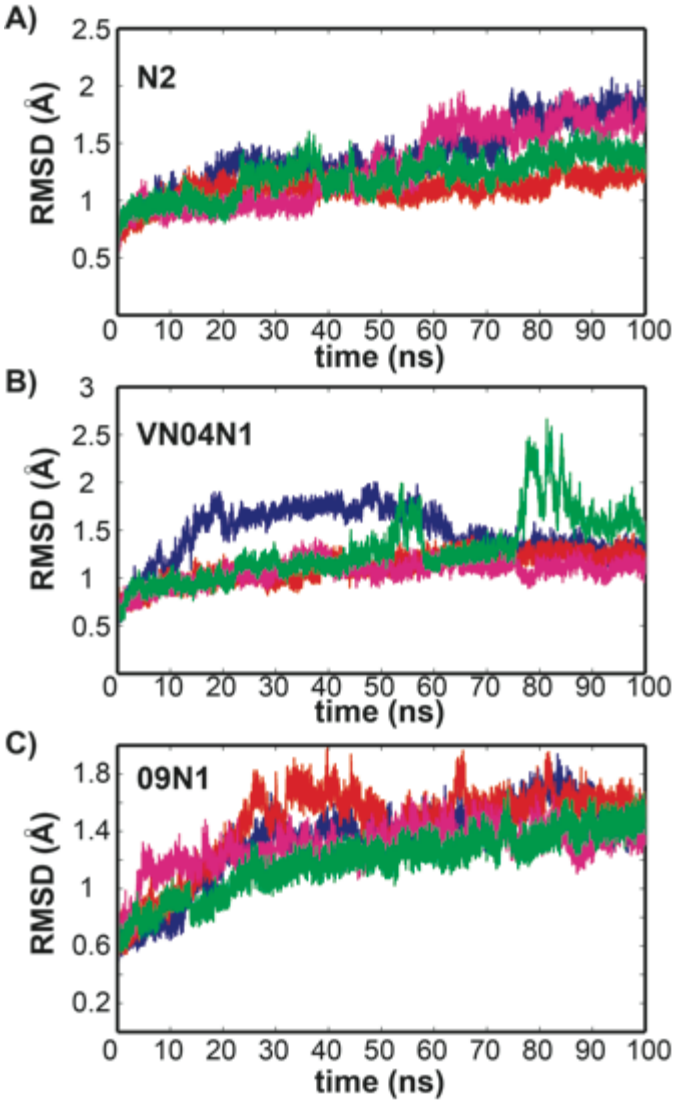

Figure 2.5: Time series for the per-monomer RMSD
The time series for the per-monomer RMSD, as computed over the alpha-carbons, is shown for the N2, VN04N1, and 09N1 systems. For all systems, chain A is shown in blue, chain B in red, chain C in fuschia, and chain D in green.

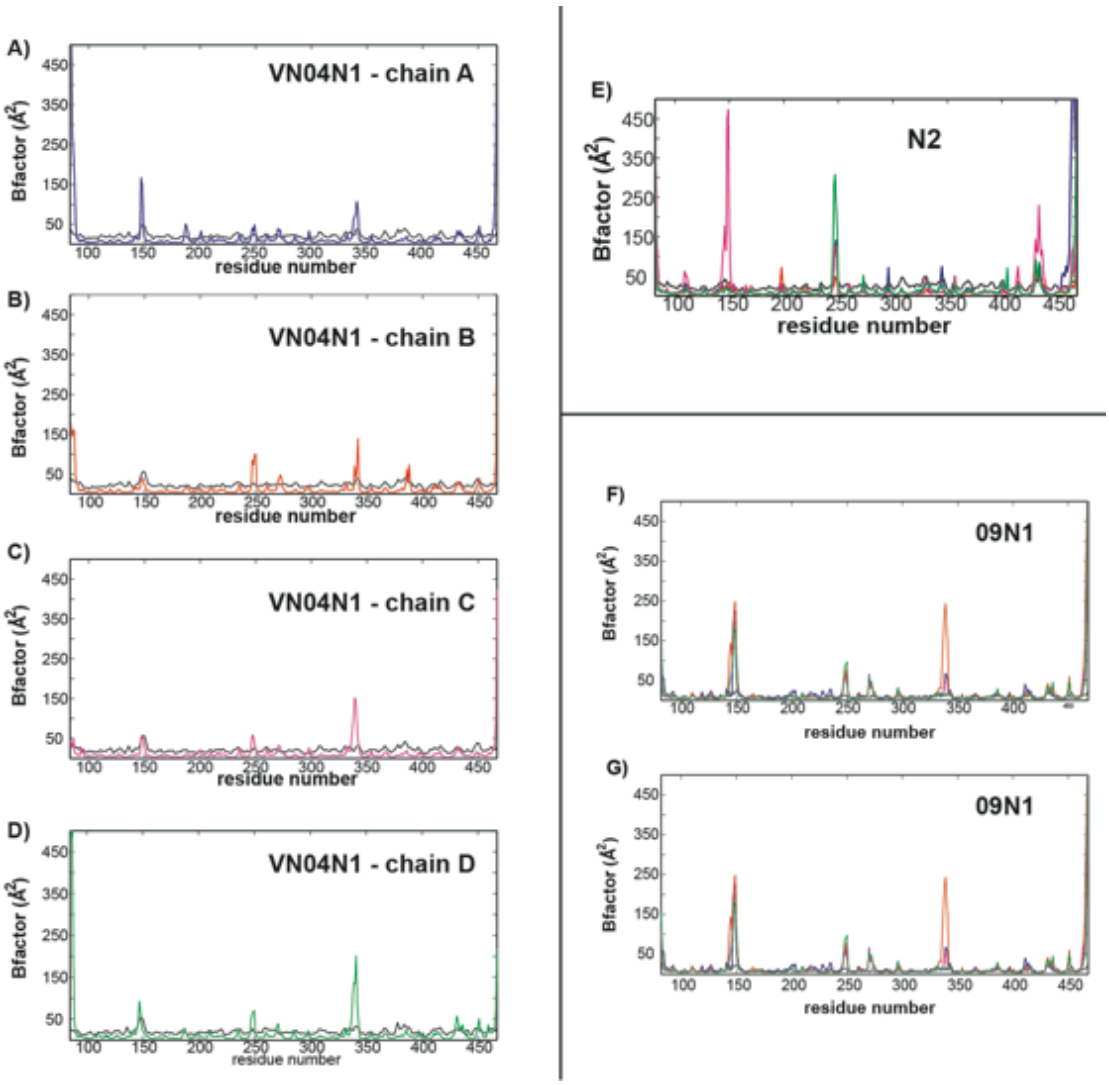

Figure 2.6: Experimental and simulation-derived B-factors
Crystal B-factors shown in black throughout. MD-derived B-factors shown in blue, orange, fuchsia, and green for tetramer chains A, B, C, D, respectively. A-D) VN04N1 PDB 2HTY is a tetramer; B-factors for each chain are compared individually. E) N2 PDB 1NN2 has one chain; all 4 MD chains are shown, F-G) 09N1 PDB 3NSS has 2 chains; all 4 MD monomers are shown against chain A (in F) and chain B (in G).

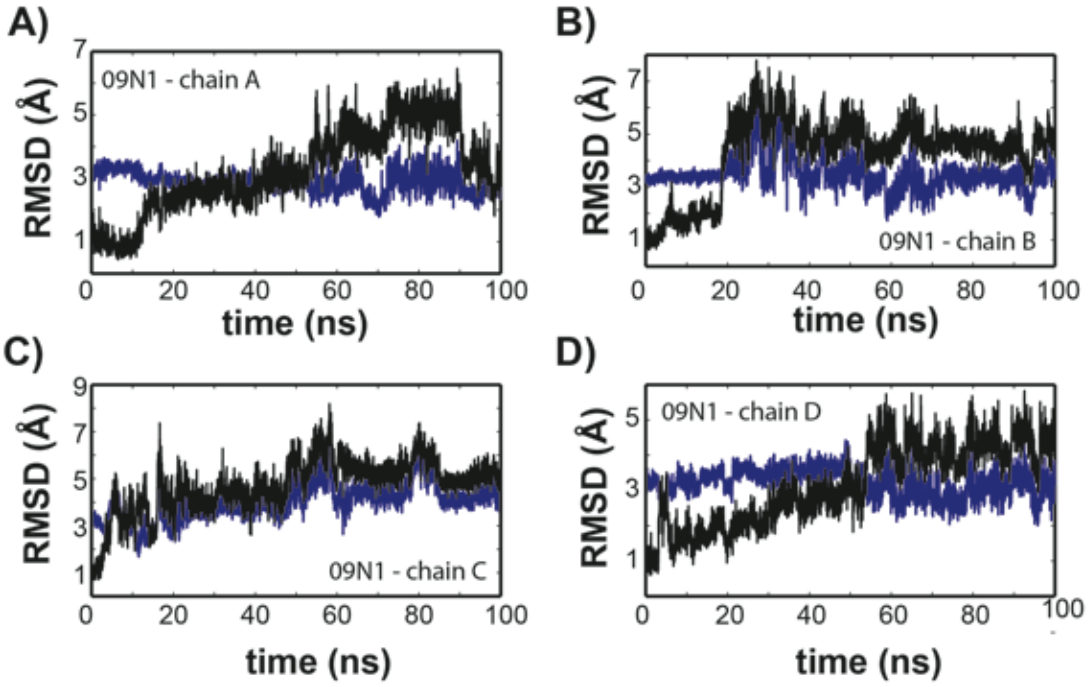

Figure 2.7: Structural deviations in the 150-loop residues in the 09N1 system

The structural deviations in the 150-loop (residues 146-152) are shown for the 09N1 system. 150-loop RMSD from the open- (shown in blue) and closed- (shown in black) 150-loop crystal structures (2HTY and 2HU4, respectively) indicate structural deviations from the experimentally resolved data. Locations where the two lines cross over indicate a loop "switching event" (from open to closed, or closed to open).

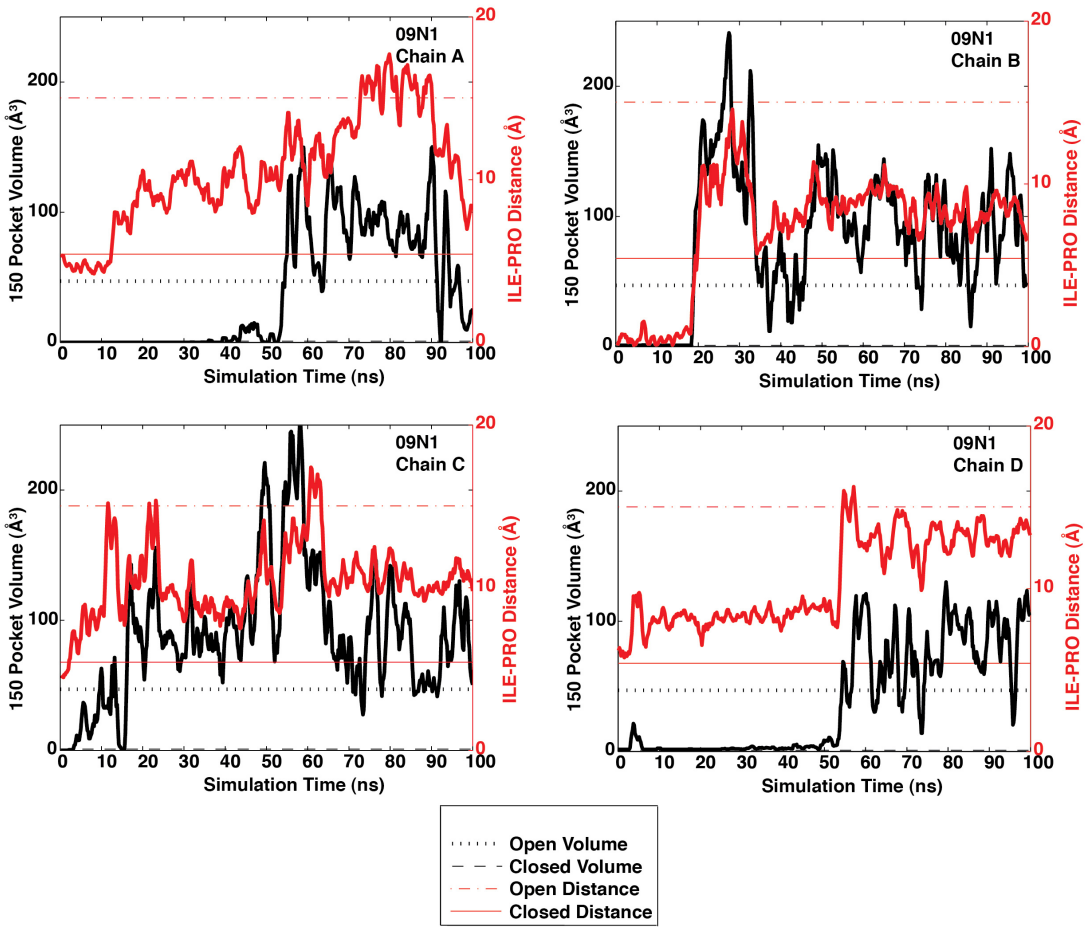

Figure 2.8: Time series analysis of 150-cavity volume and width for 09N1 system
The distance between alpha carbon of residue 431 (Pro431 in 09N1) and the closest sidechain carbon of residue 149 (Ile149 in 09N1) is computed and shown in red and the right-side y-axis. On the left side y-axis, the volume of the 150-cavity is computed over the course of simulation. Each plot represents a chain of the neuraminidase tetramer (A-D).

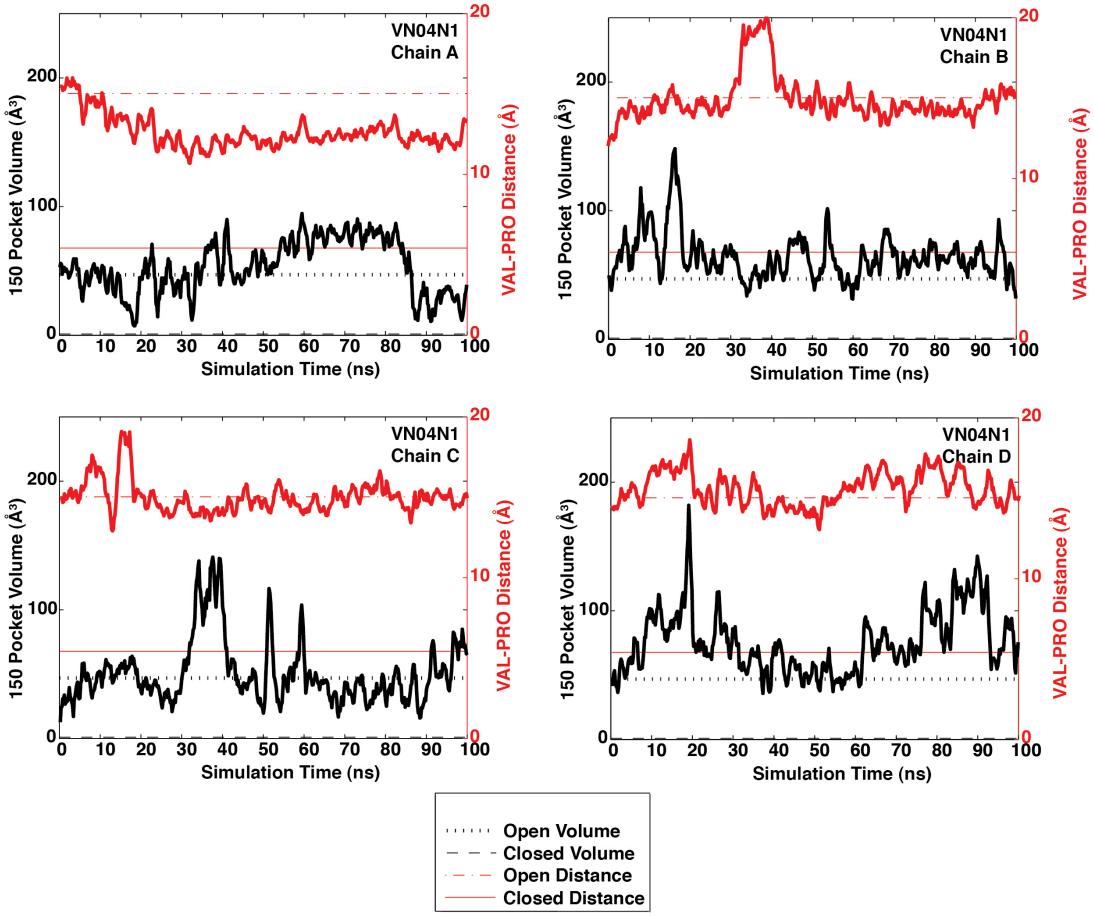

Figure 2.9: Time series analysis of 150-cavity volume and width for VN04N1 system

The distance between alpha carbon of residue 431 (Pro431 in VN04N1) and the closest sidechain carbon of residue 149 (Val149 in VN04N1) is computed and shown in red and the right-side y-axis. On the left side y-axis, the volume of the 150-cavity is computed over the course of simulation. Each plot represents a chain of the neuraminidase tetramer (A-D).

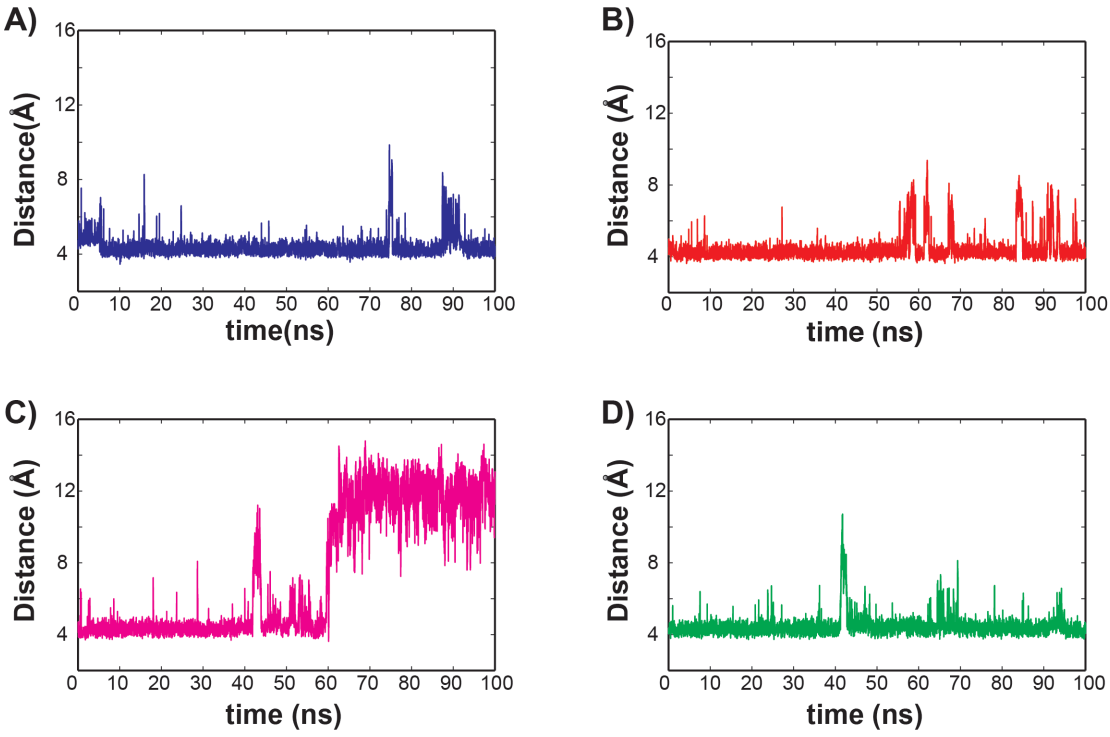

Figure 2.10: Time series plots the key salt bridge that controls 150-cavity formation in the neuraminidase enzymes

The heavy atom distance between residues Asp147 and His150 is shown for all monomers (chains A-D) of the N2 simulation. In Chain C, the salt bridge breaks and does not reform after 60 ns.

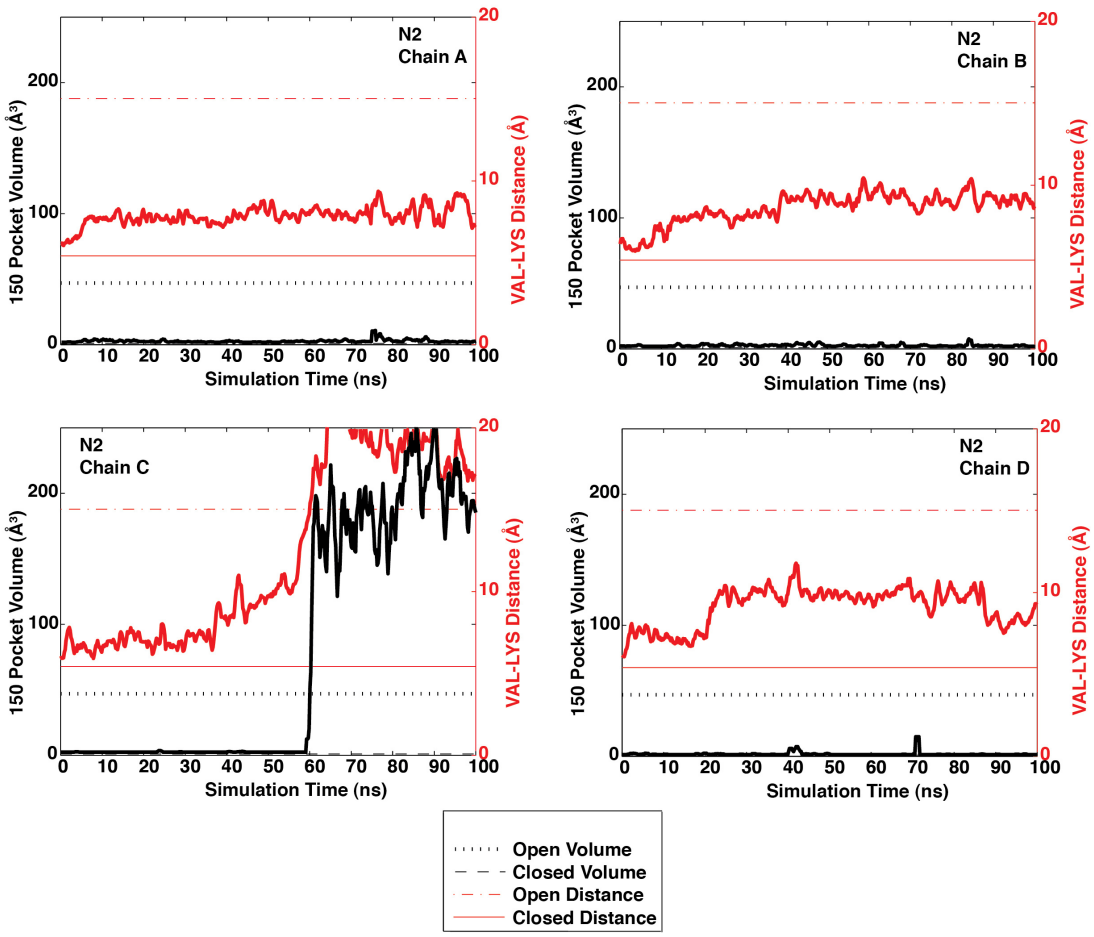

Figure 2.11: Time series analysis of 150-cavity volume and width for N2 system

The distance between alpha carbon of residue 431 (Lys431 in N2) and the closest sidechain carbon of residue 149 (Val149 in N2) is computed and shown in red and the right-side y-axis. On the left side y-axis, the volume of the 150-cavity is computed over the course of simulation. Each plot represents a chain of the neuraminidase tetramer (A-D).

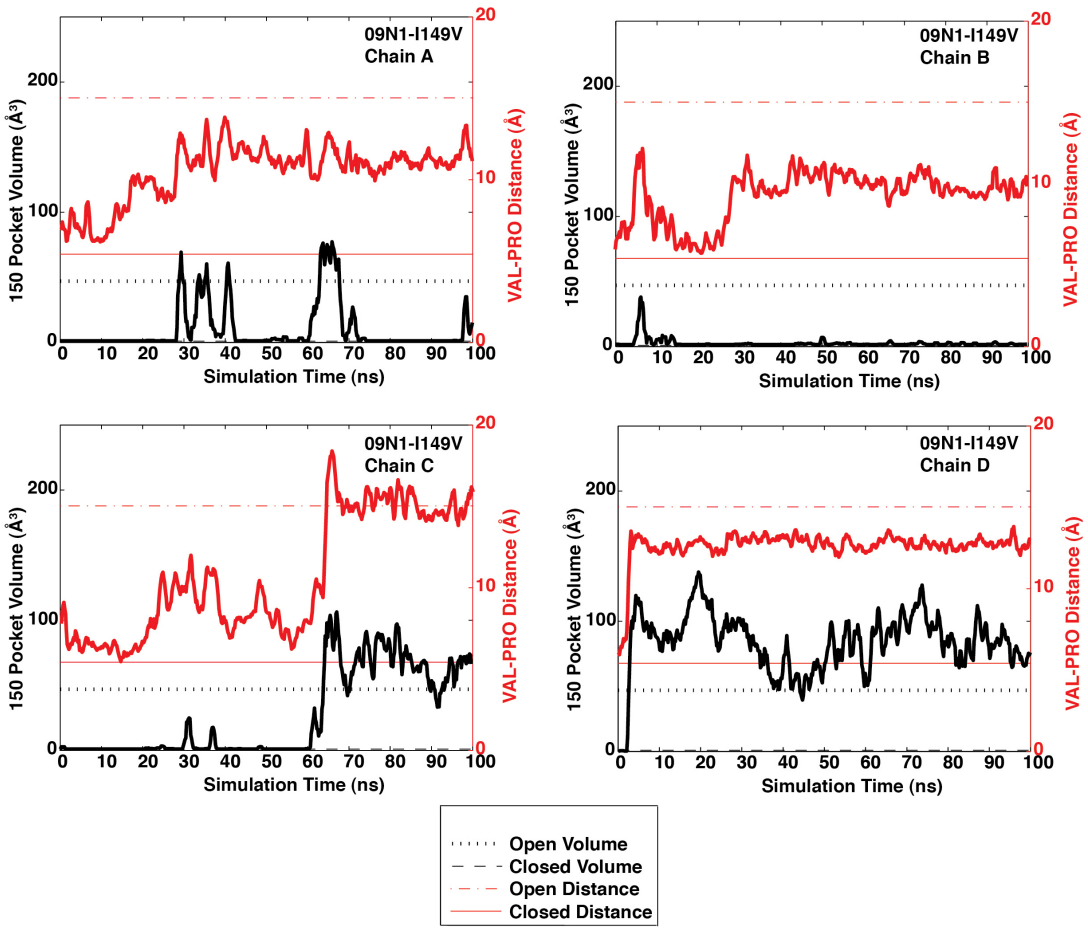

Figure 2.12: Time series analysis of 150-cavity volume and width for 09N1_I149V system

The distance between alpha carbon of residue 431 (Pro431 in 09N1_I149V) and the closest sidechain carbon of residue 149 (Val149 in 09N1_I149V) is computed and shown in red and the right-side y-axis. On the left side y-axis, the volume of the 150-cavity is computed over the course of simulation. Each plot represents a chain of the neuraminidase tetramer (A-D).

2.5.2 Supplementary Tables

Table 2.3: Description of simulated systems
System name, crystal structure PDB identifier, human strain isolate description, colloquial system description, total simulation time for tetramer simulation, and total number of atoms are shown for each system.

system name	crystal struct-ure	strain	Initial 150-loop state	simulation time (ns)	no. of atoms
N2	**1nn2**	**A/Tokyo/3/67**	**Closed**	**100**	**151,895**
09N1	**3nss**	**A/California/04/2009**	**Closed**	**100**	**165,171**
09N1-I149V	**3nss***	**A/California/mutant**	**Closed**	**100**	**156,338**
VN04N1	**2hty**	**A/Vietnam/1203/04**	**Open**	**100**	**112,311**

Table 2.4: RMSD-based clustering results
System name, percent population in the three most dominant clusters and open or closed 150-cavity designation, and the total number of clusters are shown.

system	cluster 1	cluster 2	cluster 3	no. of clusters
N2	**58.8% (closed)**	**16.7% (closed)**	**6.8% (open)**	**24**
VN04N1	**68.9% (open)**	**19.4% (open)**	**11.6% (open)**	**14**
09N1	**33.4% (closed)**	**22.4% (open)**	**13.7% (open)**	**25**
09N1_I149V	**46.9% (closed)**	**24.8% (open)**	**15.2% (closed)**	**14**

Table 2.5: 150-cavity volume for all NA crystal structures (with hydrogen atoms added) and the 3 most dominant cluster structures from the simulations
Structures and volumes correspond to Figure 1 in the main text.

System	Volume (Å^3)	Open/Closed characterization
N2 – crystal structure	**0**	**Closed**
N2 MD cluster 1	**0**	**Closed**
N2 MD cluster 2	**0**	**Closed**
N2 MD cluster 3	**284**	**Open**
VN04N1 – crystal structure	**36**	**Open**
VN04N1 MD cluster 1	**53**	**Open**
VN04N1 MD cluster 2	**136**	**Open**
VN04N1 MD cluster 3	**65**	**Open**
09N1 crystal structure	**0**	**Closed**
09N1 MD cluster 1	**0**	**Closed**
09N1 MD cluster 2	**77**	**Open**
09N1 MD cluster 3	**143**	**Open**
09N1_I149V minimized	**0**	**Closed**
09N1_I149V MD Cluster 1	**0**	**Closed**
09N1_I149V MD Cluster 2	**146**	**Open**
09N1_I149V MD Cluster 2	**1**	**Closed**

Chapter 2, in full, is a reprint of "Mechanism of 150-Cavity Formation in Influenza Neuraminidase", which was published in 2011 in Nature Communications, volume 2, page 388, by Rommie E. Amaro, Robert V. Swift, Lane W. Votapka,

Wilfred W. Li and Robin M. Bush. The dissertation author was the third investigator and author of this paper.

Chapter 3: DelEnsembleElec: Computing Ensemble-Averaged Electrostatics Using DelPhi

A new VMD plugin that interfaces with DelPhi to provide ensemble-averaged electrostatic calculations using the Poisson-Boltzmann equation is presented. The general theory and context of this approach are discussed, and examples of the plugin interface and calculations are presented. This new tool is applied to systems of current biological interest, obtaining the ensemble-averaged electrostatic properties of the two major influenza virus glycoproteins, hemagglutinin and neuraminidase, from explicitly solvated all-atom molecular dynamics trajectories. The differences between the ensemble-averaged electrostatics and those obtained from a single structure are examined in detail for these examples, revealing how the plugin can be a powerful tool in facilitating the modeling of electrostatic interactions in biological systems.

3.1 Introduction

Electrostatic interactions play an essential role in the dynamics of biological systems. These forces are the predominant long-range interactions influencing the dynamics within and between biomolecules, thus the accurate treatment of electrostatics is necessary for detailed models of these systems. The electrostatic interactions associated with ionic and polar chemical groups found in proteins, nucleic acids, lipids, and other biomolecular systems are essential to both their structure and function[36].

A variety of methods exist for computing the electrostatic interactions within the context of classical Molecular Mechanics (MM) models. The electrostatic energy and potential may be obtained by computing the pairwise Coulomb interaction between all atoms, and many approximate methods exist to efficiently estimate these interactions. While the computation of pairwise interactions is computationally expensive, scaling as the square of the number of atoms in a system, methods such as the Particle-Mesh Ewald (PME) [37] algorithm for periodic systems and the Fast Multipole Method (FMM) [38] can significantly reduce the complexity of computation by introducing well-controlled approximations.

The Poisson-Boltzmann (PB) equation is an attractive method for computing the mean-field potential of biomolecules [39]. By modeling the solvent environment as a continuum dielectric and salt ion distribution, the electrostatic potential and electrostatic free energy of a biomolecule may be estimated under the given conditions. In most implementations, the solvent provides dielectric screening of the biomolecule potential, represented as a uniform dielectric constant, while the salt is modeled based on the potential values within the continuum solvent by a nonlinear term in the PB equation. A linearized PB treatment of salt effects is accurate for proteins with modest net charges in monovalent salt solutions, while the more sophisticated treatment of salt with the nonlinear form of the PB equation is often used in treating multivalent salt solutions and in modeling highly charged polyelectrolytes such as nucleic acids[40]. Many methods of solving the Poisson-Boltzmann equation such as presented here in DelPhi utilize the finite difference method to efficiently solve the electrostatic potential on a discretized grid[41].

Realistic treatment of biomolecular dynamics requires ensemble sampling to understand the conformational flexibility of these molecules within their chemical environment. All-atom explicit solvent Molecular Dynamics (MD) is a widely accepted method for generating a canonical distribution of states for biomolecules modeled using Molecular Mechanics force fields such as AMBER[42] and CHARMM[43]. By averaging over the trajectory of states produced in these simulations, ensemble properties such as the electrostatic potential may be obtained, giving insight into thermodynamic properties of these systems. This approach has been useful in previous studies of biomolecular systems, such as in understanding the electrostatic environment of protein-bound drugs[44] and in understanding the transport of ions and biomolecules through membrane pore proteins[45]. Notably, the use of ensemble information in particular, as opposed to a single static structure, when computing the electrostatic properties of dynamic biomolecules has been shown to increase agreement of the theoretical values with experiment [44].

To facilitate the calculation of ensemble-averaged electrostatic properties with the Poisson-Boltzmann equation, we have developed a plugin interface for the visualization software package VMD[30], which interfaces with the DelPhi numerical PB equation-solving software package[41,46] . This plugin computes the ensemble electrostatic potential and free energy of biomolecules from their trajectories in VMD-compatible formats such as those used in the MD packages AMBER and CHARMM. A graphical user frontend provides a simplified interface for specifying all the individual options supported by DelPhi, and the plugin creates the proper biomolecule coordinate inputs for DelPhi from the system partial charges and radii

obtained by VMD from the input structure file, as well as automating the process of ensemble averaging. The plugin executes DelPhi and then visualizes the electrostatic potential or ensemble-averaged potential from the DelPhi grid outputs in a variety of VMD formats, such as color-coded surface, 2D plane projection, and isopotential surface plots. The plugin supports both the linear and non-linear forms of the PB equation provided in DelPhi.

The application of these Poisson-Boltzmann calculations with the DelPhi Electrostatics plugin to biomolecular systems of current research interest such as the influenza virus major glycoproteins, hemagglutinin[21,29] and neuraminidase[47-49], are presented in this work as examples of the versatility of calculations that can be performed with this convenient new tool. The ensemble-averaged mean-field potential of these proteins are obtained in identical salt and solvent conditions as the corresponding MD simulation trajectories. The electrostatic potential of any AMBER or CHARMM system may be computed from its structure file and coordinates with this plugin, importing the parameters into the DelPhi calculation in Protein Databank (PDB) format. As the software plugin and source code is freely available under the GNU Public License (GPL), it is anticipated that this tool will be of general interest to the physics-based modeling research community.

3.2 Methods

3.2.1 Molecular Dynamics Simulations

Trajectories for two biomolecules were generated using explicitly solvated, all-atom molecular dynamics (MD) simulations. We chose two proteins important in influenza pathogenesis, neuraminidase and hemagglutinin, as subject molecules on which to perform the electrostatic calculations. Three different commonly used MD programs were used for the simulation of each molecule, as described below.

3.2.2 Influenza Hemagglutinin

Atomic coordinates were taken from accession number 1HGF in the Protein Data Bank (PDB). The protonation states for histidine and other titratable residues were determined at pH 7.0 using the PDB2PQR[19] server using PROPKA[20]. Crystallographic water molecules were retained. The AMBER Tools11[21] program sLeap was used to connect disulfide bridges and also to add a TIP3P water box with a 10 Å box spacing along each edge beyond the dimensions of the protein and 20mM NaCl to act as an explicit solvent along with K+ counterions to neutralize the system. The composite system contained 198,533 atoms. The hemagglutinin trimer was then minimized for 45,000 steps and equilibrated using four stages of harmonic constraints at 250,000 steps each, starting at 4 kcal $mol^{-1}Å^{-2}$ reducing by 1 kcal mol^{-1} $Å^{-2}$ each time. The systems were then simulated using NAMD2.8[50] with the AMBER FF99SB[22] force field under periodic boundary conditions with the isothermal-isobaric (NPT) ensemble at a temperature of 310 K. Pressure was maintained at 1 atm using a Nosé-Hoover Langevin Piston[51] and Particle-Mesh Ewald[25] was used to treat long-

range electrostatic interactions. Bonds involving hydrogen positions were constrained using the RATTLE algorithm[24]. Ranger, a massively parallel Teragrid computing platform, was utilized to perform the calculations (benchmark: 3.9 ns/day using 512 cores). The final trajectory contained 100 nanoseconds of simulation, which was reduced to 500 frames with a stride of 200 picosecond between each frame for analysis of ensemble electrostatics.

3.2.3 Influenza Neuraminidase

Details for the preparation of this system has been described previously [52]. Atomic coordinates were taken from PDB accession code 3NSS. The protonation states for histidines and other titratable groups at pH 6.5 were determined using the PDB2PQR[19] server. The system was prepared using the Ambertools11 program xLeap with a padding of 10-12 Å of water molecules prepared in a similar fashion to the hemagglutinin system mentioned above. The neuraminidase tetramer was then energy minimized using the PMEMD module from AMBER11 with 2000 steps of steepest descent, followed by 5000 steps of conjugate gradient minimization with 5.0 kcal mol^{-1} $Å^{-2}$ harmonic restraints. Then 25,000 more conjugate gradient steps were performed without restraints. The neuraminidase system was then also equilibrated by gradual heating to 310 K in the isothermal/constant volume (NVT) ensemble using a Langevin thermostat with a collision frequency of 5.0 ps^{-1}. Three subsequent 250 ps runs were performed at 310 K in the isothermal/isobaric (NPT) ensemble, with 4 kcal $mol^{-1}Å^{-2}$ restraints being reduced by 1 kcal mol^{-1} $Å^{-2}$ in each consecutive run. A Berendsen barostat[23] with a coupling constant of 1 ps and a target pressure of 1 atm

was used to maintain pressure, followed by a final 250 ps segment of NPT dynamics without restraints. Production dynamics was then performed for 100 ns with conditions similar to that of the hemagglutinin system above. Minimization, equilibration, and production were all performed on the NCIS Cray XT4 and SDSC Trestles high performance compute platforms (benchmark: 10.1 ns/ day using 256 cores [1 core per node] on NICS Athena). We pruned the final trajectory to 500 frames with a stride of 200 ps between each frame for analysis of ensemble electrostatics.

3.2.4 Poisson-Boltzmann Ensemble-Averaged Electrostatics

The PB equation (Eq. 3.1) uses an implicit and continuum-based model of the solvent and counterion environment surrounding a biomolecule to give a detailed description of its electrostatics.[53] Though many variations exist, it assumes a spatially varying dielectric constant and takes into account the shape and irregular charge distribution of the biomolecule.[54] A general form is given below[46]:

$$\nabla \cdot [\epsilon(x)\nabla\phi(x)] + \frac{e}{\epsilon_0}\sum_j c_j z_j \exp\left(-\frac{ez_j\phi(x)}{kT}\right) = -\frac{e}{\epsilon_0}\sum_i q_i \delta(x - x_i) \qquad \text{Eq. 3.1}$$

Here, $\phi(\mathrm{x})$ represents the electrostatic potential at position x, $\epsilon(\mathrm{x})$ is the spatially-varying value of the dielectric constant (in units relative to ϵ_0), q_i are the individual N_i partial charges associated with the atoms of the biomolecule (with positions specified by x_i), and e is the elementary electronic charge. The value kT is the Boltzmann factor (Boltzmann constant times temperature) and $\delta(\mathrm{x})$ is the Dirac delta function. The second term $\frac{e}{\epsilon_0}\sum_j c_j z_j \exp\left(-\frac{ez_j\phi(\mathrm{x})}{kT}\right)$ is associated with the N_j -

component ion density distribution with components having concentration c_j at a distance infinity away from the biomolecule and having valence z_j, at the specified salt conditions. Introducing the approximation $\frac{e}{\epsilon_0}\sum_j c_j z_j \exp\left(-\frac{ez_j\phi(\mathrm{x})}{kT}\right) = -\kappa^2\phi(\mathrm{x})$ (where $\kappa^2 = \frac{e^2}{kT\epsilon_0}\sum_j c_j z_j^2$) into equation 3.1 yields the linear PB equation, which is accurate for conditions such as modestly charged molecules (having relatively low electrostatic potentials in the surrounding medium) in a 1:1 monovalent salt environment.

$$\nabla\cdot[\epsilon(\mathrm{x})\nabla\phi(\mathrm{x})] - \kappa^2\phi(\mathrm{x}) = -\frac{e}{\epsilon_0}\sum_i q_i\delta(\mathrm{x}-\mathrm{x}_i) \qquad \text{Eq. 3.2}$$

Typically a computational method, such as implemented in the program DelPhi, is used to solve the equation numerically in a system of biological scale, as its canonical form is a nonlinear partial differential equation. In systems without regions of large potential values, the equation can be simplified to a linear form (Eq. 3.2). The variety of approaches to approximating the PB equation allow differing degrees of accuracy and efficiency[53], usually one at the expense of the other.

3.3 User Interface

DelEnsembleElec provides a graphical interface that can be accessed within the "Extensions" menu of VMD, simplifying the process of specifying DelPhi options and then computing single-point and ensemble electrostatic calculations, using the many VMD display drawing methods to visualize the resulting potential grids. The code is written in the Tcl script language and is compatible with Microsoft Windows

and UNIX-based platforms (Mac, Linux). When DelEnsembleElec is started, the main window appears (Figure 3.1), which contains the most basic customizable options. Once one or more trajectories or structures are loaded into VMD, or if they have been loaded already, a menu button allows the user to select the molecule on which to perform the run. Entry fields allow the user to specify the atom selection of the molecule, as well as an option to write a Gaussian cube file upon completion of the run. Two additional checkboxes allow the user to specify whether and how the completed map data will be loaded into VMD for visualization.

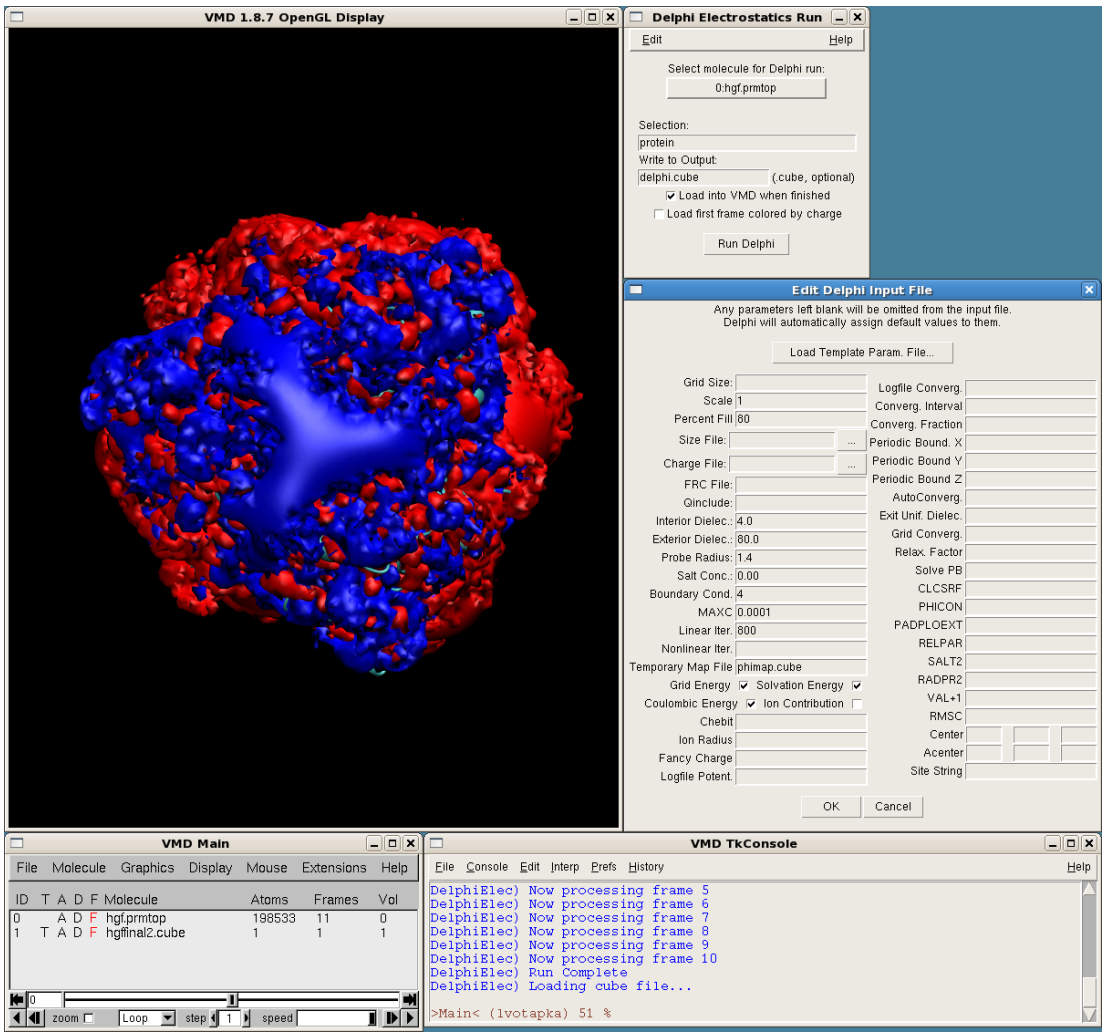

Figure 3.1: DelEnsembleElec plugin interface
A screenshot of the DelEnsembleElec plugin is shown; the top right window is the main DelEnsembleElec window. The window immediately below it is for setting the parameters in the DelPhi input file. The remaining windows are VMD windows; main display window at the top left, molecule loader and animation control at the bottom left, and console at the bottom right.

In addition, two dropdown menu options provide additional functionality. Under the menu Edit>Settings, a window appears where one can specify the working directory of the Delphi runs(s) (set automatically to the operating system's temporary directory) as well as the location of the Delphi program. The user may optionally

create the setup file only, but not actually initiate a Delphi run. Under the menu Edit>Input File, a window enables the user to customize most of the Delphi input parameters, either by entering the fields manually or by loading a template Delphi parameter file.

Once satisfied with the run specifications, the user may click the "Run Delphi" button at the bottom of the Main Window to execute the calculation. The time taken for execution depends on several factors, including the number of atoms in the system, the number of frames in the trajectory, as well as the amount of processor speed dedicated to the calculation. A benchmark on a single Intel Xeon 2.67 GHz processor takes approximately 15 minutes per frame to solve the linear PB equation for a system containing 20,000 atoms on a 169x169x169 static grid. Upon completion of the run, depending on the options selected by the user, the map data may be saved as a Gaussian cube file or loaded as a new molecule into the VMD main window, where the VMD representation window allows one to manipulate the graphical representation in a variety of formats. The plugin also has the capability to compute and display "difference maps" between the ensemble and single frame electrostatic grids (ensemble-averaged grid potential minus a single-frame grid potential).

When a single structure or a trajectory is loaded that does not contain atomic charge or radius information in a way that VMD can recognize (such as a pdb trajectory), this information must be provided to Delphi as force field parameter files. The paths to these files may be specified in the Edit>Input File window under the "Size File" and "Charge File" fields. If the user neglects to specify a path to a size file, these parameters are assigned automatically using typical CHARMM radii (e.g.

carbon: 2 Å, hydrogen: 0.7 Å, oxygen: 1.7 Å, nitrogen: 1.8 Å, sulfur: 2.0 Å, phosphorous: 2.15 Å). If the user desires different atomic radii, a radii size file must be specified in the input parameters.

3.4 Results

3.4.1 Influenza Hemagglutinin Electrostatics

Explicitly solvated all-atom MD simulations were performed on the hemagglutinin trimer protein for 100 ns using NAMD2.8. Afterwards, the protein snapshots were extracted at 200 ps intervals and aligned. Using DelEnsembleElec, a MD trajectory containing 500 frames of approximately 20,000 atoms required nearly 50 hours to complete using a single Intel Xeon: 2.67 GHz processor, as Delphi must calculate the PB equation for each individual frame of the trajectory. Once complete the Gaussian cube is loaded into VMD, it is represented by the VMD graphical representations isosurface and solvent-accessible surface area colored by loaded data (Figs. 3.2 – 3.4). The grid scale DelPhi parameter was set to 1.0 Å. The exterior dielectric was set to 80.0, with the interior set to 2.0. The salt concentration in the continuum solvent was set to 20 mM; consistent with the concentration in the MD simulations. The probe radius defining the dielectric boundary was 1.4 Å. Since the solvent was a 1:1 monovalent salt distribution, the PB equation was solved linearly by reaching a convergence of less than a 0.0001 kT/e change of potential. For comparison, DelEnsembleElec was also used to perform the same calculation on only

a single snapshot (frame 0, representing the equilibrated hemagglutinin).

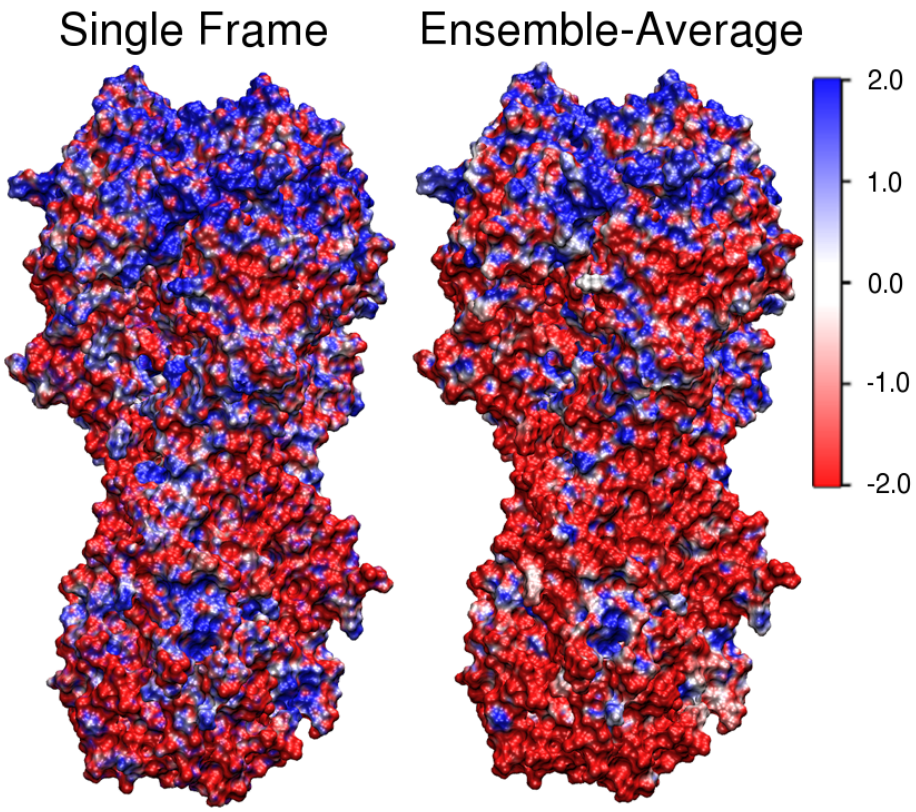

Figure 3.2: Hemagglutinin stalk electrostatics
Side view of the hemagglutinin trimer depicted by the solvent-accessible surface area colored by electrostatic potential calculated on a single frame (left panel) and on the ensemble-averaged 500-frame trajectory (right panel). Units are in $\frac{kT}{e}$ at $T = 300\text{K}$; these units are used in all of the presented results.

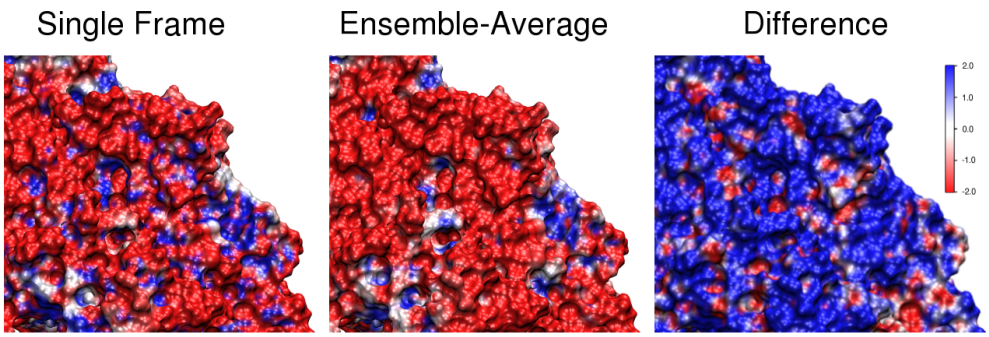

Figure 3.3: Close-up of hemagglutinin stalk electrostatics
Close-up of the hemagglutinin stalk solvent-accessible surface area colored by electrostatic potential. The left panel depicts electrostatics calculated on a single frame. The center panel depicts the same surface for the ensemble-averaged 500-frame trajectory. The right panel depicts the difference in potential as the single-frame potential subtracted from the ensemble-averaged potential. The smoother distribution of surface potential values from the ensemble-averaged electrostatics more accurately represents the effective potential encountered over the trajectory of sampled protein fluctuations. Units are in $\frac{kT}{e}$ at $T = 300\text{K}$.

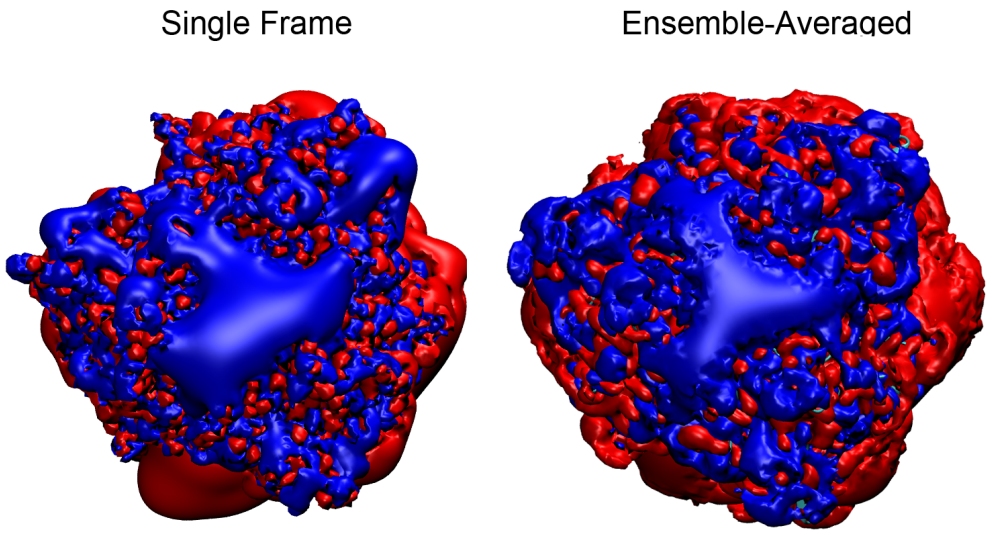

Figure 3.4: Hemagglutinin receptor binding site electrostatics
Top view of hemagglutinin with potential fields at positive (blue) and negative (red) $2\frac{kT}{e}$. The left panel shows the result when the calculation is performed with only a single frame. The right panel shows the result for the ensemble-averaged 500-frame trajectory.

The hemagglutinin glycoprotein contains a long "stalk" region, which anchors the protein to the membrane at one end. The receptor-binding domain, which is the major antigenic area, resides at the opposite end and controls the entry of viral particles through productive binding events between sialic acid receptors on the host cell. In both the single frame and ensemble-averaged electrostatic maps, it is apparent that hemagglutinin exhibits a dipole, with the receptor-binding domain having a more positive potential and the bottom of the stalk having a more negative potential (Figure 3.2). The ensemble-averaged potentials projected onto the surface of the protein indicate an increased polarity, as indicated by the increased regions of

negative (red) charge colored on the surface. In general, the ensemble averaging tends to "blur out" small pockets of variable charges that are present due to specific point charges on the protein surface (Figure 3.3). Such effects are indicated by larger patches of unobstructed electrical charge on the surface of the protein. When the values present in the single-frame grid are subtracted from the ensemble-averaged grid, the resulting "difference map" highlights the dramatic differences between the two (Figure 3.3C). Grid points located near atomic centers have very extreme values in the single frame, and these points are greatly dampened in the ensemble-average. This would account for the large difference between equivalent grid points in the difference map.

The ensemble-averaged electrostatic potentials also indicate a more symmetric electric field, as shown by isopotential values, at the receptor binding domain end of hemagglutinin (Figure 3.4). This dynamic electrostatic information may be of critical value in the rational design of improved vaccines or better understanding the interactions of hemagglutinin with glycan receptors on the surface of the host cell.[21,29]

3.4.2 Influenza Neuraminidase Electrostatics

One hundred nanoseconds of explicitly solvated, all-atom MD simulations of the tetrameric neuraminidase protein were performed using AMBER11. Protein snapshots were again extracted at 200 ps intervals and aligned based on alpha carbons to remove rotational and translational motion. DelEnsembleElec was used to compute the electrostatics for a single frame (frame 0, representing the equilibrated

protein) and the MD trajectory (500 frames). The DelPhi parameters used for neuraminidase were the same as those used for the hemagglutinin system.

The influenza neuraminidase protein controls viral particle exit from the host cell by cleaving the terminal sialic acid linkage on the host cell glycan receptors, and as such, is currently the major target for small molecule antiviral compounds.[6] Although most small molecule drug discovery efforts have focused exclusively on optimizing ligand-protein interactions within the sialic acid binding site, a secondary sialic acid binding site, whose exact function is yet unclear, exists on the periphery of the neuraminidase active site. It was recently shown through Brownian dynamics (BD) simulations that this secondary sialic acid site may affect the association kinetics (rate) of both sialic acid and the current clinically-used drug, oseltamivir (Tamiflu, Roche).[49] The BD calculations utilized a single crystallographic snapshot. Using DelEnsembleElec, it is clear that the mean field electrostatic potential exhibited through the ensemble-based approach substantially affects the surface potential at this secondary site (Figure 3.5). Although it remains to be seen what effects would result by repeating the BD calculations using the ensemble averaged electrostatic potential values as opposed to the single static structure, we hypothesize that the ensemble-based environment would more closely align with the actual electrostatic conditions *in vivo*. Such claims have already been substantiated for other systems through work shown in Ref. [44]. Notably, the general trend of the charge fields in the region of the secondary sialic binding site of neuraminidase can be more clearly identified in the ensemble-averaged view than the single frame view, where the influence of individual atoms obscure the electrical characteristic of the areas of the active site

(Figure 3.5). Looking more globally at the neuraminidase electrostatics, the ensemble-average electrostatic field again becomes more symmetric with the dynamic structural information (Figure 3.6). Interestingly, it also appears to become slightly dampened, as compared to the single frame electrostatics calculation.

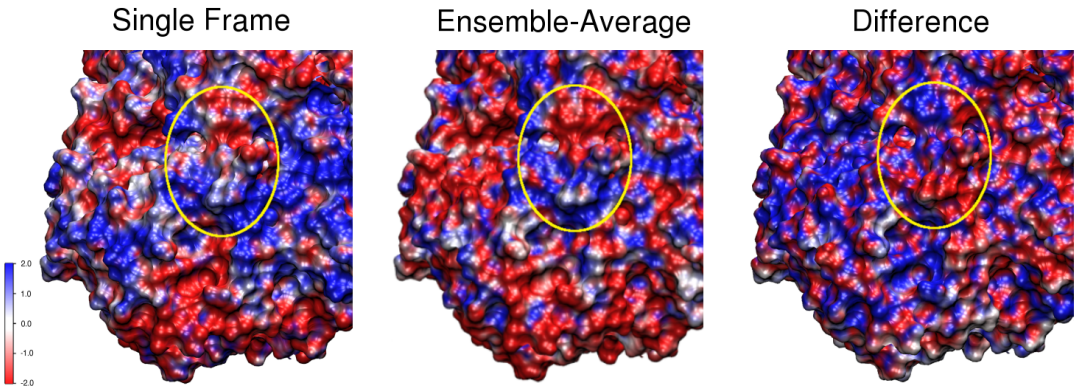

Figure 3.5: Neuraminidase secondary sialic acid binding site electrostatics
Close-up of the secondary sialic acid binding site of neuraminidase[49] represented as the solvent-accessible surface area and colored by electrostatic potential. The left panel depicts the surface area using potentials calculated using a single frame. The center panel depicts the result for the ensemble-averaged 500-frame trajectory. The right panel depicts the difference in potential as the single-frame potential subtracted from the ensemble-averaged potential. Units are in $\frac{kT}{e}$ at $T = 300\text{K}$.

3.5 Discussion

The electrostatics of the two major influenza glycoproteins, hemagglutinin and neuraminidase, are of general interest to public health as they are both vaccine and small molecule drug targets, respectively. In this work, we present ensemble-averaged electrostatics calculations for both proteins using the newly developed DelEnsembleElec plugin for VMD. These ensemble-averaged calculations showcase

the utility of utilizing ensemble-based structural information in adding insight to outstanding biological questions.

Performing ensemble-averaged electrostatics on the hemagglutinin and neuraminidase trajectory with DelEnsembleElec and comparing the results with a single-frame DelPhi run shows that the ensemble-averaged potential surfaces are generally more symmetrical, with areas of consistent charge bias more easily visible. Additionally, the instantaneous locations of the residues in a single frame appear to cause odd shapes in the potential surface (Figs. 3.4 and 3.6, left panels). This effect is alleviated in the ensemble-averaged electrostatic calculations, since random outlier charge values often exist within individual frames of a trajectory. Though some "noise" still remains from these outlier values, including more protein frames in the averaging can potentially decrease such effects. By loading the data into a surface representation of the molecule, it becomes readily apparent that ensemble averaging filters out the influence of individual point charges. Such results provide a clearer indication of the general charge of a specific area on the molecule.

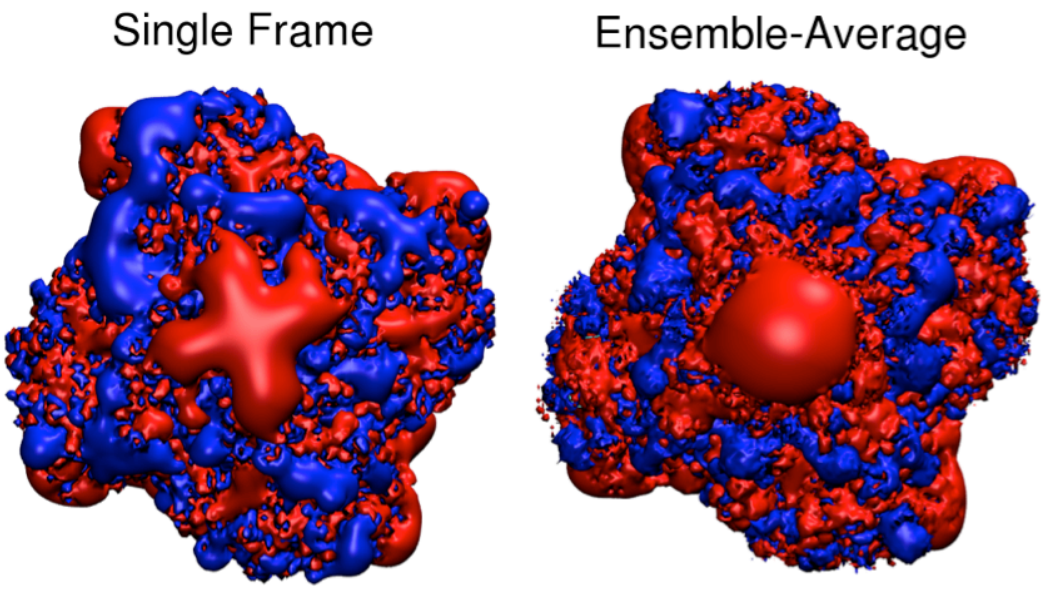

Figure 3.6: Overall electrostatics of neuraminidase
Top view of the neuraminidase tetramer with potential fields at positive (blue) and negative (red) $2\frac{kT}{e}$. The left panel depicts the electrostatics of neuraminidase when calculated using a single frame. The right panel depicts the electrostatics of the ensemble-averaged 500-frame trajectory.

The DelEnsembleElec plugin presented in this work allows interactive and customizable preparation for running ensemble-averaged PB electrostatics using the program DelPhi. It eases the preparation of trajectories and parameters with a graphical user interface, while maintaining much of the functionality of a command-line invocation of DelPhi. The increased accuracy of the PB equation combined with the benefit of ensemble averaging allows great precision in the prediction of a biomolecule's overall electrostatics. Though we only present protein examples here, we have verified that DelEnsembleElec can handle nucleic acid- and lipid-containing systems as well, as long as the charge and radius parameters are available within the loaded structure or additional parameter files are specified.

DelEnsembleElec is freely available under the Gnu Public License. The language, applications, and libraries on which it depends are also freely available. Download instructions and a tutorial can be found at http://amarolab.ics.uci.edu/delensembleelec.html . A link to DelEnsembleElec is provided at the Delphi webpage as well, http://www.ces.clemson.edu/compbio/tools.html .

Chapter 3, in full, is a reprint of "DelEnsembleElec: Computing Ensemble-Averaged Electrostatics Using DelPhi", which was published in 2013 in Communications in Computational Physics, volume 13, issue 1, pages 256-268, by Lane W. Votapka, Luke Czapla, Maxim Zhenirovskyy, and Rommie E. Amaro. The dissertation author was the primary investigator and author of this paper.

Chapter 4: Multistructural Hot Spot Characterization with FTProd

Computational solvent fragment mapping is typically performed on a single structure of a protein to identify and characterize binding sites. However, the simultaneous analysis of several mutant structures or frames of a molecular dynamics simulation may provide more realistic detail about the behavior of the sites. Here we present a plugin for VMD that streamlines the comparison of the binding configurations of several FTMAP-generated structures.

4.1 Introduction

The identification and characterization of ligand binding sites in proteins is of utmost importance for research into drug discovery, and biomolecular function. The experimental determination of regions on the surface of the protein with high recurrence of bound probes correlate well with the locations of drug-binding sites [55]. Interested readers are referred to the following reviews: [56,57]. One popular method for experimental determination of such druggable "hot spots" involves the process of Multiple Solvent Crystal Structures (MSCS) [58,59]. During MSCS, the protein is solvated within various probe compounds. The structure determined using X-ray crystallography indicates probe binding locations.

X-ray crystallization of multiple structures is expensive, and computational fragment mapping can emulate this process to identify binding sites [60]. Various computational methods for binding site identification are compared here: [61]. The FTMAP algorithm [62] seeks to mimic the MSCS method and has been shown to predict the analogous binding of probe molecules with a high degree of success. To gain a comprehensive understanding about the ligand-binding characteristics of a protein, structural knowledge alone is often insufficient. A single structure ignores the dynamics of a protein, which may create variation in probe-binding location, number, and capacity[63].

Here we present FTProd, a program capable of clustering hot spots spanning multiple structures, and allows for the ease of identification and characterization of those hot spots with a Graphical User Interface (GUI). FTProd is a plugin for Visual Molecular Dynamics (VMD)[64], a molecular visualization program free for academic use.

4.2 Methods

FTProd analyzes structures that have been processed with FTMAP, which contain a series of small molecular probes indicating the location of potentially druggable Consensus Sites (CS). When run, FTProd utilizes one of several available cross-structural clustering methods, and are described in detail in the SI section 4.5.

Depending on which method the user specifies, the algorithm selects CSs that are the most spatially similar, grouping them together into a cluster. Several hierarchical clustering methods are implemented in FTProd, as well as the “greedy

clustering" method employed in FTMAP. FTProd can cluster sites within the same structure, but also provides the option to cluster CSs that only exist within separate structures.

FTProd integrates with and utilizes VMD with the goal of providing a smooth, easy-to-use GUI, through which researchers can visualize, identify, and characterize cross-structural hot spots in proteins. Upon running FTProd on loaded and selected structures, the plugin creates a Table widget (Figure 4.1c); which tabulates every structure and CS that exists within its respective structure(s). Upon selecting one or multiple CSs, FTProd draws the relevant site and associated probe fragments in VMD's viewer. Additional FTProd features are detailed in SI section 4.5.

To demonstrate the utility of FTProd, we performed cross-structural analysis over several strains of the influenza neuraminidase (NA) glycoprotein. We chose NA for its well-understood binding sites and relatively high flexibility[63,65]). In this example, average-link agglomerative clustering was used with an inter-CS cutoff of 8.0Å.

4.3 Results

We demonstrate FTProd's ability to characterize and display cross-structural ligand-binding sites by examining four x-ray crystal apo structures of NA obtained from various influenza strains downloaded from the Protein Databank (PDB). The strains we used were PDB id's: 1MWE[66], 2HU0[5], 2HU4[5], and 3NSS[67]. The primary role of NA in influenza pathogenesis is the cleavage of sialic acid after binding to the active site. Another binding site, the secondary sialic acid site, is also partially

responsible for substrate affinity. Depending on the strain, neuraminidase may possess a so-called 150 pocket, a highly variable site[68] which may present a feasible target for drug design efforts. FTProd successfully identifies important binding sites across the structures, ranking them by decreasing predicted binding ability. The sialic acid binding site is correctly identified as the predominant binding location. PDB structure 2HU0 docks more than twice as many probes as the 150 sites in any other structure (Figure 4.1a). This is consistent with the structural understanding of 2HU0, which exhibits an open "150 pocket" [5]. One site identified for 3NSS also corresponds to a location where an acetate ion has been resolved in the 3NSS crystal structure (Figure 4.4). Two additional examples of proteins examined with FTProd are provided in SI section 4.5.

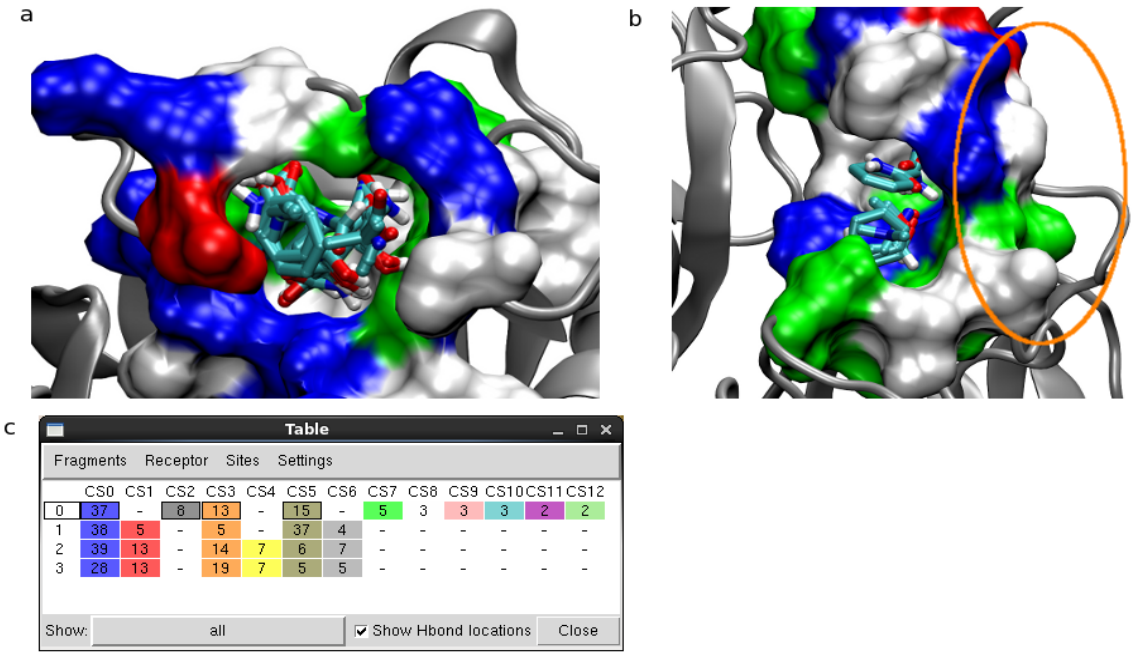

Figure 4.1: FTProd hot spot visualization and interface
(a) Probes docked inside the 150-cavity of structure 2HU0. (b) Probes docked into pocket beneath the secondary sialic acid site (circled in orange) in structure 1MWE. (c) Table widget allows user to select and view multiple cross-structural consensus sites. Surfaces in (a) and (b) colored by residue type: blue indicates positive residue; red, negative; green, polar; white, hydrophobic.

4.4 Discussion

The determination of potentially druggable sites on the surface of a protein represents an area of intense interest to drug discovery and other applications. FTProd provides the capability to compare the characteristics of pockets between crystal structures of structurally similar proteins. The burden is placed on the user to determine whether two structures ought to be compared. RMSD-based clustering of MD trajectories could be one of many methods that may be used to identify input for FTProd; along with binding site similarity, analogous structures, similar substrates, or

any other structural similarity metric. FTProd's utility is completely extensible toward the analysis of the frames of a simulation trajectory, as may be generated by a molecular dynamics simulation. To our knowledge, FTProd is the only existing tool that integrates protein structural dynamics data for the purpose of binding site characterization.

The inclusion of cross structural or dynamic information in the analysis of these "hot spots" is likely to increase the predictive accuracy and scope of these computational methods by providing a more realistic picture of protein activity. Given the high success of the FTMAP algorithm, we expect that FTProd will greatly aid researchers in the analysis of protein pockets by streamlining interstructural consensus site comparison.

FTProd is presented as a plugin for the molecular viewer program VMD, and is freely available under the GNU Public License. The language, applications, and libraries on which it depends are also freely available. Download instructions and a tutorial can be found at http://amarolab.ucsd.edu/ftprod.

4.5 Supplementary Information

4.5.1 Description of FTProd algorithm

Each member of the set of aligned structures to be compared is run through FTMAP (http://ftmap.bu.edu/)[62]. The output .pdb files from FTMAP now contains the original structures as well as the docked probe molecules. The user then loads these files into FTProd. Each consensus site(CS) identified by FTMAP is then clustered according to a method selected by the user. The clustering method options include:

Average-link hierarchical agglomerative, single-link hierarchical agglomerative, complete-link hierarchical agglomerative, and the simple greedy method employed in FTMAP[62,69] and FTSite[69]. Each clustering method is described in detail in the "Clustering Methods" section below. The user has the option to retain FTMAP's CS differentiation by toggling an option that will disable any clustering of CSs within the same structure. Once the clustering is complete, each structure along with its CSs is displayed according to the user's preference. Cross-structural visualization is described in the "FTProd Interface" section below.

4.5.2 Clustering Methods

Hierarchical Agglomerative Clustering[70]

The geometric centroid of each CS in every aligned structure is calculated where each atom of the probe molecules is given equal weight. The distance between each centroid is then measured to every other centroid and the results are placed into the upper echelon of a pairwise centroid distance matrix, where each axis represents the complete set of CSs. If the option to trust FTMAP clusters is enabled, then distances between CSs within the same structure is set to infinity. The smallest value (not along the main diagonal) within the centroid matrix is identified. If the value is below a user-specified CS distance cutoff, then the indices of the row and column correspond to which two CSs will be clustered. The row and column corresponding to both CS are removed and combined, making a cluster. The new, combined centroid cluster is added as a row and column to the centroid matrix, which is updated with new distances to the other CSs according to the specified link method:

Single-link: The distance from the two closest centroids within two clusters.

Average-link: The distance between the average of the centroids within a cluster to the average of the centroids in other clusters.

Complete-link: The distance from the two farthest centroids within two clusters.

This process is iterated until the smallest member of the centroid matrix exceeds the CS distance cutoff.

Greedy Clustering[62]

The most dominant CS is selected along with any other CS that falls within the distance cutoff range. This entire selection becomes the first cluster, and is removed from further consideration. The process is repeated for the most dominant remaining CS and each subsequent cluster.

Each clustering method may provide very different results on a set of structures. The Average-link clustering method is selected by default, since it is most likely to combine CSs into clusters that are regularly shaped. The Average-link method should be sufficient for most applications. However, literature suggests that Average-link and Complete-link may produce misleading results when clustering similar entries due to computational rounding errors and statistical sampling errors[70]. Since FTProd is not intended to process or produce highly robust statistical data, but merely to provide qualitative hints as to the characteristics of a binding site based on solvent probes' predicted global energy minima, we believe the statistical issue is trivial in this case. We also do not anticipate that rounding error will have a

significant effect in this application. Users concerned by these issues should restrict themselves to using the Single-link method. Qualitatively, the Single-link method will tend to produce clusters that are long and thin, because the method is susceptible to 'chaining' of its entries. Conversely, the Complete-link method will produce clusters that tend to be short and wide. The sizes and shapes of the active sites should be taken into account when choosing a proper clustering method. The Greedy method is the fastest and was provided for consistency: it is the method used initially by FTMAP to cluster probes into CSs. The nature of the greedy method causes the most dominant CS to determine the shape and size of all subsequent ones. CSs identified by the greedy method are also likely to be irregularly shaped. Therefore, it is probable that the CSs of two similar structures found using the greedy method will be highly dissimilar. We recommend that users utilize the greedy method only if they are concerned with being consistent by using the same clustering method used in FTMAP.

4.5.3 FTProd Interface and Additional Features

The FTProd algorithm has been implemented as a plugin for the molecular viewing program VMD[64]. Figure 4.2 shows the FTProd interface. The user selects the relevant loaded structures to include with the calculation as well as the CS distance cutoff for the cross-structural CS clustering. The user may also open the Settings window and modify a variety of options, including whether to perform intrastructural CS clustering, the clustering method, the probe-residue cutoff distance for each CS, the method to sort sites in the table, the Gaussian sigma variable for hydrophobic

coloring, the cutoff for grouping atoms as hydrogen bond (H-bond) donors/acceptors, and the factor to classify H-bond combinations.

Table CS sorting options

Four options are available for sorting CSs in the Table window. FTMAP output ranks each probe cluster, and FTProd attempts to retain this ranking as much as possible; since ranked probe clusters from multiple FTMAP output molecules may be included in the same FTProd CS.

Score Sum: The CS is ranked by the sum of all participating probe cluster scores.

Highest Total Score: The CS is ranked by the highest scoring probe cluster that it contains.

Mean Score: The CS is ranked by the mean of all participating probe cluster scores.

Mean of Each Cluster in Highest Score: The CS is ranked using the following process: in the case when a CS contains two or more probe clusters from the same FTMAP output structure, only the score of the highest scoring probe cluster is considered. The mean for all FTMAP probe clusters in the CS is given as the CS score.

Once run, the Table window appears and the graphical representations of the relevant molecule are changed. To make it easy for the user to globally change the representations, various menu options at the top of the Table window allow the user to change the drawing and coloring methods of the probes, receptor, and CSs of the

relevant molecules. Below the menu, a matrix is displayed where each CS, if found within that structure, has been tabulated into the same column as the corresponding CS for each other structure. This allows for easy comparison of cross-structural CSs. Each cell of the matrix also displays the total number of probes for the corresponding CS in its respective structure. The Table window also allows the user to display probe fragments by classification. The classification options include hydrophobic, polar, aromatic, positive, H-bond donor, and H-bond acceptor probes. The fragments, their abbreviations, and their respective classifications are listed in Table 4.1. Any residue is considered a member of a CS if it sits within the probe-residue cutoff of a probe in any structure that contains that CS.

FTProd also allows the option to display locations where two or more H-bond contributing groups overlap. If activated, an icosahedron will be drawn at any H-bond contributing groups within the H-bond grouping cutoff of each other. If the proportion of H-bond accepting groups exceeds the H-bond combination factor specified in the Settings window, the entire icosahedron will be colored red. Conversely, if the proportion of H-bond donating groups exceeds the H-bond combination factor, the icosahedron will be colored blue. An H-bond accepting or donating proportion in between these values will have an icosahedron with alternative facets colored blue and red. The H-bond accepting and donating groups of each probe are listed in Tables 4.2 and 4.3.

The user may also color the probes by hydrophobicity. The metric for hydrophobicity is determined by generating a summation of Gaussian functions centered at every hydrophobic probe atom as summarized in Table 4.4 with a sigma

value with a default of 1.0 Å. The value of this summation is calculated for every probe atom and is stored in each probe atom's user3 variable; a field that VMD reserves for user-defined values. An external program, fast_gaussian, has been developed which performs the Gaussian summation quickly, and if present in the same directory as the FTProd script, the hydrophobic calculations will complete much quicker. If fast_gaussian is not present, FTProd will calculate the Gaussian summation at a lower precision with TCL scripts, which, despite the lesser precision, takes significantly more computation time.

4.5.4 Supplementary Results

In addition to neuraminidase, we also used FTProd to examine RNA editing ligase 1 (REL1) and RET2 proteins, both of which are essential enzymes in Trypanosoma brucei, the causative agent for African sleeping sickness.

4.5.4.1 REL1

The top three RMSD clusters of the frames of a molecular dynamics (MD) simulation prepared according to reference: [71-73] and were run through FTMAP and visualized with VMD and FTProd.

The REL1 X-ray crystal structure resolved with bound ATP (PDB ID: 1XDN)[72] shows highly similar functional group interaction with FTMAP probes docked into the predominant cluster. Figure 4.4A shows ATP bound to the binding site of REL1 and interactions according to[71]. Similarly, Figure 4.4B shows aromatic probes bound to the active site of the predominant RMSD cluster as predicted by

FTMAP. The location of these aromatic probes correspond highly with the placement of the purine group of the bound ATP, which makes hydrophobic and aromatic π-π interactions with PHE 209. Similarly, Figure 4.4C shows the polar FTMAP probes making similar interactions as the ATP in the crystal structure with residues ARG111, ASN92, GLU86, and ARG288.

The third CS is a previously unresolved site that opens during the MD simulation of [71,73],who anticipated that it may provide a new opportunity for drug design & discovery. The site remains open in all three clusters and does not undergo significant structural changes. The FTMAP probes bound to this site can offer hints to the characteristics of a drug that may bind and inhibit REL1 activity. The characteristics of the site and bound probes can be easily compared with FTProd.

In all three clusters, FTMAP has placed hydrophobic and aromatic probes with similar orientations in the center of the site(Figure 4.5). In an examination of polar probes, many hydrogen bonds also exist between the probes and the residues of the active site. However, with the exception of a possible interaction between the probes and the sidechain of ASN92 in clusters 1 & 2, all polar interactions are between the probes and backbone functional groups. This insight suggests that it may be difficult to design a drug with high specificity against this site, because analogous proteins in humans have a similar backbone configuration[74]

4.5.4.2 RET2

RMSD cluster centroids with a minimum difference of 1.75 Å were extracted from an MD simulation on the apo RET2 structure[75], and the top three were selected

for examination with FTProd. Similarly, the top 3 RMSD cluster centroids were also extracted from an MD simulation of the RET2 structure using an RMSD cutoff of 1.55 Å solved in a solution containing 2 mM UTP (PDB ID: 2B51)[76]. All non-protein components were extracted and the six structures along with the original UTP-bound crystal structure were run through FTMAP. All these were clustered with the average-link method with a cutoff of 8.0Å and compared simultaneously using FTProd (Figure 4.6).

One of the most apparent results that FTProd indicates is that a very predominant binding site exists solely in the third apo cluster centroid (Figure 4.7). The fact that it appears only once in the third cluster would indicate that this site appears relatively infrequently during only the apo simulation, and is rare or absent from the UTP simulations and the crystal structure.

CS1 and CS2, however, exist in all seven structures. Comparison of the size, shape, and probe bindings of these CSs between the seven structures shows that the site remains largely unchanged between them all. This is consistent with the function of the CS, as this is where the single strand RNA (ssRNA) would bind to RET2 for modification. In UTP-bound crystal structure, which served as the initial structure for the UTP-bound MD simulation providing three of the clusters, two of the UTPs are bound in each of these CSs(Figure 4.8).

FTProd also highlights other similarities and differences between the CSs found for each type of simulation. For instance, CS3, another region of the active site, also exists across all structures. CS4 is only found in the crystal structure and one of

the apo simulation clusters, but is absent from the UTP-bound simulation. Similarly CS5 appears only in the apo simulations, CS6, CS7 and CS9 are only found in the UTP-bound simulations. CS8 and CS10 are found in both the apo simulations as well as the crystal structure. A comprehensive analysis of RET2 using FTMAP can be found here:[75].

4.5.5 Supplementary Discussion

Although FTProd aids the determination of characteristics and transience of CSs across multiple structures, the identification of the CSs must be pre-determined by FTMAP or, potentially, another program. The quality of FTProd's output is highly dependent on the quality of FTMAP's output. Many studies of FTMAP in its relation to other methods have been performed[56,61]. While a full comparison of FTProd & FTMAP with the multitude of available tools is beyond the scope of this paper, we believe that FTProd will greatly serve the drug discovery and computational biophysics community because, to our knowledge, it is the only available tool that allows the incorporation of dynamics and multistructural data in the characterization of protein binding sites and hot spots. We want to accentuate the fact that FTProd is not a new hot-spot identifier, but merely builds off the hot-spot identifying power of FTMAP, and merely allows users to easily compare FTMAP output on multiple structures and to help in its interpretation.

The quality of FTProd's output is also highly dependent on the employed clustering and scoring methods. Because of the diversity of binding sites and the qualitative nature of their characterization, the analysis of CSs can be a bit of an art.

There is no hard-and-fast clustering or scoring method that will give the "correct" results every time. For this reason, we have provided multiple clustering and scoring options with which the user may experiment. We also provide a great deal of flexibility and streamlining in the visualization of the receptor and its CSs to make analysis quick and easy for the user.

Table 4.1: Probe fragments and their classifications

Probe Abbrev.	Probe Name	Classifications
ACD	Acetamide	Polar; H-bond donor; H-bond acceptor
ACN	Acetonitrile	Polar; H-bond acceptor
ACT	Acetone	Polar; H-bond acceptor
ADY	Acetaldehyde	Polar; H-bond acceptor
AMN	Methylamine	Polar; positive; H-bond donor; H-bond acceptor
BDY	Benzaldehyde	Polar; aromatic; H-bond acceptor
BEN	Benzene	Hydrophobic; aromatic
BUT	Isobutanol	Polar; H-bond donor; H-bond acceptor
CHX	Cyclohexane	Hydrophobic
DFO	N,N-dimethylforamide	Polar; H-bond acceptor
DME	Dimethyl ether	Polar; H-bond acceptor
EOL	Ethanol	Polar; H-bond donor; H-bond acceptor
ETH	Ethane	Hydrophobic
PHN	Phenol	Polar; aromatic; H-bond donor; H-bond acceptor
THS	Isopropanol	Polar; H-bond acceptor
URE	Urea	Polar; positive; H-bond donor; H-bond acceptor

Table 4.2: Hydrogen bond accepting groups of probe molecules

Probe Abbrev.	Accepting atoms
ACD	O4
ACN	N
ACT	O
ADY	O1
AMN	N1
BDY	O1
BUT	O
DFO	O4
DME	O1
EOL	O
PHN	O
THS	OT
URE	O

Table 4.3: Hydrogen bond donating groups of probe molecules

Probe Abbrev.	Donating Atoms
ACD	N1
AMN	N1
BUT	O
EOL	O
PHN	O
THS	OT
URE	N1; N2

Table 4.4: Hydrophobic atoms of probe molecules

Probe Abbrev.	Hydrophobic Atoms
ACD	C3
ACN	C1
ACT	CH1; CH2
ADY	C1
AMN	C1
BDY	C1; C2; C3; C4; C5; C6
BEN	CG; CD1; CD2; CE1; CE2; CZ
BUT	C; C1; C2; C3
CHX	C1; C2; C3; C4; C5; C6
DFO	C1; C3
DME	C1; C2
EOL	C1; C2
ETH	C1; C2
PHN	C1; C2; C3; C4; C5; C6
THS	CH; CH1; CH2

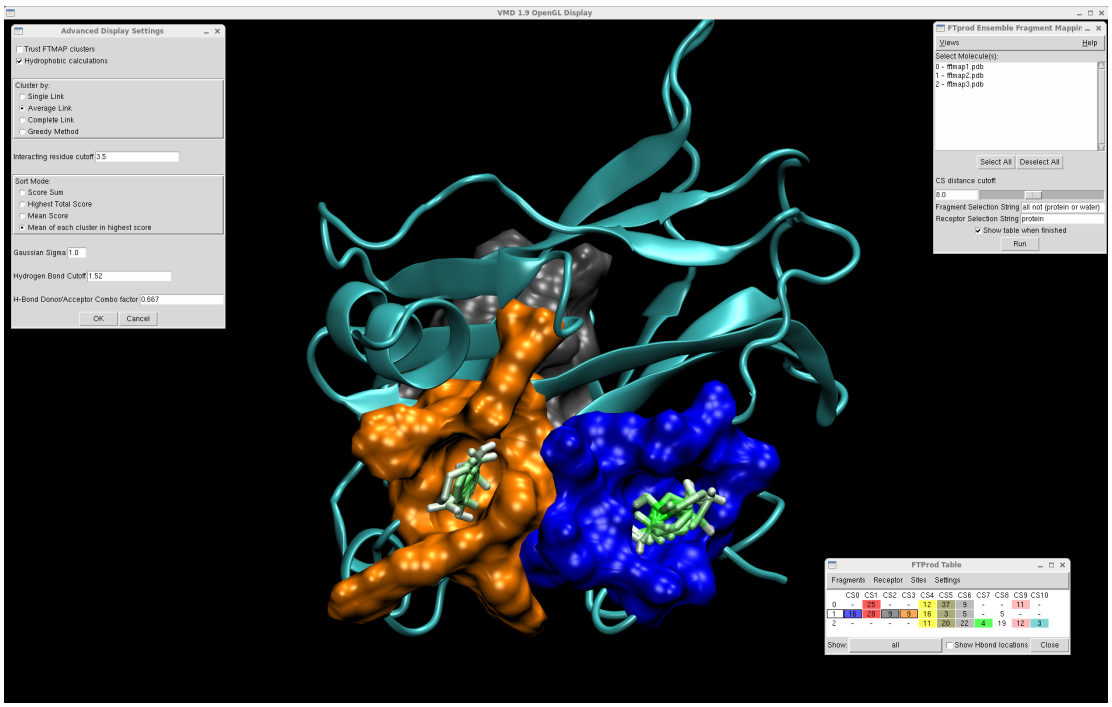

Figure 4.2: FTProd Interface

FTProd interface with Main Menu window (top right), Settings window (top left), and Table window (bottom left). The protein visualized in NewCartoon representation is a frame from a molecular dynamics simulation of p53. Consensus sites are represented as an MSMS[77] surface and the probes are depicted with the licorice representation and are colored by hydrophobicity.

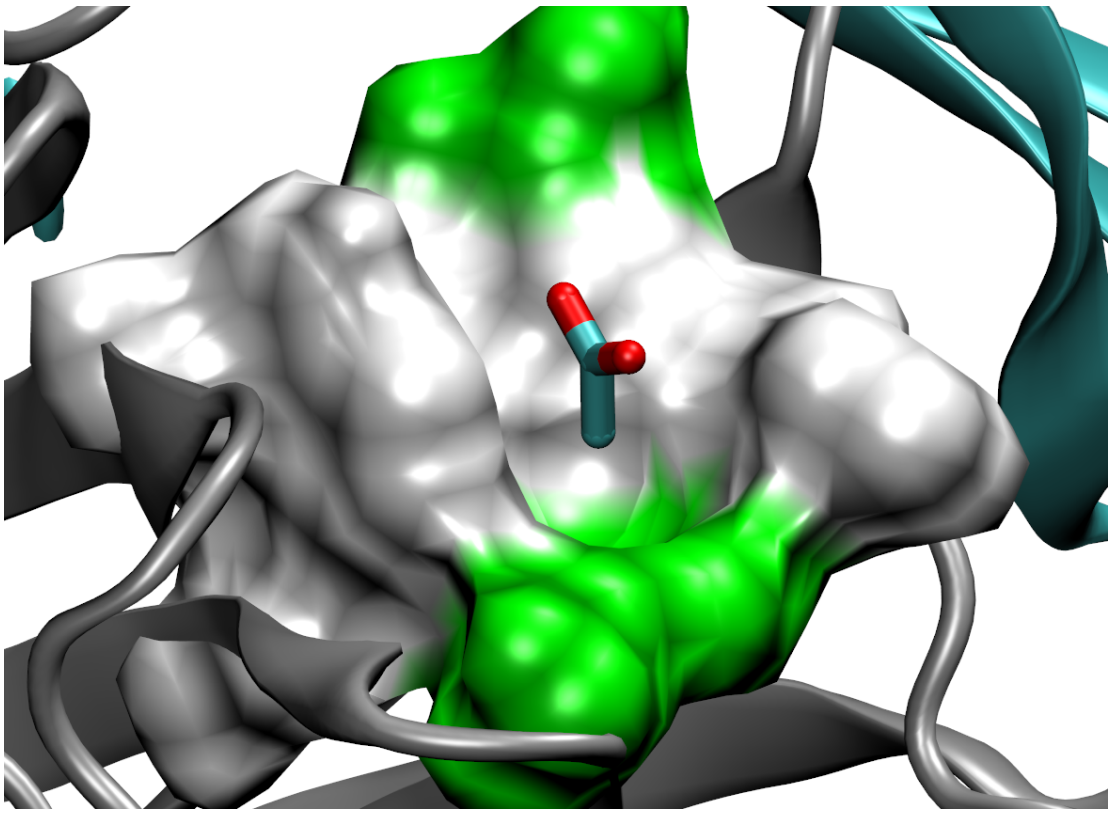

Figure 4.3: Resolved acetate ion located within the X-ray crystal structure of 2009 N1

Structure isolated from epidemic influenza virus[67] (PDB ID: 3NSS). FTMAP identified this site as the second most predominant CS on the surface of the protein. The CS surface is colored by residue type: green indicates a polar residue; white, hydrophobic.

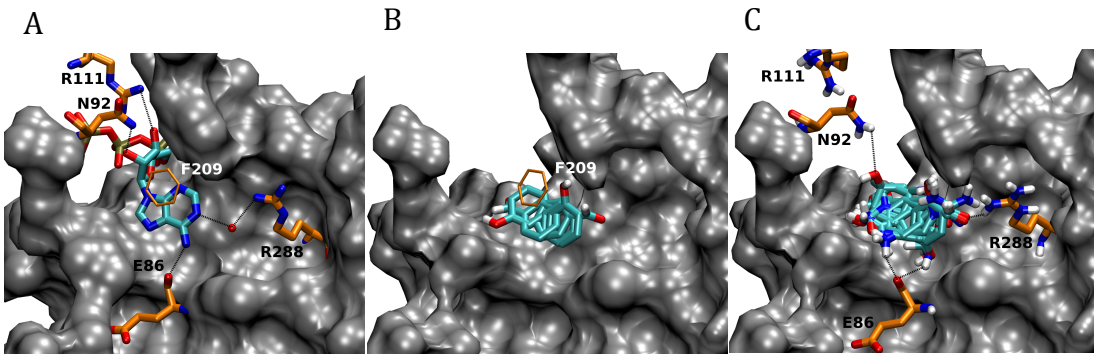

Figure 4.4: Probe clusters bound to REL1 active site

A) X-ray crystal structure of REL1, PDB ID: 1XDN (ref). Bound ATP shown in licorice representation with cyan-colored carbon atoms. Residues ARG111, ASN92, GLU86, and ARG288 are shown in licorice with orange-colored carbon atoms. Aromatic carbon ring of PHE 209 also shown in thinner licorice representation. B) Cluster 1 with aromatic FTMAP probes interacting with PHE 209. C) Cluster 1 with polar FTMAP probes interacting with residues ARG111, ASN92, GLU86, and ARG288.

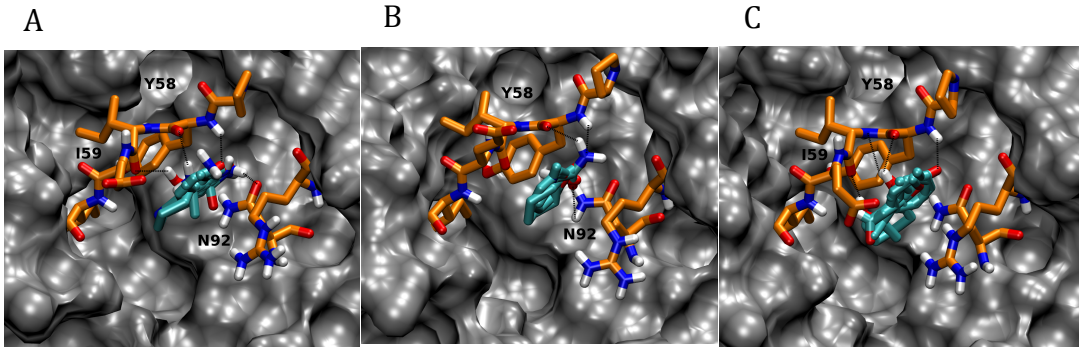

Figure 4.5: Top three clusters for REL1 showing third CS in licorice representations with orange carbon atoms

FTMAP probes colored with cyan carbon atoms. All polar interactions except hydrogen bonds in panels A & B with ASN92 are made with backbone functional groups.

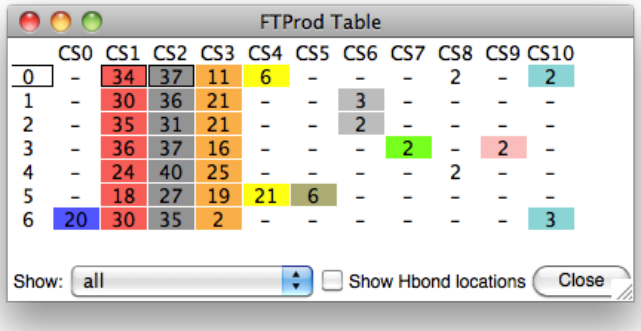

Figure 4.6: FTProd Table window showing RET2 CSs clustered using the average-link method with an 8.0Å cutoff

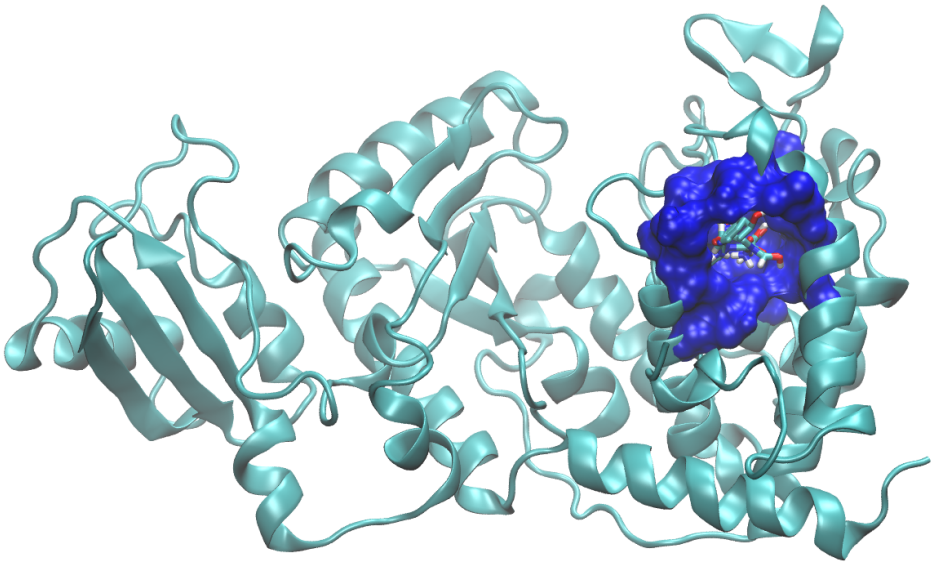

Figure 4.7: RET2 structure from third cluster of apo simulation

The CS ranked highest by FTProd is a large pocket (blue) that opened during this simulation. The pocket is closed in the crystal structure and other top cluster structures.

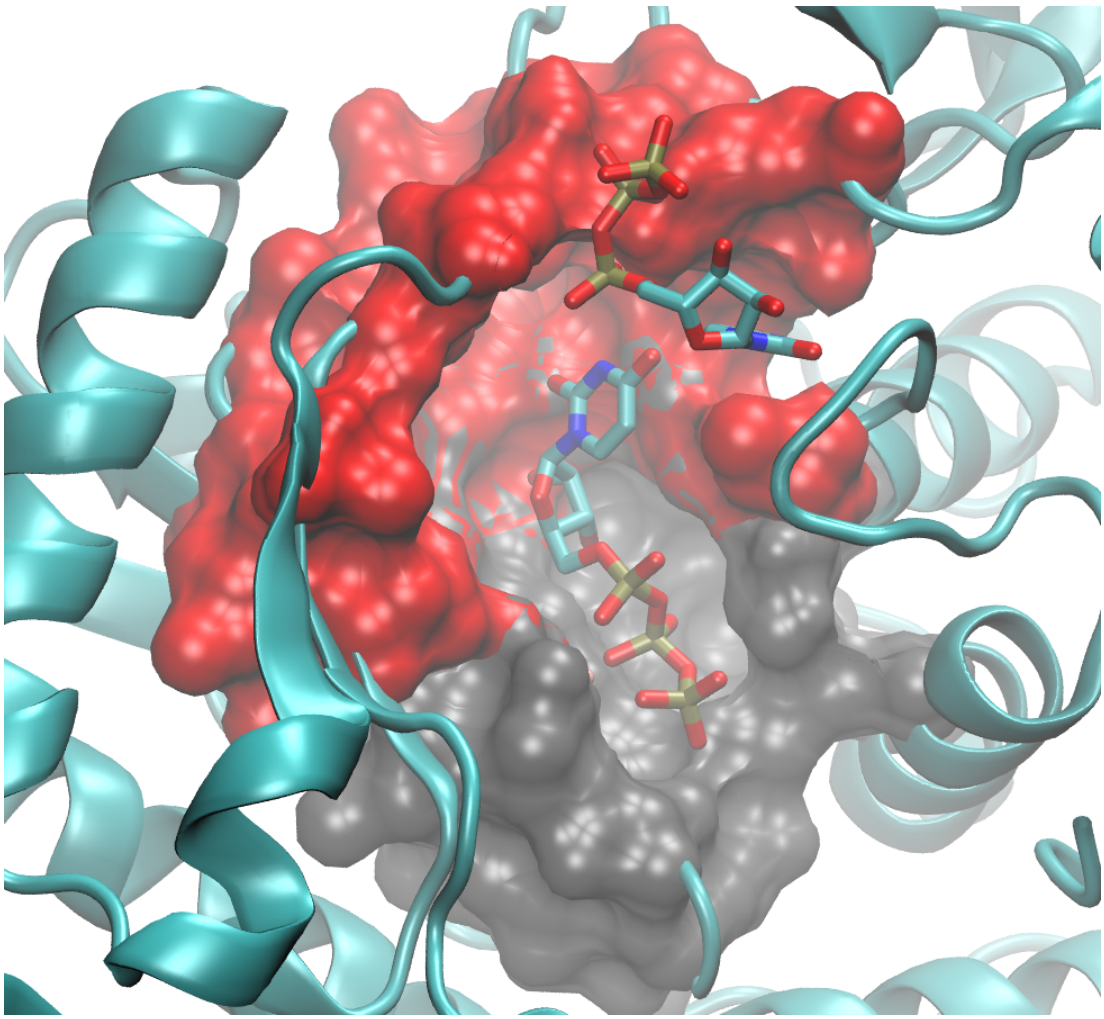

Figure 4.8: RET2 crystal structure with UTP binding sites
Sites shown in red and grey surface representation. These sites were identified using FTMAP and FTProd. The location of the two CSs correlates well with the locations of the resolved UTP molecules (shown in licorice representation) in the crystal structure.

Chapter 4, in full, is a reprint of "Multistructural Hot Spot Characterization with FTProd", which was published in 2013 in Oxford Bioinformatics, volume 29, issue 3, pages 393-394, by Lane W. Votapka and Rommie E. Amaro. The dissertation author was the primary investigator and author of this paper.

Chapter 5: Variable Ligand- and Receptor-Binding Hot Spots in Key Strains of Influenza Neuraminidase

Influenza A continues to be a major public health concern due to its ability to cause epidemic and pandemic disease outbreaks in humans. Computational investigations of structural dynamics of the major influenza glycoproteins, especially the neuraminidase (NA) enzyme, are able to provide key insights beyond what is currently accessible with standard experimental techniques. In particular, all-atom molecular dynamics simulations reveal the varying degrees of flexibility for such enzymes. Here we present an analysis of the relative flexibility of the ligand- and receptor-binding area of three key strains of influenza A: highly pathogenic H5N1, the 2009 pandemic H1N1, and a human N2 strain. Through computational solvent mapping, we investigate the various ligand- and receptor-binding "hot spots" that exist on the surface of NA which interacts with both sialic acid receptors on the host cells and antiviral drugs. This analysis suggests that the variable cavities found in the different strains and their corresponding capacities to bind ligand functional groups may play an important role in the ability of NA to form competent reaction encounter complexes with other species of interest, including antiviral drugs, sialic acid receptors on the host cell surface, and the sister protein hemagglutinin. Such considerations may be especially useful for the prediction of how such complexes form and with what binding capacity.

5.1 Introduction

The influenza A virus is a persistent public health threat that has the potential to cause human disease through both epidemic and pandemic events. The two major glycoproteins on the surface of the influenza virus particle, neuraminidase (NA) and hemagglutinin (HA), have been well studied due to their involvement as major antiviral and vaccine targets, respectively. Yet, despite decades of investigation, many intriguing questions related to their structural dynamics and biophysical interactions during infection and treatment remain unanswered.

X-ray crystallographic structures provide critical information regarding the three dimensional structure(s) of the neuraminidase enzyme. However, they typically only provide one average snapshot of the protein among the ensemble of possible substrates that may be sampled. NA in particular has been shown to be an extraordinarily flexible enzyme, especially in the 150- and 430-loop regions [5,10,47]. Although these two loops are not directly within the active site (i.e. where sialic acid (SA), the natural substrate of NA, binds and is cleaved), they line two other potentially important locations: the 150-cavity and the secondary sialic acid binding site. The composite "binding face" of NA is therefore comprised of the SA, secondary SA, and the 150- and 430-cavities (Figure 5.1). This region of NA is believed to complex with HA, host cell receptors that contain its natural substrate, and small molecule drugs. Four copies of the binding face would be exposed to incoming binding partners due to the tetramer oligomerization state of NA.

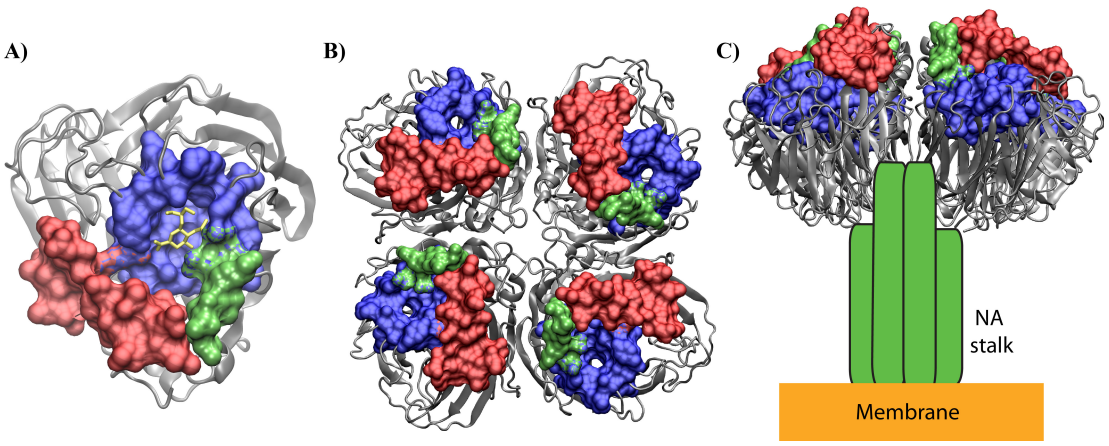

Figure 5.1: NA binding face
A) The neuraminidase binding face is shown for a single monomer, with oseltamivir (yellow) bound in the active site. Blue indicates sialic acid binding pocket, green indicates 150-loop region, red indicates secondary sialic acid binding and 430-cavities. B) The tetramer configuration of the NA binding face, shown looking down at the top of the NA protein. C) Side view of the tetramer, with stalk region (green) and membrane (orange) indicated in cartoon.

The 150-cavity adjacent to the SA binding site has been recently shown to be accessible to small molecule compounds [7]. As some strains of NA, in particular VN04N1 and the 2009 pandemic H1N1, have a high predominance of this cavity in their native structural ensemble, it has been proposed that this area may be targeted in the development of drugs that preferentially target those strains. This could be especially important for drug-resistant strains of N1, which periodically emerge in the population and are a constant public health threat. A more complete understanding of the binding site pockets available to different strains of NA may provide key strategic insights into the development of such compounds. Part of this understanding is evaluating the binding capacity of such newly revealed sites, which may be predictively assessed through computational solvent mapping experiments. In this work we employ FTMAP [62] to carry out the solvent mapping experiments. Favorable

binding regions of small organic probe molecules are determined *via* the following steps: (1) rigid body fragment docking using a fast Fourier transform approach, (2) minimization and rescoring of fragment-protein complexes, (3) clustering and ranking of low-energy fragment-protein complexes, (4) consensus site determination. Populated consensus sites found by FTMap, and its predecessor CSMap [78], have been shown to agree with ligand binding sites identified using experimental methods [62,79,80].

In this work, we evaluate and compare the relative flexibility of three different N1 and N2 strains (A/Tokyo/3/67, A/Vietnam/1204/04, and A/California/04/2009) through all-atom molecular dynamics (MD) simulations, and investigate how such flexibility affects the binding site capabilities in different regions of the NA binding face, with particular emphasis on the 150-cavity and secondary sialic acid binding site area. Structural clustering performed on the residues that line the NA binding face provides information regarding the relative flexibility of this region among the strains. Subsequent computational solvent mapping experiments assess the capacity of these regions to bind to various interaction partners for NA, including host cell receptors, sialic acid molecules, drugs, and potentially HA.

5.2 Materials and Methods

5.2.1 System Setup

The system setup was performed as follows for all simulated systems (Table 5.1). Atomic coordinates were taken from 2HTY for A/Vietnam/1203/04 (VN04N1)[5], 3NSS for A/California/04/2009 (09N1) [9], and 1NN2 for A/Aichi/3/67 (N2) [13].

Protonation states for histidines and other titratable groups were determined at pH 6.5 by the PDB2PQR [19,81] web server using PROPKA [20] and manually verified. All crystallographically resolved water molecules and calcium ions were retained where possible. The system was setup using the program xLEaP from AMBERTools 1.5 [82] using the AMBER99SB force field [22]. Disulphide bonds were enforced using the CYX residue notation in AMBER with the S-S bonds manually added in xLEaP. Each system was solvated in a orthorhombic box containing sufficient TIP3P [83] water molecules to provide a minimum distance of 10 Å between any solute atom and the edge of the box. Each system was neutralized by addition of Na+ or Cl- counter ions as appropriate and then additional Na+ and Cl- ions were added to reproduce experimental assay conditions of 20 mM NaCl.

Table 5.1: Description of simulated systems
System name, crystal structure PDB identifier, human strain isolate description, total simulation time for tetramer simulation, and total number of atoms, including solvent, are shown for each system.

System Name	Crystal Structure	Strain	Simulation Length (ns)	No. of Atoms
N2	1nn2	A/Tokyo/3/67	100	133,049
VN04N1	2hty	A/Vietnam/1203/04	100	112,311
09N1	3nss	A/California/04/2009	100	165,171

5.2.2 Molecular Dynamics Simulations

MD simulations were performed using a version of the PMEMD module from AMBER 11 that was custom performance tuned for each specific simulation and the NICS Cray XT4 and SDSC Trestles supercomputers by SDSC under the NSF's TeraGrid and XSEDE Advanced User Support Programs. Each complex was minimized and equilibrated as follows: steric clashes caused by the addition of

hydrogen atoms, water and ions were alleviated prior to performing molecular dynamics by minimization in a series of stages. Harmonic restraints, with an initial 5 kcal mol^{-1} $Å^{-2}$ force constant, on all non-hydrogen protein atoms, were slowly reduced over ~40,000 combined steepest descent and conjugate gradient minimization steps.

Following minimization, the systems were linearly heated to 310 K in the canonical NVT ensemble (constant number of particles, N; constant volume, V; constant temperature, T) using a Langevin thermostat, with a collision frequency of 5.0 ps^{-1}, and harmonic restraints of 4 kcal mol^{-1} $Å^{-2}$ on the backbone atoms. This was followed by three sequential 250 ps long runs at 310 K in the NPT ensemble, in which the restraint force constant was reduced by 1 kcal mol^{-1} $Å^{-2}$ each run. The pressure was controlled using a Berendsen barostat [23] with a coupling constant of 1 ps and a target pressure of 1 atm. A final equilibration was carried out with 250 ps of NPT dynamics at 310 K without restraints and a Langevin collision frequency of 2 ps^{-1}. Production runs of 100 ns were then conducted in the NVT ensemble at 310 K. As with the heating, the temperature was controlled with a Langevin thermostat (but with a 1.0 ps^{-1} collision frequency). The time step used for all stages was 2 fs and all bonds to hydrogen atoms were constrained using the SHAKE algorithm [24]. Long-range electrostatics were included on every step using the Particle Mesh Ewald algorithm [25] with a 4th order B-spline interpolation, a grid spacing of < 1.0 Å, and a direct space cutoff of 8 Å. For all trajectories, the random number stream was seeded using the wall clock time in microseconds. The production trajectories for each monomer of the

tetramers were extracted and concatenated to approximate 400 ns of monomer sampling.

5.2.3 Clustering

RMSD-based clustering was performed identically for each strain using the clustering algorithms implemented in the rmsdmat2 and cluster2 programs of the GROMOS++ analysis software [26]. Tetramer conformations were sampled at 200 ps intervals yielding a total of 500 conformations. Monomer conformations were then concatenated together, giving 2,000 conformations for each strain. To remove external translation and rotation, an alpha-carbon atom RMSD alignment to the first sampled conformation of chain A was performed for each sampled monomer conformation. Following alignment, clustering was carried out using the GROMOS++ clustering algorithm [27], implemented in GROMACS, using a cutoff of 2.2 Å [28,84] on the alpha carbon atoms of the following 70 binding-site residues: 117 to 119, 133 to 138, 146 to 152, 156, 178 to 180, 196 to 200, 223 to 227, 243 to 247, 276 to 278, 293 to 295, 325, 346 to 350, 368 to 371, 403 to 406, and 426 to 441.

5.2.4 Computational Fragment Mapping

Computational fragment mapping was performed on the cluster centroids of each strain, using the free FTmap web service (FT-Map http://ftmap.bu.edu) [62]. The resulting output includes the structures with ranked consensus sites (CSs). An automated algorithm to assess the differences among the ensemble of structures was implemented in a script that takes these CSs as input. Overlapping CSs were grouped

with increasing distance until reaching a user-specified cutoff. Initially, the CSs in each frame are placed along the axes of a two-dimensional, symmetric matrix. The cells of the matrix are populated with the measured distance between the centroids of each CS. The lowest value in the matrix not located in the main diagonal is iteratively extracted, thereby identifying the two closest CSs. Rows that contain each of these CSs are merged, as are the columns. Afterwards, new distances are calculated between the centroid of the new CS and the rest of the CSs. This process iterates until the lowest distance in the matrix is higher than the user-specified cutoff. The columns of the matrix correspond to non-overlapping CSs. These CSs are ranked in accordance with the ranking provided by FT-map.

5.3 Results

5.3.1 Different Strains of N1 and N2 Exhibit Varying Degrees of Flexibility

The three strains studied here: N2, VN04N1, and 09N1 (Table 5.1), exhibit varying degrees of flexibility in the region constituting the NA binding face (Figure 5.1). RMSD-based clustering on all of the atoms lining the binding face region was utilized as an indicator of flexibility. The results of this analysis indicate that the N2 strain was the most flexible, with a total of 17 clusters required to represent the structural ensemble sampled. 09N1 was the second most flexible, with 12 clusters, and the apo VN04N1 strain was the least flexible overall, requiring only 8 clusters.

5.3.2 Hot Spots in N2

The N2 strain is among the most flexible based on the clustering algorithm, and exhibits three predominant clusters that represent 44%, 18%, and 16% of the sampled ensemble, respectively. These clusters have variable hot spots that highlight how such flexibility can impact ligand and receptor binding (Figure 5.2). In the most predominant cluster, although the 150-loop is in the closed conformation (indicated in green in Figure 5.2), there are shallow ligand-binding hot spots that persist.

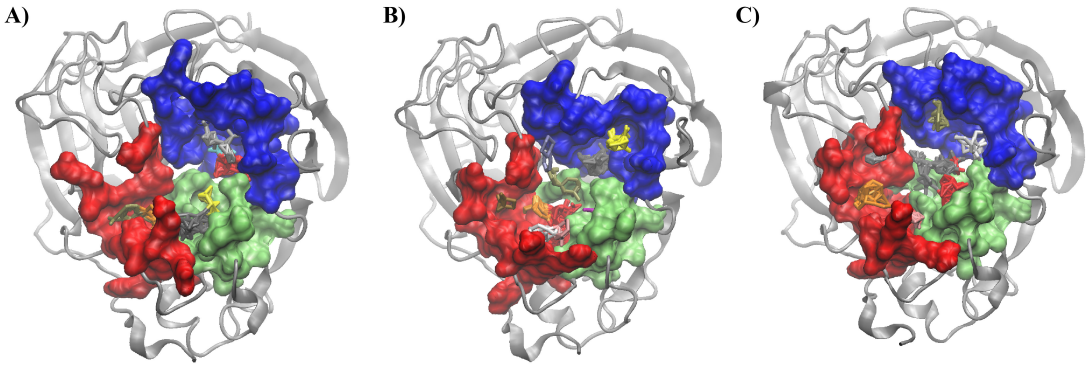

Figure 5.2: N2 hot spots
Ligand- and receptor-binding hot spots are shown for the sialic acid cavity (blue), 150-cavity (green), and secondary sialic acid and 430-cavities (red) for the most predominant (A), second most predominant (B), and third most predominant (C) ensemble structures. Clusters of organic probes indicating actual hot spot / probe binding locations are shown in various colors in stick representation.

5.3.3 Hot Spots in VN04N1

The VN04N1 strain shows remarkable rigidity in the overall NA binding face region, as compared to both N2 and 09N1. In thhe dominant first cluster of this strain, the 150-loop is in an open conformation; 88% of the ensemble falls into this first cluster (Table 5.2). This open cluster exhibits ligand and receptor binding hot spots in the secondary SA site, SA site, and the 150-loop region impling many

favorable ligand binding hot spots. Cluster 2, which represents only ~5% of the structural ensemble, has a closed 150-loop and therefore no cavity or ligand-binding hot spot is exhibited, in marked contrast to the predominant conformation. Cluster 3, representing ~4% of the trajectory ensemble, lacks a hot spot area in the lower 430-cavity, but exhibits hot spots in both the 150-cavity and the secondary SA binding site.

Table 5.2: Cluster results from molecular dynamics simulations
RMSD-based clustering results are presented for each system. System name, number of clusters total, and the individual percentages of each cluster are listed. Highlighted cluster representative structures are depicted in the accompanying figures.

System Name	N2	VN04N1	09N1
No. of Clusters	17	8	12
Cl1 %	44.35	87.85	41.3
Cl2 %	17.7	4.9	28.8
Cl3 %	16.35	4.15	17.45
Cl4 %	8.7	1.45	6.2
Cl5 %	5.2	1.3	2.15
Cl6 %	3.9	0.2	1.7
Cl7 %	0.75	0.1	0.9
Cl8 %	0.7	0.05	0.55
Cl9 %	0.5	//	0.3
Cl10 %	0.5	//	0.25
Cl11 %	0.5	//	0.25
Cl12 %	0.25	//	0.015
Cl13 %	0.2	//	//
Cl14 %	0.15	//	//
Cl15 %	0.15	//	//
Cl16 %	0.05	//	//
Cl17 %	0.05	//	//

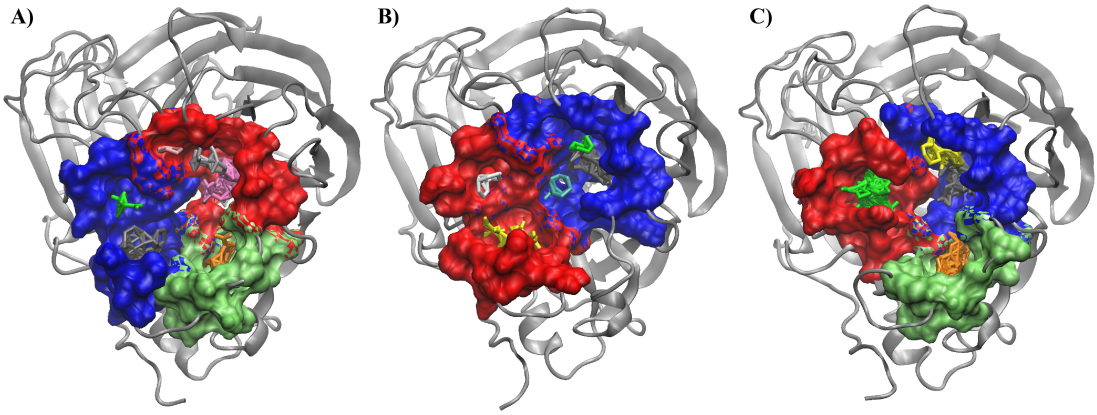

Figure 5.3: 2004 Vietnam highly pathogenic H5N1 (VN04N1) hot spots
Ligand- and receptor-binding hot spots are shown for the sialic acid cavity (blue), 150-cavity (green), and secondary sialic acid and 430-cavities (red) for the most predominant (A), second most predominant (B), and third most predominant (C) ensemble structures of the VN04N1 strain. Clusters of organic probes indicating actual hot spot / probe binding locations are shown in various colors in stick representation.

5.3.4 Hot Spots in 09N1

The 2009 pandemic H1N1 strain exhibits an intermediate degree of flexibility, as indicated by the total number of clusters (Table 5.2). Hot spots from the first three clusters, which represent 41%, 29%, and 17%, respectively, are presented in figure 5.4. While this strain exhibited a closed 150-loop configuration in the crystal structure, all-atom MD simulations indicated that it would open to adopt a loop conformation similar to the VN04N1 strain [52]. In the most predominant cluster conformation, there is no hot spot in the 150-loop region.

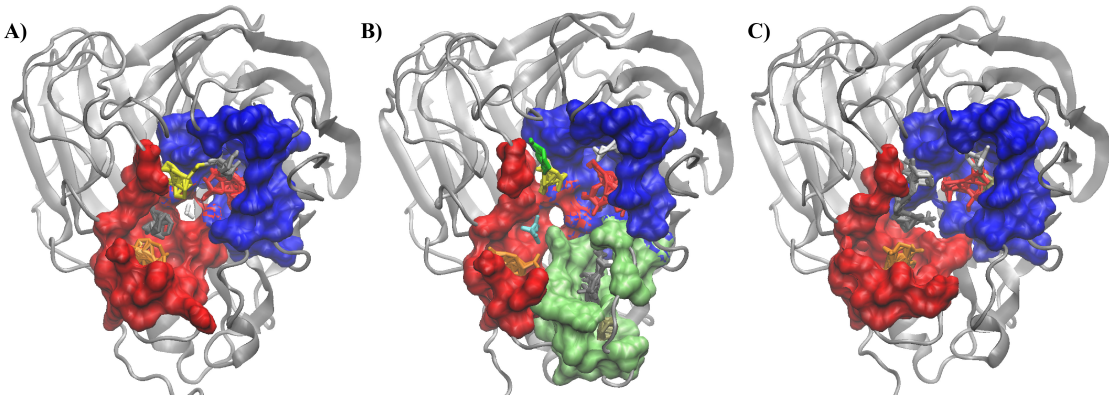

Figure 5.4: 2009 pandemic H1N1 (09N1) hot spots
Ligand- and receptor-binding hot spots are shown for the sialic acid cavity (blue), 150-cavity (green), and secondary sialic acid and 430-cavities (red) for the most predominant (A), second most predominant (B), and third most predominant (C) ensemble structures of the 09N1 strain. Clusters of organic probes indicating actual hot spot / probe binding locations are shown in various colors in stick representation.

5.4 Discussion

The current study, in which we analyze the structural dynamics and ligand-binding capacities of the NA binding face, presents several novel insights regarding recognition events in NA. All of the dominant conformations for N2, VN04N1, and 09N1 exhibit 150-cavity hot spots. However, there are subtle differences in their exact position relative to each other. Despite the fact that the 150-loop of the N2 strain does open in a small fraction of the trajectory, the ligand binding hot spots it presents in the 150-cavity is much more shallow, compared to VN04N1, which has a wide open 150-loop and a very deep cavity for ligand binding. The additional 150-cavity depth presented by VN04N1, which is indicated here to be amenable to ligand-binding through consensus sites found through computational solvent mapping experiments, may allow for additional or modified interactions with host cell

receptors or inhibitors. This finding may help rationalize how VN04N1 has been able to infect humans in some rare cases.

The 09N1 strain was recently reported to adopt a open conformation for the 150-loop over the course of 100 ns MD simulations [52]. Despite this, shallow or deep 150-cavity hot spots are completely lacking in clusters 1 and 3 for the 2009 pandemic strain (Figure 5.4A, C). This may suggest reduced capacity of this strain to bind to more varied host cell receptors, as compared to the VN04N1 strain; although the second cluster structure for this strain does indeed indicate a deep pocket available for binding in the 150-cavity, similar to the VN04N1 strain. The relative population of this deep-150-cavity structure is much less (~17%), compared to the VN04N1 deep cavities (present in over 90% of the ensemble), which may also influence such recognition events.

The secondary SA binding site was consistently presented in all of the cluster representative frames investigated, across all of the NA subtypes studied here. Subtle differences in exact location, size, and binding capacity were indicated (Figures 2-4), but for the most part, this site was highly conserved. It therefore seems very likely that this region participates in the host cell sialic acid receptor recognition events. Unfortunately, the exact glycosaccharide composition of these receptors and their configurations is not well understood; such receptors are likely to be highly complex in terms of glycosaccharide content and include branched topologies, which makes predictive models of so-called “encounter complexes” more challenging. Yet, these studies indicate the areas with which such molecules favorably interact. We propose

such information could be utilized in complex modeling studies with either sialic acid receptor models

5.5 Conclusions

- Considering the structural ensemble for NA enzymes reveals important insights relevant to its function. Analysis of x-ray crystal structures alone may be insufficient for a full understanding of these dynamic enzymes.
- The 150-cavity exhibits varying topology and frequency among the three strains studied here: N2 from Tokyo in 1967, the highly pathogenic H5N1 from Vietnam in 2004, and the 2009 pandemic H1N1 from California; ligand-binding hot spots vary correspondingly.
- The 09N1 strain exhibits an open 150-loop, but the 150-cavity it presents is much more shallow than VN04N1, which has a deep cavity. Both shallow and deep cavities exhibit persistent ligand-binding hot spots for these two strains.
- The secondary sialic acid binding site persists in throughout the structural ensembles for all of the strains studied. This suggests that this secondary site may play an important role in the complexation of NA with sialic acid receptors on the host cell.

Chapter 5, in full, is a reprint of "Variable Ligand- and Receptor-Binding Hot Spots in Key Strains of Influenza Neuraminidase", which was published in 2012 in the Journal of Molecular and Genetic Medicine, volume 6, page 293, by Lane W. Votapka, Özlem Demir, Robert V. Swift, Ross C. Walker, and Rommie E. Amaro. The dissertation author was the primary investigator and first author of this paper.

Chapter 6: Weighted Implementation of Suboptimal Paths (WISP): An Optimized Algorithm and Tool for Dynamical Network Analysis

Allostery can occur by way of subtle cooperation among protein residues (e.g., amino acids) even in the absence of large conformational shifts. Dynamical network analysis has been used to model this cooperation, helping to computationally explain how binding to an allosteric site can impact the behavior of a primary site many angstroms away. Traditionally, computational efforts have focused on the most optimal path of correlated motions leading from the allosteric to the primary active site. We present a program called Weighted Implementation of Suboptimal Paths (WISP) capable of rapidly identifying additional suboptimal pathways that may also play important roles in the transmission of allosteric signals. Aside from providing signal redundancy, suboptimal paths traverse residues that, if disrupted through pharmacological or mutational means, could modulate the allosteric regulation of important drug targets.

To demonstrate the utility of our program, we present a case study describing the allostery of HisH-HisF, an amidotransferase from *T. maritima thermotiga*. WISP and its VMD-based graphical user interface (GUI) can be downloaded from http://www.nbcr.net/wisp.

6.1 Introduction

Allosteric regulation is a key mechanism whereby proteins respond to environmental stimuli that modulate their activity[85-89]. Classic models of allostery (e.g., the MWC[90] and KNF[91] models) suggest that a binding event at an allosteric site induces substantial conformational changes in the primary catalytic site. However, allostery has since been observed in the absence of large-scale conformational changes,[92,93] suggesting that subtle alterations in protein dynamics can induce a population shift in the conformational ensemble without substantially changing the mean conformation of the protein. This subtle form of allosteric communication can be modeled by dynamical network analysis.

Recent advances in both correlated-residue clustering and dynamical network analysis have helped computationally quantify allosteric states.[94-103] Dynamical network models of allostery often focus on the single most direct path of residues leading from the allosteric to the primary active site. However, few researchers have considered the state changes of slightly longer (suboptimal) allosteric pathways. The statistical distribution of these additional pathways may be useful for locating accessible residues that, if disrupted via pharmacological or mutational means, could modulate the allosteric regulation of important drug targets.

In this paper, we introduce Weighted Implementation of Suboptimal Paths (WISP), a tool that compliments current dynamical network models of allostery by rapidly calculating the primary communicating path between two residues as well as the slightly longer suboptimal paths. To facilitate use, we have also created a WISP plugin for the popular Visual Molecular Dynamics (VMD) package[64]. WISP has

been specifically tested on several operating systems, using several versions of Python, NumPy, SciPy, and NetworkX.[104-109] The program is open source and can be downloaded from http://www.nbcr.net/wisp.

6.2 Materials and Methods

6.2.1 Molecular-Dynamics Trajectory Input

As input, WISP accepts an aligned molecular dynamics trajectory in the common multi-frame PDB format[110]. Trajectory post-processing is necessary prior to WISP analysis, as most trajectories are not initially aligned or PDB formatted. We often use the freely available Visual Molecular Dynamics (VMD) software package[64] to perform the necessary alignment and conversion.

6.2.2 Generating the Correlation Matrix

WISP, similar to other dynamical network analysis tools[111], is based on the dynamic interdependence among protein constituents (e.g., amino acids). A protein system is first simplified by representing each constituent as a single node. For example, depending on user-specified WISP parameters, an amino acid can be represented by a node positioned at the residue center of mass, the side-chain center of mass, the backbone center of mass, or the alpha-carbon. As a default, the residue center of mass is used.

The interdependence among nodes is represented as a connecting edge with an associated numeric value that reflects its strength. There are numerous methods for describing the interdependence among nodes in a protein network. Typically, this

interdependence is represented by a matrix C with values corresponding to the weights of each edge. By default, WISP generates a N^2 matrix C by calculating the correlated motion among node-node pairs as shown in Eq. 6.1 and 6.2:

$$C_{ij} = \frac{\left\langle \Delta\vec{r}_i(t) \cdot \Delta\vec{r}_j(t) \right\rangle}{\left(\left\langle \Delta\vec{r}_i(t)^2 \right\rangle \left\langle \Delta\vec{r}_j(t)^2 \right\rangle \right)^{1/2}} \qquad \text{Eq. 6.1}$$

$$\Delta\vec{r}_i(t) = \vec{r}_i(t) - \left\langle \vec{r}_i(t) \right\rangle \qquad \text{Eq. 6.2}$$

where N is the number of nodes, i and j are indices corresponding to individual nodes, $r_i(t)$ is the location of node i at time t, and C_{ij} is the matrix element at position (i, j).

The absolute value of C_{ij} is larger when the motions of two nodes are highly correlated or anticorrelated. In order to compute signaling pathways, it is useful to construct a matrix where the opposite is true, i.e., where small values indicate highly correlated or anti-correlated motions. Consequently, the correlation matrix is functionalized according to Eq. 6.3, as outlined in previous works.[96,97]

$$w_{ij} = -\log\left(\left|C_{ij}\right|\right) \qquad \text{Eq. 6.3}$$

As a point of clarification, each w_{ij} can be thought of as a "distance" in functionalized correlation space. Throughout the remainder of this paper, concepts like length and distance will refer to spans in this space, unless specifically described as "Cartesian" or "physical." We further note that, while WISP's default

functionalized correlation matrix is generally useful, any user-specified matrix that defines signaling strength as inversely proportional to edge length can be used.

6.2.3 Reducing the Complexity of the Functionalized Correlation Matrix

In order to improve the speed of subsequent path-finding steps, the complexity of the functionalized correlation matrix W must be reduced. To this end, two techniques are used. First, a contact-map matrix $M_{contact}$ is used to separate entries in W that are physically distant from entries in W that exhibit physical interaction through contact. By default, $M_{contact}$ is constructed using p_{cutoff}, a user-specified Cartesian cutoff distance that captures physical proximity.

The average location of each atom over the course of the aligned molecular dynamics trajectory is first calculated, followed by a pairwise Cartesian distance comparison. Two nodes are considered to be in physical contact if the average locations of any of their associated residue atoms come within p_{cutoff} of one another. $M_{contact}$ entries are set to zero for all node-node pairs that are not in physical contact. A simplified, functionalized correlation matrix W_{simp} is then constructed by multiplying W and $M_{contact}$ element-wise. The entries of W_{simp} that equal zero represent node-node interactions that are subsequently ignored. Alternatively, users can provide their own $M_{contact}$ if desired.

Second, to further reduce the complexity of the functionalized correlation matrix W, a pruning algorithm identifies nodes that only participate in pathways having lengths in network space that are greater than another user-defined cutoff (d_{cutoff}). As the ultimate goal is to identify suboptimal paths with lengths less than

d_{cutoff}, these nodes can be effectively discarded as well. To identify these nodes, we first generate the set of all forced-node paths (FNPs). An FNP is the optimal pathway between two user specified nodes n_a and n_b that is forced to pass through a given third node n_i. For any two fixed nodes n_a and n_b, each third node n_i is associated with a single FNP. The set of all FNPs can therefore be generated by iterating over all the nodes, n_i, of the system.

To calculate an FNP, Dijkstra's algorithm, included in NetworkX,[104] is first used to identify the optimal paths between $n_a \rightarrow n_i$, and $n_b \rightarrow n_i$, respectively. The FNP has a length equal to the sum of these two constituent paths. Any path between n_a and n_b that passes through n_i must have a length equal to or greater than that of the associated FNP. Consequently, if the length of the FNP is greater than d_{cutoff}, all entries in W_{simp} associated with n_i are set to zero, so that n_i is effectively ignored.

6.2.4 Calculating Suboptimal Pathways

Having generated W_{simp}, we are now ready to search for both the single optimal and multiple suboptimal paths between n_a and n_b. Fortunately, the optimal path is fairly easy to identify using Dijkstra's algorithm, mentioned above. In contrast, identifying all suboptimal paths is difficult because the number of possible pathways between n_a and n_b grows rapidly as the total number of nodes increases.

To identify suboptimal paths, a recursive, bidirectional approach is employed. Simultaneous searches start from n_a and n_b (Figure 1, in blue and red, respectively) and recursively traverse the nodes of the dynamical network. The recursive algorithm ignores the connections/correlations between nodes that are physically distant (Figure

6.1, grey lines). Additionally, nodes eliminated using the FNP technique described above are likewise ignored (Figure 6.1, grey circles). As soon as any of the lengthening paths grows longer than d_{cutoff}, that branch of the recursion is killed (Figure 6.1, red 'X').

At each recursive step, all branches originating from n_a and n_b are compared for common nodes (Figure 6.1, the node marked with an asterisk). If a common node exists, the two paths are joined at this node. If the length of this composite path is less than d_{cutoff}, a suboptimal path has been identified. As WISP has been developed to take advantage of multiple processors, running the program on a multi-core system can lead to further speedups beyond the software optimizations described above.

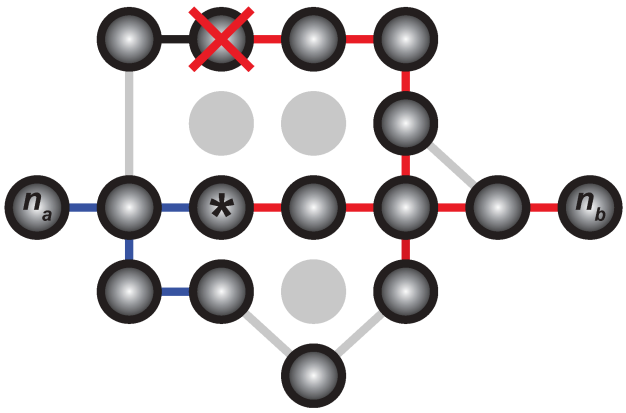

Figure 6.1: A schematic for path Identification
Simultaneous searches start from n_a and n_b (blue and red, respectively) and recursively traverse the nodes of the dynamical network. Connections/correlations between nodes that are physically distant are ignored (grey lines). Nodes eliminated using the FNP technique are also ignored (grey circles). As soon as any of the lengthening paths grows too long, that branch of the recursion is killed (red 'X'). At each recursive step, all branches originating from n_a and n_b are compared for common nodes (asterisk). If a common node exists, the two paths are joined. If the length of this composite path is sufficiently short, a suboptimal path has been identified.

6.2.5 Program Output

The program output is a directory containing multiple files, including the specific W and $M_{contact}$ matrices used. The primary output file is a TCL script that, when loaded into VMD, draws three-dimensional splines representative of the optimal and suboptimal paths. User defined parameters control the relationship between spline thickness, color, opacity and path length. Useful information is also given as comments in the TCL file, including path lengths and participating protein residues.

6.2.6 Graphical User Interface

In addition to the command-line program, we have also developed a Visual Molecular Dynamics[64] (VMD) plugin and TCL-based GUI for easy preparation and visualization of WISP results. The plugin can be accessed through the VMD "Extensions" menu. The main window of the WISP GUI (Figure 6.2) allows the user to specify the molecular trajectory and the allosteric-signal source and sink residues. Several additional window interfaces allow the user to modify more advanced program options if needed. All options available through the WISP command-line interface are available to users of the GUI.

Once satisfied with the run specifications, the user may click the "Run WISP" button at the bottom of the WISP main window to execute the job. The plugin loads the visualization of the allosteric pathways into the main VMD window, where the appearance can be further modified according to the user's preferences.

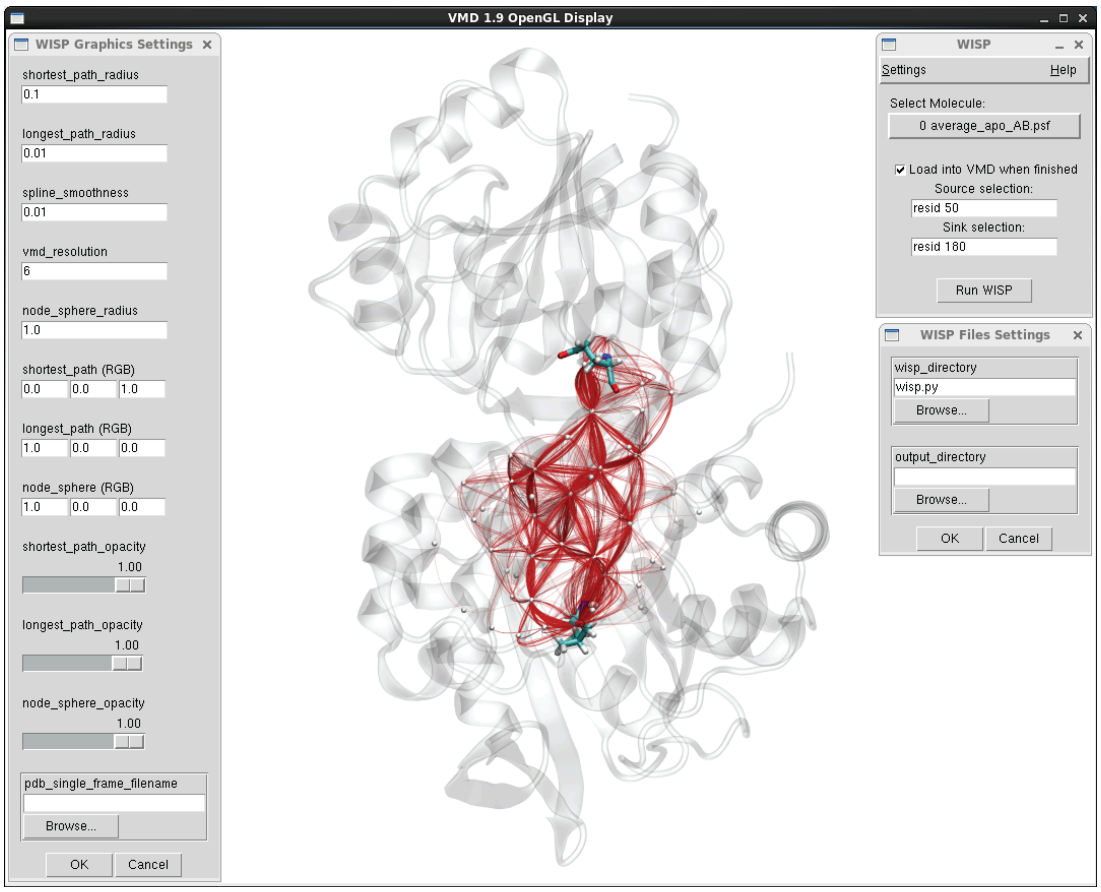

Figure 6.2: WISP Graphical User Interface (GUI)
In this demonstration, the GUI is used to visualize the allosteric pathways between Leu50:HisF and Glu180:HisH. In the main window (top left), the user selects the relevant molecule, and which residues to use as the source and sink. The user may also select to load the visualization into VMD upon job completion. The setting option windows (left and bottom right) allow the user to specify additional WISP arguments.

6.2.7 HisH-HisF Details

The molecular dynamics simulations of HisH-HisF used in the current study have been described previously.[97] In brief, a model of the HisH−HisF apo dimer was prepared from the 1GPW[112] crystal structure (*Thermotoga maritima*). To generate the corresponding holo structure, the 1OX5[113] crystal structure (*Saccharomyces*

cerevisiae), which contains a co-crystallized PRFAR allosteric effector molecule, was aligned to the apo model, effectively positioning PRFAR within the 1GPW:HisF allosteric site. The aligned 1OX5 PRFAR was then merged with the 1GPW-based apo model to yield the corresponding holo structure. Following solvation and equilibration, 20 ns of production dynamics were run for both the apo and holo systems using NAMD[114], the CHARMM27 force field[115], and the same PRFAR parameterization used previously.[116]

6.3 Results/Discussion

Allosteric regulation is crucial to many biological processes. Consequently, one natural strategy for rational drug design is to impede or agonize protein function via allosteric modulation. Classic views of allostery suggest that the binding of an effector molecule at an allosteric site induces large conformational shifts that alter the activity of the primary site. However, as allostery is not necessarily limited to large shifts, this reasoning does not explain some examples of regulation at a distance. For instance, Chung-Jung Tsai et al.[93] recently showed that significant backbone deformations are not required for an allosteric effect; rather, in the absence of large conformational changes, subtle shifts in local dynamics driven by entropic effects[92] govern certain types of allostery.

Quasi-harmonic analysis (e.g., like that used by software packages such as CARMA[117,118] to calculate entropy) is commonly used to build dynamical network models that quantify signaling pathways among protein constituents. Optimal and suboptimal pathways are calculated that connect protein constituents believed to be

important for allostery (i.e., “sources” and “sinks”). An optimal pathway is the shortest distance traversed between source and sink along weighted edges (e.g., as determined by correlated motions), and suboptimal pathways are those closest in length to, but not including, the optimal path. Existing tools can compute optimal and suboptimal pathways between residues[119]; however, these programs lack the speed required to compute more than fifty suboptimal pathways within a reasonable amount of time (several hours or days). As statistics related to suboptimal pathways may provide important insights that cannot be gleaned from the single optimal pathway, faster algorithmic advances must be made.

WISP was designed to facilitate the calculation of hundreds of suboptimal pathways in minutes, thereby permitting fast and robust statistical analysis of biological systems modeled as dynamical networks. For example, using a modern workstation with 24 cores, we recently used a 20,000-frame trajectory to identify 750 pathways. WISP loaded and analyzed the trajectory, generated the functionalized correlation matrix, and identified the 750 pathways in 21 min and 52 seconds. When the calculation was repeated using a copy of the functionalized correlation matrix saved from the first run, the 750 pathways were identified in only 5 minutes and 44 seconds.

To demonstrate the utility of the WISP algorithm, we used it to study HisH-HisF, a multidomain globular protein known to exhibit allostery. The activity of HisH-HisF, which regulates the fifth step of the histidine biosynthetic pathway in plants, fungi, and microbes, is substantially altered by the allosteric effector N1-[(5'-phosphoribulosyl)-formimino]-5-aminoimidazole-4-carboxamide ribonucleotide

(PRFAR).[120] Guided by previous work,[97] we investigated the suboptimal pathways between residues Leu50:HisF and Glu180:HisH using 20-ns molecular dynamics simulations of both apo and holo HisH-HisF.

A total of 700 pathways (Figure 6.3) between Leu50:HisF and Glu180:HisH were calculated using WISP's default correlation (Eqs. 6.1-6.3) and contact-map matrices, described in the Materials and Methods. Had only the two optimal pathways (apo vs. holo) been considered, we would have concluded that communication between the allosteric and primary site is fundamentally different in the presence and absence of the PRFAR effector molecule (Figure 6.3,6.4). The optimal pathway between Leu50:HisF and Glu180:HisH in the apo state was LEU50:HisF → PHE49:HisF → PHE77:HisF → PRO76:HisF → LYS181:HisH → GLU180:HisH. In contrast, the optimal pathway with PRFAR bound was LEU50:HisF → GLY80:HisF → VAL79:HisF → LYS99:HisF → ASP98:HisF → LYS181:HisH → GLU180:HisH.

However, when multiple suboptimal paths were considered, it became apparent that the allosteric mechanism is in fact far more intricate. The optimal path in the apo simulation is the shortest suboptimal path in the holo simulation (top 0.3%), and the optimal path in the holo simulation is the 13th shortest suboptimal path in the apo simulation (top 2.0%). In light of this multi-pathway analysis, the idea that PRFAR binding fundamentally alters a solitary line of communication between the allosteric and primary site becomes less tenable. Rather, the binding of the effector molecule likely has small effects on multiple pathways, both optimal and suboptimal, that when taken together yield a substantial allosteric effect.

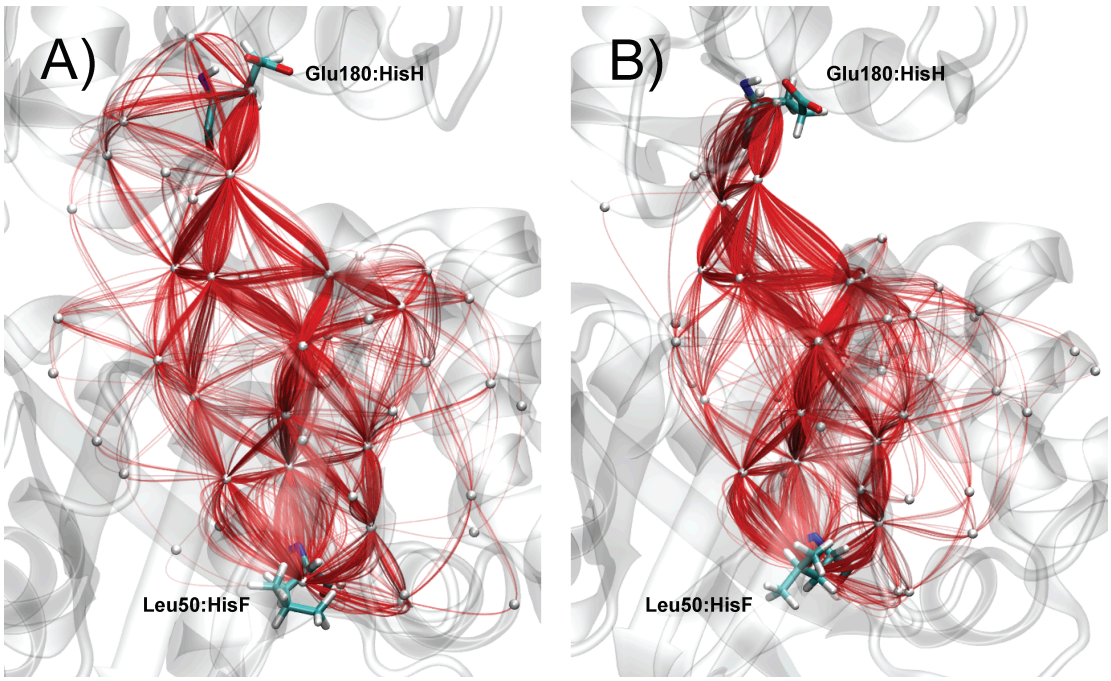

Figure 6.3: WISP generated signaling pathways
The 700 shortest paths between Leu50:HisF and Glu180:HisH, shown as red splines, derived from A) the apo trajectory, and B) the holo trajectory. Wisp allows the user to choose between a number of graphical settings to better visualize signaling among nodes.

We next sought to characterize the strength of the allosteric effect. The lengths of the two optimal pathways did not differ substantially (apo: 2.97; holo: 2.84). Consequently, had only these two pathways been considered, some might have mistakenly concluded that the allosteric consequences of PRFAR binding are minor. In contrast, when hundreds of suboptimal paths were also considered, a large PRFAR-dependent shift in communication between the allosteric and primary site became apparent. To demonstrate this shift, we generated a histogram of all path lengths for both the holo and apo simulations (Figure 6.4). The distribution derived from the holo trajectory is substantially skewed towards shorter path lengths, suggesting that the motions of the residues connecting the allosteric and primary sites

are more tightly correlated when PRFAR is bound. A loss of entropy along the pathways may therefore explain the allosteric signal.

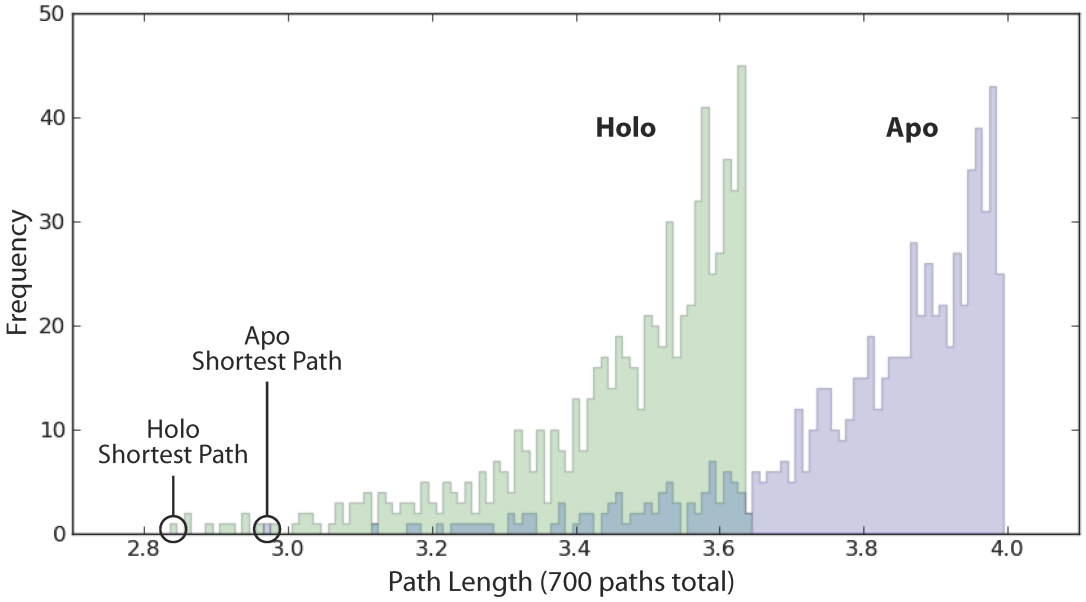

Figure 6.4: Statistical distribution of signaling pathways
A histogram of the 700 path lengths associated with the apo and holo trajectories are shown. The optimal paths are denoted "Shortest Path." The path distribution is largely shifted to the left for the holo (allosteric) state. This shift likely results from a more coherent signal in the holo simulation, indicating a possible decrease in the entropy along the pathways due to PRFAR binding.

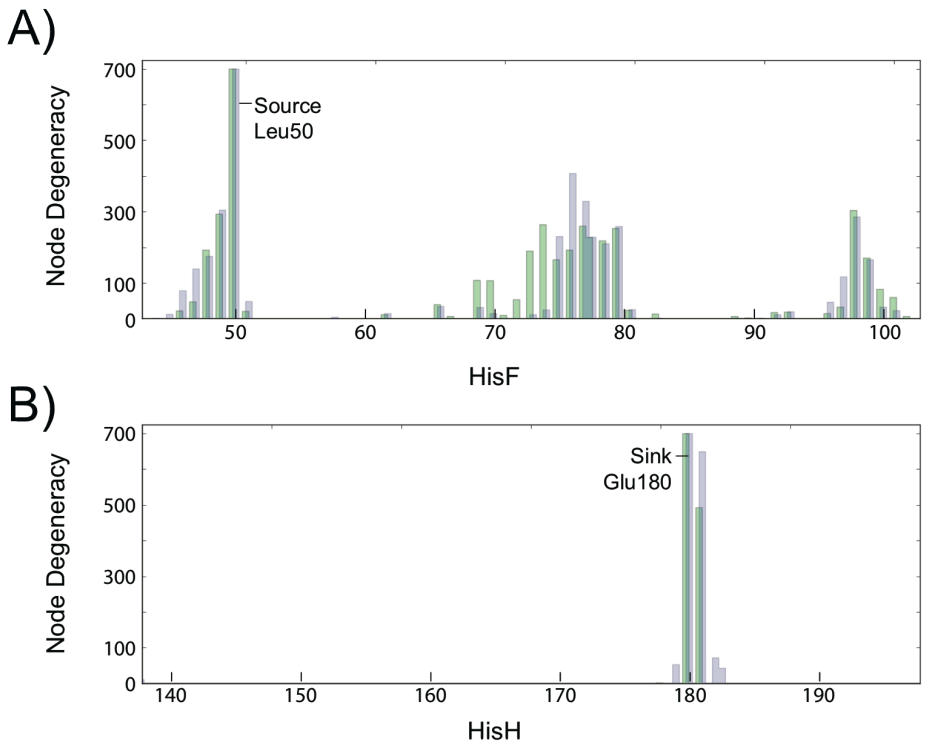

Figure 6.5: Node degeneracy in signaling pathways
The total number of times a given residue participates in any of the 700 paths (i.e., node degeneracy) is shown in A) HisF and B) HisH. Green indicates the holo state, blue indicates the apo state, and cyan indicates an overlap. Note that Leu50:HisF and Glu180:HisH are in all 700 paths.

To identify protein residues critical for allosteric transmission, we counted the number of times each residue appeared in any of the 700 paths associated with the apo and holo trajectories, respectively (i.e., the degeneracy of each node, Figure 6.5). Notably, a number of residues had large effector-molecule dependent shifts in degeneracy (i.e., HisF: LEU47 (shifts down), VAL69 (shifts up), ALA70 (shifts up), ILE73 (shifts up), ASP74 (shifts up), PRO76 (shifts down), and ALA97 (shifts down); HisH: LYS181 (slight shift down) as seen from Table 1). Importantly, these residues, which may be crucial for the regulation of protein activity, did not all appear

in the optimal apo and holo paths and so would not have been identified had the suboptimal paths been ignored. Previous studies in evolutionary conservation have shown HisF: LEU47, VAL69, ALA70, and ILE73 to be partially or strongly conserved and HisF: PRO76, ALA97, HisH: LYS181 to be strictly conserved across the Glutmine Amidotransferase family.[121] No conservation exists in HisF: ASP74, but this amino acid is still predicted to play a role in allostery.[121] Compounds that target these residues may serve as useful precursors to future allosteric modulating drugs.

The analysis of suboptimal pathways in the dynamical network of a protein may prove powerful, however there is an issue that should be mentioned. The number of sub-optimal pathways generated by WISP is currently an arbitrary value that depends on the users choice of d_{cutoff}. Future work that quantifies criteria for the best number of pathways to calculate is greatly welcome.

6.4 Conclusion

We present WISP, a program that rapidly calculates both optimal and suboptimal communication pathways between distinct protein residues. The program is available as a VMD plugin or a standalone command-line script. WISP outputs path members and lengths that can be subsequently used in the analysis of path distributions, node degeneracy, etc.

To demonstrate the utility of our program, we presented a dynamical analysis of the HisH-HisF protein. Allosteric modulation, in our test case, was likely the result of subtle changes in multiple suboptimal pathways, rather than large changes in

a single optimal path. Additionally, we showed that PRFAR binding causes a large shift towards shorter path lengths (i.e., more correlated motions) in 700 communication pathways between residues HisF:Leu50 and HisH:Glu180. This shift conveys the strong allosteric effects of the PRFAR modulator (Figure 6.4). The multiple suboptimal pathways are dominated by a few select residues, as indicated by the shift in node degeneracy between the apo and holo states (Figure 6.5 and Table 6.1).

WISP has been successfully tested on a number of platforms (Table 6.2). We are hopeful that the program will be a useful tool for the computational-biology community.

Table 6.1: Node Degeneracy table
A numerical representation of the same data from Fig. 6.5. The comparison between the apo and holo states shows that certain residues are more sensitive to the allosteric effector PRFAR than others (shaded columns).

700 Paths									
HisF Nodes	**LEU** 47	**VAL** 48	**PHE** 49	**LEU** 50	**VAL** 69	**ALA** 70	**ILE** 73	**ASP** 74	**ILE** 75
Degeneracy Apo	139	175	304	700	31	14	11	24	230
Degeneracy Holo	47	192	293	700	107	106	189	263	165

HisF Nodes	**PRO** 76	**PHE** 77	**THR** 78	**VAL** 79	**GLY** 80	**ALA** 97	**ASP** 98	**LYS** 99
Degeneracy Apo	407	329	228	210	258	117	285	165
Degeneracy Holo	192	259	227	218	253	32	303	170

HisH Nodes	**GLU** 180	**LYS** 181
Degeneracy Apo	700	649
Degeneracy Holo	700	492

Table 6.2: WISP operating specifications
WISP has been tested on a number of operating systems, using various versions of NumPy, SciPy, and NetworkX. We note that installation under Windows was difficult; however, the command-line version of the program was successfully executed after installing the appropriate dependencies using the ActivePython software package.

Operating System	Python	NumPy	SciPy	NetworkX
Scientific Linux 6.4	2.6	1.7	0.9.0	1.7
Mac OSX 10.6	2.7.2	1.6.1	0.9.0	1.8.1
Ubuntu 12.04	2.7.5	1.7.1	0.12.0	1.8.1
Windows XP	2.7.3	1.7.0rc1	0.11.0	1.8.1

Chapter 6, in full, is a reprint of "Weighted Implementation of Suboptimal Paths (WISP): an Optimized Algorithm and Tool for Dynamical Network Analysis", which was published in 2014 in the Journal of Chemical Theory and Computation, volume 10, issue 2, pages 511-517, by Adam T. Van Wart, Jacob D. Durrant, Lane W. Votapka and Rommie E. Amaro. The dissertation author was the third investigator and author of this paper.

Chapter 7: POVME 2.0: An Enhanced Tool for Determining Pocket Shape and Volume Characteristics

Analysis of macromolecular/small-molecule binding pockets can provide important insights into molecular recognition and receptor dynamics. Since its release in 2011, the POVME (POcket Volume MEasurer) algorithm has been widely adopted as a simple-to-use tool for measuring and characterizing pocket volumes and shapes. We here present POVME 2.0, which is an order of magnitude faster, has improved accuracy, includes a graphical user interface, and has new features for improved pocket analysis. To demonstrate the utility of the algorithm, we use it to analyze the binding pocket of RNA editing ligase 1 from the unicellular parasite *Trypanosoma brucei*, the etiological agent of African sleeping sickness. The POVME analysis characterizes the full dynamics of a potentially druggable transient binding pocket and so may guide future antitrypanosomal drug-discovery efforts. We are hopeful that this new version will be a useful tool for the computational- and medicinal-chemist community.

7.1 Introduction

Binding-pocket analysis is an active area of research that includes pocket detection and characterization, druggability prediction, and the study of binding-site flexibility.[122] The advent of the Protein Data Bank (PDB[123]) spurred the creation of a number of software packages aimed at facilitating the analysis of macromolecular pockets.[124-127] In recent years, additional programs have been developed with improved accuracy and increasingly advanced pocket-characterization algorithms,[15,62,128-135] as recently reviewed by Zheng et al.[136]

Pocket analysis is useful for studying receptor dynamics.[52,137-158] One can get a good sense of the full gamut of possible binding-pocket conformational states by obtaining multiple structures from X-ray crystallography, NMR spectroscopy, or molecular dynamics (MD) simulations and comparing pocket volumes and, in particular, shapes. These comparisons facilitate the identification of novel, pharmacologically relevant binding-pocket conformations, as well as transient binding pockets that are not evident when a limited number of static structures are considered.

Additionally, pocket analysis can also be applied to computer-aided drug discovery (CADD). Among the many complex factors that govern molecular recognition,[159,160] pocket volume and shape are perhaps the most straightforward. Simply put, a ligand will not generally bind to a receptor if it cannot physically fit within the confines of the binding pocket, and receptor/ligand shape complementarity plays a key role in molecular recognition.[161] Consequently, pocket characterization has been used to inform CADD efforts aimed at predicting ligand binding, whether

through virtual screening, QSAR, or volumetric similarity searching.[162-164] Given the astounding variety of pocket geometries possible,[165] this characterization is no trivial task.

To address this challenge, both ligand- and receptor-centric approaches have been developed. Ligand-based methods such as OpenEye's Rapid Overlay of Chemical Structures (ROCS) algorithm[166] seek to identify novel small-molecule binders by querying a compound database for entries with three-dimensional shapes that are similar to that of a known template ligand,[167] as assessed by the degree of volume-overlap mismatch. These techniques perform comparably to more traditional virtual-screening methods[168] and have been used to successfully identify a number of experimentally validated ligands (see, for example, ref. [169-171]).

While ligand-based approaches will certainly continue to have high utility,[172] a more receptor-centric methodology is sometimes advantageous to consider. Bound ligands often occupy only a portion of their respective pockets,[161,165] on average perhaps as little as a third of the total space available.[161] Analysis of ligand volume and shape alone cannot account for potential interactions with pocket regions that are not occupied by the template ligand itself. In contrast, receptor-based pocket analysis elucidates the volume and shape of the entire cavity, including regions not yet exploited by existing pharmacophores.

Receptor-centric techniques can also be used to select diverse pocket shapes for use in subsequent virtual-screening efforts. It is often advantageous to dock a library of small molecules into multiple receptor conformations in order to account for receptor flexibility. Carefully selecting conformations with unique pocket

geometries has been shown enhance hit rates and subsequent ligand diversity.[173-176]

To simplify binding-pocket characterization, we recently developed an algorithm called POVME (POcket Volume MEasurer).[15] POVME floods a pocket-encompassing region with equidistant points, removes those points that are near receptor atoms, and calculates the volume from the remaining points. The points can themselves be saved, providing a specific description of the pocket shape as well. Inspired by the fairly widespread adoption of our program (43 citations in Google Scholar as of April 2014), we have now created a second, much improved version. POVME 2.0 is over an order of magnitude faster than POVME 1.0, includes a graphical user interface (Figure 7.1) that greatly improves usability, and incorporates new features that improve accuracy and facilitate analysis.

POVME 2.0 has been tested on all major operating systems with various versions of python, *numpy*, and *scipy* (Table 7.1).[107-109,177,178] A copy of the program, which is released under the terms of the GNU General Public License, can be obtained from http://nbcr.ucsd.edu/POVME. We are hopeful that POVME will be a useful tool for the computational- and medicinal-chemist community.

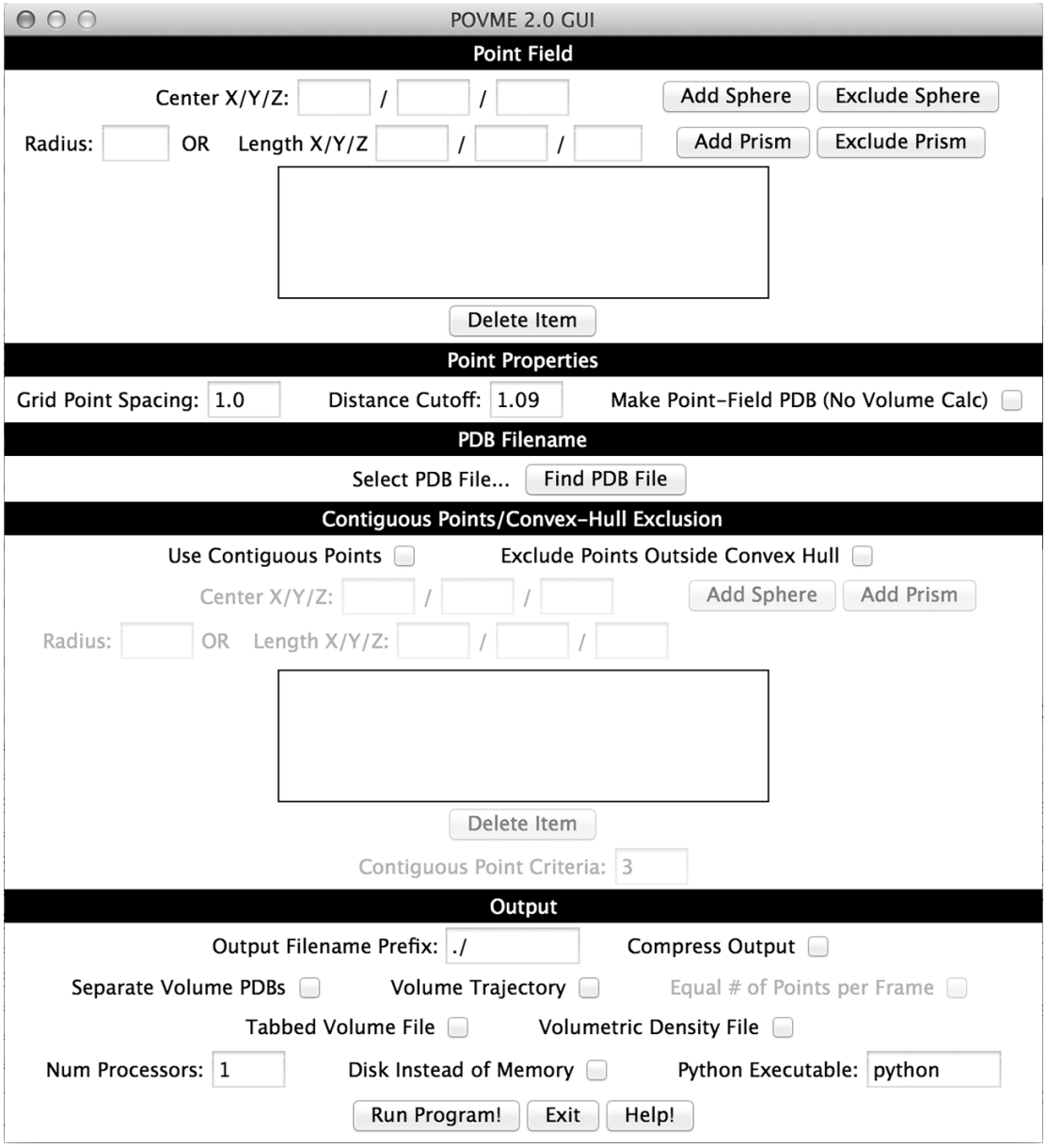

Figure 7.1: The POVME 2.0 graphical user interface

Table 7.1: **Operating-system compatibility**
POVME 2.0 has been successfully tested on all major operating systems with various versions of python, *numpy*, and *scipy*.

Operating System	**Python Version**	***Numpy* Version**	***Scipy* Version**
Scientific Linux 6.2	2.6.6	1.6.2	0.11.0
OS X 10.9.1	2.7.5	1.6.2	0.11.0
Windows 7 Home Premium	2.7.6	1.8.0	0.13.3

7.2 Materials and Methods

7.2.1 The POVME Algorithm

Successful POVME use includes three required and two optional steps. Trajectory alignment, the construction of a pocket-encompassing region, and the subsequent identification of the pocket-occupying space are required. Optionally, the user can also instruct POVME to eliminate subregions that fall outside the receptor's convex hull and/or are non-contiguous with the primary pocket. A detailed description of each of these steps follows.

1) Aligning the trajectory. POVME accepts a multi-frame PDB (Protein Data Bank) file as input. We expect that MD simulations will be the most common source of these files, but multiple crystal structures or NMR conformations can also be used. We have found that the computer program Visual Molecular Dynamics (VMD)[30] is useful for aligning trajectories and converting files to the PDB format, but other software packages can also be used for this purpose. Alignment is necessary because the POVME algorithm assumes the pocket being measured does not translate or rotate in space. Different alignment methodologies can subtly alter how this requirement is

met, as discussed in the Results and Discussion. We note also that single-frame PDB files can likewise serve as POVME input if the user wishes only to characterize a single pocket.

2) Defining a region that encompasses all trajectory binding pockets. The user must next define "inclusion" (Figure 7.2A) and "exclusion" (Figure 7.2B) regions, respectively. Both of these regions are constructed from a combination of user-specified spheres and rectangular prisms. The required inclusion region should entirely encompass all the binding-pocket conformations of the trajectory. The optional exclusion region defines portions of the inclusion region that should be ignored, perhaps because they are not truly associated with the pocket. To generate a field of equidistant points that encompasses all the binding-pocket conformations of the trajectory, POVME first floods the user-specified inclusion region with points and then removes any points also contained in the optional exclusion region (Figure 2C).

3) Removing points that are near receptor atoms. As the purpose of POVME is to measure the volume of the binding-pocket cavity, the program next removes any points that are close to receptor atoms, leaving only those points that are likely to be located within the binding pocket itself (Figure 2D).

4) Removing points outside the receptor's convex hull. POVME 2.0 introduces an optional new feature for removing points that lie entirely outside the binding pocket. Specifically, the gift-wrapping algorithm is used in combination with the Akl-Toussaint heuristic[179] to define the convex hull of receptor atoms near the user-defined inclusion region. As the gift-wrapping algorithm runs in $O(n^2)$ time, where n is the number of atoms in the receptor structure, it is not necessarily the fastest

algorithm for computing the convex hull. However, by coupling it with the Alk-Toussaint heuristic, the expected running time is lowered to O(n). Ultimately, any points that fall outside the convex hull are removed (Figure 2E). This feature is particularly useful when the user defines an inclusion region that protrudes into the surrounding solvent-occupying space.

5) Removing points that are not contiguous with the primary pocket. Like the original POVME program, version 2.0 retains the optional ability to remove isolated patches of points that are not contiguous with the primary binding pocket. This feature requires that the user define a third region, again using spheres and rectangular prisms, that always falls within the primary binding-pocket region, regardless of the trajectory frame considered (Figure 2F). All pocket-occupying points within or contiguous to this region are retained, but isolated patches of points that are not directly connected are deleted (Figure 2G).

POVME output. By default, POVME writes a number of files to the disk. The calculated pocket volumes, as well as user-defined parameters and progress messages, are saved to a simple text-based log file. POVME can also be instructed to save the volume measurements to a second file in a simple tabular format that can be easily pasted into popular spreadsheet programs. Pocket-occupying points are equidistant (1.0 Å by default), so each point is associated with an identical cubical volume (e.g. 1.0 $Å^3$). The volume of a whole pocket is calculated by simply summing the individual volumes associated with each unique point.

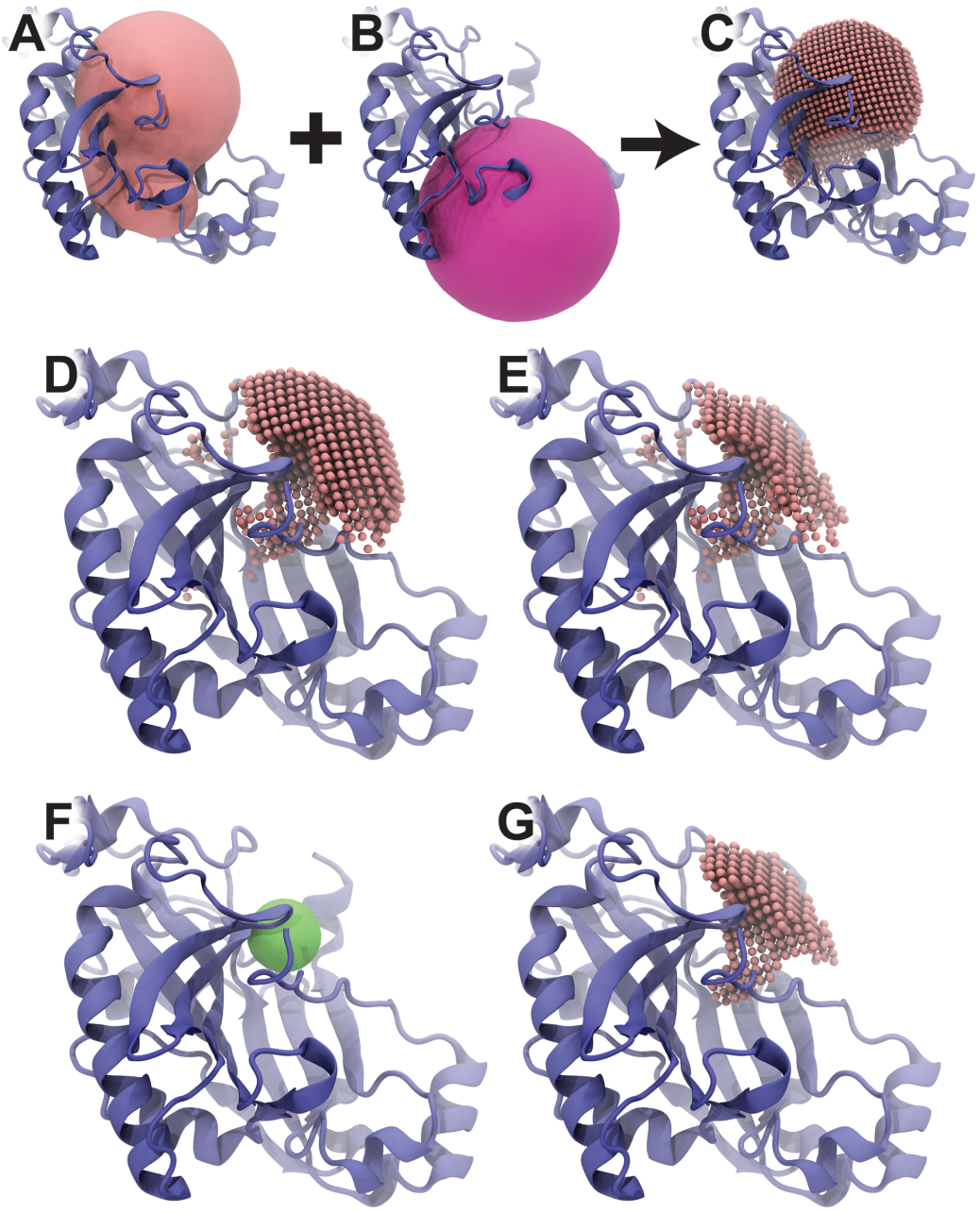

Figure 7.2: A graphical summary of the POVME 2.0 algorithm
A) The user defines an inclusion region. B) The user defines an exclusion region. C) The portion of the inclusion region that is not also in the exclusion region is flooded with equidistant points. D) Any of the points that are close to receptor atoms are deleted. E) Any points outside the convex hull are optionally deleted. F) The user can optionally define a contiguous-points region. G) All points that are not contiguous with that region are similarly deleted.

POVME also optionally saves the pocket-occupying points of each frame to PDB file(s) on the disk. The user can instruct the program to save these points to separate files and/or to a single PDB trajectory. Some visualization programs (e.g. VMD) are only compatible with trajectories that have the same number of atoms in each frame. POVME can optionally write extra points to the origin (0.0, 0.0, 0.0) on a frame-by-frame basis to satisfy this requirement. As these POVME frames are formatted similarly to those produced by SiteMap, they are also compatible with the pocket-shape volumetric overlap clustering tools produced by Schrödinger.[133,134,173,174]

Finally, POVME also optionally saves a volumetric density map in the Data Explorer (DX) format, similar to the MDpocket algorithm.[132] A volumetric density value is associated with each of the pocket-occupying points by calculating the fraction of all trajectory pockets that include the given point. If the density map is displayed as an isosurface, the value of the isosurface expresses the fraction of time (e.g. over the course of the simulation) that the pocket included the displayed volume.

7.2.2 Test System: RNA Editing Ligase 1

We obtained the 1XDN crystal structure,[72] which includes enzyme residues 52-365 as well as an ATP molecule and a magnesium ion bound in the active site, from the Protein Data Bank.[123] Selenomethionine residues were replaced with methionine. All crystallographic water molecules were retained. The AMBER LEaP module was used to submerge the protein in a rectangular box of water molecules that extended 10 Å beyond the system atoms in all three Cartesian dimensions.

Monovalent ions were added to neutralize the system and to bring it to a 0.1 M salt concentration. The protein and water atoms were parameterized using the Amber99SB force field[22] and the TIP4P-ew water model,[180] respectively. Additionally, we used the parameters for ATP, magnesium, and monovalent ions developed by Meagher et al.,[181] Allner et al.,[182] and Joung and Cheatham,[183] respectively.

The REL1 system was subjected to five 5000-step energy minimizations using the NAMD molecular-dynamics simulation package[50,184] to gradually introduce full flexibility. We first allowed only hydrogen atoms to move, second released all water molecules, third released ions and ATP, fourth released the protein amino-acid side chains, and fifth removed all constraints. The system was then heated from 0 to 310 K in an NVT ensemble for 500 ps, with the protein backbone restrained. Equilibration was achieved in two segments, each consisting of a 250-ps simulation in the NPT ensemble. In the first segment, the protein backbone was restrained; in the second, no restraints were applied.

Five production simulations were performed, starting from the fully equilibrated structure. A total of 650 ns were simulated (one simulation of 250 ns, and four of 100 ns). Different random seeds were used for each productive simulation to generate different starting velocities.

To study the flexibility of the *Tb*REL1 active site, we extracted 6,500 frames from the simulations, evenly spaced 100 ps apart. All waters, counterions, ATP molecules, and magnesium ions were removed. VMD's RMSD Trajectory Tool[30] was used to align the extracted frames. In order to determine how differing alignment

methodologies would impact the POVME analysis, we used several different protocols. The extracted frames were concatenated and aligned by 1) the atoms of the bound ATP ligand; 2) the atoms of the active-site residues (e.g. any residue within 5 Å of the crystallographic ligand); 3) the alpha-carbon atoms (C_α) of the active-site residues; and 4) the C_α of the entire protein. Each of these four aligned trajectories were saved as separate multi-frame PDB files.

Separate POVME analyses were performed for each aligned trajectory. In each case, we characterized the combined ATP/transient pockets using an inclusion region defined by 10 carefully positioned spheres. This region was filled with equidistant points spaced 1.0 Å apart. No exclusion regions were required. Points that were not contiguous with those contained within a small sphere centered at the opening of the ATP-binding pocket were discarded. The new convex-hull feature was enabled.

To benchmark POVME 1.0 and POVME 2.0, we further considered the REL1 trajectory aligned by all C_α. Additional analyses of this trajectory were performed using POVME 1.0 and POVME 2.0 with the new convex-hull feature disabled.

7.3 Results/Discussion

As pocket volume and shape play critical roles in determining small-molecule binding, they are often the focus of computer-docking campaigns, QSAR studies, and molecular-dynamics analyses. We previously created a novel algorithm for characterizing macromolecular pockets called POVME (POcket Volume MEasurer) that has been widely adopted.[15] We here present a much-improved version of the

algorithm, POVME 2.0.

POVME 2.0 has four primary advantages over previous versions. First, it is an order of magnitude faster because it relies on the *numpy* and *scipy* python modules to perform matrix-based calculations at nearly the speed of compiled C programs.[107-109,177,178] Additionally, the user can instruct POVME 2.0 to take advantage of multiple processors to further improve the speed of the calculation.

Second, POVME 2.0 comes with an optional graphical user interface (GUI) to facilitate usability (Figure 7.1). The GUI requires that Tkinter,[185] a python binding to the Tk GUI toolkit,[186] be installed. Fortunately, Tkinter is included in the standard Windows and OS X python distributions, as well as many Linux distributions.

Third, POVME 2.0 includes a new convex-hull-clipping option that improves the accuracy of the volume calculation. Portions of the binding pocket that fall outside the convex hull of nearby receptor atoms are discarded; consequently, only portions of the pocket that are truly interior to the protein surface are considered.

Fourth, unlike the original version, POVME 2.0 can analyze entire trajectories in addition to single protein conformations. With POVME 1.0, users were required to save each trajectory frame to a separate PDB file in order to study changes in pocket volume and shape over the course of a MD trajectory. In contrast, POVME 2.0 can read multi-frame trajectory files without requiring that each frame be saved separately. When analyzing MD trajectories, POVME outputs both frame-by-frame and whole-trajectory analyses. For frame-by-frame analysis, POVME saves the individual pocket shapes in the PDB format. For whole-trajectory analysis, POVME creates a volumetric density map showing the frequency with which different regions

of the protein are included in the pocket over the course of the trajectory.

7.3.1 Test Case: *Trypanosoma brucei* RNA Editing Ligase 1 (TbREL1)

To demonstrate the utility of this new POVME implementation, we used it to analyze an MD simulation of RNA editing ligase 1 (REL1) from the parasite *Trypanosoma brucei*, the etiological agent of African sleeping sickness. REL1 is a critical component of the *T. brucei* editosome, which edits transcriptional RNA prior to translation. This extensive RNA-editing process is essential for trypanosomatid survival, and REL1 has been shown to be a viable drug target.[187,188] Indeed, REL1 inhibitors have been identified that kill the whole-cell parasite.[189]

Previous studies of related crystal structures have hinted at the existence of a transient subpocket connected to the distal portion of the primary ATP-binding site that may provide unique opportunities for drug discovery.[139] Compounds that bind to the REL1 primary site may also target other ATP-binding proteins with structurally similar pockets; however, compounds that bind to the unique transient pocket may prove more target specific.

To better characterize the dynamics of the REL1 pockets, we characterized 6,500 combined ATP-transient pockets extracted from 650 ns of MD simulations. We first aligned the trajectory to ensure that the binding pocket was consistently in the same location. As with other pocket-analysis programs,[132,136,190] simulation-trajectory alignment impacts the calculation of the average volumetric density maps. To demonstrate this sensitivity, we performed four separate POVME analyses, aligning the REL1 trajectory by 1) all ATP-ligand atoms; 2) all the atoms of the active-site

residues; 3) the alpha-carbon atoms (C_α) of the active-site residues; and 4) the C_α of the entire protein. Volumetric density maps were calculated for each of these aligned trajectories and were visualized superimposed on the receptor structure using VMD. When displayed as an isosurface, these density maps show the fraction of frames with measured pockets that included the displayed volume.

For the purposes of comparison, we judged the utility of each alignment protocol by how consistently the associated POVME analysis captured the ATP-binding-pocket region over the course of the entire trajectory. As our simulations included a bound ATP ligand, the ATP-binding subpocket should always be open (i.e. the region of the volumetric map corresponding to ATP in our simulations should have a high density, in excess of 95%).

When the trajectory was aligned to all active-site C_α, the POVME-identified pocket consistently included the ATP-binding region (Figure 7.3B). We also found that aligning by all active-site atoms or even the atoms of the bound ligand itself led to similar POVME results (Figure 7.4). In contrast, the pocket analysis was less than optimal when the trajectory was aligned by the C_α of the whole receptor (Figure 7.3C), likely because substantial protein motions distant from the active site led to poor binding-pocket alignment. Consequently, the transient pocket was identified as open only half as often when the trajectory was aligned by all C_α vs. active-site C_α.

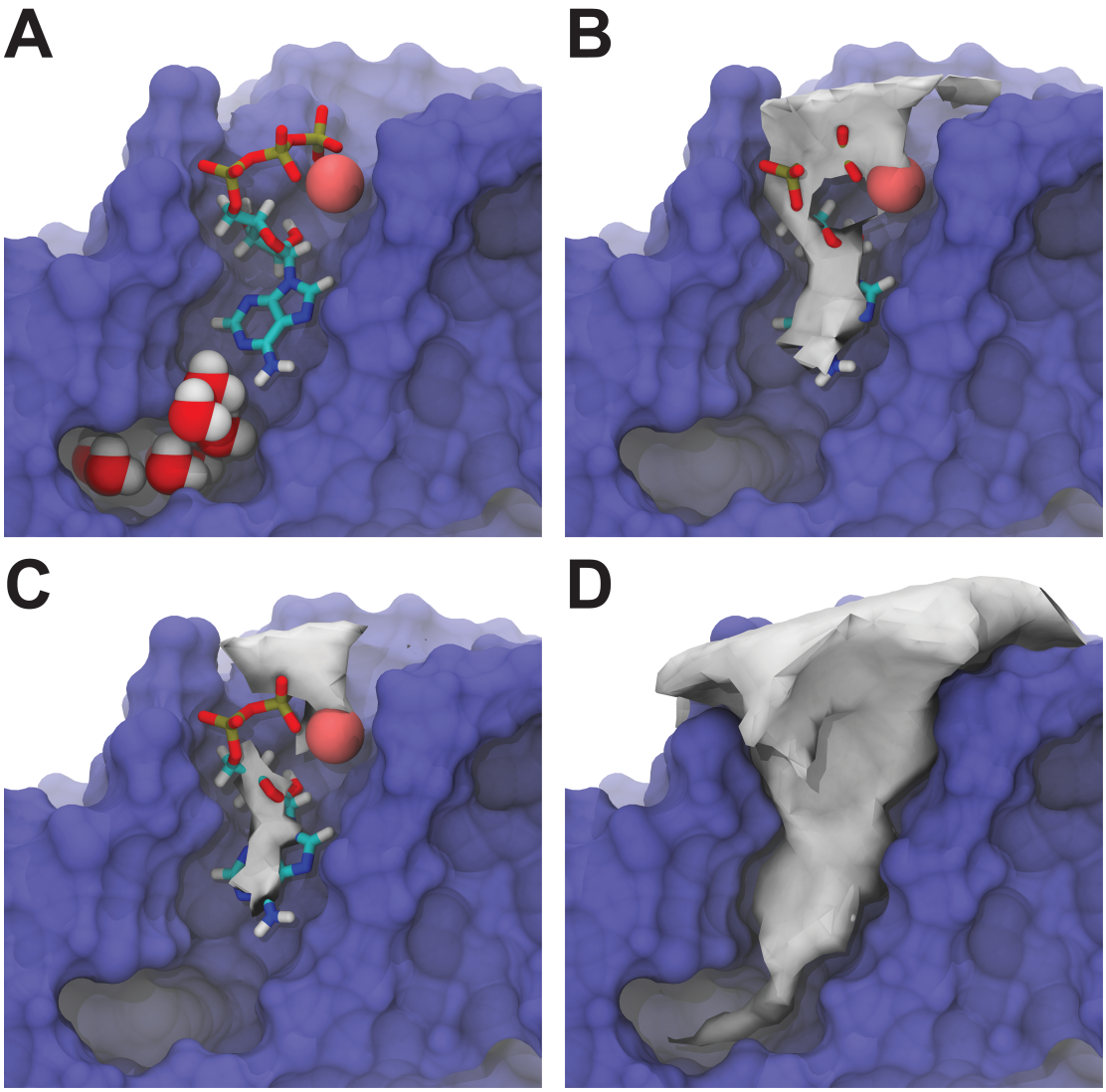

Figure 7.3: Volumetric density maps of the *Tb*REL1 active site
Some regions of the protein have been removed to facilitate visualization. A) The crystallographic pose of the bound ATP molecule. Crystallographic water molecules indicate the location of a secondary binding pocket that is transiently accessible from the ATP-binding pocket. B) The region of the binding pocket identified as "open" at least 95% of the time when the trajectory was aligned by the active-site C_α. C) The same region when the trajectory was aligned by the C_α of the whole protein. D) The region of the binding pocket identified as "open" at least 25% of the time when the active-site-C_α alignment was again used.

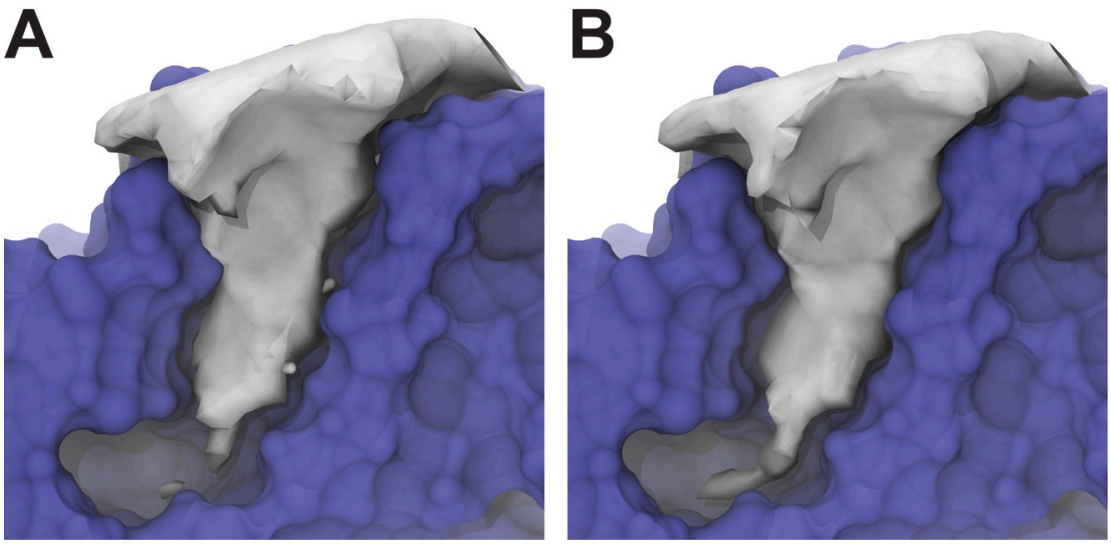

Figure 7.4: *Tb*REL1 volumetric density maps
Some regions of the protein have been removed to facilitate visualization. The region of the binding pocket identified as "open" at least 25% of the time is shown, as determined when the trajectory was A) aligned by all the atoms of the ATP ligand, and B) aligned by all the atoms of the binding-pocket residues.

While the best protocol to use is likely system dependent, based on these REL1 results we concur with others in recommending that trajectories be aligned by active-site C_α.[190] When the binding pocket is partly composed of flexible loops, aligning by pocket C_α that belong to stable secondary-structure elements may be appropriate.

Having considered four different alignment strategies, we ultimately chose the active-site-C_α aligned trajectory. POVME analysis revealed the full dynamics of the transient REL1 pocket, as indicated by the density maps in Figures 7.3B and 7.3D at isovalues of 95% and 25%, respectively. As expected given that we simulated the holo protein, the primary ATP-binding pocket was persistently open throughout the entire simulation (Figure 7.3D, 95% isovalue). The intermittent transient pocket was

open at least 25% of the time (Figure 7.3D), suggesting a persistence sufficient to support our hypothesis of druggability.

7.3.2 Benchmarking

To judge POVME 2.0 performance, we similarly analyzed a *Tb*REL1 trajectory using POVME 1.0. When the convex-hull algorithm was disabled, both POVME 1.0 and 2.0 gave nearly identical volume measurements (Figure 7.5 graph, in black), but POVME 2.0 completed the pocket-volume calculation over 35 times faster (5.0 vs. 175.4 processor-hours). When the new convex-hull feature was enabled, POVME 2.0 required 32.8 processor-hours (Figure 7.5 graph, in gray). Although the convex-hull feature does add computational expense, it leads to more accurate characterizations that do not include regions outside the confines of the pocket (Figure 7.5, bottom panel).

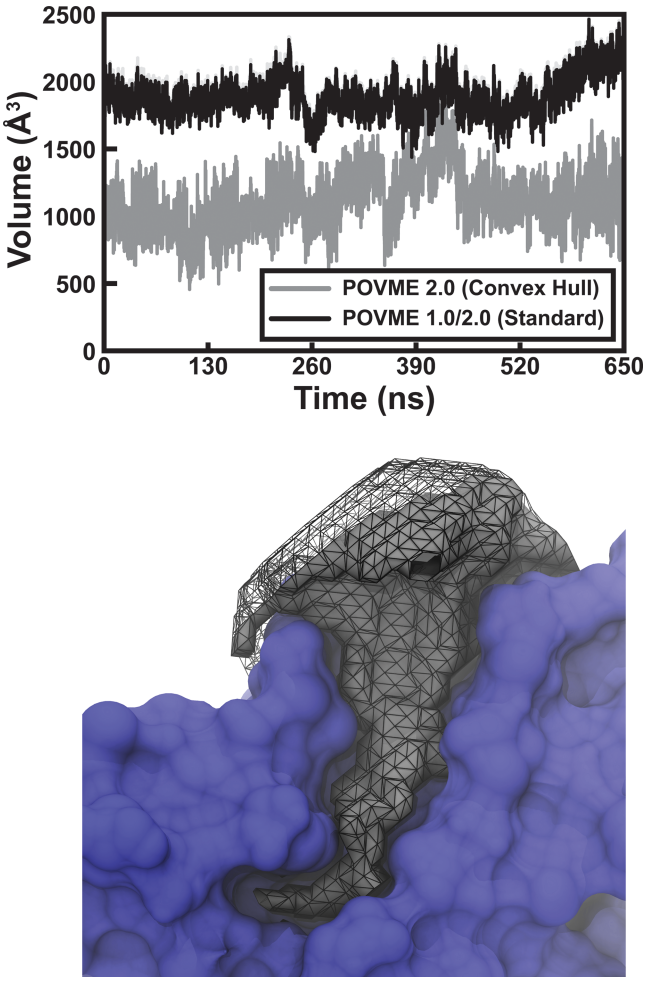

Figure 7.5: POVME 1.0 and 2.0 benchmarks
The graph shows benchmark REL1 pocket volumes as a function of simulation time. POVME 1.0 and 2.0 give nearly the same volume measurements (in black). When the POVME 2.0 convex-hull option is enabled, the volumes are smaller (in grey). The bottom panel, generated using the 1XDN crystal structure, illustrates the difference. When the convex-hull option is enabled, the region of the binding pocket is more accurately captured (solid grey) than when it is deactivated (black wireframe). Some portions of the protein have been removed to facilitate visualization.

7.4 Conclusion

POVME 2.0 is a much improved version of our popular algorithm for characterizing the volumes and shapes of macromolecular (e.g., protein) binding pockets. Version 2.0 implements a number of enhancements, including speed

improvements due to *numpy*/*scipy* integration and the optional use of multiple processors; better accuracy due to an optional convex-hull implementation; additional volumetric-analysis tools (i.e. volumetric density maps); and a graphical user interface that improves usability.

Although pocket-shape and volumetric analyses are not novel, factors such as the high computational cost of most algorithms have discouraged widespread adoption. POVME 2.0 significantly reduces the amount of time required, allowing users to more rapidly analyze large ensembles of pocket shapes derived from multiple experimental structures or simulation methods, such as MD. The added volumetric-density-map analysis feature provides a pocket-centric view of receptor flexibility with potentially useful drug-discovery applications. Indeed, others have shown that docking into structurally distinct binding pockets can lead to enhanced hit rates and chemical diversity.[173-176]

To demonstrate how POVME 2.0 can provide pharmacologically relevant information about pocket flexibility, we used it to analyze the dynamics of an essential, ATP-binding component of the *T. brucei* editosome, *Tb*REL1.[187,188] Given that ATP-binding pockets are ubiquitous, small-molecule inhibitors that bind exclusively to the primary REL1 pocket may also bind to the ATP pockets of critical human enzymes, leading to undesirable side effects. Consequently, we considered a unique secondary binding pocket that is transiently accessible from the primary REL1 ATP-binding pocket. POVME suggests this transient pocket assumes an open conformation roughly 25% of the time. Identifying less promiscuous REL1 inhibitors that exploit this unique pocket is an important component of our ongoing efforts to

target this crucial enzyme.

Chapter 7, in full, is a reprint of “POVME 2.0: An Enhanced Tool for Determining Pocket Shape and Volume Characteristics”, which was published in 2014 in the Journal of Chemical Theory and Computation, volume 10, issue 11, pages 5047-5056, by Jacob D. Durrant, Lane W. Votapka, Jesper Sørensen, and Rommie E. Amaro. The dissertation author was the secondary investigator and author of this paper.

Chapter 8: Multiscale Estimation of Binding Kinetics Using Brownian Dynamics, Molecular Dynamics and Milestoning

The kinetic rate constants of binding were estimated for four biochemically relevant molecular systems by a method that combines Brownian dynamics simulations with more detailed molecular dynamics simulations using milestoning theory. The rate constants found using this method were in good agreement with experimentally and theoretically obtained values. We predicted the association rate of a small charged molecule toward both a charged and an uncharged sphere and verified the estimated value with Smoluchowski theory. We also calculated the k_{on} rate constant for superoxide dismutase with its natural substrate, O_2^-, in a validation of a previous experiment using similar methods but with a number of important improvements. We also calculated the k_{on} for a new system: the N-terminal domain of Troponin C with its natural substrate Ca^{2+}. The k_{on} calculated for both systems closely resemble experimentally obtained values. This novel multiscale approach is computationally cheaper and more parallelizable compared to other methods of similar accuracy. We anticipate that this methodology will be useful for predicting kinetic rate constants and for understanding the process of binding between a small molecule and a protein receptor.

8.1 Introduction

Estimating kinetics is an important and challenging task in computational biophysics. The kinetic rate constants of ligand-receptor interactions, in particular the k_{on} and k_{off} values, play an important role in enzymology[191] and drug discovery[192]. Kinetic rate constants of ligand-receptor association and dissociation are important determinants of drug efficacy[192], and the optimization of these quantities is an important problem in medicinal chemistry. Although these values may often be measured experimentally, an accurate computational estimate would be attractive in cases where experimental measurement is expensive or difficult. In addition, advances in computational power, particularly in parallel computing, offer great potential for methods that take advantage of the vast and increasing power of computation.

As indicated in Eq. 8.1, ligands typically bind to receptors according to a second order reaction process with a rate constant of k_{on}. Unless a nonreversible reaction occurs, ligands typically unbind from their receptors according to a first order process with a rate constant of k_{off}.

$$\boldsymbol{R + L} \underset{\boldsymbol{k_{off}}}{\overset{\boldsymbol{k_{on}}}{\rightleftarrows}} \boldsymbol{RL} \qquad \text{Eq. 8.1}$$

A number of computational techniques exist to predict rate constants. The timescale of kinetic events vary wildly in biomolecular systems, and can extend between 10^8 events per second to less than 1 event per hour[191] for a single reaction event at physiological concentrations of reactants. For computational methods that estimate kinetic quantities, there is typically a high correlation between accuracy and

computational cost. Explicit all-atom molecular dynamics (MD) is one approach to estimate the k_{on} between a protein and a small molecule[193-196]. Though it offers a relatively high degree of accuracy, this technique involves extensive cyberinfrastructure overhead or access to specialized hardware such as the Anton machine[197]. To our knowledge, the longest MD simulations to date are limited to the low millisecond range[198].

Various simplification theories and algorithms offer cheaper alternatives to making kinetic approximations using brute-force, all-atom explicit MD simulations. Examples include two closely related techniques: Markov state models (MSM)[199-207] and milestoning[1,2,208-211] among many others.

Brownian dynamics (BD) is a simulation method used to model macromolecular diffusion in an aqueous solvent[212]. Compared to MD simulations of intermolecular encounters, BD simulations typically require far less computation to simulate an association event. Due to various approximations, including rigid body dynamics, reduced point-charge interactions, implicit solvent, and relatively large timestep, millions of protein/small molecule binding or association events can be simulated in 24 hours using modest parallelization. However, the approximations and assumptions made when using BD to simulate molecular binding can also introduce inaccuracies. BD can be used alone to model ligand association[213]. However, an accurate recovery of experimentally determined observables related to a binding process frequently requires additional models to approximate physical effects due to solvation shells & polarization, solvent entropic effects, and solute internal degrees of

freedom. Schemes to include these factors in BD simulations have been implemented[214-217].

Methods for combining the speed of rigid body BD simulations with the precision of all-atom MD simulations to predict kinetics have been used in the past. In a technique invented by Luty, El Amrani, & McCammon, the k_{on} of superoxide dismutase (SOD) with its natural substrate O_2^- was estimated by partitioning space into a region close to the binding site for simulation with MD, and a region far from the binding site where simulation with BD was more appropriate[218,219]. The statistics of each were combined into a k_{on} estimate using a MSM.

Although Luty et. al.'s original method dramatically decreased the cost to estimate binding kinetics compared to brute-force MD, a number of optimizations can be made to the procedure. Though proportionally smaller, the MD regime was disproportionally more expensive than the BD in Luty et. al.'s initial implementation. In this work, we used milestoning theory instead of an MSM to utilize the transition probabilities and incubation times between states. We modified Luty et. al.'s method by further partitioning the MD regime into additional milestones. We also used a first hitting point distribution (FHPD) as the starting phase space points for the milestoning trajectories rather than an equilibrium distribution[208,210], a required procedure in milestoning theory. It is interesting to note that Luty et. al.'s method was remarkably similar to milestoning. Their use of surface states in phase space and a transition matrix to represent traversal between the states was somewhat prescient. However, Luty et. al. did not go so far as to integrate time information into the method to estimate mean first passage times (MFPT), nor did they did use FHPDs.

Milestoning proper came later[1] and the formalism has since been extensively developed by others[1,2,208-210]. A milestoning model is very similar to an MSM; so much so that milestoning techniques have been used to perform MSM calculations[3], and a number of papers provide extensive comparisons of the two approaches[2,220,221].

In addition to repeating the analysis of SOD made by Luty et. al. with our new method, we also estimated k_{on} values for three additional systems. We calculated k_{on}s for two simple, analytically verifiable "spherical receptor" systems: the rate that a Na^+ particle crosses an uncharged sphere of radius 6.0, and the rate the same particle crosses a charged sphere of radius 6.0 (Figure 8.1). We also estimated the k_{on} of binding between the N-domain of Troponin C (TnC) and its natural substrate Ca^{2+}. Since experimentally measured k_{on}s existed for each of the two protein systems mentioned above, we attempted to closely recreate the experimental conditions within our simulations and subsequently recapture the correct k_{on}s to validate our methods. Armed with this technique, one can make new attempts to estimate kinetic values for biologically or pathogenically interesting systems.

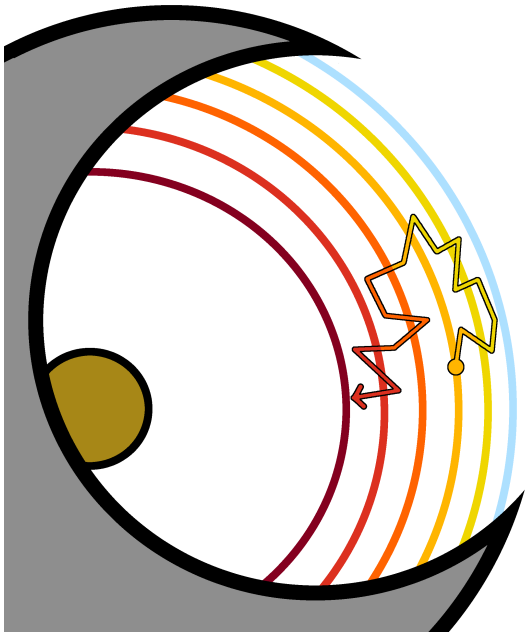

Figure 8.1: A cartoon depiction of a hypothetical path taken by a ligand as it diffuses in the vicinity of its binding site in the MD simulation regime
As the ligand travels, it crosses a series of milestones. Upon crossing, the ligand is considered to be in the crossed milestone's state until it diffuses across a different milestone. The trajectory is terminated when the ligand crosses the "binding surface", where it is considered bound, or when it crosses the BD surface, thus existing the MD simulation regime.

8.2 Theory

8.2.1 Molecular dynamics

MD is a simulation technique that uses Newton's or Langevin's equations of motion in combination with a specified molecular bond structure, parametrized force fields, and a starting conformation of atomic positions and velocities in order to propagate the dynamics of atoms within a molecular system. Ensembles of conformations or trajectories can be sampled to estimate thermodynamic or kinetic quantities[208,222,223].

8.2.2 Brownian dynamics

In addition to MD, BD simulation is another technique can be used to model macromolecular diffusion in an aqueous solvent[212,213,224,225]. BD can also be used to model the association of biomolecules in solution[226]. BD simulations rely on the assumptions inherent in the theory of Brownian motion[212,226,227]. In its simplest form, these assumptions include: a solvent whose atoms may be approximated by a dielectric and ionic continuum and whose hydrodynamic properties can be described using diffusion coefficients or tensors, solute molecules that can be adequately represented as rigid bodies, and forces that can be reduced to electrostatics, steric hindrances, and other inter-solute interactions. BD simulations are propagated according to the general equation of Brownian motion[215] (Eq. 2) which has been derived from the N-particle Fokker-Planck Equation[228,229].

$$\boldsymbol{d}\begin{pmatrix}\boldsymbol{x_i}\\ \boldsymbol{\varphi_i}\end{pmatrix} = \frac{\boldsymbol{dt}}{\boldsymbol{k_B T}}\mathbf{D}\cdot\begin{pmatrix}\mathbf{F}_i\\ \mathbf{T}_i\end{pmatrix} + \sqrt{2\boldsymbol{dt}}\mathbf{S}\cdot\boldsymbol{w} + \boldsymbol{\nabla}\cdot\mathbf{D}\boldsymbol{dt} \qquad \text{Eq. 8.2}$$

Where i is the index of a particle in the system. The values $\boldsymbol{x_i}$, $\boldsymbol{\varphi_i}$, $\boldsymbol{F_i}$, $\boldsymbol{T_i}$ are the position, rotation, force, and torque of particle i respectively, **D** is the diffusion tensor, **S** is the matrix square root of **D**, **w** is a random vector whose components are Gaussian variables with unit variance and zero mean. The Northrup-Allison-McCammon algorithm[213] solves these equations numerically, and can be used to propagate timesteps within a BD simulation.

When BD is used to estimate the k_{on} of a ligand-receptor association reaction, the k_{on} rate constant can be split into two terms. (Eq. 8.3)

$$\boldsymbol{k_{on} = k_b \beta} \qquad \text{Eq. 8.3}$$

Where k_b is the rate of diffusion of the ligand to a spherical surface of radius b (b-surface) centered on the ligand. The k_b can be analytically calculated by using Eq. 8.4.

$$\boldsymbol{k_b = 4\pi \left[\int_{\mathrm{b}}^{\infty} \frac{\exp\left(U(r)/k_B T\right)}{r^2 D(r)} dr \right]^{-1}} \qquad \text{Eq. 8.4}$$

Where $U(r)$ is the effective potential energy between the sphere and the substrate at a distance r from the center of the sphere, k_B is Boltzmann's constant, and T is temperature, and $D(r)$ is the spatially-varying diffusion coefficient. β is the probability that a ligand located on the b-surface will continue on to react with the enzyme rather than escaping to an infinite distance. Normally, β can be determined by running BD simulations started from random locations on the b-surface and then counting the proportion of trajectories that lead to binding. In this study, β was determined by combining BD with MD using milestoning.

8.2.3 Milestoning Theory

Milestoning computationally models the kinetics as well as the thermodynamics of chemical processes, with the benefit of extensive parallelizability [1,2,210,230]. Using milestoning techniques, the stationary flux distribution **q** and the probability distribution **p** can be found across a reaction coordinate along which a number of milestones have been defined. Milestoning can also be used to find the mean first passage time (MFPT) of a transport process starting from one milestone and ending at another. The methods within milestoning theory provide a flexible

approach to investigate a wide range of dynamics, including non-equilibrium conditions [209,211] and has been applied in a variety of contexts [3,211,220,231,232]. Milestoning does not rely on any assumption concerning system damping [1], and thus can be applied to Newtonian, Langevin, and Brownian systems alike [1,2,210].

In our implementation, we defined a number of concentric spherical surfaces in phase space that encircle the binding site on each receptor. These surfaces in phase space are termed "milestones" and are roughly perpendicular to the reaction coordinate (Figure 8.1).

In a typical milestoning procedure, unbiased simulations are initiated from a set of equilibrium distributions along the milestones, which one obtains using umbrella sampling. Each of these simulations is independent from the others, and the ligand center of mass is positioned at or very near the milestoning surface. As the simulations progress, transitions between milestones are recorded to construct a proper FHPD across the milestones, and then used to construct a transition kernel matrix with elements **K**, whose entries describe the probability that a ligand in one of the milestones will subsequently transition to another. An incubation time vector $\langle \boldsymbol{t} \rangle$ is also obtained by determining the average time the system takes to transition from each milestone to an adjacent one. Given these quantities, stationary fluxes **q** across each milestone, the probability distribution **p**, and MFPT $\langle \tau \rangle$ are found using eqs. 8.5-8.7.

$$\boldsymbol{q}_{stat}(\mathbf{I}-\mathbf{K}) = \mathbf{0} \qquad \text{Eq. 8.5}$$

$$\boldsymbol{p}_{i,stat} = \boldsymbol{q}_{i,stat} \cdot \langle \boldsymbol{t}_i \rangle \qquad \text{Eq. 8.6}$$

$$\langle \boldsymbol{\tau} \rangle = \mathbf{p} \cdot (\mathbf{I}-\mathbf{K})^{-1} \langle \mathbf{t} \rangle \qquad \text{Eq. 8.7}$$

where **I** is the identity matrix, and *i* is the index of a particular milestone.

Milestoning theory, as well as the method employed by Luty et. al., defines states using surfaces in phase space. The current state of the simulated system is the surface that has been most recently crossed. Each of the surfaces must be sufficiently far apart from one another in order to ensure that velocity is decorrelated between transitions. In our implementation, we defined our milestones as concentric spheres in order to closely approximate isosurfaces of the committor function. For a rigorous discussion of 'surface' states and their requirements and assumptions, the reader is referred to additional publications on milestoning theory[1,2,208,210].

In our implementation, all trajectories used to populate the statistics in the milestoning model are started from FHPDs calculated on each of the surface states. The FHPD represents the distribution of system conformations that have just crossed a surface state and that were previously in a different state. The difference between the FHPD and an equilibrium distribution is that the latter also includes conformations whose last crossing event was the same as the current state. A trajectory is started from the FHPD and allowed to propagate according to the simulation dynamics, crossing surfaces as it diffuses (Figure 8.1). If the trajectory ever crosses the surface of a sink state, such as the bound state or a state leading to another simulation regime, the trajectory is halted. As surfaces are crossed, the counts are tallied to construct the transition matrix **K** and the average incubation time vector $\langle \mathbf{t} \rangle$. Error estimation of computed values were made using a Monte Carlo method to sample matrix distributions defined in Milestoning[210,221] similar to one used in MSM

theory[233]. Details of error estimation are outlined in the Supplementary Information (SI) section 8.6.3.

8.2.4 Theoretical determination of k_{on}

The flux rate *k(r)* of a particle across a sphere of radius r may be solved analytically for some simple systems. The value *k(r)* is equivalent to k_{on} if a sphere of radius *r* is modeled as a binding surface. In the uncharged spherical receptor system, there are no average forces on the substrate and the *k(r)* can be obtained by solving the Smoluchowski equation[234].

$$\boldsymbol{k(r) = 4\pi r D} \qquad \text{Eq. 8.8}$$

Where *r* is the radius of the reacting sphere, and *D* is the diffusion coefficient of the substrate. The *k(r)* can also be calculated for systems with centrosymmetric forces by solving the Smoluchowski equation in spherical coordinates[234,235]. The result is expressed as Eq. 8.4. Assuming a constant diffusion coefficient and that the effective potential energy is defined by Coulomb's Law in a uniform dielectric, Eq. 8.4 can be reduced and solved exactly for the charged spherical receptor system (Eq. 9):

$$\boldsymbol{k(r) = -\frac{DQ_cQ_s}{\left[1 - exp\left\{\frac{Q_cQ_s}{4\pi\varepsilon_0\varepsilon_r k_B T r}\right\}\right]\varepsilon_0\varepsilon_r k_B T}} \qquad \text{Eq. 8.9}$$

Where Q_s is the charge of the diffusing particle, Q_c is the charge in the center of the receptor sphere, ε_0 is the permittivity of a vacuum, and ε_r is the dielectric constant of the solvent. The derivation of Eq. 8.9 from Eq. 8.4 is described in the SI

section 8.6.2. A solution to more complicated scenario can also be derived numerically[236].

In addition to the flux rate *k(r)* across spheres, the MFPT that a particle remains within a certain domain of space can also be obtained. For a system that obeys Smoluchowski theory, Eq. 8.10 describes how the MFPT relates to a stationary distribution in that domain.

$$\langle \boldsymbol{\tau} \rangle = \frac{\boldsymbol{N}}{\boldsymbol{J}} = \frac{\int_{\boldsymbol{V}} \boldsymbol{u} \boldsymbol{dV}}{\boldsymbol{D} \sum_{\boldsymbol{i=1}}^{\boldsymbol{k}} \int_{\boldsymbol{A_i}} (\boldsymbol{\nabla u}) \boldsymbol{dA_i}} \qquad \text{Eq. 8.10}$$

Where $\langle \tau \rangle$ is the MFPT, N is the total number of particles present in the system, J is the total flux of particles across all absorbing boundaries at any given time, u is the stationary distribution of particles, V is the volume of the system, D is the diffusion coefficient of the particle, k is the number of absorbing boundaries, and i is the index of a particular absorbing boundary A_i[237].

8.3 Materials & Methods

8.3.1 Preparation of MD

All MD simulations were carried out using NAMD 2.9[50]. The MD FHPDs were made with the help of MDAnalysis[238]. All calculations were performed on the Gordon supercomputer at the San Diego Supercomputer Center, the Stampede supercomputer at the Texas Advanced Computing Center, and on local machines.

8.3.2 Spherical Receptor Systems

MD simulations of the charged and uncharged spherical receptor simulations were prepared using a simple 40 Å x 40 Å x 40 Å TIP3P[83] water box with a Cl^- placed in the center of the box for only the charged spherical receptor. Both systems contain approximately 7600 atoms. Na^+ and Cl^- parameters were obtained from the ions94 library of the AMBER ff03 forcefield[239]. The spherical receptor systems were minimized for 10000 steps to allow the water molecules to relax in relation to each other and to the Cl^-. Both systems were then equilibrated for 20 ns at a constant temperature of 300K using the Langevin thermostat and constant pressure using the Langevin piston at 1 atm with a damping coefficient of 5 ps^{-1}. The Cl^- was constrained to a stationary position in the center of the charged spherical receptor system.

Following this equilibration, four copies were made of the systems, and a Na^+ was placed at the milestones located at 7 Å, 8 Å, 9 Å and 10 Å from the center of the water box in the uncharged system (Figure 8.2), and from the Cl^- in the charged system. Two additional milestones were also placed at 6Å and 11 Å. Waters clashing with the Na^+ were removed. The system was once again allowed to minimize for another 5000 steps to relax the waters around the ions. Then the system was heated in 10 K increments up to 350 K and then reduced back to 300 K at 2 ps intervals each at constant volume. Then, in order to obtain an ensemble distribution, the systems were simulated at constant temperature at 300 K at constant volume for 20 ns. To this point, all ions have been constrained. In order to obtain a FHPD, 900 position/velocity configurations were uniformly chosen between the 2 ns and 20 ns

marks in the ensemble simulations. Velocities were reversed, and the trajectories were allowed to propagate backwards. If the trajectory struck another milestone before re-crossing the one it came from, that trajectory was considered part of the FHPD. All members of the FHPD were then allowed to proceed with their velocities in the forward direction. Each transition event was monitored for future milestoning analysis. Reverse simulations were carried out using a special plugin for NAMD 2.9 by Cameron Abrams, which allows velocities to be reversed at arbitrary timesteps.

For comparison with the milestoning results, brute-force MD simulations were run and Smoluchowski theory were calculated to obtain a β, k_{on}, and MFPT for the spherical receptor systems. All brute-force MD simulations were set up with the same parameters as for milestoning above, except that the system was equilibrated for 40 ns and 10000 frames were sampled between the 20 and 40 ns time. Each of the 10000 simulations were started with the Na^+ placed on the 10Å milestone and monitored for a crossing event at either the 6Å or the 11Å milestone. The value β was simply the number that crossed the 6Å milestone out of the total number of simulations. The MFPT was the average amount of time that all the simulations lasted before a crossing event.

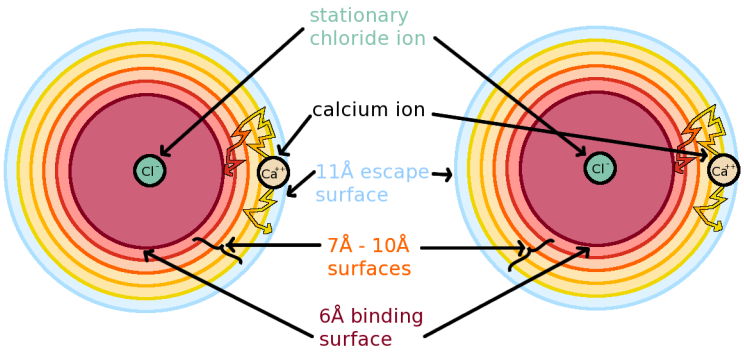

Figure 8.2: A cartoon depiction of the two spherical receptor systems drawn approximately to scale

The uncharged system in panel A has no central charged molecule. The charged system in panel B has a Cl^- constrained to the center of the spherical milestones. Both systems contain the escape milestone (light blue curves), four intermediate milestones (curves in shades of orange and yellow), and binding milestone (dark red curves). Two hypothetical paths are also depicted per system. The upper path shows a trajectory where Na^+ diffuses within the simulation region, crossing surfaces and finally reacting with the 6Å spherical milestone. The bottom path shows Na^+ diffusing across a few states before escaping to the 11Å milestone.

8.3.3 SOD System

MD force field (FF) parameters for SOD were obtained as a generous gift from Branco et. al.[240] The system was surrounded by a TIP3P[83] water box with 150 mM NaCl solution. The simulation contained approx. 44,000 atoms. The SOD system was then equilibrated for 80 ns at a constant temperature of 300 K using the Langevin thermostat and constant pressure using the Langevin piston at 1 atm using a damping coefficient of 5 ps^{-1}.

Following equilibration, ten copies were made of the apo system, and O_2^- was inserted at eight different milestones (located at 4Å-11Å in 1Å increments) from each of the two copper ions in SOD's two active sites, yielding a total of sixteen different milestones simulated (Figure 8.3). Waters clashing with O_2^- were removed. The

solvent molecules in the system were minimized for another 5000 steps to relax around the newly placed ions. Then the system was heated in 10 K increments up to 350 K and then reduced back to 295 K at 2 ps intervals each at constant volume. The protein and O_2^- atom positions were constrained during the minimizations and heating/cooling. In order to obtain an ensemble distribution, the systems were simulated at a constant temperature of 300 K and constant volume for 200 ns each with an imposed harmonic "spring" force of 300 kcal mol^{-1} $Å^{-2}$ that constrained O_2^- close to a spherical milestone at each system's proper distance from the SOD active site catalytic copper. In order to obtain a FHPD, 700 position/velocity configurations were uniformly chosen between the 60 ns and 200 ns marks in the ensemble simulations. Velocities were reversed, and the trajectories were allowed to propagate backwards in time. If the trajectory struck another milestone before recrossing the one it came from, that trajectory was considered part of the FHPD. The autoimage function in CPPTraj[241] was used to center the ligand in the waterbox before the reversal stage. All members of the FHPD were then allowed to proceed in the forward direction. Each crossing event was monitored for future analysis. The reversal phases were simulated using the custom plugin for NAMD 2.9 by Cameron Abrams.

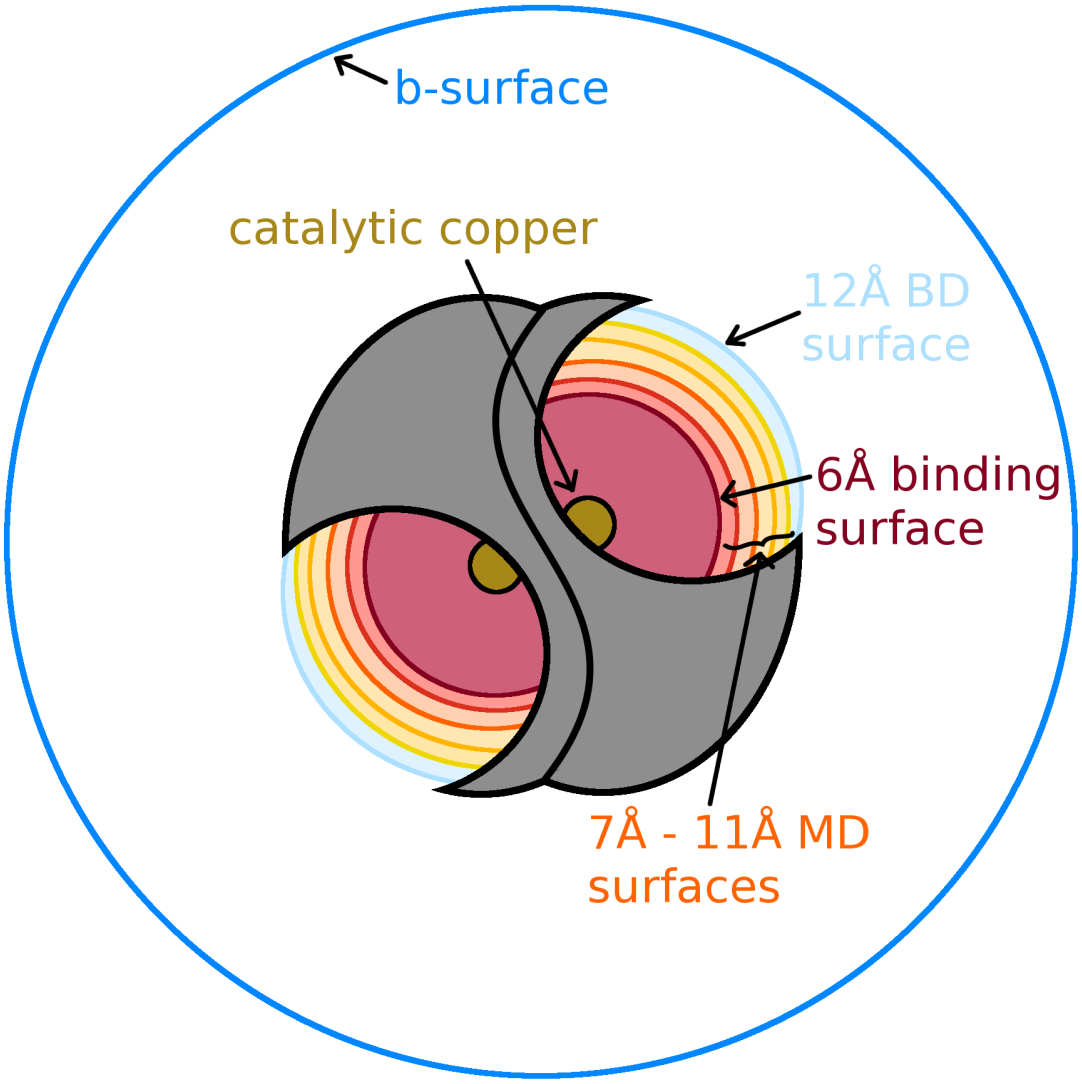

Figure 8.3: A cartoon depiction of the SOD system
The system has two binding sites, b-surface (dark blue circle), BD milestones (light blue curves), MD milestones (curves in shades of orange and yellow), and the binding milestone (dark red curves). The catalytic coppers at the center of the spherical surfaces are also depicted as tan circles in the bottom of each active site.

FF parameters for TnC were prepared according to the protocol followed by Lindert et. al.[140] The system was surrounded by a TIP3P[83] waterbox with 100 mM KCl solution. The simulation contained approximately 27,000 atoms. The TnC

system was then equilibrated for 100 ns at a constant temperature of 288 K using the Langevin thermostat and pressure using the Langevin piston at 1 atm using a damping coefficient of 5 ps^{-1}.

Following this equilibration, twelve copies were made of the systems, and the Ca^{2+} was inserted on the binding side of the TnC site II loop at 1 Å increments from 2 Å to 9 Å from the center of mass of the alpha carbons of residues ASP 65, ASP 67, SER 69, THR 71, and GLU 76 (Figure 8.4). Waters clashing with Ca^{2+} were removed. The solvent molecules in the system were minimized for another 5000 steps to relax around the newly placed ions. Then the system was heated in 10 K increments up to 350 K and then reduced back to 295 K at 2 ps intervals each at constant volume. The protein and Ca^{2+} atoms were constrained during the minimizations and heating/cooling cycles. In order to obtain an ensemble distribution, the systems were simulated at a constant temperature of 300 K and constant volume for 100 ns each with an imposed harmonic force of 300 kcal mol^{-1} $Å^{-2}$ that constrained Ca^{2+} close to the spherical surface at each system's proper distance from the active site center of mass. In order to obtain a FHPD, 700 position/velocity configurations were uniformly chosen between the 30 ns and 100 ns marks in the ensemble simulations. Velocities were reversed, and the trajectories were allowed to propagate backwards in time. If the trajectory struck another milestone before re-crossing the one it came from, that trajectory was considered part of the FHPD. CPPTraj[241] was used to center the ligand in the waterbox before the reversal stage. All members of the FHPD were then allowed to proceed in the forward direction. Each crossing event

was monitored for future analysis. The reverse phase of MD simulations were carried out using a custom plugin for NAMD 2.9 developed in Cameron Abrams' lab.

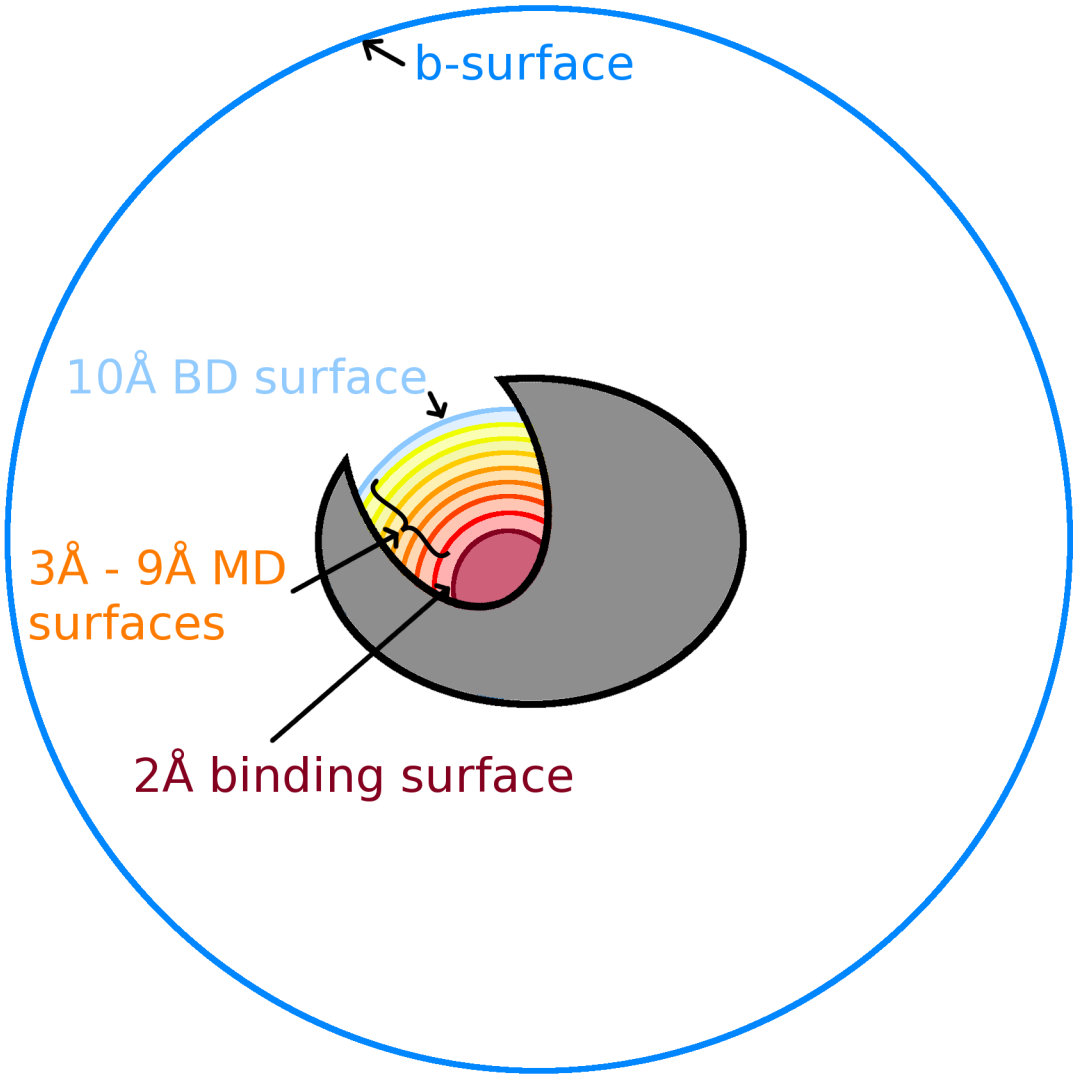

Figure 8.4. A cartoon depiction of TnC
The system contains a b-surface (dark blue circle), BD surfaces (light blue curves), MD surfaces (curves in shades of orange and yellow), and binding surface (dark red curves) all located in the Ca^{2+} binding site (site II). Each curve represents a milestone.

8.3.4 Preparation of BD

All Brownian dynamics simulations were performed using BrownDye[215] with desolvation forces and hydrodynamic interactions activated. All electrostatics calculations were performed using the Poisson-Boltzmann Equation solver APBS[242]. The solvent dielectric was left at the default of 78, the solvent viscosity was left at the default of $8.9 * 10^{-4}$ $kg*m^{-1}s^{-1}$, and the permittivity of a vacuum was left at the default of 8.854×10^{-12} $C^2N^{-1}m^{-2}$. All solute dielectrics were set to 2. A 6-12 hard sphere Lennard-Jones interaction was used. Simulations were distributed across 10 to 20 threads on a local computing node. The BrownDye program bd_top was used to prepare all systems for simulation. A phantom atom of zero charge and zero radius was placed at the center of the active sites in order to detect crossings of spherical milestones. The phantom atom has no effect on the dynamics, but is merely a convenient way to detect surface-crossing events. The BrownDye program nam_simulation was used for simulation, and the program compute_rate_constant was used to aid in the calculation of the association rate constants. Trajectories were processed using the BrownDye programs process_trajectories and xyz_trajectory in combination with in-house Python scripts.

8.3.5 BD for SOD

A PQR file for SOD was prepared from the crystal structure PDB ID: 1CBJ[243] using LEaP[244] and DelEE[245] with the AMBER forcefield[246,247] and PROPKA,[47] assigned protonation states at a pH of 7.0. A PQR file for O_2^- was made by hand, with each oxygen given a partial charge of -0.5 and a radius of 1.5 Å. APBS[242] was then

used to calculate the electrostatic field at 295 K and a NaCl concentration of 150 mM to approximate conditions used during the experimental measurement of k_{on} for SOD[248]. BrownDye was used to prepare and run 1×10^6 BD simulations at 295 K with the ligand starting from a b-surface at ~61 Å from the SOD center of mass. Based on experimentally determined diffusion coefficient[249] of 1.5×10^{-5} cm^2s^{-1}, a hydrodynamic radius of 2.01 Å was used for O_2^- in the simulations (See SI section 8.6.1). Reactions with both active sites, and also escape events were counted. 1000 configurations of ligand encounters with both active sites (12 Å from catalytic copper) were extracted to make two additional FHPD distributions. 1000 simulations were started from each configuration (2×10^6 total). These were allowed to react with a surface further down the site (11 Å from the catalytic copper) react with the surface around the other site (12 Å from the other catalytic copper) or escape to infinity. All reaction and escape events were counted to construct the statistics of the transition kernel **K** and incubation time vector $\langle \mathbf{t} \rangle$.

8.3.6 BD for Troponin C

A PQR file for TnC was prepared from the crystal structure 1SPY[250]. Partial charges were assigned according the AMBER forcefield[244] using LEaP [246]and DelEE[245] and PROPKA[247] assigned protonation states at a pH of 7.0. A PQR file for Ca^{2+} was made by hand, given a charge of 2.0 e and an atomic radius of 1.14 Å. APBS[242] was then used to calculate the electrostatic field at 288 K and a KCl concentration of 100 mM to approximate conditions used during the experimental measurement of k_{on} and k_{off} for TnC[251]. A hydrodynamic radius of 5.5 Å was assigned

based on an experimentally determined diffusion coefficient[252] of 6.73×10^{-6} cm^2s^{-1} at 291 K (See SI). BrownDye was used to prepare and run 1×10^6 BD simulations at 288K with the ligand starting from a b-surface at ~57 Å from the TnC center of mass. Diffusion to the active site surface, and escapes were counted. 1000 configurations of ligand encounters with the active site (10 Å from binding site center of mass of residues ASP 65, ASP 67, SER 69, THR 71, and GLU 76) were extracted to make a FHPD distribution. 1000 simulations were started from each configuration (1×10^6 total). These were allowed to react with a surface further down the site (7 Å from binding site center) or escape to infinity. All reaction and escape events were counted to construct the milestoning model.

8.3.7 Theoretical Calculations

For our spherical receptor calculations, we used a dielectric of 92 to mimic the dielectric of TIP3P water[253], a permittivity of 8.854×10^{12} $C^2N^{-1}m^{-2}$, and a diffusion coefficient[252] of 1.33×10^{-5} cm^2s^{-1} for Na^+.

The rate constants *k(a), k(b),* and *k(q)* were calculated using Eq. 8.8 for the uncharged spherical receptor and Eq. 8.9 for the charged spherical receptor for the reaction surface, b-surface, and q-surface, respectively. The rate constant *k(a)* is the analytic solution to the spherical receptor association. For comparison, we deduced *k(a)* using only *k(b),* and *k(q)* by using a transition matrix **K** obtained from monitoring transitions of the spherical receptor systems in a series of MD simulations. A binding probability β was calculated using Eq. 8.5 where $\beta=q_{stat,i}$

where *i* is the index of the bound milestone that has been modified to be a sink state. The k_{on} for each spherical receptor system was calculated using Eq. 11.

$$\boldsymbol{k_{on} = k(b)\left(\frac{\beta}{1-(1-\beta)\left(\frac{k(b)}{k(q)}\right)}\right)} \qquad \text{Eq. 8.11}$$

The MFPT represents the mean time taken a particle started on the b-surface and allowed to diffuse until touching either the reaction surface or the q-surface. The MFPT was calculated using Eq. 8.10. The values *k(b)* and *k(q)* are obtained using Eq. 8.8 or Eq. 8.9, depending respectively on the absence of presence of a receptor charge.

8.3.8 Milestoning Calculations

For each system, the milestoning calculations were performed using custom scripts that used Numpy 1.7, Scipy 0.9.0 and the GNU Parallel tool[254].

8.4 Results

Using Smoluchowski theory, milestoning, and brute force MD simulations, the probability β of each system starting on the b-surface and continuing on to touch the reaction surface is listed in Tables 1 and 2 along with the resulting k_{on}. The MFPT is also listed for the spherical receptor systems.

Table 8.1: Computationally and theoretically determined results for the uncharged spherical receptor system
All simulations were carried out in a dilute aqueous environment. β is the probability of a particle starting on the b-surface to reach the bound state before touching the q-surface. MFPT refers to the mean first passage time of a particle started on the b-surface to reach either the reaction surface or the q-surface.

Method	β	k_{on} ($M^{-1}s^{-1}$)	**MFPT (ps)**
Milestoning MD	0.113±0.012	$5.9\pm0.9\times10^{9}$	7.4±0.5
Analytic solution (using Eq. 8.7)	0.12	6.039×10^{9}	13.5
Brute-force MD	0.114±0.013	$5.9\pm0.9\times10^{9}$	7.2±0.3

Table 8.2: Computationally and theoretically determined results for the charged spherical receptor system
All simulations were carried out in a dilute aqueous environment. β is the probability of a particle starting on the b-surface to reach the bound state before touching the q-surface. MFPT refers to the mean first passage time of a particle started on the b-surface to reach either the reaction surface or the q-surface.

Method	**β**	k_{on} ($M^{-1}s^{-1}$)	**MFPT (ps)**
Milestoning MD	0.127±0.013	$9.1\pm1.3\times10^{9}$	7.6±0.4
Analytic Solution (using Eq. 8.9)	0.146	9.589×10^{9}	14.2
Brute-force MD	0.135±0.012	$9.3\pm1.2\times10^{9}$	8.3±0.3

Using the stationary probabilities obtained with milestoning of SOD, Eq. 8.5, Eq. 8.6, and Eq. 8.12 below, we constructed a free energy profile for the approach of O_2^- to the SOD binding site (Figure 8.5) setting that the 10Å milestone to zero energy as a reference.

$$\Delta G_i = -k_b T \ln\left(\frac{p_{i,stat}}{p_{ref,stat}}\right) \qquad \text{Eq. 8.12}$$

Where ΔG_i is the estimated free energy of milestone i, k_B is Boltzmann's constant, T is temperature, and $p_{i,stat}$ and $p_{ref,stat}$ are the stationary probabilities of milestone i and the reference milestone at 10Å, respectively, obtained using Eq. 8.6.

Luty et. al. assumed that the bound state was a spherical surface of radius 6Å centered on the catalytic copper. This location does appear to have a shallow local minimum at 6Å in the free energy as depicted in Figure 8.5. Because Luty et. al. assumed that the 6Å sphere was the bound state, and because it is the location of a shallow local minimum in the free energy profile in Figure 8.5, we assume that 6Å is the bound state in all subsequent SOD milestoning calculations.

Table 8.3: Computationally and experimentally determined k_{on}s for SOD by us and others

The experimental value that this study attempted to emulate[248] measured a k_{on} is listed along with the k_{on} that Luty et. al. determined for SOD using different simulation conditions and model setup.

Researchers	k_{on} ($M^{-1}s^{-1}$)	Temp. (K)	Ion Conc. (mM)	Method
This study	$8.8\pm0.7\times10^{8}$	295	150 NaCl	MD/BD/milestoning
Cudd, et. al.[248]	8.5×10^{8}	300	140 NaCl	Pulse-Radiolysis
Argese, et. al.[255]	1.6×10^{9}	295	160 $NaClO_4$	Polarographic method of catalytic currents & NMR
Luty, et. al.[218]	$1.62\pm0.86\times10^{9}$	300	0	MD/BD, 7-state MSM

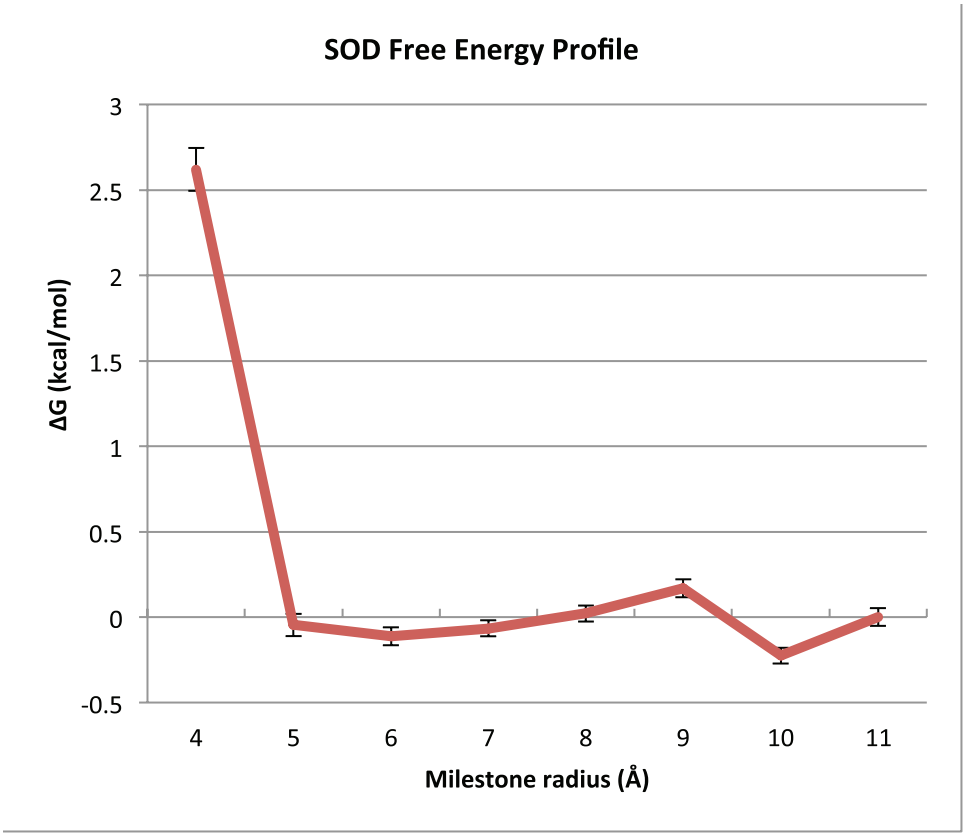

Figure 8.5: SOD system free energy profile
This plot depicts the free energy in kcal/mol at each milestone along the reaction coordinate in the SOD system relative to the 10Å milestone, the nearest to the bulk solution. These free energies were computed using milestoning theory according to Eq. 8.12. A slight local minimum occurs at 6Å and we assume this to be the bound state.

As with the SOD system, we used the stationary probabilities obtained with milestoning of TnC, Eq. 8.5, Eq. 8.6, and Eq. 8.12 to construct a free energy profile for Ca^{2+} in its approach to the TnC binding site (Figure 8.6) with the 10Å milestone free energy as the reference. According to this profile, the lowest energy state is located at 3Å from the binding site center. We assume that when the Na^{+} has reached this distance, it is in the bound state. We use a 3Å binding surface for all subsequence milestoning calculations on TnC.

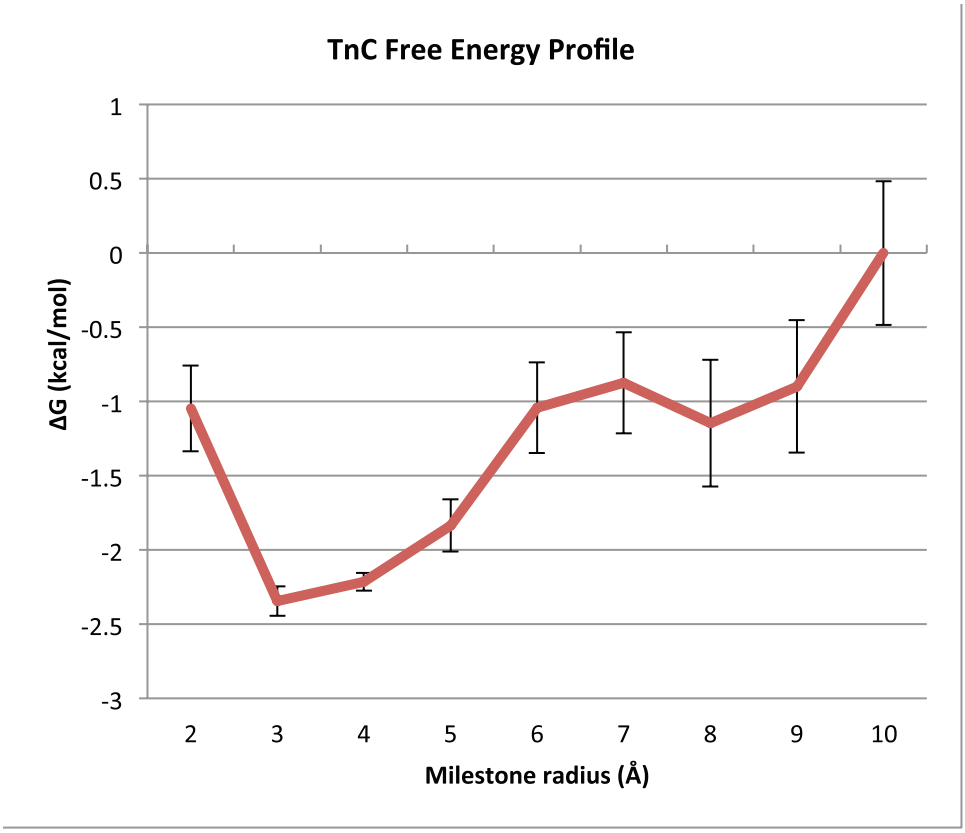

Figure 8.6: TnC system free energy profile
This plot depicts the free energy profile of TnC binding (in kcal/mol) relative to the 10Å milestone at each milestone along the reaction coordinate of the TnC system. These free energies were computed using milestoning theory according to Eq. 8.12. The lowest relative free energy is at the 3Å milestone, the location we assume to be the bound state of the TnC system.

Table 8.4: Computationally and experimentally determined k_{on}s for TnC by us and others
The k_{on} we predicted is listed along with that of the experimental value that this study attempted to emulate, along with additional experimental k_{on}s.

Researchers	K_{on} ($M^{-1}s^{-1}$)	Temp. (K)	Ion Conc. (mM)	Method
This study	$1.5\pm0.7\times10^{8}$	288	100 KCl	MD/BD/milestoning
Tikunova, et. al.[251]	$1.7\pm0.3\times10^{8}$	288	90 KCl	Stopped-flow
Hazard, et. al.[256]	$2\text{-}4\times10^{8}$	277	90 KCl	Stopped-flow
Ogawa, et. al.[257]	$>4.0\times10^{7}$	293	100 KCl	Stopped-flow

In addition to the calculation of k_{on} rate constants, the milestoning models and distributions across the states can be used to visualize the path of the ligand in its approach to association within the binding site. The FHPD for SOD at 12 Å is displayed in Figure 8.7 and the FHPD for TnC at 10 Å is displayed in Figure 8.8.

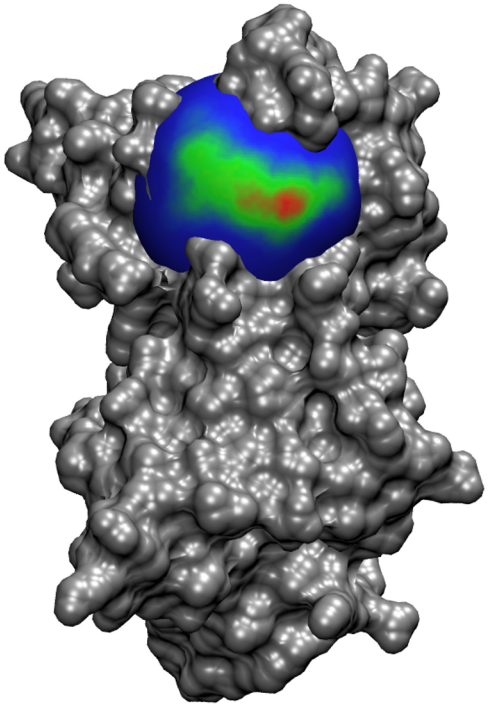

Figure 8.7: The FHPD for O_2^- encounter on the 12Å around the active site of SOD

Blue indicates zero crossing events per square Å, and the color scale increases to red, indicating up to 1.2×10^5 crossing events among all 1×10^6 simulations. The distribution suggests that O_2^- approaches directly from the solvent instead of approaching laterally from another portion of the protein surface. The image was generated using VMD[30] with an MSMS surface[77].

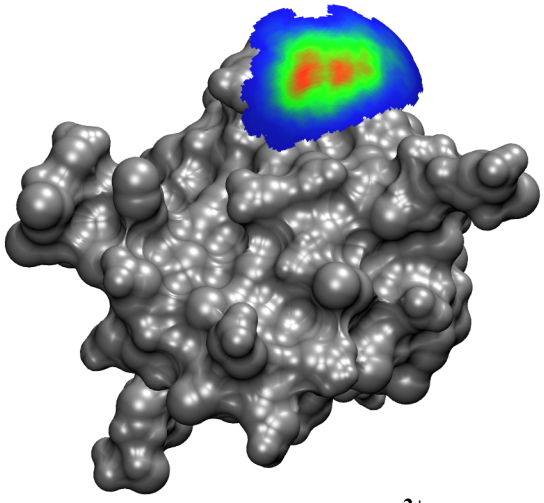

Figure 8.8. The FHPD for Ca^{2+} encounter on the 10Å around the binding site of TnC

Blue indicates zero crossing events per square Å, and the color scale increases to red, indicating up to 8.8×10^{4} crossing events among all 1×10^{6} simulations. The distribution suggests that Ca^{2+} approaches the site directly from the solvent instead of approaching laterally from another portion of the protein surface. The region of the sphere where no ligands crossed was removed to reveal the site II loop over the binding site, though the binding site itself is concealed by the FHPD. The image was generated using VMD[30] with an MSMS surface[77].

8.4.1 Computational Performance

The total computational cost of all systems simulated in this study for both MD and BD was approximately 65,000 CPU hours. Computational costs of each simulated system and simulation regime are listed in Table 8.5. The cost of performing all non-simulation calculations was negligible. Table 8.5 includes all computer time spent on the supercomputer as well as on local machines. The β, k_{on}s and error estimates for all systems were well-converged and are reported in the SI section 8.6.

Table 8.5: The computational cost of calculating kinetics for each system using milestoning
BD simulations were not run for the spherical receptor systems, so no costs are listed. Also, brute force MD simulations were not run for SOD and TnC.

System	Cost of MD (CPU-hours)	Length of MD (ns)	Cost of BD (CPU-hours)	Computer used for MD	Computer used for BD	Cost of Brute-force MD (CPU-hours)
Uncharged spherical receptor system	~600	~100	-	Linux desktop	-	~1350
Charged spherical receptor system	~600	~100	-	Linux desktop	-	~1450
SOD	~53,000	~1630	~100	Stampede Supercomputer	Linux desktop	-
TnC	~5100	~900	~100	Gordon Supercomputer	Linux desktop	-

8.5 Discussion

8.5.1 Idealized Systems

The k_{on} calculated using milestoning for the uncharged spherical receptor system matches within 3% to the theoretically determined value and 0.3% to the brute-force MD value. These estimates are well within the bounds of uncertainty introduced by the milestoning model. As a system that can diffuse freely without forces or solvation shells, it is expected that Smoluchowski theory would yield such a close result to simulation. This similarity to a value obtained using well-established theory is a good validation of our basic methodology. The large difference between the MFPT predicted by theory and the MFPTs predicted by milestoning and brute

force MD could be due to a difference between the experimentally measured diffusion coefficient of Na^+, and the diffusion coefficient that is observed in an MD simulation using the AMBER forcefield.

The k_{on} calculated using milestoning for the charged spherical receptor system differs by 13% from the k_{on} predicted by Smoluchowski theory and by only 6% from the k_{on} obtained by brute force MD simulation. This difference between the simulation-obtained values and the value obtained by theory is likely due to effects caused by the explicit solvent in our simulations, for which this simple implementation of Smoluchowski theory does not account. Very likely, solvation shells have formed around the Cl^- placed in the center of the system, as well as the diffusing Na^+. Solvation shells create unevenness in the potential of mean force and the position-dependent diffusion coefficient of Eq. 8.4. As such, using Coulomb's law for the electrostatic potential and a constant diffusion coefficient may not be sufficiently valid assumptions for ions in solution at such close proximity. Previous studies on close NaCl ion pair interactions in dilute solvent show oscillations in the mean force potential of the interionic distance that extend several molecular layers into the solvent[258-260]. Accounting for these factors and using an alternative solution to Eq. 8.4 would likely result in a calculated value much closer to what we obtained using milestoning and the brute force MD. The fact that the milestoning results and the brute-force MD results are so similar supports the validity of the milestoning methodology. Similarly, with the charged receptor, the large difference in the MFPT predicted by theory and the MFPTs predicted by milestoning and brute force MD could be due to a difference between the experimentally measured diffusion

coefficient of Na^+, and the diffusion coefficient that would be observed in an MD simulation using the AMBER forcefield. It could also be due to same effects observed on β caused by the aforementioned solvation shells.

8.5.2 Superoxide Dismutase (SOD)

SOD is an enzyme found in a wide variety of organisms[255]. It is a homodimer that makes use of a catalytic copper bound in its active site to catalyze the dismutation of the superoxide ion O_2^- into O_2 and H_2O_2 [248,255]. SOD was the subject of many early enzymology experiments[261] and ligand-receptor binding simulations[225,262].

The SOD k_{on} estimated using milestoning is within 4% of the experimentally measured k_{on}; this falls well within the uncertainty bracket calculated for the milestoning model. The k_{on} we calculated is also close to the value obtained by Luty, et. al. in their seminal study of SOD kinetics[218]. It is well understood that a higher salt concentration slows the rate of O_2^- binding to SOD[255]. Therefore, the k_{on} measured in this study is likely smaller than the value measured by Luty, et. al. because they simulated MD and BD with a solvent salt concentration of zero. The discrepancy could also be due to differences used by Luty et. al. in their implementations of atomic constraints on the protein, different boundary conditions in the MD phase, and the lack of desolvation forces in the BD phase.

While it is not clear how much error is introduced by using an equilibrium distribution across the milestones, our use of a FHPD should, theoretically, provide a more accurate treatment due to its consistency with formal milestoning theory[1,2]. An assertion reinforced by the similarity of our calculated SOD k_{on} rate to the

experimental value. The insertion of additional states in the MD region also allowed us to obtain much better sampling of transition events than would be available for a comparable computation time if the MD region had been composed of only a single milestone. The FHPD of SOD at 12Å (Figure 8.7) indicates that O_2^- approaches directly from the solvent and does not seem to sample much of the protein surface before entering the active site. Although a k_{on} has already been obtained for this system by Luty et. al. using similar methods, our approach offers a number of key improvements and very closely resembles the experimentally obtained rate constant; both insofar as the conditions that the system was exposed to, as well as the final result.

8.5.3 Troponin C (TnC)

In order to try this milestoning method on a new system, we also calculated the k_{on} of TnC. The troponin complex is a set of proteins that regulates muscle contraction in skeletal and cardiac muscles[250,251,257]. One of the subunits, TnC is attached to the thin filaments of a muscle fiber, and regulates the binding of Ca^{2+} to the N-terminal domain of TnC[263]. Ca^{2+} binding triggers changes within the complex that allow myosin to latch onto the thin filaments and induce muscle contraction. TnC has been extensively studied due to its critical involvement with heart function and failure, and has been marked as a therapeutic target in heart disease and other disorders[251].

Our method is able to determine the k_{on} to a value that is within 11% of the experimentally measured k_{on}. This discrepancy falls within both the experimental

uncertainty as well as the uncertainty of the milestoning calculation. The FHPD of TnC at 10 Å (Figure 8.8) indicates that Ca^{2+} approaches directly from the solvent, probably due to the high desolvation penalty incurred when the highly charged Ca^{2+} is removed from its aqueous environment. The surface map seems to indicate two close but distinct minima on the FHPD, suggesting that Ca^{2+} may have two possible routes to binding. The k_{on} value for the TnC system is relatively unaffected by the choice of reaction criteria (Figure 8.6); remaining within ten percent of the estimated amount even when the bound milestone was chosen to be within the 2Å to 5Å range. This insensitivity to the reaction criteria offers some tolerance when choosing the reaction criteria for this system. Tolerance of the reaction criteria was less for the SOD system (Figure 8.5), and the calculated k_{on} was more sensitive to the choice of reaction criteria. This relative intolerance was likely related to the flatness of the free energy profile, where a diffusing ligand has low barriers when traversing between the bound and unbound states.

In total, the entire project, including all simulations of all systems analyzed in this study, cost approximately 65,000 hours of CPU usage. The vast majority of this computation was spread across hundreds or thousands of cores at any one time due to the highly parallel nature of milestoning. The total length of MD simulation for our systems required anywhere between 100 and 1600 ns of total MD time each with relatively low uncertainty due to the high rate of sampling along the milestones leading to binding. The cost is significantly less per target than brute force MD simulations run in past studies to observe kinetic events while yielding similar or superior resemblance to experiment[194,195], which were indicated to require between

600 and 15000 ns of MD simulation to achieve even just a single binding event, with some simulations never even yielding a binding event.

Our multiscale MD-BD-milestoning method offers many advantages; yielding predictive k_{on} estimates for biologically relevant molecular systems within experimental error at a cost much less than brute-force MD alone and at accuracy much greater than could be obtained using BD alone. This method also benefits from high parallelism due to the spread of MD computation across multiple states. Given a large number of cores and sufficient CPU hours, the MD portion of the calculation can be completed rapidly. Another advantage of this method is its flexibility, giving the user the ability to adjust the cost-to-accuracy balance by performing additional simulation and adding trajectory samples to increase result convergence.

The main disadvantage of this method lies in its complexity of concept and implementation. However, with sufficiently robust software-based automation, the burden of maintaining many parallel instances of simulation, as well as simulation preparation and analysis, can be greatly reduced. Another disadvantage of the milestoning framework is that the simulations are still relatively expensive at this time; requiring a supercomputer or cluster to obtain sufficient sampling within a reasonable time frame, although GPU-based MD could potentially alleviate this burden.

8.5.4 Conclusions

We present a new method to estimate kinetic rates. This method uses milestoning to leverage the strengths and minimize the weaknesses of MD and BD,

thereby offering an efficient, high-accuracy estimation of k_{on} rate constants. This multiscale method has been successfully used to estimate the k_{on} rate constant for both idealized and realistically sized, biologically relevant systems. Our work demonstrates that milestoning can be used to obtain kinetic quantities of interest with a high resemblance to experiment. We anticipate that this multiscale approach can be used to determine rate constants of interest as well as system-specific binding details that are applicable to drug discovery, biomolecular modeling, and protein-ligand interactions.

8.6 Supplementary Information

8.6.1 Hydrodynamic Radius

The hydrodynamic radius *a* of a molecule relates to its diffusion coefficient *D* according to Eq. 8.13 [264]

$$D = \frac{k_B T}{6\pi\eta a} \qquad \text{Eq. 8.13}$$

Where T is the temperature, k_B is Boltzmann's constant, and η is the viscosity of the solvent.

8.6.2 Calculation of error for Milestoning

The statistical error of all milestoning calculations can be estimated by generating a distribution of rate matrices according to Eq. 8.14 [2,210].

$$p\left(\mathbf{Q}\middle|\{N_{\alpha\gamma}, \langle t\rangle_\alpha\}\right) \propto \prod_\alpha \prod_{\gamma\neq\alpha} q_{\alpha\gamma}^{N_{\alpha\gamma}} e^{-q_{\alpha\gamma} N_\alpha \langle t\rangle_\alpha} P(\mathbf{Q}) \qquad \text{Eq. 8.14}$$

Where **N** is a count matrix whose element $N_{\alpha\gamma}$ is equal to the number of times in the milestoning simulation that the system started at milestone α and ended at milestone γ, $\langle \mathbf{t} \rangle$ is the incubation time vector whose element $\langle t \rangle_\alpha$ is the average amount of time a system started at milestone α spends before crossing another milestone. *P(***Q***)* is a prior probability distribution that, in this case and typically, we set to uniform density. **Q** is a rate matrix whose nondiagonal elements $q_{\alpha\beta}$ can be used to reconstruct the transition kernel **K** and the incubation time vector $\langle \mathbf{t} \rangle$ found in Eqs. 8.5-8.7 according to Eqs. 8.15a and 8.15b

$$K_{\alpha\beta} = \frac{q_{\alpha\beta}}{\sum_{\gamma \neq \alpha} q_{\alpha\gamma}} \qquad \text{Eq. 8.15a}$$

$$\langle t \rangle_\alpha = \frac{1}{\sum_{\gamma \neq \alpha} q_{\alpha\gamma}} \qquad \text{Eq. 8.15b}$$

The diagonal elements of **Q** are defined as: $q_{\alpha\alpha} = -\sum_{\alpha \neq \beta} q_{\alpha\beta}$. All non-diagonal elements $q_{\alpha\beta} > 0$ and all diagonal elements $q_{\alpha\alpha} < 0$.

By extracting a large number (hundreds or thousands) of matrices from this distribution, and performing the necessary milestoning calculations with each of them, a distribution of any of the results can be found, giving an estimate of the error for each result by finding a standard deviation of the distribution.

We used a nonreversible element shift Monte Carlo algorithm to sample the posterior probability in Eq. 8.14 Inspired by an algorithm used to compute the error of Markov state models (MSM)[233], our algorithm is defined below:

Algorithm for sampling rate matrices. To sample the distribution Eq. 8.14, a metropolis criterion is defined that evaluates whether to take a proposed step in **Q**

space. The first rate matrix **Q*** is the matrix that maximizes the likelihood of Eq. 8.14.

$$q^*_{\alpha\beta} = N_{\alpha\beta}/(N_\alpha t_\alpha) \qquad \text{Eq. 8.16}$$

Given a proposed matrix **Q***'* and a current matrix **Q**, the probability of accepting that member of the distribution is defined as:

$$p_{accept} = \frac{p(\mathbf{Q}'|\{\mathbf{N}, \langle\mathbf{t}\rangle\})}{p(\mathbf{Q}|\{\mathbf{N}, \langle\mathbf{t}\rangle\})} \qquad \text{Eq. 8.17}$$

A proposed change Δ relates the difference between an element of **Q** and **Q***'*.

$$q'_{\alpha\beta} = q_{\alpha\beta} + \Delta \qquad \text{Eq. 8.18a}$$

$$q'_{\alpha\alpha} = q_{\alpha\alpha} - \Delta \qquad \text{Eq. 8.18b}$$

The proposed change must ensure that all non-diagonal elements of **Q***'* remain positive, and that all diagonal elements remain negative. Thus, Δ is drawn from an exponential distribution on the range:

$$\Delta \in [-Q_{\alpha\beta}, \infty) \qquad \text{Eq. 8.19}$$

With a mean value at zero. Finally,

$$p_{accept} = \frac{p(\mathbf{Q}'|\{\mathbf{N}, \langle\mathbf{t}\rangle\})}{p(\mathbf{Q}|\{\mathbf{N}, \langle\mathbf{t}\rangle\})} = \left(\frac{q_{\alpha\gamma} + \Delta}{q_{\alpha\gamma}}\right)^{N_{\alpha\beta}} \frac{e^{-(q_{\alpha\gamma}+\Delta)N_\alpha\langle t\rangle_\alpha}}{e^{-q_{\alpha\gamma}N_\alpha\langle t\rangle_\alpha}} \qquad \text{Eq. 8.20}$$

Example 1. To support that the above workflow is correct, we have constructed a simple system to validate it. Figure 8.9 compares the distributions of the off-diagonal elements of a 2x2 rate matrix computed using the count matrix and incubation time vector below:

$$\mathbf{N} = \begin{pmatrix} 0 & 12 \\ 30 & 0 \end{pmatrix} \qquad \langle \mathbf{t} \rangle = \begin{pmatrix} 500 \\ 150 \end{pmatrix}$$

A set of 1×10^6 rate matrices were generated stochastically using the algorithm outlined above and the off-diagonal elements were binned to generate the 2-dimensional histogram in panel (a) of Fig. 8.9 For comparison, an analytic probability distribution was constructed in panel (b) of Fig. 8.9 by simply plotting the likelihood L for sampling that point in **Q**-space given the count matrix **N** and incubation vector $\langle \mathbf{t} \rangle$.

$$L(q_{01}, q_{10}) = q_{01}^{N_{01}} q_{10}^{N_{10}} e^{-q_{01}N_0\langle t\rangle_0 - q_{10}N_1\langle t\rangle_1}$$

The high degree of similarity between the two plots of Figure 8.9 supports the correctness of the algorithm. These computations were done using a custom script.

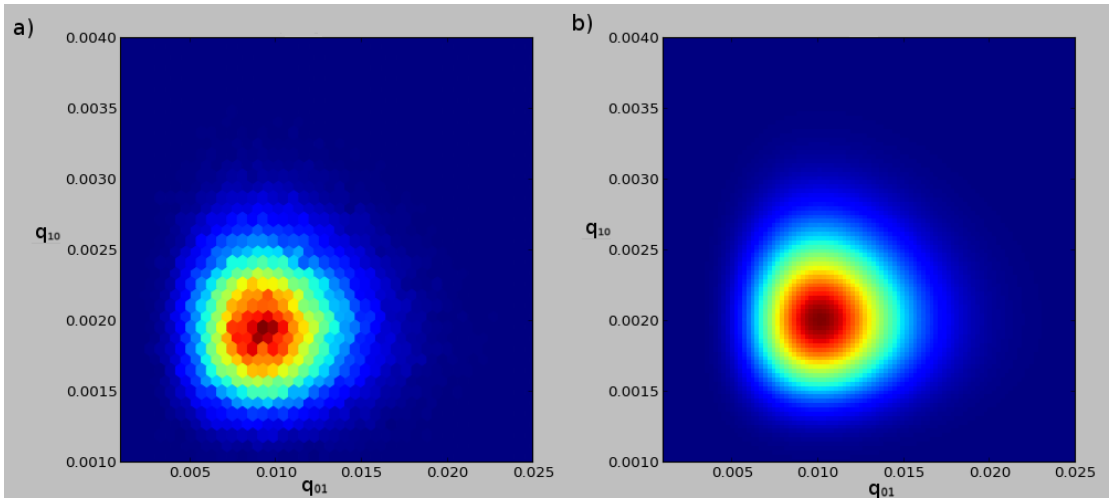

Figure 8.9: Plot illustrating the sampling of rate matrix

The rate matrix was constructed using count matrix $\mathbf{N} = \begin{pmatrix} 0 & 12 \\ 30 & 0 \end{pmatrix}$ and the incubation time vector $\langle \mathbf{t} \rangle = \begin{pmatrix} 500 \\ 150 \end{pmatrix}$. 1×10^6 **Q** matrices were sampled from Eq. 8.14 using these criteria and the off-diagonal elements were binned into a histogram to generate the plot in panel (a). The plot in panel (b) was generated analytically for comparison.

8.6.3 Error Estimate Convergence

For the spherical receptor systems, 1×10^7 matrices were generated from the distribution in Eq. 8.14 Of these, every 1000 were skipped, and the remaining 1000 were used to construct the error estimates for the spherical receptor systems. Figure 8.10 and Figure 8.11 were also constructed from those same samples to demonstrate convergence. For both SOD and TnC, only 100 matrices were skipped between samples, though 1000 total were sampled to construct Figure 8.12 and Figure 8.13

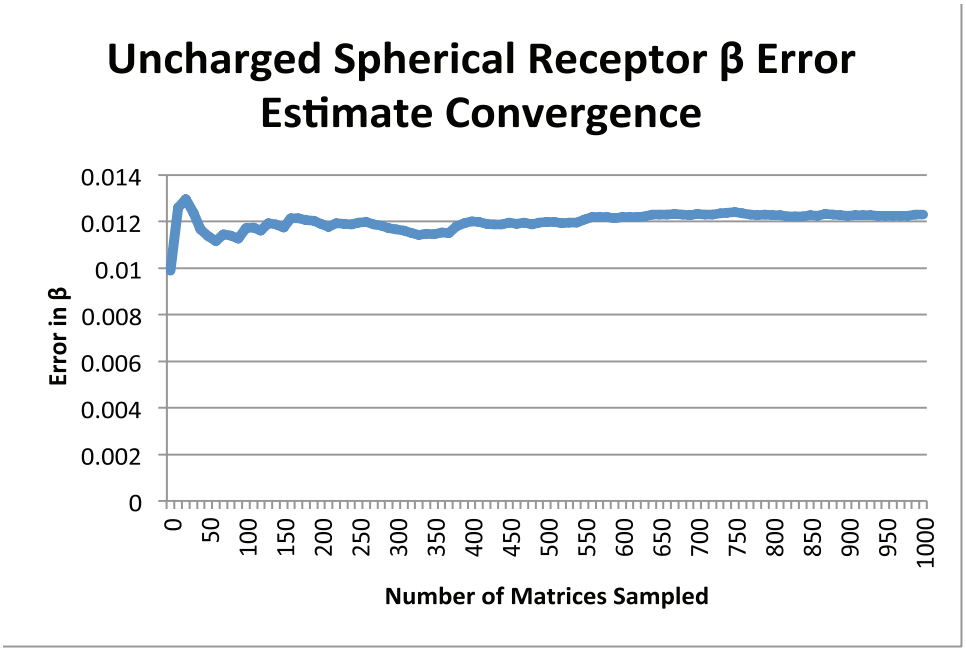

Figure 8.10 Convergence of error estimate for the β of the uncharged spherical receptor
The estimate is well converged before the full 1000 matrices have been sampled.

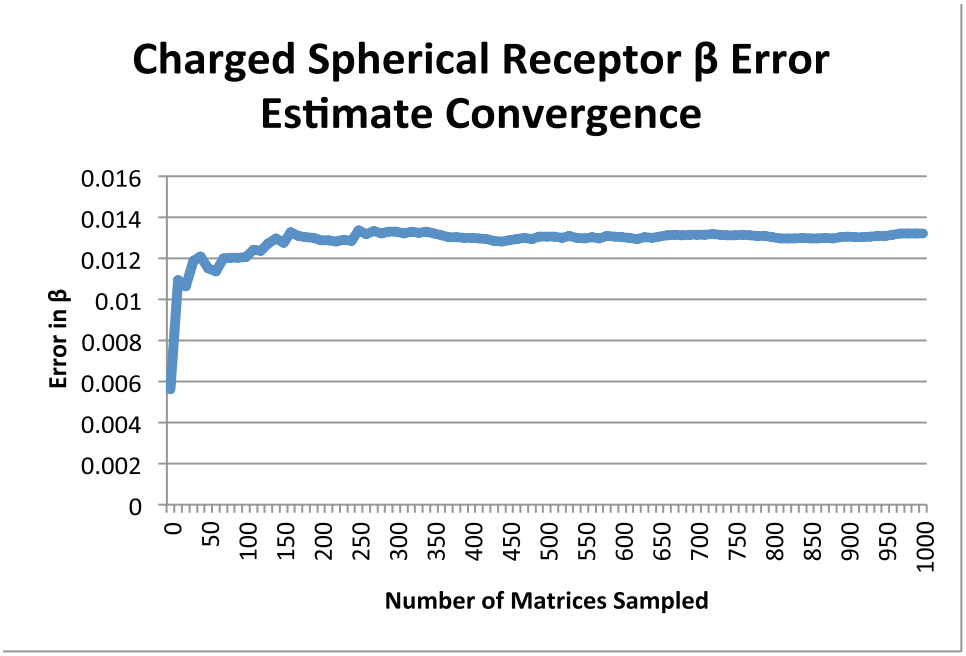

Figure 8.11 Convergence of error estimate for the β of the charged spherical receptor

The estimate is well converged before the full 1000 matrices have been sampled.

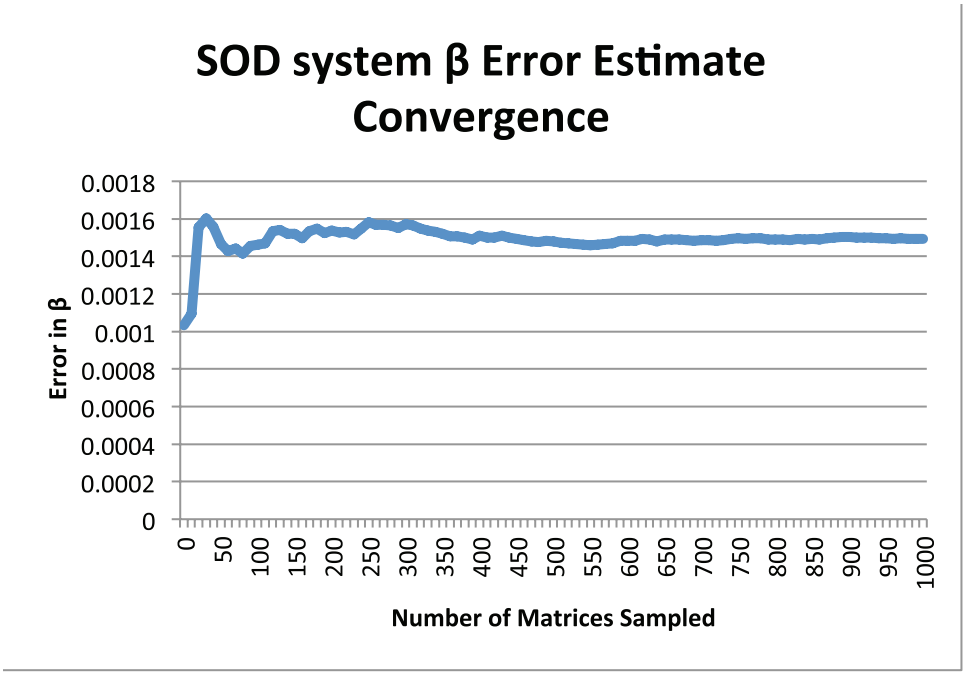

Figure 8.12 Convergence of error estimate for the β of SOD

The estimate is well converged before the full 1000 matrices have been sampled.

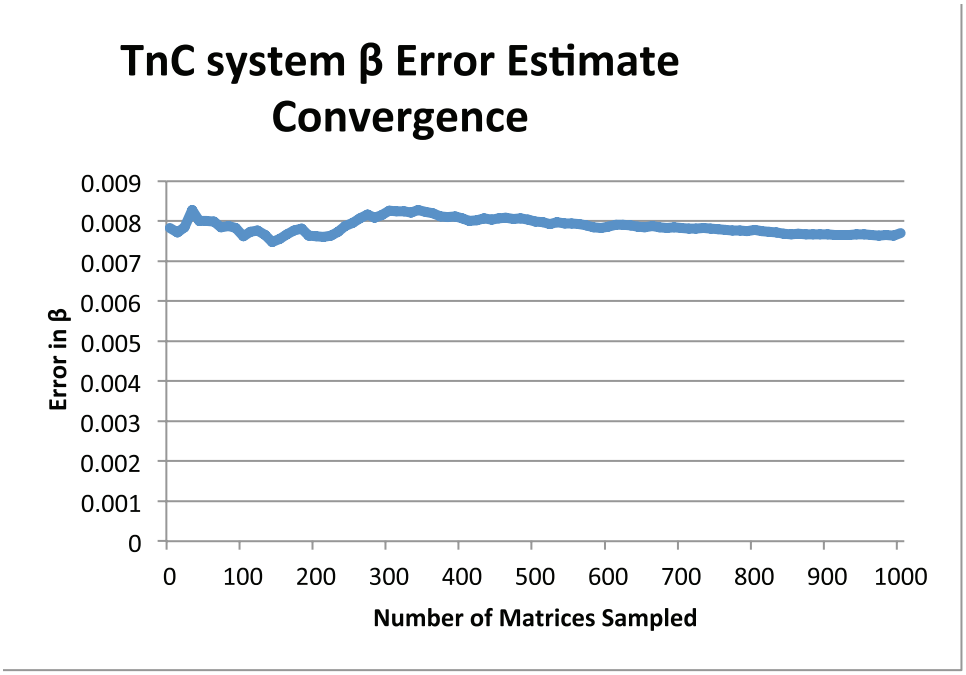

Figure 8.13 Convergence of error estimate for the β of TnC
The estimate is well converged before the full 1000 matrices have been sampled.

8.6.4 Results Convergence

The convergence of β, the mean first passage time (MFPT), and the k_{on} for each system was calculated by progressively increasing the number of MD trajectories from each milestone included in the milestoning computation (Figures 8.14 – 8.17).

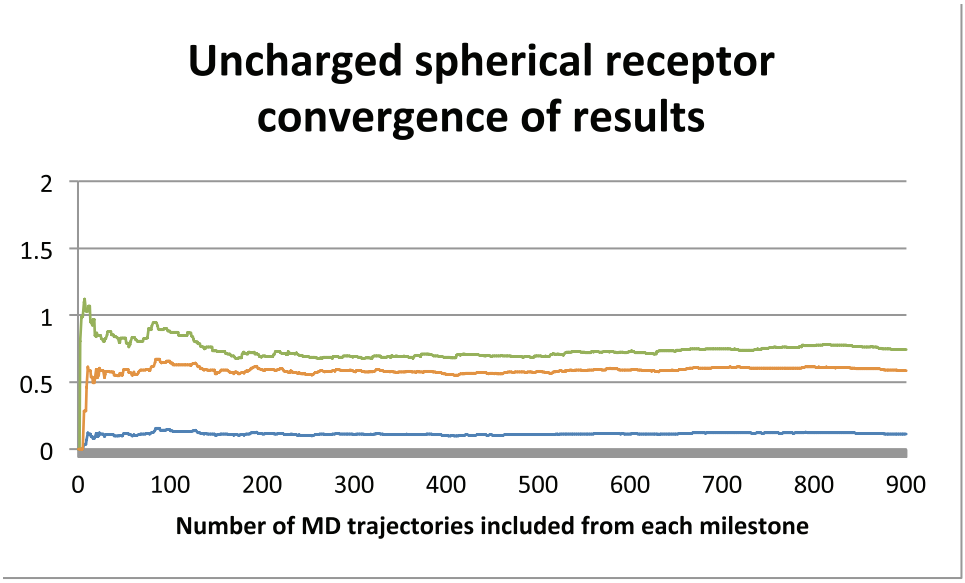

Figure 8.14 Convergence of the results of the uncharged spherical receptor system
The β convergence is displayed in blue, the MFPT ($\times 10^{11}$ s) in red, and the k_{on} ($\times 10^{-10}$ $M^{-1}s^{-1}$) in green.

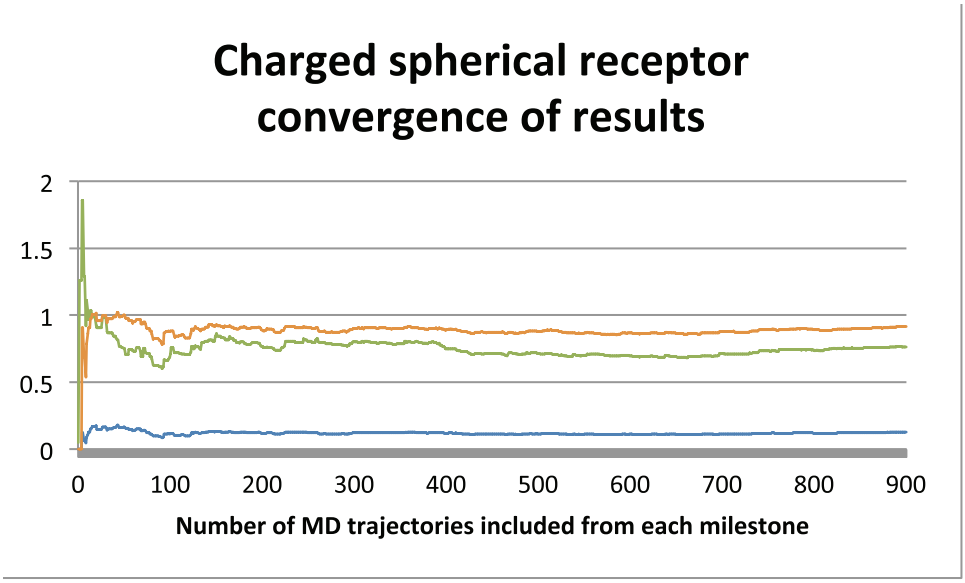

Figure 8.15 Convergence of the results of the charged spherical receptor system
The β convergence is displayed in blue, the MFPT ($\times 10^{11}$ s) in red, and the k_{on} ($\times 10^{-10}$ $M^{-1}s^{-1}$) in green.

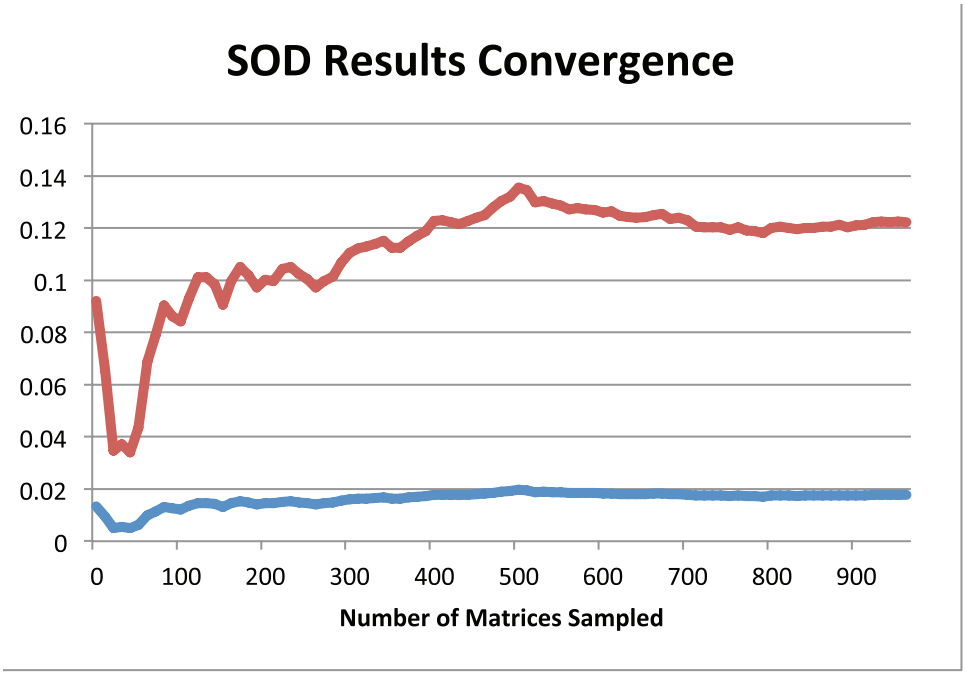

Figure 8.16 Convergence of the results of SOD system
The convergence of β is displayed in red, the k_{on} ($\times 10^{-10}$ $M^{-1}s^{-1}$) in blue.

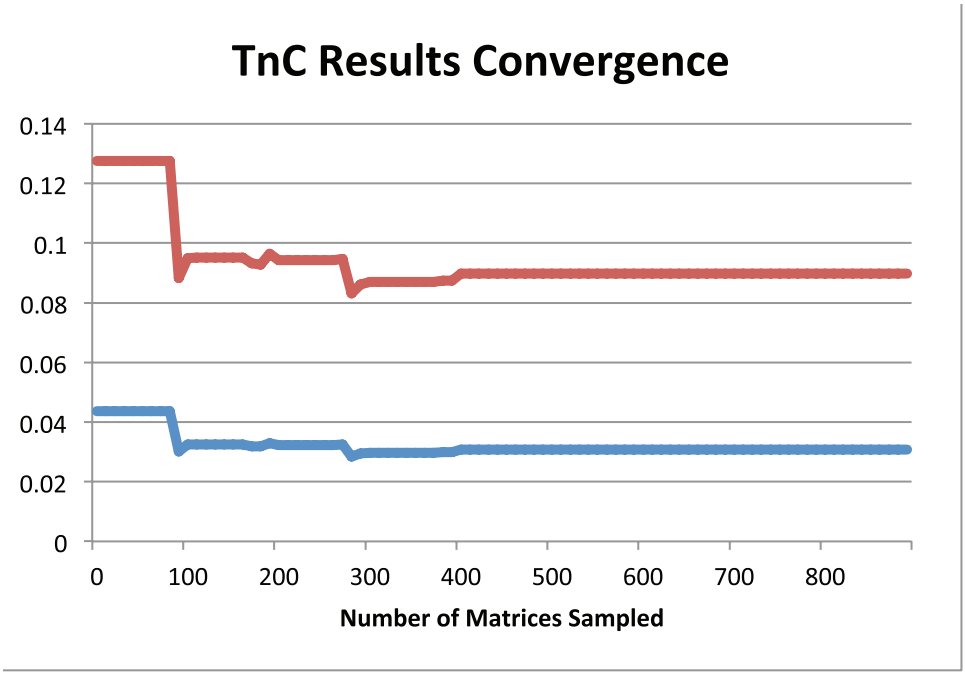

Figure 8.17 Convergence of the results of TnC system
The convergence of β is displayed in red, the k_{on} ($\times 10^{-10}$ $M^{-1}s^{-1}$) in blue.

8.6.5 Derivation of Eq. 8.9

Eq. 8.4 describes the solution to the diffusion-convection equation for a charged particle diffusing around an absorbing spherical surface surrounded by a centrosymmetric force. We assume *D(r)* is a constant *D* and *U(r)* is defined by Coulomb's law,

$$U(r) = \frac{Q_c Q_s}{4\pi\varepsilon_0\varepsilon_r r} \tag{8.21}$$

Where Q_s is the charge of the diffusing particle, Q_c is the charge in the center of the receptor sphere, ε_0 is the permittivity of a vacuum, ε_r is the dielectric constant of the solvent, and *r* is the radius from the sphere center. Note that

$$\int_b^\infty r^{-2} e^{Cr^{-1}} dr = -\frac{1}{C}\left[1 - e^{\frac{C}{b}}\right] \tag{8.22}$$

Where *C* is some constant. By assuming that

$$C = \frac{Q_c Q_s}{4\pi\varepsilon_0\varepsilon_r k_B T} \tag{8.23}$$

We obtain Eq. 8.9.

Chapter 8, in full, is a reprint of "Multiscale Estimation of Binding Kinetics Using Brownian Dynamics, Molecular Dynamics, and Milestoning", which was published in 2015 in PLOS Computational Biology, volume 11, issue 10, by Lane W. Votapka and Rommie E. Amaro. The dissertation author was the primary investigator and author of this paper.

Chapter 9: Bridging Scales Through Multiscale Modeling

The goal of multiscale modeling in biology is to use structurally based physico-chemical models to integrate across temporal and spatial scales of biology and thereby improve mechanistic understanding of, for example, how a single mutation can alter organism-scale phenotypes. This approach may also inform therapeutic strategies or identify candidate drug targets that might otherwise have been overlooked. However, in many cases, it remains unclear how best to synthesize information obtained from various scales and analysis approaches, such as atomistic molecular models, Markov state models (MSM), subcellular network models, and whole cell models. In this paper, we use protein kinase A (PKA) activation as a case study to explore how computational methods that model different physical scales can complement each other and integrate into an improved multiscale representation of the biological mechanisms. Using measured crystal structures, we show how molecular dynamics (MD) simulations coupled with atomic-scale MSMs can provide conformations for Brownian dynamics (BD) simulations to feed transitional states and kinetic parameters into protein-scale MSMs. We discuss how milestoning can give reaction probabilities and forward-rate constants of cAMP association events by seamlessly integrating MD and BD simulation scales. These rate constants coupled with MSMs provide a robust representation of the free energy landscape, enabling access to kinetic and thermodynamic parameters unavailable from current experimental data.

These approaches have helped to illuminate the cooperative nature of PKA activation in response to distinct cAMP binding events. Collectively, this approach exemplifies a general strategy for multiscale model development that is applicable to a wide range of biological problems.

9.1 Introduction

The goal of multiscale modeling is to understand how the hierarchy of biological structures integrates to produce biochemical, cellular and physiological functions. At the single cell scale, signaling networks are analyzed using system analysis methods to provide mechanistic insights into the functional interactions between proteins and second messengers. Network models of cell signaling have recently been developed for neurons [265], myocytes [266], and pancreatic beta cells [267], to name a few. These cell-scale network models are helpful to understanding normal cell physiology, pathobiology and therapeutic mechanisms. Interest in the phenomenological effects of protein mutations [268,269] are driving the development of new methods to incorporate atomic and molecular-scale models and data into whole cell simulations. To this end, advances in atomic-scale modeling, particularly molecular dynamics (MD) and Brownian dynamics (BD) simulations, have provided insights into the effects of mutations on protein folding and protein-protein interactions [270-273]. However, bridging these scales and disciplines to create models that can predict the effect of a point mutation or post-translational modification on cellular phenotypes remains a daunting task. Frequently, even nomenclature does not readily transcend disciplines, making interdisciplinary collaborations across scales

more difficult. Furthermore, understanding the limitations of models and methods at each scale to avoid error propagation is essential to obtaining physiologically meaningful solutions. In this article, we describe atomic and protein-scale Markov state modeling (MSM), as well as milestoning, which allow us to bridge atomic-scale molecular models to cell-scale signaling networks (Figure 9.1).

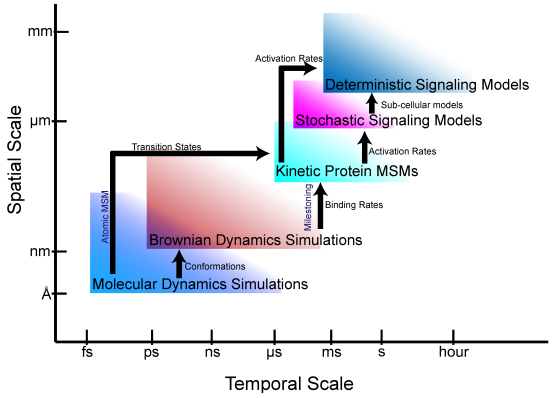

Figure 9.1: Bridging gaps through multiscale modeling
Simulation and modeling methods are limited in the spatial and temporal scales that can be represented. Arrows show the information that can be fed from one simulation regime to another.

Over the past decade, the availability of high-resolution protein structures and the capabilities of atomistic molecular modeling techniques has improved dramatically. MD and atomic-scale MSMs use atomic-resolution structural data to model the position of atoms in a protein and calculate the forces between them. This is helpful in predicting functional states and rates of conformational change. However, these methods cannot easily calculate the rates of interactions between molecules, which are needed for higher scale reaction network models.

Advances in BD simulations and milestoning have provided tools that are specialized in calculating diffusion-limited association rate constants. Previously, the data used for parameterization of the transitions in protein-scale MSM came almost exclusively from *in vitro* experiments where conditions are controlled to limit the number of potential states. These data included phosphorylation rates, k_{on}/k_{off} of binding events, and ion channel transitions [274-276]. However, many molecular events occur at time-scales that cannot be easily accessed by experiments[277]. Fortunately, computational simulations have provided alternative methods for determining parameters for whole-cell models. BD simulations rely on simplifying assumptions that allow simulations of microscopic events that span larger systems and timescales than more detailed methods, such as MD, allow. BD can be used to determine association rate constants (k_{on}) for diffusion-limited protein-protein and protein-small molecule interactions. It specifically examines how electrostatic and steric properties of molecules affect molecular encounter rates. Combining this information with *in vitro* experiments and MD-derived states will enable a new generation of protein-scale MSMs to be developed for incorporation into whole cell models.

As an example problem necessitating the integration of approaches across a broad range of spatial and temporal scales, we focus here on protein kinase A (PKA), which is activated by cAMP and is a key regulator of many cellular processes. In cardiac myocytes, for example, PKA is a critical regulator of intracellular calcium handling cycling, and its dysregulation is well known to be a contributing factor in heart failure[278]. The PKA holoenzyme consists of two regulatory (R) subunits and two catalytic (C) subunits. Each R subunit has two cAMP-binding domains (CBD), a

DD-docking domain, and a disordered linker region containing the inhibitory sequence that interacts with the C subunit. PKA is activated upon cAMP binding to the CBDs on the R subunit inducing release of the C subunit. Over the last 15 years, several whole-cell models of ventricular myocytes that incorporate calcium release and beta-adrenergic stimulation through a simplified PKA activation mechanism were developed [266,279]. More recently, a mechanistic protein-scale MSM of PKA holoenzyme activation was developed [274]. Still, incorporating an improved PKA MSM into existing whole cell models will provide a more physiological testing of PKA activation as well as the capability to predict the effects of PKA mutations on the whole cell scale.

In this review, we highlight some of the tools and techniques used to develop integrative models that span scales from the molecule to the cell, including: MD, atomic MSM, BD, milestoning models, protein MSM, and whole cell modeling. We provide the nomenclature necessary to bridge these scales and discuss the limitations of these approaches as well as ways to minimize error propagation. Finally, we show the role of MD and BD simulations have played in the development of a protein scale MSM of PKA RIα and discuss the role this new protein-scale MSM of PKA will play in existing whole cell models of cardiac function and disease states.

9.1.1 Nomenclature

This paper deals primarily with Markovian models, or models that are only dependent on the current state of the model and not the history of the states it has visited. Both MSM and milestoning models operate under a Markovian assumption.

Also, for this paper we use “atomistic” or “atomic-scale” to describe any model that treats atoms explicitly. This generally includes MD, MSM, BD, and milestoning. These models stand in contrast to “protein-scale” Markov models and cell-systems models which primarily focus on protein and cell function and general protein-protein and small molecule-protein binding events. Even though atomistic MSM and protein MSM are both Markovian models, they serve distinct purposes.

9.1.2 Accessing the Conformational Ensembles of Proteins

A protein’s function is governed by its conformational ensemble, which can be modulated though mutations and intermolecular interactions[280-285]. Therefore, to build multiscale models starting at the atomic scale, one needs to elucidate the key conformational states of a protein and the dynamics of those states from atomistic data associated with those states. This can be achieved through exploration and characterization of the protein’s conformational ensemble. In this section, we review computational methods for modeling the conformational ensembles of proteins important in cell signaling. We begin with an overview of molecular dynamics simulation methods and conclude with a discussion on the use of MSM to determine the conformational ensemble more efficiently.

9.1.3 Molecular Mechanics and Molecular Dynamics Simulations

Atomistic models of conformational ensembles can be computationally generated from molecular mechanics simulations. These simulations require two

components: a force field that describes how the atoms interact with each other and a method for exploring the conformational ensemble[286,287].

To simplify the complex quantum mechanical interactions between atoms, molecular mechanics simulations use empirical force fields to describe the interactions between atoms. These force fields are described in terms of classical mechanics[286,288,289]. For example, each atom of a system is described as a charged particle in space. Bonding interactions between atoms are described as springs using Hooke's law. Nonbinding interactions between atoms are described as Columbic and van der Waals interactions. Commonly used force fields include CHARMM[290], AMBER [291], OPLS[292] and GROMOS[293]. While a discussion on force field selection is beyond the scope of this review, it is important to understand the assumptions and performance bias of a force field used in any simulation[294,295].

The motion of the atoms resulting from the force field determines the conformational ensemble of the system. The motions of these particles are generally simulated either with Monte Carlo techniques that randomly sample conformational space, or through MD simulations, where Langevin's or Newton's laws of motion are solved over time [286,287]. While MD is more computationally expensive than MC, it retains the temporal relationship between conformations, which is advantageous when quantification of kinetic parameters is desired. Popular MD programs include AMBER[244], CHARMM[290], GROMOS[26] and NAMD[50].

Theoretically, MD simulations can sample the entire conformational ensemble of a system given infinite simulation time. While certain specialized supercomputers have been built to sample into the millisecond range[296], with current commodity-level

resources, MD simulations can only continuously sample a system for a few microseconds at most, which is insufficient to effectively sample most ensembles, including the CBD. However, with the increasing performance of supercomputers, GPU-accelerated MD simulations [297-299], and the use of highly distributed computing[300,301], multiple parallel MD simulations can achieve total non-continuous sampling time approaching the high-microsecond to low-millisecond range. MSMs can subsequently be used to stitch together the many short-timescale simulations into one cohesive framework that allows the extrapolation of longer-timescale data. This MSM framework was used for the CBD system discussed below.

9.1.4 Atomic-Scale Markov State Models of a Conformational Ensemble

An atomic-scale MSM describes the conformational ensemble of a protein as the probability of transitioning between discrete collections of conformational states at a fixed time[199,302]. This can be visualized as a bidirectional graph,(see Figure 9.2), where each node represents a cluster of similar conformations. The probability of transition between states is indicated by the thickness of the connecting lines in figure 9.2. If the conformational states and the transitions can be accurately determined, then the MSM describes the thermodynamics and the kinetics of the system's conformational ensemble. Thus one can derive the key parameters required for higher scale models with a MSM[303].

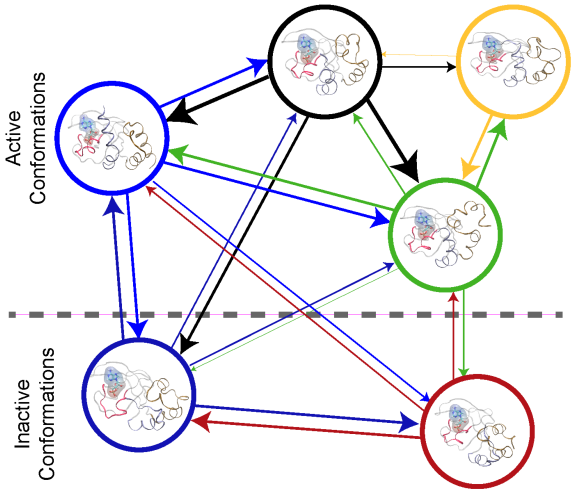

Figure 9.2: Protein Kinase A cyclic nucleotide binding domain Markov state model
This figure shows a graph repensantion the transtions between metastable states of the CBD with cAMP bound. Each node repesents the conformational state. The edges the transition between the node with their thickness being proportional to the probiblity of transtion.

Atomic-scale MSMs of the conformational ensemble of a protein are built from MD simulations. Each conformation sampled during the simulation is assigned to a discrete conformational state, usually by clustering. Then the transitions between the discrete states are determined from the MD trajectory by counting the transitions. The transition counts are then used to generate a transition probability matrix, the mathematical representation of the MSM[199,302-304]. The transition probability matrix can be analyzed to determine the equilibrium population of each confrontational state, to identify metastable conformational states, to understand the principal motions of the protein, and to study the mechanisms of conformational will change[199,302-304].

Because a MSM depends on the probabilities of transitions between discrete conformational states, the conformational ensemble of the protein can be sampled

more efficiently than with traditional MD. To effectively sample the conformational ensemble of a protein at equilibrium using traditional MD simulations requires running the simulation long enough to explore the conformational ensemble multiple times. However when building a MSM the MD simulations can be focused on the transitions between states avoiding spending time sampling stable conformations and improving the sampling of rare events. For example, a hypothetical transition between active and inactive states can be determined from multiple short simulations that explore the intervening conformations without requiring a single simulation to bridge the two states. Additionally, once a preliminary MSM is build poorly sampled transition can be additionally sampled to improve the quality of overall MSM.

Detailed methods for building MSMs for MD simulations have already been described[304,305]. Here we highlight key considerations for building a MSM that will be integrated with higher-scale models with examples from a recently developed MSM of the cyclic-nucleotide binding domain of the R subunit of PKA[306]. The overall process of building a MSM is as follows: (1) defining the conformational space; (2) initial molecular dynamics sampling; (3) iterative refinement of the MSM; and finally (4) selection of the final model for analysis.

The goal of our study was to determine the kinetics of the conformational ensemble of the CBD with and without cAMP bound. We defined the conformational space as the atomic coordinates of the alpha carbons in a protein, dividing the conformational states discrete into stats using RMSD-based clustering. We started sampling the CBD in either a crystallographic predetermined active or inactive state with and without cAMP bound. Building the final MSMs required over 70 μs of total

sampling time comprised of both long-timescale initial sampling iterative adaptive sampling to refine the models[306].

Throughout the sampling and refinement process, the quality of a MSM is judged using implied timescale plots[199,304]. Data points of the plot are constructed with eigenvalues of the transition probability matrix populated at different lag times, or times between events. The plots indicate at what lag times the models are Markovian and if the models are consistently capturing the principal conformational changes of the system. Additionally, a Chapman-Kolmogorov-test is used to validate the consistency of a MSM with molecular dynamics simulations[303]. Using these two metrics, a final model is selected, the statistics of which are sampled at a specific lag time, which represents the fastest transition within the conformational ensemble that is also Markovian. This final model can then be used to derive the parameters for the multiscale model.

As described before, a MSM consists of the equilibrium probabilities for each conformational state. These probabilities are used to derive thermodynamic properties. Spectroscopic analysis can be used to identify metastable states within the conformational ensemble that can be used to build coarse-grained models of the system[303,307]. Transition path theory[308] can be employed to approximate the kinetics of transitions between states. These rates become the parameters to feed into the multiscale model. For the CBD model we were able to obtain the rates of transitions between the active and inactive states and show how cAMP modulates the conformational ensemble, changing the function of the CBD. These rates have been

an important benchmark in understanding the dynamics of the CBD, and form the foundation for examining the total R subunit and its interactions with the C subunit.

While the use of MSMs provides to conformational ensembles, there are still several important considerations and limitations to this method that should be considered in context of integrating them into a multiscale model. First, because conformational space is discretized, all kinetic rates are artificially fast [303,304], and should be considered an upper bound, especially when applied to high scale models. Second, a recent study indicates that modern force fields used in MD simulations produce varying transition kinetics [294]. Therefore, the same force field should be used for all models of a system, and the limitations of the force fields should be understood. Thirdly, while the MSM is somewhat robust to errors in clustering, give a sufficiently fine division of conformational space (i.e. a lot of clusters) [303], the MSM is still dependent on the starting conformation used to initialize the simulations and the limitations of MD. Therefore, it is possible to not have included states important conformational states leading to an incomplete model of the conformational landscape and incorrect predictions. However, limitation can be overcome using enhanced sampling methods [309] and from understanding acquired in the large-scale models. Finally, the MSMs are computationally demanding. This cost limits their usefulness in multiscale models, as a significant amount of time can be required to describe only one state in a higher scale model. Other sampling methods may be sufficient to obtain parameters for larger models. For example, if the opening and closing of a flap on a protein is the only permutation of interest, elastic network models are more computationally efficient in estimating those rates than MSM.

9.2. Investigating Intermolecular Interactions

As we extend into larger spatial scales of modeling, the focus of our discussion shifts from intramolecular investigations with MD to the study of intermolecular encounters using BD. BD simulations are used to estimate the rate constants of second-order association events between two molecules. The output of these simulations provides kinetic on-rates used directly in higher levels of modeling. The application of BD simulations has extended beyond bi-molecular encounters in simulations of molecular crowding[310] in cellular environments. In this section, we discuss the methodology and limitations of BD simulations, what can be gained from their use, and a brief overview of their application to multiscale modeling.

9.2.1 Brownian Dynamics Simulations

In BD, molecular diffusion is modeled using the theory of Brownian motion; where internal dynamics of each molecule are frozen, constraining the molecules into rigid bodies that are free to diffuse and tumble in solution, but may not change shape. Popular programs used to carry out BD simulations include BrownDye[215], SDA[214,224], ReaDDy[311,312], Brownmove[312], and BD_BOX[313]. Similar to MD, one must choose a force field for BD simulations of the molecular system: AMBER[314], CHARMM[315], GROMOS[293], etc. However, the only force field quantities utilized in BD simulations are the partial charges and Van der Waals radii of each of the atoms of the biomolecule. In conjunction, these properties can be used to obtain the electrostatic potential from software that can solve the Poisson-Boltzmann (PB) equation. The

electrostatic potentials of the biomolecules determine the long-range forces that the molecules impose on each other. Thus, electrostatics function as one of the most important determinants of the outcomes of BD simulations. Popular software packages that solve the PB equation include APBS[242,316] and DelPhi[36,41]. Rigorous derivations and discussions of the form and proper usage of the PB equation can be found in the literature[39,317,318].

In BD simulations, the solvent is modeled as a continuum; that is, there are no water molecules or dissolved ions modeled in atomic form in the simulation. Instead, the solvent is modeled as a field that surrounds the biomolecules and can have varying degrees of physical realism. This significantly reduces the computational power necessary for BD simulations in comparison to explicit solvent MD. The user typically specifies parameters that control solvent dielectric, hydrodynamics, desolvation, and ion screening, all which affect the realism of the solvent model and the computational cost of the simulation.

In addition to the long-range forces imposed by inter-molecular electrostatics, a stochastically determined force is also imposed on the molecules in a BD simulation. This stochastic force is directed randomly with a magnitude sampled from a Gaussian distribution centered at zero whose variance depends on the simulation time-step and the molecule's diffusivity properties. The stochastic force is intended to approximate the random "kicks" that would be caused by the solvent, but are otherwise absent in the continuum model.

Finally, the simulation must ensure that the Van der Waals radii of the atoms of different molecules do not overlap; a phenomenon known as a steric clash. Often,

simulation steps that result in a steric clash are discarded and recomputed. Alternatively, many BD programs can impose a Lennard-Jones force at close molecular proximity to prevent a steric clash[215,226]. BD simulation and the theory behind it compose a rich and expansive field, and many sources exist to allow the interested reader to improve his or her knowledge and technique[212,226,227,319-322].

9.2.2 Considerations for Brownian Dynamics Simulations

A key starting point for BD simulations is the selection of the encounter complex, which describes the atomic interactions that define a reaction between molecules. Ideally, crystal structures will inform this step. If crystal structures of the encounter complex do not exist, molecular docking programs can serve as a substitute. In the case of PKA, two crystallized structures of the regulatory subunit of protein kinase A RIα show very different conformations when bound to either cAMP [323] or the catalytic subunit [324] To test the effects of structure on cAMP association with BD methods, one can use the crystal structure conformations of the regulatory subunit in separate BD simulations. Alternatively, the two different conformations can be used as starting points of separate MD simulations. A number of structures in the conformational ensemble will be generated and can serve as structures for separate BD simulations.

At the start of a simulation, the ligand is placed at a distance *b* from the receptor, at a location known as the *b surface*, which is defined as the distance where forces between the two molecules are centrosymmetric. Simulations terminate either upon the molecules reaching the predefined bimolecular encounter complex (a

binding event), or when the molecules separate beyond a greater intermolecular distance *q*. The distance *q,* the radius of the *q surface,* is typically 10 to 50nm larger than the distance *b*[224]. The probability of association versus escape is then used to calculate the association rate constant (k_{on}). This schematic, including the surfaces at the *b* and *q* distances, are depicted using PKA as the receptor and cAMP as the ligand (Figure 9.3). BD can be used to model the association of cAMP with PKA, and predict features of the binding event, including the route of approach, the encounter complex, and the rate constant of association.

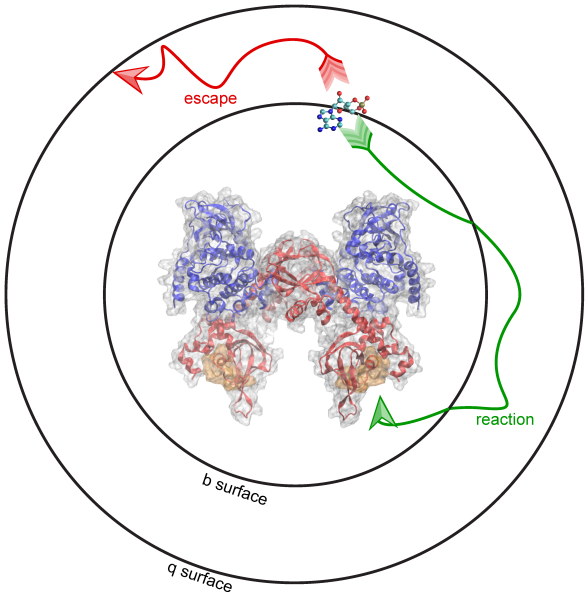

Figure 9.3: Brownian dynamics simulation method
BD simulations begin by placing molecules at a distance b from one another, shown here as a b-surface around PKA. When molecules diffuse toward the encounter complex (gold) a "reaction" (green arrow) occurs. Alternatively, molecules can "escape" (red arrow) by diffusing past a distance equal to q, shown here as the q-surface.

A second important factor in BD simulations is the structure of the molecules used in the simulations. Recall that BD simulations use a rigid-body approximation of

molecules, meaning that the conformation of the molecule will not change throughout the simulation as it does in MD. Typically, crystal structure conformations are used. Another attractive possibility is the use of conformations generated by MD as starting points for BD simulations. Using this method, the user can select meta-stable or even rare conformations of a protein generated in MD simulations and compare the association rates and probabilities with respect to structural changes in the protein. MD trajectories can also be used to generate ensemble-averaged electrostatics[245] where the simulated molecular motions are combined to form an electrostatic potential that includes the dynamic properties of the molecule. This effectively leads to a more holistic, dynamic representation of the electrostatic potential, effectively mediating some of the limitations of the rigid-body approximation of the simulations.

Solutions to the Poisson-Boltzmann equation include variations in the dipole moment and especially the charge density, with respect to how the solute affects the solvent, but also how the solvent affects *itself*. So while common implementations of the Poisson-Boltzmann equation solvers do not include many features of true aqueous solvents, it at least does assume that certain aspects of the solvent are heterogeneous. In addition, BD simulations themselves often model such things as hydrodynamics and desolvation forces, which are intended to approximate additional solvent features such as inertia and entropy at a surface, respectively.

Despite their ability to calculate association rate constants with respect to steric and electrostatic properties of molecules, BD simulations have limitations that users should know and recognize. First, the results of BD simulations depend on the encounter complex criteria. Such criteria must usually be tested and optimized in

order to reproduce a reasonable association rate constant. Incorrectly chosen encounter criteria can significantly limit the accuracy of the simulation outcomes. Second, the rigid-body approximation of molecules in BD can only represent one part of the binding process: the diffusion—meaning that the rate constant calculated by simulations is that of association and not actual binding. Nevertheless, alternative methods that combine intermolecular investigations of BD with intramolecular dynamics of MD are being developed that promise kinetic rate estimations through simulation [325]. These developments represent an approach toward spanning the MD and BD simulation regimes into a unified multiscale framework.

To our knowledge, no systematic method yet exists for estimating the true amount of error propagated by the assumptions inherent in BD. However, general consensus agrees that BD performs relatively well if the rate constants of an event it is estimating can be classified as a diffusion-limited process; that is, a process whose time to completion is primarily limited by particle diffusion. In the case of binding, the range of diffusion-limited rate constants is considered to include values of approximately 10^8-10^9 $M^{-1}s^{-1}$[1191].

Schemes do exist to approximate the precision of a k_{on} based on the statistical sampling of a binding probability versus escape. Specifically, the uncertainty of the rate constant of binding is proportional to the inverse square root of the number of trajectories in BD simulations that have completed in a binding event. Since millions or even billions of BD trajectories can usually be completed at relatively little computational cost, it is typically not difficult to obtain relatively high precision of a rate constant using BD. However, while the estimated rate constant may be precise, it

still may be inaccurate if the rigid molecules, implicit continuum solvent, or some other approximation assumed by BD do not adequately model the system. Comparison to experimental rate constants of the simulated ligand-receptor system, or perhaps of similar systems, can give an indication of the discrepancy between the "true" value, and the value obtained using BD.

9.2.3 Unifying MD and BD simulations through Milestoning

The possibility of combining the speed of rigid-body BD simulations with the flexibility of all-atom MD simulations to predict kinetic and thermodynamic quantities of interest is an attractive option. Ensembles of conformations or trajectories can be sampled from each simulation method, and statistics involving the probability and timescales of transitions between predefined states from the simulations can be combined using MSMs or the theory of milestoning[1] to model the details of intermolecular interactions.

Milestoning is a technique that is similar to the theory used in MSMs and can serve as an alternative approach to investigating biomolecular events, such as conformational sampling[209,326], diffusion[326], and membrane permeation[211], among others. Milestoning retrieves the kinetics as well as the thermodynamics of chemical processes [2,208,210], and can make use of extensive parallelization. Although similar to MSMs, milestoning models have a number of key differences, and may or may not be well suited to address a particular biophysical question. Unlike MSM states that are volumes in phase space where the system exists until it crosses into another,

milestones are surfaces in phase space that the system traverses, and where the system's current "state" is the surface that the system has most recently crossed.

To give an example, we examine the hypothetical case where the k_{on} of binding between PKA and cAMP can be predicted. In this milestoning model, we define a set of concentric spheres of different radii, all centered on the binding site of PKA (Figure 9.4). These concentric spheres define the milestones. MD simulations are started from conformations where cAMP is located on each spherical milestone, and each simulation is similarly terminated once cAMP diffuses to another surface. Thus, to the milestoning model, whichever simulation method is used to populate the transition kernels and incubation time vectors with statistics is of no consequence. The most appropriate simulation method can be chosen when cAMP is started on a particular surface.

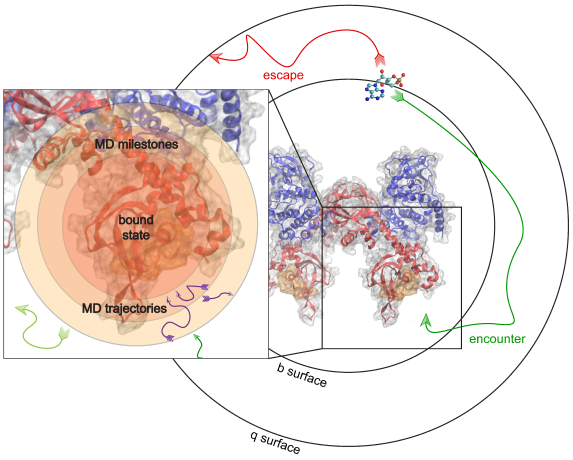

Figure 9.4: Milestoning applied to unite MD and BD
MD and BD Simulations are run to populate transition times and probilities in a milestoning model of cAMP binding to PKA. BD simulations are used to model an encounter event, and subsequent MD simulations model the details of the actual binding or reaction event.

Unlike MSMs, milestoning transitions may only occur to states that are adjacent in the positional or conformational space and lag times (or incubation times) can vary between inter-state transitions. Therefore, milestoning may be a desirable technique in situations where the system crossing surfaces would more appropriately represent transitions than the system traversing regions of space. For instance, because current implementations of BD simulations make extensive use of surfaces, such as the surfaces at the b and q distances and the encounter surfaces, milestoning is a natural choice to utilize transition statistics obtained in BD simulations. MD simulations modeling a binding event can make use of either milestoning models or MSMs, but when a combination with BD is desired, milestoning offers a promising framework to combine statistics from the two simulation methods.

Milestoning theory can be used to investigate a wide diversity of biophysical scenarios, and has been applied in a variety of contexts[1,209,211,220,232]. In some physical situations, implementations of milestoning outperformed MSMs in resemblance to experimental results[2]. The application of milestoning to intermolecular interactions is still a recent development, and many possible improvements may enhance the efficiency and accuracy of the estimation of binding rate constants. Examples of these include further discretizing the system into grid-like milestones, rotational milestones, or milestones that can represent internal degrees of freedom. Extensive derivations and discussions of milestoning theory are discussed elsewhere[1,2,208,210].

9.3 From Atomistic to Protein-Scale Models

Bridging the gap from atomic simulations to whole cell models is challenging. Protein-scale MSMs connect the atomistic scale to cell or tissue phenomena by reducing the complexity of molecular models. This enables simulations on larger time and spatial scales, while maintaining structural details required for protein function. These models simulate biological phenomena relevant to a whole-cell model, including ionic currents, fraction phosphorylated, or percent activation, and the output can be compared to *in vivo* experiments.

Protein-scale MSMs have been used to represent protein interactions since the early-1990's[327]. Several papers have been written on the development of protein-scale MSM, particularly of ion channels[328,329] [327,330]. Ion channel MSMs have been made possible by the detailed statistical data that comes from single channel patch clamp recordings[331]. These models have started to replace traditional phenomenological Hodgkin-Huxley style models of ion channel kinetics in whole cell action potential models [332]. They have been most useful when there is a need to model the effects of specific channel modifications, such as drug binding [333], gene mutation [332] or post-translational modifications [334]. But the use of protein-scale MSMs is not limited to those systems where dynamic biophysical recordings are available; instead, these models can be built from BD and MD simulations.

9.3.1 Protein-Scale MSM

The first step in model development is to determine the overall structure of the model. Unlike atomic-scale MSMs, protein-scale MSMs do not represent every

conformation of atoms as a state; instead, each state represents an ensemble of related atomic conformations that comprise a functional structure. This significantly limits the number of degrees of freedom and decreases the computational power needed, which enables multiple protein-scale MSMs to be combined into system-scale models. However, because these models are a simplification of the total potential states, the choice of which states are relevant becomes essential to making a useful model of a protein.

9.3.1.1 Functional State Discovery through MD Simulations

Frequently, several different states are captured by molecular-scale experiments, including X-ray crystallography and mass spectrometry. These experimental approaches can provide data on particular stable conformations (e.g., active or inactive states); however, due to the static nature of these tools, significantly less information is known about the transitions between states. For example, there are published structures of the R subunit of PKA bound either to cAMP or to the C subunit, but little is known about the transition between these end states. MD simulations can suggest intermediate states for incorporation into protein-scale MSMs. Similarly, atomic-scale MSMs provide insights into which states are populated and the rates of transitions between conformations.

9.3.1.2 Using BD Simulation to Inform Kinetics

For small molecule-protein interactions and protein-protein interactions, BD and experimental data can serve complementary roles in determining kinetics.

Dissociation-rates are typically slower than association rates and are therefore easier to measure experimentally using techniques such as surface plasmon resonance [335]. Additionally, most dissociation events are limited by conformational changes and not by diffusion—the latter of which BD is designed to model. For these same problems, MD simulations would be required to run for inaccessibly long periods of time (msec to sec) to register release events. Association-rates, on the other hand, tend to be orders of magnitude faster and therefore are harder to measure experimentally. BD simulations are ideal for measuring fast interaction rates on the ns to μs time scales, many of which are limited by diffusion. In combination with equilibrium data, these techniques can be used synergistically to determine rates for small molecule-protein and protein-protein interactions. By basing the ensemble of states of the model on MD simulations, and the kinetic interactions on BD simulations, it is possible to predict the effect of a mutation on protein function and, by extension, on the whole cell.

9.3.1.3 Testing with Empirical Data

Data from experiments, MD simulations, and BD simulations can be integrated into a simplified protein-scale MSM with states and interactions relevant to protein function. Frequently, these combined methods will suggest several possible functional state ensembles. Competing models are generated, with different states or different relationships between the states. Subsequently, the resulting models are tested to determine their ability to fit relevant experimental data. (Boras 2014) For protein-scale MSMs, the data used for fitting most often comes from *in vitro*

experiments. Ideally, the data used to differentiate between competing models is collected under conditions that are most relevant to a whole cell. For example, in the PKA-RIα model developed by Boras, et. al.,[274] all of the data used for fitting was collected in the presence of excess Mg^{2+} and ATP, both of which have been shown to affect PKA activation [336]. These conditions are similar to what is found in a cell; however, recently published data has also highlighted the role of ADP in PKA activation[337], which could affect the role of PKA in metabolism but is absent in the current MSM.

The accuracy of each theoretical model is determined using an error function based on the weighted sum of squares difference between the model's predictions and the available experimental data. Minimizing this error function optimizes unknown parameters. If the MSM are nested (all possible states in a model with fewer degrees of freedom can be represented in the model with more degrees of freedom) then a statistical F-test can be performed to determine if the added degrees of freedom significantly improve the fit[338]. This ensures that MSMs do not become needlessly complex without an improvement in the accuracy of the model's predictions.

Frequently, data acquired with mutant proteins that cannot reach specific states is used to differentiate competing models. The MSMs are altered slightly by removing those states, without refitting any parameters, and the output is compared against the experimental results [274]. For example Clancy et. al., used MSM of a cardiac sodium channel to show that a mutation in its C terminus can lead to long-QT syndrome, which causes life-threatening arrhythmias (Clancy 2002). This highlights

how protein-scale MSM based on atomistic data can predict the effect a mutation will have on the whole cell and eventually on the organ scale as well.

To mitigate error propagation when the protein-scale MSM is added to whole cell models, a sensitivity analysis can be performed to test the robustness of the solution[275]. In this process each rate is perturbed to determine its effect on the desired output of the model. States can also be removed to see how essential they are to the final result. The objective is to quantify how much the final result relies on any individual rate or state and compare that to the uncertainty in the experimental measurements. This technique can also highlight which states predicted from atomic scale modeling would have the greatest effect if mutated or pharmaceutically targeted. This is especially useful in quantifying the potential effect of rare conformations. Due to sampling bias they may not be captured in MD simulations but by adding them to the model their potential effect can be determined even if precise kinetic parameters are not known.

9.3.2 Applying to MD and BD modeling to Protein Scale PKA-RI MSM

These techniques have been applied to the development of a novel PKA protein scale MSM [274](Figure 9.5). First, the effects of cAMP binding on CBD-A of the regulatory subunit of PKA were examined. (Malmstrom 2015) Using extensive all atom molecular dynamics simulations integrated with atomic-MSM, the conformation of the CBD with and without cAMP bound was determined. Conformational selection was identified as the general mechanism of allostery within a single CBD, which transitions between an active and an inactive conformation whether or not cAMP is

bound. cAMP was found to regulate the function of the CBD by deepening the free energy landscape and selecting conformational states that favor the active conformation. Interestingly, cAMP modifies the transition rate between the active and inactive conformation and not the transition between the inactive and active conformations. Additionally, the roles of each of the signaling motifs in the CBD were elucidated.

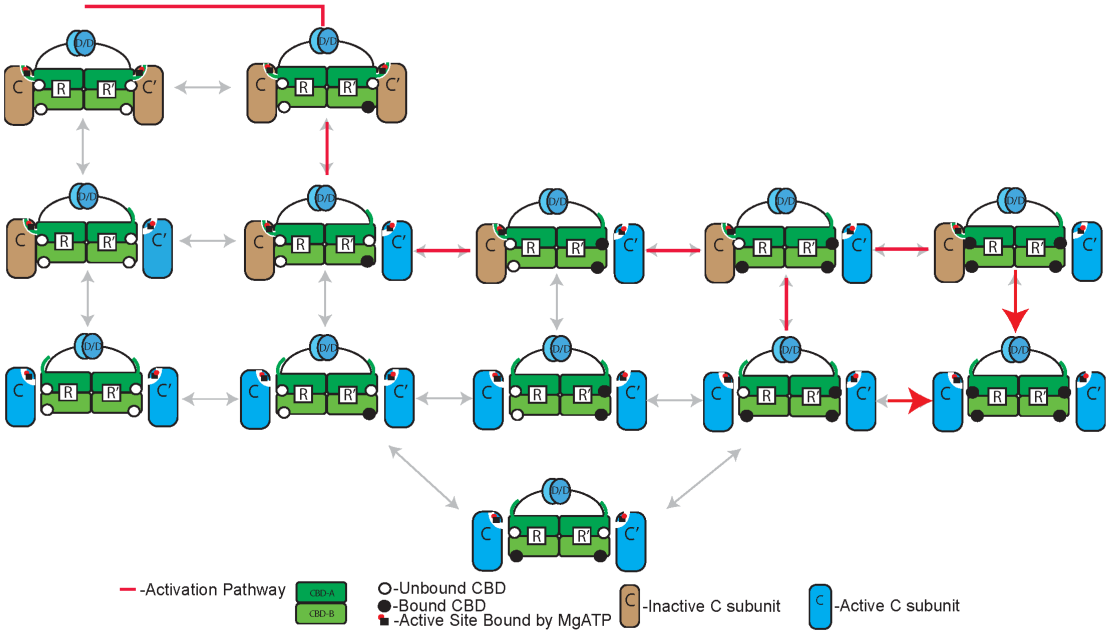

Figure 9.5: The Markov State Model of PKA-RIα R_2C_2 holoenzyme
A representation of MSM states for the activation of PKA-RIα R_2C_2 holoenzyme by cAMP first published in JBC [274]. The red arrows represent the dominant pathway during activation. The two R- and C-subunits are identical but for simplicity of naming the first R-subunit to bind C-subunit is named R, while the first R-subunit to bind cAMP is R'.

Based on these findings and crystal structure data, five nested protein scale MSM were considered. Each model was structured to test competing theories of PKA R_2C_2 activation based on MD simulations. The crystal structures suggested that a model that treated each R-C heterodimer as independent would be insufficient to fit

the data, due to the compactness of the R_2C_2 holoenzyme. Atomistic MSM predicted that a conformational selection mechanism would most accurately fit the data for isolated CBDs. The models were developed in the Virtual Cell computational environment[339] before being translated into MATLAB to take advantage of optimization programs[340].The models were fitted to kinase activity and cAMP binding data under physiological conditions[341]. Additionally, the various models were then compared to mutant PKA experimental data with either an inhibited CBD-A or CBD-B binding site. One model was shown to fit the wild type and predict the experimental results better than any other. This model validated the atomistic MSM by showing that CBD-B binding leads to release of the C-subunit prior to CBD-A binding similar to a conformational selection mechanism and created a thermodynamic protein-scale MSM of PKA activation.

However, since the fitting data, as well as the mutant data, were collected at long enough intervals that an assumption of thermodynamic equilibrium was valid, the resulting MSM could only reproduce equilibrium behavior. From a cellular perspective, PKA's response to a stimulus over time is essential to understanding PKA function. The single turnover rate in response to a stimulus has been implicated in activation due to A kinase anchoring proteins that bind PKA near one target [342]. Therefore, the addition of kinetic rates would significantly increase the utility of the MSM a whole cell model.

Owing to the fast rate of activation of PKA in the presence of cAMP, the amount of experimental kinetic data on PKA activation, particularly on-rates, is limited. This is a problem ideally suited to solving using atomic simulations. The

atomic scale MSM suggested that cAMP binding only affects the rate of transition from active to inactive states but not the reverse. BD simulations can be combined with experimental data to suggest binding and release rates for R-C and R-cAMP interactions. In conjunction with *in-vitro* experimental data this type of data will allow the thermodynamic MSM to become a kinetic MSM better suited to whole cell scale analysis of signaling network properties

9.4.0 Integrating Protein scale MSM into Whole Cell Models

The potential of molecular and protein-scale models culminate in whole-cell and tissue-scale models that can predict phenotypes and mechanistically explain disease states. These models combine several MSMs to predict cellular responses to either internal or external stimuli by tracing behavior down to molecular interactions. When developing these protein-scale MSMs, it is best to keep in mind what broader biological function will be modeled at a larger scale since this will determine not only what states are relevant but also what type of model is best for a given phenomena.

9.4.1 Stochastic and Deterministic MSM in the Whole Cell

Some cell functions are best simulated using a continuum of species concentrations and smooth probability distributions, while statistically rare events are better modeled when individual molecules are tracked and the stochasticity of interactions is accounted for. Correspondingly, protein-scale MSM can be either stochastic or deterministic in nature. The stochastic models, like the atomic-scale MSM described earlier, are based on Monte Carlo simulations where the probability

of transitioning between states is dependent on the kinetics of the binding and/or the conformational shift that each transition represents. This is the most accurate representation, where each event is dependent on the chance that two molecules will interact or that a conformation will be sampled based on random motion.

Many biological processes, such as calcium sparks in cardiac myocytes, can be explained with stochastic simulations. Calcium sparks occur when calcium is released from the sarcoplasmic reticulum via an isolated cluster of ryanodine receptor calcium release channels in the absence of a depolarizing event. In other words, a single cleft or a cluster of clefts acts differently than the rest of the cell. Whole-cell deterministic models require that every channel of a given type are identical and therefore every channel could be fractionally open but no one channel could be fully open while the others were fully closed without changing the conditions. Therefore, to model phenomena like this, a stochastic model is necessary. When translating these models up to the whole cell, the stochastic models are ideal for agent based spatial modeling tools, such as MCell[343], where each molecule is tracked and diffusion is represented by a random walk; although, it is worth noting that a whole-cell model can consist of a continuum diffusion approximation but still contain stochastic protein-scale MSMs. Agent-based models are ideal for small numbers of molecules or short time and spatial scales, where tracking each molecule is computationally reasonable or average approximations may be invalid.

Over a long enough time-scale or a large enough population of molecules, the Monte Carlo simulation will approach the deterministic solution. The deterministic solution is represented by a system of ordinary differential equations, instead of being

represented by a transition matrix of probabilities. In these models, the states of the MSM are frequently populated by concentrations instead of a specific number of molecules.

Many biological processes can be represented deterministically, most often when the system has a large number of molecules, or covers a long time and spatial scales. Models of the calcium concentration in a cell, for example, would require a deterministic model because computationally there are too many molecules to follow and the simulation becomes intractable. However, even on a small scale a deterministic approximation can be valid. For example, Hake *et al.* [344] showed that for a single dyadic cleft in a cardiac myocyte, the random walk and the deterministic continuum approximation gives the same result for a calcium induced calcium release event, even though a continuum approximation of the calcium in the cleft is unrealistic due to the scarcity of calcium ions. By treating the continuum as deterministic but the protein-scale MSMs as stochastic we can reproduce the stochastic sparks while limiting the required computational power.

9.4.2 Advantages in Whole Cell Modeling

The potential of molecular and protein-scale models culminate in whole-cell and tissue-scale models that can predict phenotypes and mechanistically explain disease states. These models combine several MSMs to predict cellular responses to either internal or external stimuli by tracing behavior down to molecular interactions. The power of building atomic-scale and protein-scale MSMs for wild type and disease mutants comes from their integration into whole cell models. At the whole

cell scale, differences in sub-cellular dynamics of protein mutations can be studied comparatively with their wild-type counterparts. Several disease states come from known protein mutations. For example, in the case of PKA-RIα, 117 polymorphisms and mutations have already been discovered[345]. Owing to the complexity of signaling pathways, how these mutations affect cell function is frequently unclear but by creating a whole cell model from molecular mechanisms it is possible to predict how a given mutation will lead to a particular a cellular phenotype.

Whole cell models based on atomic resolution information have opened entirely new avenues of research into drug discovery. In addition to suggesting which protein is a viable target, mechanistic whole cell models can suggest which protein conformation is most favorable and even the chemical shape of a small molecule necessary to inhibit/promote activation. This allows a scale of specificity that could decrease toxicity and limit side effects.

Cardiac arrhythmias are a prime example of the potential relevance of whole cell models. Currently, one of the most commonly prescribed classes of drugs to treat arrhythmias are beta-blockers. Beta blockers bind the beta adrenergic receptors to inhibit epinephrine and norepinephrine binding to reduce the chance of a second heart attack[346]. However, this inhibits the entire beta-adrenergic pathway. By combining this new PKA protein-scale MSM with previously published adrenergic signaling models of the heart[266,279], it is possible to suggest drug targets and even specific binding pockets to inhibit parts of the pathway while limiting their effect on the rest of the cell.

9.5 Conclusions

For years, atomic-resolution protein structures have aided our understanding of protein function not only through static structures provided by NMR and crystallographic experiments, but also through the prediction of dynamic properties with MD simulations. MD has revealed ensembles of structures that comprise the conformational landscape of a protein. Due to computational limitations, classical MD simulations are only able to generate microseconds (or less) of simulated time. This significantly limits the extent of the protein conformational ensemble sampled. However, information generated by MD simulations can be integrated into atomic scale MSMs, which are used to link states generated in a conformational ensemble through a kinetic scheme. The outputs of structures from MD simulations are analyzed according to a chosen conformational state description and are discretized into microstates. The MD trajectory informs which transitory states are most favorable and calculates the transition rates between these states to be used in different scales of modeling.

MD simulations subsequently inform both the atomic scale MSMs and BD simulations. BD simulations typically use rigid-body representations; therefore selected conformations are important for understanding the effect of different structures on association probability. MD simulations provide relevant conformations for BD by generating stable conformations. In addition, ensemble-averaged electrostatics can be generated from the MD trajectories, reflecting the dynamic properties of a molecule in a static electrostatic potential map. Finally, MD and BD

simulations can be directly integrated through milestoning to derive association rate constants (k_{on}) of diffusion-limited processes; a process that combines the two distinct simulation regimes—utilizing the advantages and minimizing the disadvantages of each. Such a scheme can vastly expand the time and length scales accessible in the simulation of multimolecular interactions between proteins and small molecules and/or other proteins to be combined with experimental data in protein-scale MSMs.

Protein-scale MSMs draw on every facet of the atomistic models to bridge the atomic and cellular scales. MD simulations and atomistic-scale MSMs suggest which ensemble of states will reproduce a molecular function. BD simulations combined with milestoning predict association rate constants that would be difficult to experimentally reproduce. This information, when combined with *in vitro* experimental data and statistical analysis tools, leads to the development of protein-scale MSM's for incorporation into whole cell models. Whole cell models based on atomic level details provide a new scale of specificity. The ability to scale up the effects of a protein mutation on a cellular level function is the ideal goal of a robust MSM of this kind.

As discussed throughout this paper, during the process of multiscale modeling it is essential to consider error propagation, or the effect of inaccuracies in small-scale models and the translation of this error into higher levels. For example, conditions such as molecular crowding in the cell likely affect the energetic landscape of a protein, a phenomenon not explicitly represented in an MD simulation. The limited sampling time of MD simulations can bias the conformational landscape of the protein, affecting the kinetic rates determined by the MSM. Structures and kinetics

abstracted from biased simulations can further limit the accuracy of BD simulations and protein-scale MSM, respectively. Furthermore, coarser-grained simulations such as BD and protein-scale MSMs are not free of their own inaccuracies. The sources of errors in many modeling methods may or may not be easy to recognized and the best practices for quantifying the errors is still an active area of research. Iterating through the multiscale modeling process is extremely time consuming, since frequently MSMs must be fully recreated when new constrains are added. With ample computational resources, a multiscale modeler can incorporate recursive feedback loops from multiple scales to converge to a steady solution of the represented system.

Building a multiscale model, despite inaccuracies, is extremely useful. The findings of larger scale models can be used to inform the finer scales, like the identification of unknown conformational states in protein-small molecule energetic landscape. Carefully considering the whole-cell constraints of a given disease state or drug target before creating the initial model can allow these multiscale models to be a powerful and efficient tool for understanding the mechanisms behind some of the most intriguing biological questions.

Chapter 9, in full, is a reprint of “Bridging Scales Through Multiscale Modeling: A Case Study on Protein Kinase A”, which was published in 2015 in Frontiers in Physiology, volume 6, by Britton W. Boras, Sophie P. Hirakis, Lane W. Votapka, Robert D. Malmstrom, and Rommie E. Amaro. The dissertation author was the third investigator and author of this paper.

Conclusion of the Dissertation

In this dissertation, I outline several projects, all of which make extensive use of dynamic structural information obtained using one or more computer simulation methods. Multiple algorithms and techniques were utilized and developed to perform the necessary preparation, execution, and analysis of the simulations and the resulting trajectories.

I presented four tools that are variously used to perform ensemble-averaged electrostatics (DelEE), multistructural hot-spot determination using computational solvent fragment mapping (FTProd), allosterical communication pathway determination (WISP), multistructural pocket volume determination (POVME). All of these tools allow the user to utilize multistructural or dynamic information; that is, they are not bound by a single static structure. Proteins and other biomolecules are almost never static crystals. They are dynamic nanomachines that vibrate, fluctuate, breathe, tumble and drift. Therefore, a multistructural approach can offer superior insight into biomolecular characteristics compared to a single static structure. Through the use of simulation and other computational methods, the understanding of a protein, and its relation to other molecules and to its cellular environment, can be enhanced by including information about its thermodynamic and kinetic properties.

In addition to the tools mentioned in the paragraph above, I also present a tool that performs prediction of rate constants of binding by combining MD with BD using the theory of milestoning (SEEKR). By utilizing multiple simulation methods of differing accuracy and computational cost and combining the trajectories using

appropriate mathematical techniques, the strengths of the simulation methods can be utilized while simultaneously diminishing their weaknesses. This can allow for vastly larger time and space scales to be made available in the investigation of a biomolecular system, and thereby allows one to predict interesting thermodynamic or kinetic quantities that may otherwise be more difficult or expensive to compute using only a single simulation method.

For many of these projects, particularly in the case where molecular visualization or a graphical user interface was desirable, I utilized VMD as a convenient platform. Since VMD is widely used and contains frameworks to crowd-source improvements and additions to its plugin library, I wrote several plugins for VMD that are open to free use and improvement by the scientific community. Other tools were developed as standalone programs that, either because they require no graphical visualization or for speed or other reasons do not work inside any another program's framework.

Computer software has immense potential to be used to compute solutions to the numerous equations used in science for which an analytic solution is difficult or impossible. Open-source computational scientific tools provide free use and improvement of these resources, and therefore function as important assets for streamlining and expediting scientific advancement. Therefore, one general goal for this body of work was to generate automated and interactive tools that allow myself and others to easily perform calculations that are useful for biophysical and biomedical research.

References

1 Faradjian, A. K. & Elber, R. Computing time scales from reaction coordinates by milestoning. *The Journal of chemical physics* **120**, 10880-10889, doi:10.1063/1.1738640 (2004).

2 Vanden-Eijnden, E., Venturoli, M., Ciccotti, G. & Elber, R. On the assumptions underlying milestoning. *J Chem Phys* **129**, doi:Artn 174102 Doi 10.1063/1.2996509 (2008).

3 Schutte, C., Noe, F., Lu, J., Sarich, M. & Vanden-Eijnden, E. Markov state models based on milestoning. *J Chem Phys* **134**, 204105, doi:10.1063/1.3590108 (2011).

4 Maragliano, L., Vanden-Eijnden, E. & Roux, B. Free Energy and Kinetics of Conformational Transitions from Voronoi Tessellated Milestoning with Restraining Potentials. *J Chem Theory Comput* **5**, 2589-2594, doi:Doi 10.1021/Ct900279z (2009).

5 Russell, R. J., Haire, L. F., Stevens, D. J., Collins, P. J., Lin, Y. P., Blackburn, G. M., Hay, A. J., Gamblin, S. J. & Skehel, J. J. The structure of H5N1 avian influenza neuraminidase suggests new opportunities for drug design. *Nature* **443**, 45-49, doi:10.1038/nature05114 (2006).

6 von Itzstein, M. The war against influenza: discovery and development of sialidase inhibitors. *Nature reviews. Drug discovery* **6**, 967-974, doi:10.1038/nrd2400 (2007).

7 Rudrawar, S., Dyason, J. C., Rameix-Welti, M.-A., Rose, F. J., Kerry, P. S., Russell, R. J. M., van der Werf, S., Thomson, R. J., Naffakh, N. & von Itzstein, M. Novel sialic acid derivatives lock open the 150-loop of an influenza A virus group-1 sialidase. *Nat Commun* **1**, 113 (2011).

8 Dharan, N. J., Gubareva, L. V., Meyer, J. J., Okomo-Adhiambo, M., McClinton, R. C., Marshall, S. A., St George, K., Epperson, S., Brammer, L., Klimov, A. I., Bresee, J. S. & Fry, A. M. Infections with oseltamivir-resistant influenza A(H1N1) virus in the United States. *Jama* **301**, 1034-1041, doi:10.1001/jama.2009.294 (2009).

9 Li, Q., Qi, J., Zhang, W., Vavricka, C. J., Shi, Y., Wei, J., Feng, E., Shen, J., Chen, J., Liu, D., He, J., Yan, J., Liu, H., Jiang, H., Teng, M., Li, X. & Gao, G. F. The 2009 pandemic H1N1 neuraminidase N1 lacks the 150-cavity in its active site. *Nat Struct Mol Biol* **17**, 1266-1268, doi:http://www.nature.com/nsmb/journal/v17/n10/abs/nsmb.1909.html - supplementary-information (2010).

10 Amaro, R. E., Minh, D. D., Cheng, L. S., Lindstrom, W. M., Jr., Olson, A. J., Lin, J. H., Li, W. W. & McCammon, J. A. Remarkable loop flexibility in avian influenza N1 and its implications for antiviral drug design. *J Am Chem Soc* **129**, 7764-7765, doi:10.1021/ja0723535 (2007).

11 Li, J. & Cardona, C. J. Adaptation and transmission of a wild duck avian influenza isolate in chickens. *Avian Dis* **54**, 586-590 (2010).

12 Garten, R. J., Davis, C. T., Russell, C. A., Shu, B., Lindstrom, S., Balish, A., Sessions, W. M., Xu, X., Skepner, E., Deyde, V., Okomo-Adhiambo, M., Gubareva, L., Barnes, J., Smith, C. B., Emery, S. L., Hillman, M. J., Rivailler, P., Smagala, J., de Graaf, M., Burke, D. F., Fouchier, R. A., Pappas, C., Alpuche-Aranda, C. M., Lopez-Gatell, H., Olivera, H., Lopez, I., Myers, C. A., Faix, D., Blair, P. J., Yu, C., Keene, K. M., Dotson, P. D., Jr., Boxrud, D., Sambol, A. R., Abid, S. H., St George, K., Bannerman, T., Moore, A. L., Stringer, D. J., Blevins, P., Demmler-Harrison, G. J., Ginsberg, M., Kriner, P., Waterman, S., Smole, S., Guevara, H. F., Belongia, E. A., Clark, P. A., Beatrice, S. T., Donis, R., Katz, J., Finelli, L., Bridges, C. B., Shaw, M., Jernigan, D. B., Uyeki, T. M., Smith, D. J., Klimov, A. I. & Cox, N. J. Antigenic and genetic characteristics of swine-origin 2009 A(H1N1) influenza viruses circulating in humans. *Science* **325**, 197-201, doi:10.1126/science.1176225 (2009).

13 Varghese, J. N. & Colman, P. M. Three-dimensional structure of the neuraminidase of influenza virus A/Tokyo/3/67 at 2.2 A resolution. *J Mol Biol* **221**, 473-486, doi:0022-2836(91)80068-6 [pii] (1991).

14 Caves, L. S., Evanseck, J. D. & Karplus, M. Locally accessible conformations of proteins: multiple molecular dynamics simulations of crambin. *Protein Sci* **7**, 649-666, doi:10.1002/pro.5560070314 (1998).

15 Durrant, J. D., de Oliveira, C. A. & McCammon, J. A. POVME: An algorithm for measuring binding-pocket volumes. *J Mol Graph Model* **29**, 773-776, doi:S1093-3263(10)00155-5 [pii] 10.1016/j.jmgm.2010.10.007 (2011).

16 Sanner, M. F., Olson, A. J. & Spehner, J. C. Reduced surface: an efficient way to compute molecular surfaces. *Biopolymers* **38**, 305-320 (1996).

17 Amaro, R. E., Cheng, X., Ivanov, I., Xu, D. & McCammon, J. A. Characterizing loop dynamics and ligand recognition in human- and avian-type influenza neuraminidases via generalized born molecular dynamics and end-point free energy calculations. *J Am Chem Soc* **131**, 4702-4709, doi:10.1021/ja8085643 (2009).

18 Wagner, R., Matrosovich, M. & Klenk, H. D. Functional balance between haemagglutinin and neuraminidase in influenza virus infections. *Rev Med Virol* **12**, 159-166, doi:10.1002/rmv.352 (2002).

19 Dolinsky, T., Nielsen, J., McCammon, J. & Baker, N. PDB2PQR: an automated pipeline for the setup, execution, and analysis of Poisson-Boltzmann electrostatics calculations. *Nucleic Acids Res* **32**, W665-W667 (2004).

20 Li, H., Robertson, A. D. & Jensen, J. H. Very fast empirical prediction and rationalization of protein pKa values. *Proteins* **61**, 704-721, doi:10.1002/prot.20660 (2005).

21 Newhouse, E. I. Mechanism of Glycan Receptor Recognition and Specificity Switch for Avian, Swine, and Human Adapted Influenza Virus Hemagglutinins: A Molecular Dynamics Perspective. *Journal of the American Chemical Society* **131**, 17430-17442 (2009).

22 Hornak, V., Abel, R., Okur, A., Strockbine, B., Roitberg, A. & Simmerling, C. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins* **65**, 712-725, doi:10.1002/prot.21123 (2006).

23 Berendsen, H. J. C. Molecular Dynamics with coupling to an external bath. *J Chem Phys* **81**, 3684 (1984).

24 Andersen, H. RATTLE: A "Velocity" version of the SHAKE algorithm for molecular dynamics calculations. *J Comput Phys* **52**, 24-34 (1983).

25 Darden, T. Particle mesh Ewald: An N [center-dot] log(N) method for Ewald sums in large systems. *J Chem Phys* **98**, 10089-10092 (1993).

26 Christen, M., Hunenberger, P. H., Bakowies, D., Baron, R., Burgi, R., Geerke, D. P., Heinz, T. N., Kastenholz, M. A., Krautler, V., Oostenbrink, C., Peter, C., Trzesniak, D. & Van Gunsteren, W. F. The GROMOS software for biomolecular simulation: GROMOS05. *J Comput Chem* **26**, 1719-1751, doi:Doi 10.1002/Jcc.20303 (2005).

27 Daura, X., Jaun, B., Seebach, D., van Gunsteren, W. F. & Mark, A. E. Reversible peptide folding in solution by molecular dynamics simulation. *J Mol Biol* **280**, 925-932 (1998).

28 Lindahl, E., Hess, B. & van der Spoel, D. GROMACS 3.0: A package for molecular simulation and trajectory analysis. *J Mol Mod* **7**, 306-317 (2001).

29 Xu, D., Newhouse, E. I., Amaro, R. E., Pao, H. C., Cheng, L. S., Markwick, P. R., McCammon, J. A., Li, W. W. & Arzberger, P. W. Distinct glycan

topology for avian and human sialopentasaccharide receptor analogues upon binding different hemagglutinins: a molecular dynamics perspective. *J Mol Biol* **387**, 465-491, doi:10.1016/j.jmb.2009.01.040 (2009).

30 Humphrey, W., Dalke, A. & Schulten, K. VMD: visual molecular dynamics. *J Mol Graph* **14**, 33-38, doi:0263785596000185 [pii] (1996).

31 Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F. & Higgins, D. G. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic acids research* **25**, 4876-4882 (1997).

32 Ronquist, F. & Huelsenbeck, J. P. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572-1574 (2003).

33 Posada, D. jModelTest: phylogenetic model averaging. *Mol Biol Evol* **25**, 1253-1256, doi:msn083 [pii] 10.1093/molbev/msn083 (2008).

34 Rambaut, A. & Drummond, A. J. *Tracer v1.4, Available from* http://beast.bio.ed.ac.uk/Tracer, 2007).

35 PAUP*: phylogentic analysis using parsimony (Sinauer Associates, Sunderland, MA, 1998).

36 Honig, B. & Nicholls, A. Classical Electrostatics in Biology and Chemistry. *Science* **268**, 1144-1149, doi:Doi 10.1126/Science.7761829 (1995).

37 York, D. M., Darden, T. A. & Pedersen, L. G. The Effect of Long-Range Electrostatic Interactions in Simulations of Macromolecular Crystals - a Comparison of the Ewald and Truncated List Methods. *J Chem Phys* **99**, 8345-8348 (1993).

38 Greengard, L. & Rokhlin, V. A Fast Algorithm for Particle Simulations. *J Comput Phys* **73**, 325-348 (1987).

39 Fogolari, F., Brigo, A. & Molinari, H. The Poisson-Boltzmann equation for biomolecular electrostatics: a tool for structural biology. *J Mol Recognit* **15**, 377-392, doi:Doi 10.1002/Jmr.577 (2002).

40 Harris, R. C., Bredenberg, J. H., Silalahi, A. R. J., Boschitsch, A. H. & Fenley, M. O. Understanding the physical basis of the salt dependence of the electrostatic binding free energy of mutated charged ligand-nucleic acid complexes. *Biophysical Chemistry* **156**, 79-87, doi:10.1016/j.bpc.2011.02.010 (2011).

41 Rocchia, W., Sridharan, S., Nicholls, A., Alexov, E., Chiabrera, A. & Honig, B. Rapid grid-based construction of the molecular surface and the use of

induced surface charge to calculate reaction field energies: applications to the molecular systems and geometric objects. *J Comput Chem* **23**, 128-137, doi:10.1002/jcc.1161 (2002).

42 Case, D. A., Cheatham, T. E., 3rd, Darden, T., Gohlke, H., Luo, R., Merz, K. M., Jr., Onufriev, A., Simmerling, C., Wang, B. & Woods, R. J. The Amber biomolecular simulation programs. *J Comput Chem* **26**, 1668-1688, doi:10.1002/jcc.20290 (2005).

43 Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S. & Karplus, M. Charmm - a Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *J Comput Chem* **4**, 187-217 (1983).

44 Suydam, I. T. Electric fields at the active site of an enzyme: direct comparison of experiment with theory (vol 313, pg 200, 2006). *Science* **313**, 1887-1887 (2006).

45 Aksimentiev, A. & Schulten, K. Imaging alpha-hemolysin with molecular dynamics: ionic conductance, osmotic permeability, and the electrostatic potential map. *Biophys J* **88**, 3745-3761, doi:10.1529/biophysj.104.058727 (2005).

46 Rocchia, W., Alexov, E. & Honig, B. Extending the applicability of the nonlinear Poisson-Boltzmann equation: Multiple dielectric constants and multivalent ions. *J Phys Chem B* **105**, 6507-6514 (2001).

47 Amaro, R. E., Cheng, X., Ivanov, I., Xu, D. & McCammon, J. A. Characterizing Loop Dynamics and Ligand Recognition in Human- and Avian-type Influenza Neuraminidases via Generalized Born Molecular Dynamics and End-point Free Energy Calculations. *J Am Chem Soc* **131**, 4702-4709 (2009).

48 Lawrenz, M., Wereszczynski, J., Amaro, R., Walker, R., Roitberg, A. & McCammon, J. A. Impact of calcium on N1 influenza neuraminidase dynamics and binding free energy. *Proteins* **78**, 2523-2532, doi:10.1002/prot.22761 (2010).

49 Sung, J. C., Van Wynsberghe, A. W., Amaro, R. E., Li, W. W. & McCammon, J. A. Role of secondary sialic acid binding sites in influenza N1 neuraminidase. *Journal of the American Chemical Society* **132**, 2883-2885, doi:10.1021/ja9073672 (2010).

50 Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R. D., Kale, L. & Schulten, K. Scalable molecular dynamics

with NAMD. *J Comput Chem* **26**, 1781-1802, doi:Doi 10.1002/Jcc.20289 (2005).

51 Scott, E. F., Yuhong, Z., Richard, W. P., Bernard, R. B. Constant pressure molecular dynamics simulation: the Langevin piston method. *J Chem Phys* **103**, 4613-4621 (1995).

52 Amaro, R. E., Swift, R. V., Votapka, L., Li, W. W., Walker, R. C. & Bush, R. M. Mechanism of 150-cavity formation in influenza neuraminidase. *Nat Commun* **2**, 388, doi:10.1038/ncomms1390 ncomms1390 [pii] (2011).

53 Baker, N. A. Poisson-Boltzmann methods for biomolecular electrostatics. *Methods Enzymol* **383**, 94-118, doi:10.1016/S0076-6879(04)83005-2 (2004).

54 Baker, N. A. & McCammon, J. A. *Structural Bioinformatics*. 427-440 (John Wiley & Sons, 2003).

55 Hajduk, P. J., Huth, J. R. & Fesik, S. W. Druggability indices for protein targets derived from NMR-based screening data. *Journal of Medicinal Chemistry* **48**, 2518-2525, doi:Doi 10.1021/Jm049131r (2005).

56 Vajda, S. & Guarnieri, F. Characterization of protein-ligand interaction sites using experimental and computational methods. *Current Opinion in Drug Discovery & Development* **9**, 354-362 (2006).

57 Sperandio, O., Miteva, M. A., Segers, K., Nicolaes, G. A. & Villoutreix, B. O. Screening Outside the Catalytic Site: Inhibition of Macromolecular Interactions Through Structure-Based Virtual Ligand Screening Experiments. *The open biochemistry journal* **2**, 29-37, doi:10.2174/1874091X00802010029 (2008).

58 Mattos, C. & Ringe, D. Locating and characterizing binding sites on proteins. *Nature biotechnology* **14**, 595-599, doi:10.1038/nbt0596-595 (1996).

59 Allen, K. N., Bellamacina, C. R., Ding, X. C., Jeffery, C. J., Mattos, C., Petsko, G. A. & Ringe, D. An experimental approach to mapping the binding surfaces of crystalline proteins. *J Phys Chem-Us* **100**, 2605-2611 (1996).

60 Vajda, S. & Guarnieri, F. Characterization of protein-ligand interaction sites using experimental and computational methods. *Current opinion in drug discovery & development* **9**, 354-362 (2006).

61 Morrow, J. K. & Zhang, S. X. Computational Prediction of Protein Hot Spot Residues. *Current Pharmaceutical Design* **18**, 1255-1265 (2012).

62 Brenke, R., Kozakov, D., Chuang, G. Y., Beglov, D., Hall, D., Landon, M. R., Mattos, C. & Vajda, S. Fragment-based identification of druggable 'hot spots'

of proteins using Fourier domain correlation techniques. *Bioinformatics* **25**, 621-627, doi:10.1093/bioinformatics/btp036 (2009).

63 Landon, M. R., Amaro, R. E., Baron, R., Ngan, C. H., Ozonoff, D., McCammon, J. A. & Vajda, S. Novel druggable hot spots in avian influenza neuraminidase H5N1 revealed by computational solvent mapping of a reduced and representative receptor ensemble. *Chemical biology & drug design* **71**, 106-116, doi:10.1111/j.1747-0285.2007.00614.x (2008).

64 Humphrey, W., Dalke, A. & Schulten, K. VMD: Visual molecular dynamics. *Journal of Molecular Graphics & Modelling* **14**, 33-38 (1996).

65 Votapka L, D. O., Swift RV, Walker RC, Amaro RE. Variable Ligand- and Receptor-Binding Hot Spots in Key Strains of Influenza Neuraminidase. *Journal of Molecular and Genetic Medicine* **6** (2012).

66 Varghese, J. N., Colman, P. M., vanDonkelaar, A., Blick, T. J., Sahasrabudhe, A. & McKimmBreschkin, J. L. Structural evidence for a second sialic acid binding site in avian influenza virus neuraminidases. *P Natl Acad Sci USA* **94**, 11808-11812 (1997).

67 Li, Q., Qi, J., Zhang, W., Vavricka, C. J., Shi, Y., Wei, J., Feng, E., Shen, J., Chen, J., Liu, D., He, J., Yan, J., Liu, H., Jiang, H., Teng, M., Li, X. & Gao, G. F. The 2009 pandemic H1N1 neuraminidase N1 lacks the 150-cavity in its active site. *Nature structural & molecular biology* **17**, 1266-1268, doi:10.1038/nsmb.1909 (2010).

68 Amaro, R. E., Swift, R. V., Votapka, L., Li, W. W., Walker, R. C. & Bush, R. M. Mechanism of 150-cavity formation in influenza neuraminidase. *Nature communications* **2**, 388, doi:10.1038/ncomms1390 (2011).

69 Ngan, C. H., Hall, D. R., Zerbe, B., Grove, L. E., Kozakov, D. & Vajda, S. FTSite: high accuracy detection of ligand binding sites on unbound protein structures. *Bioinformatics* **28**, 286-287, doi:Doi 10.1093/Bioinformatics/Btr651 (2012).

70 Jardine, N. & Sibson, R. *Mathematical taxonomy*. (Wiley, 1971).

71 Durrant, J. D., Hall, L., Swift, R. V., Landon, M., Schnaufer, A. & Amaro, R. E. Novel naphthalene-based inhibitors of Trypanosoma brucei RNA editing ligase 1. *PLoS neglected tropical diseases* **4**, e803, doi:10.1371/journal.pntd.0000803 (2010).

72 Deng, J., Schnaufer, A., Salavati, R., Stuart, K. D. & Hol, W. G. High resolution crystal structure of a key editosome enzyme from Trypanosoma

brucei: RNA editing ligase 1. *J Mol Biol* **343**, 601-613, doi:10.1016/j.jmb.2004.08.041 (2004).

73 Amaro, R. E., Swift, R. V. & McCammon, J. A. Functional and structural insights revealed by molecular dynamics simulations of an essential RNA editing ligase in Trypanosoma brucei. *PLoS neglected tropical diseases* **1**, e68, doi:10.1371/journal.pntd.0000068 (2007).

74 Shuman, S. & Lima, C. D. The polynucleotide ligase and RNA capping enzyme superfamily of covalent nucleotidyltransferases. *Current opinion in structural biology* **14**, 757-764, doi:10.1016/j.sbi.2004.10.006 (2004).

75 Demir, O. & Amaro, R. E. Elements of nucleotide specificity in the Trypanosoma brucei mitochondrial RNA editing enzyme RET2. *Journal of chemical information and modeling* **52**, 1308-1318, doi:10.1021/ci3001327 (2012).

76 Deng, J., Ernst, N. L., Turley, S., Stuart, K. D. & Hol, W. G. Structural basis for UTP specificity of RNA editing TUTases from Trypanosoma brucei. *The EMBO journal* **24**, 4007-4017, doi:10.1038/sj.emboj.7600861 (2005).

77 Sanner, M. F., Olson, A. J. & Spehner, J.-C. in *Proceedings of the eleventh annual symposium on Computational geometry* 406-407 (ACM, Vancouver, British Columbia, Canada, 1995).

78 Kortvelyesi, T., Dennis, S., Silberstein, M., Brown, L., 3rd & Vajda, S. Algorithms for computational solvent mapping of proteins. *Proteins* **51**, 340-351, doi:10.1002/prot.10287 (2003).

79 Landon, M., Lancia, D., Yu, J., Thiel, S. & Vadja, S. Identification of hot spots within druggable binding regions by computational solvent mapping of proteins. *J Med Chem* **50**, 1231 - 1240 (2007).

80 Landon, M., Lieberman, R., Hoang, Q., Ju, S., Caaveiro, J., Orwig, S., Kozakov, D., Brenke, R., Chuang, G.-Y., Beglov, D., Vajda, S., Petsko, G. & Ringe, D. Detection of ligand binding hot spots on protein surfaces via fragment-based methods: application to DJ-1 and glucocerebrosidase. *Journal of Computer-Aided Molecular Design* **23**, 491-500, doi:10.1007/s10822-009-9283-2 (2009).

81 Dolinsky, T. J., Czodrowski, P., Li, H., Nielsen, J. E., Jensen, J. H., Klebe, G. & Baker, N. A. PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. *Nucleic Acids Res* **35**, W522-525, doi:gkm276 [pii] 10.1093/nar/gkm276 (2007).

82 AMBER11 (2010).

83 Johnson, J. E., Jorgensen, J. H., Crawford, S. A., Redding, J. S. & Pruneda, R. C. Comparison of two automated instrument systems for rapid susceptibility testing of gram-negative bacilli. *Journal of clinical microbiology* **18**, 1301-1309 (1983).

84 Hess, B., Kutzner, C., van der Spoel, D. & Lindahl, E. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *Journal of Chemical Theory and Computation* **4**, 435-447, doi:10.1021/ct700301q (2008).

85 Nussinov, R., Tsai, C. J. & Ma, B. Y. The Underappreciated Role of Allostery in the Cellular Network. *Annu Rev Biophys* **42**, 169-189, doi:Doi 10.1146/Annurev-Biophys-083012-130257 (2013).

86 Cui, Q. & Karplus, M. Allostery and cooperativity revisited. *Protein Sci* **17**, 1295-1307, doi:Doi 10.1110/Ps.03259908 (2008).

87 Berezovsky, I. N. Thermodynamics of allostery paves a way to allosteric drugs. *Bba-Proteins Proteom* **1834**, 830-835, doi:Doi 10.1016/J.Bbapap.2013.01.024 (2013).

88 Nussinov, R. & Tsai, C. J. Allostery in Disease and in Drug Discovery. *Cell* **153**, 293-305, doi:Doi 10.1016/J.Cell.2013.03.034 (2013).

89 Laine, E., Martinez, L., Ladant, D., Malliavin, T. & Blondel, A. Molecular Motions as a Drug Target: Mechanistic Simulations of Anthrax Toxin Edema Factor Function Led to the Discovery of Novel Allosteric Inhibitors. *Toxins* **4**, 580-604, doi:Doi 10.3390/Toxins4080580 (2012).

90 Monod, J., Wyman, J. & Changeux, J. P. On the Nature of Allosteric Transitions: A Plausible Model. *J. Mol. Biol.* ***12***, 88-118 (1965).

91 Koshland, D. E., Jr., Nemethy, G. & Filmer, D. Comparison of experimental binding data and theoretical models in proteins containing subunits. *Biochemistry* ***5***, 365-385 (1966).

92 Cooper, A. & Dryden, D. T. F. Allostery without Conformational Change - a Plausible Model. *Eur. Biophys. J. Biophys. Lett.* ***11***, 103-109 (1984).

93 Tsai, C. J., del Sol, A. & Nussinov, R. Allostery: absence of a change in shape does not imply that allostery is not at play. *J. Mol. Biol.* ***378***, 1-11, doi:10.1016/j.jmb.2008.02.034 (2008).

94 Atilgan, A. R. & Atilgan, C. Local motifs in proteins combine to generate global functional moves. *Briefings in functional genomics* **11**, 479-488, doi:10.1093/bfgp/els027 (2012).

95 Gasper, P. M., Fuglestad, B., Komives, E. A., Markwick, P. R. & McCammon, J. A. Allosteric networks in thrombin distinguish procoagulant vs. anticoagulant activities. *Proc Natl Acad Sci U S A* **109**, 21216-21222, doi:10.1073/pnas.1218414109 (2012).

96 Sethi, A., Eargle, J., Black, A. A. & Luthey-Schulten, Z. Dynamical networks in tRNA:protein complexes. *Proc. Natl. Acad. Sci. U. S. A.* ***106***, 6620-6625, doi:10.1073/pnas.0810961106 (2009).

97 VanWart, A. T., Eargle, J., Luthey-Schulten, Z. & Amaro, R. E. Exploring Residue Component Contributions to Dynamical Network Models of Allostery. *J Chem Theory Comput* **8**, 2949-2961, doi:Doi 10.1021/Ct300377a (2012).

98 Miao, Y., Nichols, S. E., Gasper, P. M., Metzger, V. T. & McCammon, J. A. Activation and dynamic network of the M2 muscarinic receptor. *Proc Natl Acad Sci U S A* **110**, 10982-10987, doi:10.1073/pnas.1309755110 (2013).

99 Rivalta, I., Sultan, M. M., Lee, N. S., Manley, G. A., Loria, J. P. & Batista, V. S. Allosteric pathways in imidazole glycerol phosphate synthase. *Proc Natl Acad Sci U S A* **109**, E1428-1436, doi:10.1073/pnas.1120536109 (2012).

100 Ghosh, A., Sakaguchi, R., Liu, C., Vishveshwara, S. & Hou, Y. M. Allosteric communication in cysteinyl tRNA synthetase: a network of direct and indirect readout. *The Journal of biological chemistry* **286**, 37721-37731, doi:10.1074/jbc.M111.246702 (2011).

101 Boehr, D. D., Schnell, J. R., McElheny, D., Bae, S. H., Duggan, B. M., Benkovic, S. J., Dyson, H. J. & Wright, P. E. A Distal Mutation Perturbs Dynamic Amino Acid Networks in Dihydrofolate Reductase. *Biochemistry* **52**, 4605-4619, doi:Doi 10.1021/Bi400563c (2013).

102 Long, D. & Bruschweiler, R. Structural and Entropic Allosteric Signal Transduction Strength via Correlated Motions. *Journal of Physical Chemistry Letters* **3**, 1722-1726, doi:Doi 10.1021/Jz300488e (2012).

103 Liu, M. S., Todd, B. D. & Sadus, R. J. Allosteric Conformational Transition in Adenylate Kinase: Dynamic Correlations and Implication for Allostery. *Aust J Chem* **63**, 405-412, doi:Doi 10.1071/Ch09449 (2010).

104 Hagberg, A. A., Schult, D. A. & Swart, P. J. 11-15.

105 Dubois, P. F. & Yang, T. Y. Extending Python with Fortran. *Comput Sci Eng* **1**, 66-73, doi:Doi 10.1109/5992.790589 (1999).

106 Peterson, P. F2PY: a tool for connecting Fortran and Python programs. *International Journal of Computational Science and Engineering* **4**, 296-305 (2009).

107 SciPy: Open Source Scientific Tools for Python v. 0.11.0 (2001).

108 Ascher, D., Dubois, P. F., Hinsen, K., Hugunin, J. J. & Oliphant, T. *Numerical Python*. (Lawrence Livermore National Laboratory, 1999).

109 Oliphant, T. E. *Guide to NumPy*. (Brigham Young University, 2006).

110 Couch, G. S., Pettersen, E. F., Huang, C. C. & Ferrin, T. E. Annotating Pdb Files with Scene Information. *J Mol Graphics* **13**, 153-158, doi:Doi 10.1016/0263-7855(95)00003-O (1995).

111 Eargle, J. & Luthey-Schulten, Z. NetworkView: 3D display and analysis of protein.RNA interaction networks. *Bioinformatics* **28**, 3000-3001, doi:10.1093/bioinformatics/bts546 (2012).

112 Douangamath, A., Walker, M., Beismann-Driemeyer, S., Vega-Fernandez, M. C., Sterner, R. & Wilmanns, M. Structural evidence for ammonia tunneling across the (beta alpha)(8) barrel of the imidazole glycerol phosphate synthase bienzyme complex. *Structure* **10**, 185-193, doi:Doi 10.1016/S0969-2126(02)00702-5 (2002).

113 Chaudhuri, B. N., Lange, S. C., Myers, R. S., Davisson, V. J. & Smith, J. L. Toward understanding the mechanism of the complex cyclization reaction catalyzed by imidazole glycerolphosphate synthase: Crystal structures of a ternary complex and the free enzyme. *Biochemistry* **42**, 7003-7012, doi:Doi 10.1012/Bi034320h (2003).

114 Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R. D., Kale, L. & Schulten, K. Scalable molecular dynamics with NAMD. *J. Comput. Chem.* ***26***, 1781-1802, doi:10.1002/jcc.20289 (2005).

115 Brooks, B. R., Brooks, C. L., 3rd, Mackerell, A. D., Jr., Nilsson, L., Petrella, R. J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S., Caflisch, A., Caves, L., Cui, Q., Dinner, A. R., Feig, M., Fischer, S., Gao, J., Hodoscek, M., Im, W., Kuczera, K., Lazaridis, T., Ma, J., Ovchinnikov, V., Paci, E., Pastor, R. W., Post, C. B., Pu, J. Z., Schaefer, M., Tidor, B., Venable, R. M., Woodcock, H. L., Wu, X., Yang, W., York, D. M. & Karplus, M. CHARMM: the biomolecular simulation program. *J. Comput. Chem.* ***30***, 1545-1614, doi:10.1002/jcc.21287 (2009).

116 Amaro, R. & Luthey-Schulten, Z. Molecular dynamics simulations of substrate channeling through an α–β barrel protein. *Chem Phys* ***307***, 147-155, doi:10.1016/j.chemphys.2004.05.019 (2004).

117 Andricioaei, I. & Karplus, M. On the calculation of entropy from covariance matrices of the atomic fluctuations. *J. Chem. Phys.* ***115***, 6289-6292 (2001).

118 Glykos, N. M. Software news and updates. Carma: a molecular dynamics analysis program. *J. Comput. Chem.* ***27***, 1765-1768, doi:10.1002/jcc.20482 (2006).

119 Floyd, R. W. Algorithm-97 - Shortest Path. *Commun Acm* ***5***, 345-345 (1962).

120 Alifano, P., Fani, R., Lio, P., Lazcano, A., Bazzicalupo, M., Carlomagno, M. S. & Bruni, C. B. Histidine biosynthetic pathway and genes: structure, regulation, and evolution. *Microbiological reviews* ***60***, 44-69 (1996).

121 Amaro, R. E., Sethi, A., Myers, R. S., Davisson, V. J. & Luthey-Schulten, Z. A. A network of conserved interactions regulates the allosteric signal in a glutamine amidotransferase. *Biochemistry* ***46***, 2156-2173, doi:10.1021/bi061708e (2007).

122 Perot, S., Sperandio, O., Miteva, M. A., Camproux, A. C. & Villoutreix, B. O. Druggable pockets and binding site centric chemical space: a paradigm shift in drug discovery. *Drug Discov Today* **15**, 656-667, doi:Doi 10.1016/J.Drudis.2010.05.015 (2010).

123 Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. The Protein Data Bank. *Nucl. Acids Res.* **28**, 235-242 (2000).

124 Levitt, D. G. & Banaszak, L. J. Pocket - a Computer-Graphics Method for Identifying and Displaying Protein Cavities and Their Surrounding Amino-Acids. *J Mol Graphics* **10**, 229-234, doi:Doi 10.1016/0263-7855(92)80074-N (1992).

125 Smart, O. S., Goodfellow, J. M. & Wallace, B. A. The Pore Dimensions of Gramicidin-A. *Biophys J* **65**, 2455-2460 (1993).

126 Kleywegt, G. J. & Jones, T. A. Detection, Delineation, Measurement and Display of Cavities in Macromolecular Structures. *Acta Crystallogr D* **50**, 178-185, doi:Doi 10.1107/S0907444993011333 (1994).

127 Laskowski, R. A. Surfnet - a Program for Visualizing Molecular-Surfaces, Cavities, and Intermolecular Interactions. *J Mol Graphics* **13**, 323-&, doi:Doi 10.1016/0263-7855(95)00073-9 (1995).

128 Chovancova, E., Pavelka, A., Benes, P., Strnad, O., Brezovsky, J., Kozlikova, B., Gora, A., Sustr, V., Klvana, M., Medek, P., Biedermannova, L., Sochor, J. & Damborsky, J. CAVER 3.0: A Tool for the Analysis of Transport Pathways in Dynamic Protein Structures. *Plos Comput Biol* **8**, doi:Artn E1002708 Doi 10.1371/Journal.Pcbi.1002708 (2012).

129 Eyrisch, S. & Helms, V. Transient pockets on protein surfaces involved in protein-protein interaction. *J Med Chem* **50**, 3457-3464, doi:Doi 10.1021/Jm070095g (2007).

130 Brady, G. P. & Stouten, P. F. W. Fast prediction and visualization of protein binding pockets with PASS. *J Comput Aid Mol Des* **14**, 383-401, doi:Doi 10.1023/A:1008124202956 (2000).

131 Le Guilloux, V., Schmidtke, P. & Tuffery, P. Fpocket: An open source platform for ligand pocket detection. *Bmc Bioinformatics* **10**, doi:Artn 168 Doi 10.1186/1471-2105-10-168 (2009).

132 Schmidtke, P., Bidon-Chanal, A., Luque, F. J. & Barril, X. MDpocket: open-source cavity detection and characterization on molecular dynamics trajectories. *Bioinformatics* **27**, 3276-3285, doi:Doi 10.1093/Bioinformatics/Btr550 (2011).

133 Halgren, T. A. Identifying and Characterizing Binding Sites and Assessing Druggability. *J Chem Inf Model* **49**, 377-389, doi:Doi 10.1021/Ci800324m (2009).

134 Halgren, T. New method for fast and accurate binding-site identification and analysis. *Chem Biol Drug Des* **69**, 146-148, doi:Doi 10.1111/J.1747-0285.2007.00483.X (2007).

135 Votapka, L. & Amaro, R. E. Multistructural hot spot characterization with FTProd. *Bioinformatics* **29**, 393-394, doi:Doi 10.1093/Bioinformatics/Bts689 (2013).

136 Zheng, X. L., Gan, L. F., Wang, E. K. & Wang, J. Pocket-Based Drug Design: Exploring Pocket Space. *Aaps J* **15**, 228-241, doi:Doi 10.1208/S12248-012-9426-6 (2013).

137 Baron, R. & Vellore, N. A. LSD1/CoREST is an allosteric nanoscale clamp regulated by H3-histone-tail molecular recognition. *P Natl Acad Sci USA* **109**, 12509-12514, doi:Doi 10.1073/Pnas.1207892109 (2012).

138 Fuchs, J. E., Huber, R. G., Von Grafenstein, S., Wallnoefer, H. G., Spitzer, G. M., Fuchs, D. & Liedl, K. R. Dynamic Regulation of Phenylalanine

Hydroxylase by Simulated Redox Manipulation. *Plos One* **7**, doi:ARTN e53005 DOI 10.1371/journal.pone.0053005 (2012).

139 Sinko, W., de Oliveira, C., Williams, S., Van Wynsberghe, A., Durrant, J. D., Cao, R., Oldfield, E. & McCammon, J. A. Applying Molecular Dynamics Simulations to Identify Rarely Sampled Ligand-bound Conformational States of Undecaprenyl Pyrophosphate Synthase, an Antibacterial Target. *Chem Biol Drug Des*, doi:10.1111/j.1747-0285.2011.01101.x (2011).

140 Lindert, S., Kekenes-Huskey, P. M. & McCammon, J. A. Long-Timescale Molecular Dynamics Simulations Elucidate the Dynamics and Kinetics of Exposure of the Hydrophobic Patch in Troponin C. *Biophysical journal* **103**, 1784-1789, doi:Doi 10.1016/J.Bpj.2012.08.058 (2012).

141 Boechi, L., de Oliveira, C. A., Da Fonseca, I., Kizjakina, K., Sobrado, P., Tanner, J. J. & McCammon, J. A. Substrate-dependent dynamics of UDP-galactopyranose mutase: Implications for drug design. *Protein Sci*, doi:10.1002/pro.2332 (2013).

142 Wu, Y., Qin, G. R., Gao, F., Liu, Y., Vavricka, C. J., Qi, J. X., Jiang, H. L., Yu, K. Q. & Gao, G. F. Induced opening of influenza virus neuraminidase N2 150-loop suggests an important role in inhibitor binding. *Sci Rep-Uk* **3**, doi:Artn 1551 Doi 10.1038/Srep01551 (2013).

143 Han, N. Y. & Mu, Y. G. Plasticity of 150-Loop in Influenza Neuraminidase Explored by Hamiltonian Replica Exchange Molecular Dynamics Simulations. *Plos One* **8**, doi:ARTN e60995 DOI 10.1371/journal.pone.0060995 (2013).

144 Schultes, S., Nijmeijer, S., Engelhardt, H., Kooistra, A. J., Vischer, H. F., de Esch, I. J. P., Haaksma, E. E. J., Leurs, R. & de Graaf, C. Mapping histamine H-4 receptor-ligand binding modes. *Medchemcomm* **4**, 193-204, doi:Doi 10.1039/C2md20212c (2013).

145 Li, P., Chen, Z., Xu, H., Sun, H., Li, H., Liu, H., Yang, H., Gao, Z., Jiang, H. & Li, M. The gating charge pathway of an epilepsy-associated potassium channel accommodates chemical ligands. *Cell Res*, doi:10.1038/cr.2013.82 cr201382 [pii] (2013).

146 Kekenes-Huskey, P. M., Metzger, V. T., Grant, B. J. & McCammon, J. A. Calcium binding and allosteric signaling mechanisms for the sarcoplasmic reticulum Ca2+ATPase. *Protein Sci* **21**, 1429-1443, doi:Doi 10.1002/Pro.2129 (2012).

147 Bung, N., Pradhan, M., Srinivasan, H. & Bulusu, G. Structural Insights into E. coli Porphobilinogen Deaminase during Synthesis and Exit of 1-

Hydroxymethylbilane. *Plos Comput Biol* **10**, e1003484, doi:10.1371/journal.pcbi.1003484 (2014).

148 Torres, R., Swift, R. V., Chim, N., Wheatley, N., Lan, B. S., Atwood, B. R., Pujol, C., Sankaran, B., Bliska, J. B., Amaro, R. E. & Goulding, C. W. Biochemical, Structural and Molecular Dynamics Analyses of the Potential Virulence Factor RipA from Yersinia pestis. *Plos One* **6**, doi:ARTN e25084 DOI 10.1371/journal.pone.0025084 (2011).

149 Grant, B. J., Lukman, S., Hocker, H. J., Sayyah, J., Brown, J. H., McCammon, J. A. & Gorfe, A. A. Novel Allosteric Sites on Ras for Lead Generation. *Plos One* **in press** (2011).

150 Mowrey, D., Cheng, M. H., Liu, L. T., Willenbring, D., Lu, X. H., Wymore, T., Xu, Y. & Tang, P. Asymmetric Ligand Binding Facilitates Conformational Transitions in Pentameric Ligand-Gated Ion Channels. *J Am Chem Soc* **135**, 2172-2180, doi:Doi 10.1021/Ja307275v (2013).

151 Yi-Xin, A., Jun-Rui, L., Chun-Wei, X., Jiang-Bei, M., Xu-Yun, Y. & He, Z. Simulated Mechanism of Triclosan in Modulating the Active Site and Loop of FabI by Computer. *Acta Physico-Chimica Sinica* **30**, 559-568, doi:10.3866/PKU.WHXB201401132 (2014).

152 Blachly, P. G., de Oliveira, C. A. F., Williams, S. L. & McCammon, J. A. Utilizing a Dynamical Description of IspH to Aid in the Development of Novel Antimicrobial Drugs. *Plos Comput Biol* **9**, doi:Artn E1003395 Doi 10.1371/Journal.Pcbi.1003395 (2013).

153 Demir, O. & Amaro, R. Dynamical Insights into the Essential Editosome Enzymes of Trypanosoma brucei. *Protein Sci* **21**, 214-214 (2012).

154 Mowrey, D. D., Liu, Q., Bondarenko, V., Chen, Q., Seyoum, E., Xu, Y., Wu, J. & Tang, P. Insights into Distinct Modulation of alpha 7 and alpha 7 beta 2 Nicotinic Acetylcholine Receptors by the Volatile Anesthetic Isoflurane. *J Biol Chem* **288**, 35793-35800, doi:Doi 10.1074/Jbc.M113.508333 (2013).

155 Bustamante, J. P., Abbruzzetti, S., Marcelli, A., Gauto, D., Boechi, L., Bonamore, A., Boffi, A., Bruno, S., Feis, A., Foggi, P., Estrin, D. A. & Viappiani, C. Ligand Uptake Modulation by Internal Water Molecules and Hydrophobic Cavities in Hemoglobins. *J Phys Chem B* **118**, 1234-1245, doi:Doi 10.1021/Jp410724z (2014).

156 Selvam, B., Porter, S. L. & Tikhonova, I. G. Addressing Selective Polypharmacology of Antipsychotic Drugs Targeting the Bioaminergic Receptors through Receptor Dynamic Conformational Ensembles. *J. Chem. Inf. Model.* **53**, 1761-1774 (2013).

157 Weinreb, V., Li, L., Chandrasekaran, S. N., Koehl, P., Delarue, M. & Carter, C. W., Jr. Enhanced Amino Acid Selection in Fully Evolved Tryptophanyl-tRNA Synthetase, Relative to Its Urzyme, Requires Domain Motion Sensed by the D1 Switch, a Remote Dynamic Packing Motif. *J Biol Chem* **289**, 4367-4376, doi:10.1074/jbc.M113.538660 (2014).

158 Li, J. N., Jonsson, A. L., Beuming, T., Shelley, J. C. & Voth, G. A. Ligand-Dependent Activation and Deactivation of the Human Adenosine A(2A) Receptor. *J Am Chem Soc* **135**, 8749-8759, doi:Doi 10.1021/Ja404391q (2013).

159 Baron, R. & McCammon, J. A. Molecular Recognition and Ligand Association. *Annu Rev Phys Chem* **64**, 151-175, doi:Doi 10.1146/Annurev-Physchem-040412-110047 (2013).

160 Ariga, K., Ito, H., Hill, J. P. & Tsukube, H. Molecular recognition: from solution science to nano/materials technology. *Chem Soc Rev* **41**, 5800-5835, doi:Doi 10.1039/C2cs35162e (2012).

161 Kahraman, A., Morris, R. J., Laskowski, R. A. & Thornton, J. M. Shape variation in protein binding pockets and their ligands. *J Mol Biol* **368**, 283-301, doi:Doi 10.1016/J.Jmb.2007.01.086 (2007).

162 Seddon, G., Lounnas, V., McGuire, R., van den Bergh, T., Bywater, R. P., Oliveira, L. & Vriend, G. Drug design for ever, from hype to hope. *J Comput Aid Mol Des* **26**, 137-150, doi:Doi 10.1007/S10822-011-9519-9 (2012).

163 Meng, X. Y., Zhang, H. X., Mezei, M. & Cui, M. Molecular Docking: A Powerful Approach for Structure-Based Drug Discovery. *Curr Comput-Aid Drug* **7**, 146-157 (2011).

164 Golbraikh, A., Wang, X. S., Zhu, H. & Tropsha, A. in *Handbook of Computational Chemistry* (ed Jerzy Leszczynski) Ch. 37, 1309-1342 (Springer, 2012).

165 Liang, J., Edelsbrunner, H. & Woodward, C. Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Protein Sci* **7**, 1884-1897 (1998).

166 Rush, T. S., Grant, J. A., Mosyak, L. & Nicholls, A. A shape-based 3-D scaffold hopping method and its application to a bacterial protein-protein interaction. *J Med Chem* **48**, 1489-1495, doi:Doi 10.1021/Jm040163o (2005).

167 Wirth, M., Volkamer, A., Zoete, V., Rippmann, F., Michielin, O., Rarey, M. & Sauer, W. H. B. Protein pocket and ligand shape comparison and its

application in virtual screening. *J Comput Aid Mol Des* **27**, 511-524, doi:Doi 10.1007/S10822-013-9659-1 (2013).

168 Hawkins, P. C. D., Skillman, A. G. & Nicholls, A. Comparison of shape-matching and docking as virtual screening tools. *J Med Chem* **50**, 74-82, doi:Doi 10.1021/Jm0603365 (2007).

169 Distinto, S., Esposito, F., Kirchmair, J., Cardia, M. C., Gaspari, M., Maccioni, E., Alcaro, S., Markt, P., Wolber, G., Zinzula, L. & Tramontano, E. Identification of HIV-1 reverse transcriptase dual inhibitors by a combined shape-, 2D-fingerprint- and pharmacophore-based virtual screening approach. *Eur J Med Chem* **50**, 216-229, doi:Doi 10.1016/J.Ejmech.2012.01.056 (2012).

170 LaLonde, J. M., Elban, M. A., Courter, J. R., Sugawara, A., Soeta, T., Madani, N., Princiotto, A. M., Do Kwon, Y., Kwong, P. D., Schon, A., Freire, E., Sodroski, J. & Smith, A. B. Design, synthesis and biological evaluation of small molecule inhibitors of CD4-gp120 binding based on virtual screening. *Bioorgan Med Chem* **19**, 91-101, doi:Doi 10.1016/J.Bmc.2010.11.049 (2011).

171 Tuccinardi, T., Ortore, G., Santos, M. A., Marques, S. M., Nuti, E., Rossello, A. & Martinelli, A. Multitemplate Alignment Method for the Development of a Reliable 3D-QSAR Model for the Analysis of MMP3 Inhibitors. *J Chem Inf Model* **49**, 1715-1724, doi:Doi 10.1021/Ci900118v (2009).

172 Nicholls, A., McGaughey, G. B., Sheridan, R. P., Good, A. C., Warren, G., Mathieu, M., Muchmore, S. W., Brown, S. P., Grant, J. A., Haigh, J. A., Nevins, N., Jain, A. N. & Kelley, B. Molecular Shape and Medicinal Chemistry: A Perspective. *J Med Chem* **53**, 3862-3886, doi:Doi 10.1021/Jm900818s (2010).

173 Osguthorpe, D. J., Sherman, W. & Hagler, A. T. Exploring Protein Flexibility: Incorporating Structural Ensembles From Crystal Structures and Simulation into Virtual Screening Protocols. *J Phys Chem B* **116**, 6952-6959, doi:Doi 10.1021/Jp3003992 (2012).

174 Osguthorpe, D. J., Sherman, W. & Hagler, A. T. Generation of Receptor Structural Ensembles for Virtual Screening Using Binding Site Shape Analysis and Clustering. *Chem Biol Drug Des* **80**, 182-193, doi:Doi 10.1111/J.1747-0285.2012.01396.X (2012).

175 Ben Nasr, N., Guillemain, H., Lagarde, N., Zagury, J. F. & Montes, M. Multiple Structures for Virtual Ligand Screening: Defining Binding Site Properties-Based Criteria to Optimize the Selection of the Query. *J Chem Inf Model* **53**, 293-311, doi:Doi 10.1021/Ci3004557 (2013).

176 Nichols, S. E., Swift, R. V. & Amaro, R. E. Rational Prediction with Molecular Dynamics for Hit Identification. *Curr Top Med Chem* **12**, 2002-2012 (2012).

177 Dubois, P. F. Extending Python with Fortran. *Computing Science and Engineering* **1**, 66-73 (1999).

178 Peterson, P. F2PY: a tool for connecting Fortran and Python programs. *International Journal of Computational Science and Engineering* **4**, 296-305 (2009).

179 Akl, S. G. & Toussaint, G. T. in *Proc. 4th. Int. Joint Conf. on Pattern Recognition (Kyoto, Japan)*. 483-487.

180 Horn, H. W., Swope, W. C., Pitera, J. W., Madura, J. D., Dick, T. J., Hura, G. L. & Head-Gordon, T. Development of an improved four-site water model for biomolecular simulations: TIP4P-Ew. *J Chem Phys* **120**, 9665-9678, doi:Doi 10.1063/1.1683075 (2004).

181 Meagher, K. L., Redman, L. T. & Carlson, H. A. Development of polyphosphate parameters for use with the AMBER force field. *J Comput Chem* **24**, 1016-1025, doi:10.1002/jcc.10262 (2003).

182 Allner, O., Nilsson, L. & Villa, A. Magnesium Ion-Water Coordination and Exchange in Biomolecular Simulations. *J Chem Theory Comput* **8**, 1493-1502, doi:Doi 10.1021/Ct3000734 (2012).

183 Joung, I. S. & Cheatham, T. E. Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *J Phys Chem B* **112**, 9020-9041, doi:Doi 10.1021/Jp8001614 (2008).

184 Kale, L., R. Skeel, M. Bhandarkar, R. Brunner, A. Gursoy, N. Krawetz, J. Phillips, A. Shinozaki, K. Varadarajan, and K. Schulten. NAMD2: greater scalability for parallel molecular dynamics. *J. Comput. Phys.* **151**, 283-312 (1999).

185 Shipman, J. W. (New Mexico Tech Computer Center, Socorro, New Mexico, 2010).

186 Welch, B. B. *Practical programming in Tcl/Tk*. 4th edn, (Prentice Hall/PTR, 2003).

187 Schnaufer, A., Panigrahi, A. K., Panicucci, B., Igo, R. P., Jr., Salavati, R. & Stuart, K. An RNA Ligase Essential for RNA Editing and Survival of the Bloodstream Form of Trypanosoma brucei. *Science* **291**, 2159-2162 (2001).

188 Rusche, L. N., Huang, C. E., Piller, K. J., Hemann, M., Wirtz, E. & Sollner-Webb, B. The two RNA ligases of the Trypanosoma brucei RNA editing complex: cloning the essential band IV gene and identifying the band V gene. *Mol.Cell.Biol.* **21**, 979-989 (2001).

189 Durrant, J. D., Hall, L., Swift, R. V., Landon, M., Schnaufer, A. & Amaro, R. E. Novel Naphthalene-Based Inhibitors of Trypanosoma brucei RNA Editing Ligase 1. *PLoS Negl Trop Dis* **4**, e803 (2010).

190 Craig, I. R., Pfleger, C., Gohlke, H., Essex, J. W. & Spiegel, K. Pocket-Space Maps To Identify Novel Binding-Site Conformations in Proteins. *J Chem Inf Model* **51**, 2666-2679, doi:Doi 10.1021/Ci200168b (2011).

191 Bar-Even, A., Noor, E., Savir, Y., Liebermeister, W., Davidi, D., Tawfik, D. S. & Milo, R. The Moderately Efficient Enzyme: Evolutionary and Physicochemical Trends Shaping Enzyme Parameters. *Biochemistry-Us* **50**, 4402-4410, doi:Doi 10.1021/Bi2002289 (2011).

192 Copeland, R. A., Pompliano, D. L. & Meek, T. D. Drug-target residence time and its implications for lead optimization (vol 5, pg 730, 2006). *Nat Rev Drug Discov* **6**, 249-249 (2007).

193 Jorgensen, W. L. Foundations of Biomolecular Modeling. *Cell* **155**, 1199-1202, doi:Doi 10.1016/J.Cell.2013.11.023 (2013).

194 Shan, Y. B., Kim, E. T., Eastwood, M. P., Dror, R. O., Seeliger, M. A. & Shaw, D. E. How Does a Drug Molecule Find Its Target Binding Site? *J Am Chem Soc* **133**, 9181-9183, doi:Doi 10.1021/Ja202726y (2011).

195 Shan, Y. B., Eastwood, M. P., Zhang, X. W., Kim, E. T., Arkhipov, A., Dror, R. O., Jumper, J., Kuriyan, J. & Shaw, D. E. Oncogenic Mutations Counteract Intrinsic Disorder in the EGFR Kinase and Promote Receptor Dimerization. *Cell* **149**, 860-870, doi:Doi 10.1016/J.Cell.2012.02.063 (2012).

196 Dror, R. O., Pan, A. C., Arlow, D. H., Borhani, D. W., Maragakis, P., Shan, Y. B., Xu, H. F. & Shaw, D. E. Pathway and mechanism of drug binding to G-protein-coupled receptors. *P Natl Acad Sci USA* **108**, 13118-13123, doi:Doi 10.1073/Pnas.1104614108 (2011).

197 Shaw, D. E., Deneroff, M. M., Dror, R. O., Kuskin, J. S., Larson, R. H., Salmon, J. K., Young, C., Batson, B., Bowers, K. J., Chao, J. C., Eastwood, M. P., Gagliardo, J., Grossman, J. P., Ho, C. R., Ierardi, D. J., Kolossvary, I., Klepeis, J. L., Layman, T., McLeavey, C., Moraes, M. A., Mueller, R., Priest, E. C., Shan, Y. B., Spengler, J., Theobald, M., Towles, B. & Wang, S. C. Anton, a Special-Purpose Machine for Molecular Dynamics Simulation. *Conf Proc Int Symp C*, 1-12 (2007).

198 Shaw, D. E., Dror, R. O., Salmon, J. K., Grossman, J. P., Mackenzie, K. M., Bank, J. A., Young, C., Deneroff, M. M., Batson, B., Bowers, K. J., Chow, E., Eastwood, M. P., Ierardi, D. J., Klepeis, J. L., Kuskin, J. S., Larson, R. H., Lindorff-Larsen, K., Maragakis, P., Moraes, M. A., Piana, S., Shan, Y. B. & Towles, B. Millisecond-Scale Molecular Dynamics Simulations on Anton. *Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis* (2009).

199 Pande, V. S., Beauchamp, K. & Bowman, G. R. Everything you wanted to know about Markov State Models but were afraid to ask. *Methods* **52**, 99-105, doi:Doi 10.1016/J.Ymeth.2010.06.002 (2010).

200 Buchete, N. V. & Hummer, G. Coarse master equations for peptide folding dynamics. *J Phys Chem B* **112**, 6057-6069, doi:Doi 10.1021/Jp0761665 (2008).

201 Held, M. & Noe, F. Calculating kinetics and pathways of protein-ligand association. *Eur J Cell Biol* **91**, 357-364, doi:Doi 10.1016/J.Ejcb.2011.08.004 (2012).

202 Prinz, J. H., Wu, H., Sarich, M., Keller, B., Senne, M., Held, M., Chodera, J. D., Schutte, C. & Noe, F. Markov models of molecular kinetics: generation and validation. *J Chem Phys* **134**, 174105, doi:10.1063/1.3565032 (2011).

203 Swope, W. C., Pitera, J. W. & Suits, F. Describing protein folding kinetics by molecular dynamics simulations. 1. Theory. *J Phys Chem B* **108**, 6571-6581, doi:Doi 10.1021/Jp037421y (2004).

204 Sarich, M., Prinz, J. H. & Schutte, C. Markov model theory. *Advances in experimental medicine and biology* **797**, 23-44, doi:10.1007/978-94-007-7606-7_3 (2014).

205 Schutte, C., Fischer, A., Huisinga, W. & Deuflhard, P. A direct approach to conformational dynamics based on hybrid Monte Carlo. *J Comput Phys* **151**, 146-168, doi:Doi 10.1006/Jcph.1999.6231 (1999).

206 Noe, F., Horenko, I., Schutte, C. & Smith, J. C. Hierarchical analysis of conformational dynamics in biomolecules: Transition networks of metastable states. *J Chem Phys* **126**, doi:Artn 155102 10.1063/1.2714539 (2007).

207 Chodera, J. D., Singhal, N., Pande, V. S., Dill, K. A. & Swope, W. C. Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics. *J Chem Phys* **126**, 155101, doi:10.1063/1.2714538 (2007).

208 Kirmizialtin, S. & Elber, R. Revisiting and computing reaction coordinates with Directional Milestoning. *The journal of physical chemistry. A* **115**, 6137-6148, doi:10.1021/jp111093c (2011).

209 West, A. M., Elber, R. & Shalloway, D. Extending molecular dynamics time scales with milestoning: example of complex kinetics in a solvated peptide. *The Journal of chemical physics* **126**, 145104, doi:10.1063/1.2716389 (2007).

210 Majek, P. & Elber, R. Milestoning without a Reaction Coordinate. *J Chem Theory Comput* **6**, 1805-1817, doi:Doi 10.1021/Ct100114j (2010).

211 Cardenas, A. E., Jas, G. S., DeLeon, K. Y., Hegefeld, W. A., Kuczera, K. & Elber, R. Unassisted transport of N-acetyl-L-tryptophanamide through membrane: experiment and simulation of kinetics. *The journal of physical chemistry. B* **116**, 2739-2750, doi:10.1021/jp2102447 (2012).

212 Ermak, D. L. & Mccammon, J. A. Brownian Dynamics with Hydrodynamic Interactions. *J Chem Phys* **69**, 1352-1360, doi:Doi 10.1063/1.436761 (1978).

213 Northrup, S. H., Allison, S. A. & Mccammon, J. A. Brownian Dynamics Simulation of Diffusion-Influenced Bimolecular Reactions. *J Chem Phys* **80**, 1517-1526, doi:Doi 10.1063/1.446900 (1984).

214 Gabdoulline, R. R. & Wade, R. C. Simulation of the diffusional association of Barnase and Barstar. *Biophysical journal* **72**, 1917-1929 (1997).

215 Huber, G. A. & McCammon, J. A. Browndye: A software package for Brownian dynamics. *Comput Phys Commun* **181**, 1896-1905, doi:Doi 10.1016/J.Cpc.2010.07.022 (2010).

216 Still, W. C., Tempczyk, A., Hawley, R. C. & Hendrickson, T. Semianalytical Treatment of Solvation for Molecular Mechanics and Dynamics. *J Am Chem Soc* **112**, 6127-6129, doi:Doi 10.1021/Ja00172a038 (1990).

217 Greives, N. H.-X. Z. BDflex: A method for efficient treament of molecular flexibility in calculating protein-ligand binding rate constants from Brownian dynamics simulations. *J. Chem. Phys.* **137** (2012).

218 Luty, B. A., Elamrani, S. & Mccammon, J. A. Simulation of the Bimolecular Reaction between Superoxide and Superoxide-Dismutase - Synthesis of the Encounter and Reaction Steps. *J Am Chem Soc* **115**, 11874-11877 (1993).

219 Luty, B. A. & Mccammon, J. A. Simulation of Bimolecular Reactions - Synthesis of the Encounter and Reaction Steps. *Mol Simulat* **10**, 61-65 (1993).

220 Elber, R. & West, A. Atomically detailed simulation of the recovery stroke in myosin by Milestoning. *P Natl Acad Sci USA* **107**, 5001-5005, doi:Doi 10.1073/Pnas.0909636107 (2010).

221 Vanden-Eijnden, E. & Venturoli, M. Markovian milestoning with Voronoi tessellations. *J Chem Phys* **130**, doi:Artn 194101 Doi 10.1063/1.3129843 (2009).

222 Pedersen L., D., T. Molecular Dynamics: Techniques and Applications to Proteins. *Encyclopedia of Computational Chemistry* (2002).

223 Karplus, M. & McCammon, J. A. Molecular dynamics simulations of biomolecules (vol 9, pg 646, 2002). *Nat Struct Biol* **9**, 788-788, doi:Doi 10.1038/Nsb1002-788a (2002).

224 Gabdoulline, R. R. & Wade, R. C. Brownian dynamics simulation of protein-protein diffusional encounter. *Methods* **14**, 329-341, doi:10.1006/meth.1998.0588 (1998).

225 Sines, J. J., Allison, S. A. & Mccammon, J. A. Point-Charge Distributions and Electrostatic Steering in Enzyme Substrate Encounter - Brownian Dynamics of Modified Copper-Zinc Superoxide Dismutases. *Biochemistry-Us* **29**, 9403-9412, doi:Doi 10.1021/Bi00492a014 (1990).

226 Elcock, A. H. Molecular simulations of diffusion and association in multimacromolecular systems. *Numerical Computer Methods, Pt D* **383**, 166-198 (2004).

227 Madura J.D., B. J. M., Wade R.C., Gabdoulline R.R. Brownian Dynamics. *Encyclopedia of Computational Chemistry* (2002).

228 Murphy, T. J. & Aguirre, J. L. Brownian Motion of N Interacting Particles .1. Extension of Einstein Diffusion Relation to N-Particle Case. *J Chem Phys* **57**, 2098-&, doi:Doi 10.1063/1.1678535 (1972).

229 Wilemski, G. Derivation of Smoluchowski Equations with Corrections in Classical-Theory of Brownian-Motion. *J Stat Phys* **14**, 153-169, doi:Doi 10.1007/Bf01011764 (1976).

230 Kirmizialtin, S. & Elber, R. Revisiting and Computing Reaction Coordinates with Directional Milestoning. *J Phys Chem A* **115**, 6137-6148, doi:Doi 10.1021/Jp111093c (2011).

231 Kreuzer, S. M., Moon, T. J. & Elber, R. Catch bond-like kinetics of helix cracking: Network analysis by molecular dynamics and Milestoning. *J Chem Phys* **139**, doi:Artn 121902 Doi 10.1063/1.4811366 (2013).

232 Elber, R. A milestoning study of the kinetics of an allosteric transition: atomically detailed simulations of deoxy Scapharca hemoglobin. *Biophysical journal* **92**, L85-87, doi:10.1529/biophysj.106.101899 (2007).

233 Noe, F. Probability distributions of molecular observables computed from Markov models. *J Chem Phys* **128**, doi:Artn 244103 Doi 10.1063/1.2916718 (2008).

234 Mccammon, J. A., Northrup, S. H. & Allison, S. A. Diffusional Dynamics of Ligand Receptor Association. *J Phys Chem-Us* **90**, 3901-3905, doi:Doi 10.1021/J100408a015 (1986).

235 Calef, D. F. & Deutch, J. M. Diffusion-Controlled Reactions. *Annu Rev Phys Chem* **34**, 493-524, doi:Doi 10.1146/Annurev.Pc.34.100183.002425 (1983).

236 Song, Y. H., Zhang, Y. J., Shen, T. Y., Bajaj, C. L., McCammon, A. & Baker, N. A. Finite element solution of the steady-state Smoluchowski equation for rate constant calculations. *Biophysical journal* **86**, 2017-2029 (2004).

237 Hardt, S. L. The Diffusion Transit-Time - a Simple Derivation. *B Math Biol* **43**, 89-99, doi:Doi 10.1007/Bf02460942 (1981).

238 Michaud-Agrawal, N., Denning, E. J., Woolf, T. B. & Beckstein, O. Software News and Updates MDAnalysis: A Toolkit for the Analysis of Molecular Dynamics Simulations. *J Comput Chem* **32**, 2319-2327, doi:Doi 10.1002/Jcc.21787 (2011).

239 Duan, Y., Wu, C., Chowdhury, S., Lee, M. C., Xiong, G. M., Zhang, W., Yang, R., Cieplak, P., Luo, R., Lee, T., Caldwell, J., Wang, J. M. & Kollman, P. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J Comput Chem* **24**, 1999-2012, doi:Doi 10.1002/Jcc.10349 (2003).

240 Branco, R. J. F., Fernandes, P. A. & Ramos, M. J. Molecular dynamics simulations of the enzyme Cu, Zn superoxide dismutase. *J Phys Chem B* **110**, 16754-16762, doi:Doi 10.1021/Jp0568551 (2006).

241 Roe, D. R. & Cheatham, T. E. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J Chem Theory Comput* **9**, 3084-3095, doi:Doi 10.1021/Ct400341p (2013).

242 Baker, N. A., Sept, D., Joseph, S., Holst, M. J. & McCammon, J. A. Electrostatics of nanosystems: Application to microtubules and the ribosome. *P Natl Acad Sci USA* **98**, 10037-10041, doi:Doi 10.1073/Pnas.181342398 (2001).

243 Hough, M. A. & Hasnain, S. S. Crystallographic structures of bovine copper-zinc superoxide dismutase reveal asymmetry in two subunits: Functionally important three and five coordinate copper sites captured in the same crystal. *J Mol Biol* **287**, 579-592, doi:Doi 10.1006/Jmbi.1999.2610 (1999).

244 Pearlman, D. A., Case, D. A., Caldwell, J. W., Ross, W. S., Cheatham, T. E., Debolt, S., Ferguson, D., Seibel, G. & Kollman, P. Amber, a Package of Computer-Programs for Applying Molecular Mechanics, Normal-Mode Analysis, Molecular-Dynamics and Free-Energy Calculations to Simulate the Structural and Energetic Properties of Molecules. *Comput Phys Commun* **91**, 1-41, doi:Doi 10.1016/0010-4655(95)00041-D (1995).

245 Votapka, L. W., Czapla, L., Zhenirovskyy, M. & Amaro, R. E. DelEnsembleElec: Computing Ensemble-Averaged Electrostatics Using DelPhi. *Communications in Computational Physics* **13**, 256-268, doi:Doi 10.4208/Cicp.170711.111111s (2013).

246 Luo, R., Wang, J. M. & Kollman, P. A. Development of a more accurate amber united-atom force field for protein folding and large-scale biomolecular simulations. *Abstr Pap Am Chem S* **224**, U470-U471 (2002).

247 Bas, D. C., Rogers, D. M. & Jensen, J. H. Very fast prediction and rationalization of pK(a) values for protein-ligand complexes. *Proteins* **73**, 765-783, doi:Doi 10.1002/Prot.22102 (2008).

248 Cudd, A. & Fridovich, I. Electrostatic Interactions in the Reaction-Mechanism of Bovine Erythrocyte Superoxide-Dismutase. *J Biol Chem* **257**, 1443-1447 (1982).

249 Divisek, J. & Kastening, B. Electrochemical Generation and Reactivity of Superoxide Ion in Aqueous-Solutions. *J Electroanal Chem* **65**, 603-621, doi:Doi 10.1016/S0022-0728(75)80153-7 (1975).

250 Spyracopoulos, L., Li, M. X., Sia, S. K., Gagne, S. M., Chandra, M., Solaro, R. J. & Sykes, B. D. Calcium-induced structural transition in the regulatory domain of human cardiac troponin C. *Biochemistry-Us* **36**, 12138-12146, doi:Doi 10.1021/Bi971223d (1997).

251 Tikunova, S. B. & Davis, J. P. Designing calcium-sensitizing mutations in the regulatory domain of cardiac troponin C. *J Biol Chem* **279**, 35341-35352, doi:Doi 10.1074/Jbc.M405413200 (2004).

252 Li, Y. H. & Gregory, S. Diffusion of Ions in Sea-Water and in Deep-Sea Sediments. *Geochim Cosmochim Ac* **38**, 703-714 (1974).

253 Lamoureux, G., MacKerell, A. D. & Roux, B. A simple polarizable model of water based on classical Drude oscillators. *J Chem Phys* **119**, 5185-5197, doi:Doi 10.1063/1.1598191 (2003).

254 Tange, O. GNU Parallel - The Command-Line Power Tool. *The USENIX Magazine* **February**, 42-47 (2011).

255 Argese, E., Orsega, E. F., Decarli, B., Scarpa, M. & Rigo, A. Application of Short Controlled Drop-Time Polarography to the Study of Superoxide Ion Dismutation in Aqueous-Solutions - Determination of the Activity of Superoxide Dismutases. *Bioelectroch Bioener* **13**, 385-392, doi:Doi 10.1016/0302-4598(84)87039-7 (1984).

256 Hazard, A. L., Kohout, S. C., Stricker, N. L., Putkey, J. A. & Falke, J. J. The kinetic cycle of cardiac troponin C: Calcium binding and dissociation at site II trigger slow conformational rearrangements. *Protein Sci* **7**, 2451-2459 (1998).

257 Ogawa, Y. Calcium-Binding to Troponin C and Troponin - Effects of Mg-2+, Ionic-Strength and Ph. *J Biochem-Tokyo* **97**, 1011-1023 (1985).

258 Berkowitz, M., Karim, O. A., Mccammon, J. A. & Rossky, P. J. Sodium-Chloride Ion-Pair Interaction in Water - Computer-Simulation. *Chem Phys Lett* **105**, 577-580, doi:Doi 10.1016/0009-2614(84)85660-2 (1984).

259 Karim, O. A. & Mccammon, J. A. Rate Constants for Ion-Pair Formation and Dissociation in Water. *Chem Phys Lett* **132**, 219-224, doi:Doi 10.1016/0009-2614(86)80111-7 (1986).

260 Guardia, E., Rey, R. & Padro, J. A. Potential of Mean Force by Constrained Molecular-Dynamics - a Sodium-Chloride Ion-Pair in Water. *Chem Phys* **155**, 187-195, doi:Doi 10.1016/0301-0104(91)87019-R (1991).

261 Mccord, J. M. & Fridovic.I. Superoxide Dismutase-an Enzymic Function for Erythrocuprein. *Fed Proc* **28**, 346-& (1969).

262 Wong, Y. T., Clark, T. W., Shen, J. & Mccammon, J. A. Molecular-Dynamics Simulation of Substrate-Enzyme Interactions in the Active-Site Channel of Superoxide-Dismutase. *Mol Simulat* **10**, 277-&, doi:Doi 10.1080/08927029308022169 (1993).

263 Li, M. X., Wang, X. & Sykes, B. D. Structural based insights into the role of troponin in cardiac muscle pathophysiology. *J Muscle Res Cell M* **25**, 559-579, doi:Doi 10.1007/S10974-004-5879-2 (2004).

264 Mccrackin, F. L., Guttman, C. M. & Akcasu, A. Z. Monte-Carlo Calculations of the Hydrodynamic Radii of Polymers in Theta and Good Solvents. *Macromolecules* **17**, 604-610, doi:Doi 10.1021/Ma00134a015 (1984).

265 Cowan, A. E., Moraru, II, Schaff, J. C., Slepchenko, B. M. & Loew, L. M. Spatial modeling of cell signaling networks. *Methods Cell Biol* **110**, 195-221, doi:10.1016/B978-0-12-388403-9.00008-4 (2012).

266 Bondarenko, V. E. A Compartmentalized Mathematical Model of the beta(1)-Adrenergic Signaling System in Mouse Ventricular Myocytes. *Plos One* **9**, doi:ARTN e89113 DOI 10.1371/journal.pone.0089113 (2014).

267 Wang, Y. F., Khan, M. & van den Berg, H. A. Interaction of fast and slow dynamics in endocrine control systems with an application to beta-cell dynamics. *Math Biosci* **235**, 8-18, doi:10.1016/j.mbs.2011.10.003 (2012).

268 Cong, S., Ma, X. T., Li, Y. X. & Wang, J. F. Structural basis for the mutation-induced dysfunction of human CYP2J2: a computational study. *J Chem Inf Model* **53**, 1350-1357, doi:10.1021/ci400003p (2013).

269 Kirchner, F., Schuetz, A., Boldt, L. H., Martens, K., Dittmar, G., Haverkamp, W., Thierfelder, L., Heinemann, U. & Gerull, B. Molecular insights into arrhythmogenic right ventricular cardiomyopathy caused by plakophilin-2 missense mutations. *Circ Cardiovasc Genet* **5**, 400-411, doi:10.1161/CIRCGENETICS.111.961854 (2012).

270 Cregut, D. & Serrano, L. Molecular dynamics as a tool to detect protein foldability. A mutant of domain B1 of protein G with non-native secondary structure propensities. *Protein Sci* **8**, 271-282 (1999).

271 Koukos, P. I. & Glykos, N. M. Folding Molecular Dynamics Simulations Accurately Predict the Effect of Mutations on the Stability and Structure of a Vammin-Derived Peptide. *J Phys Chem B* **118**, 10076-10084, doi:Doi 10.1021/Jp5046113 (2014).

272 Kozack, R. E. & Subramaniam, S. Brownian Dynamics Simulations of Molecular Recognition in an Antibody Antigen System. *Protein Sci* **2**, 915-926 (1993).

273 De Rienzo, F., Gabdoulline, R. R., Menziani, M. C., De Benedetti, P. G. & Wade, R. C. Electrostatic analysis and Brownian dynamics simulation of the association of plastocyanin and cytochrome F. *Biophys J* **81**, 3090-3104 (2001).

274 Boras, B. W., Kornev, A., Taylor, S. S. & McCulloch, A. D. Using Markov state models to develop a mechanistic understanding of protein kinase A regulatory subunit RIalpha activation in response to cAMP binding. *J Biol Chem* **289**, 30040-30051, doi:10.1074/jbc.M114.568907 (2014).

275 Campbell, S. G., Lionetti, F. V., Campbell, K. S. & McCulloch, A. D. Coupling of adjacent tropomyosins enhances cross-bridge-mediated cooperative activation in a markov model of the cardiac thin filament. *Biophys J* **98**, 2254-2264, doi:10.1016/j.bpj.2010.02.010 (2010).

276 Clancy, C. E. & Rudy, Y. Linking a genetic defect to its cellular phenotype in a cardiac arrhythmia. *Nature* **400**, 566-569, doi:10.1038/23034 (1999).

277 Zhou, H. X. & Bates, P. A. Modeling protein association mechanisms and kinetics. *Curr Opin Struc Biol* **23**, 887-893, doi:Doi 10.1016/J.Sbi.2013.06.014 (2013).

278 Bers, D. M. *Excitation-contraction coupling and cardiac contractile force*. 2nd edn, (Kluwer Academic Publishers, 2001).

279 Saucerman, J. J., Brunton, L. L., Michailova, A. P. & McCulloch, A. D. Modeling beta-adrenergic control of cardiac myocyte contractility in silico. *Journal of Biological Chemistry* **278**, 47997-48003, doi:DOI 10.1074/jbc.M308362200 (2003).

280 Tsai, C. J., Kumar, S., Ma, B. Y. & Nussinov, R. Folding funnels, binding funnels, and protein function. *Protein Sci* **8**, 1181-1190 (1999).

281 Boehr, D. D., Nussinov, R. & Wright, P. E. The role of dynamic conformational ensembles in biomolecular recognition (vol 5, pg 789, 2009). *Nat Chem Biol* **5**, 954-954, doi:Doi 10.1038/Nchembio1209-954d (2009).

282 Motlagh, H. N., Wrabl, J. O., Li, J. & Hilser, V. J. The ensemble nature of allostery. *Nature* **508**, 331-339, doi:Doi 10.1038/Nature13001 (2014).

283 Marsh, J. A., Teichmann, S. A. & Forman-Kay, J. D. Probing the diverse landscape of protein flexibility and binding. *Curr Opin Struc Biol* **22**, 643-650, doi:Doi 10.1016/J.Sbi.2012.08.008 (2012).

284 Teilum, K., Olsen, J. G. & Kragelund, B. B. Protein stability, flexibility and function. *Bba-Proteins Proteom* **1814**, 969-976, doi:Doi 10.1016/J.Bbapap.2010.11.005 (2011).

285 Henzler-Wildman, K. & Kern, D. Dynamic personalities of proteins. *Nature* **450**, 964-972, doi:Doi 10.1038/Nature06522 (2007).

286 Adcock, S. A. & McCammon, J. A. Molecular dynamics: survey of methods for simulating the activity of proteins. *Chemical reviews* **106**, 1589-1615, doi:10.1021/cr040426m (2006).

287 Karplus, M. & McCammon, J. A. Molecular dynamics simulations of biomolecules. *Nat Struct Biol* **9**, 646-652, doi:Doi 10.1038/Nsb0902-646 (2002).

288 Ponder, J. W. & Case, D. A. Force fields for protein simulations. *Adv Protein Chem* **66**, 27-+ (2003).

289 Wang, W., Donini, O., Reyes, C. M. & Kollman, P. A. Biomolecular simulations: Recent developments in force fields, simulations of enzyme catalysis, protein-ligand, protein-protein, and protein-nucleic acid noncovalent interactions. *Annu Rev Bioph Biom* **30**, 211-243, doi:Doi 10.1146/Annurev.Biophys.30.1.211 (2001).

290 Brooks, B. R., Brooks, C. L., Mackerell, A. D., Nilsson, L., Petrella, R. J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S., Caflisch, A., Caves, L., Cui, Q., Dinner, A. R., Feig, M., Fischer, S., Gao, J., Hodoscek, M., Im, W., Kuczera, K., Lazaridis, T., Ma, J., Ovchinnikov, V., Paci, E., Pastor, R. W., Post, C. B., Pu, J. Z., Schaefer, M., Tidor, B., Venable, R. M., Woodcock, H. L., Wu, X., Yang, W., York, D. M. & Karplus, M. CHARMM: The Biomolecular Simulation Program. *J Comput Chem* **30**, 1545-1614, doi:Doi 10.1002/Jcc.21287 (2009).

291 Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W. & Kollman, P. A. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules (vol 117, pg 5179, 1995). *J Am Chem Soc* **118**, 2309-2309, doi:Doi 10.1021/Ja955032e (1995).

292 Kaminski, G. A., Friesner, R. A., Tirado-Rives, J. & Jorgensen, W. L. Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *J Phys Chem B* **105**, 6474-6487, doi:Doi 10.1021/Jp003919d (2001).

293 Oostenbrink, C., Villa, A., Mark, A. E. & Van Gunsteren, W. F. A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets 53A5 and 53A6. *J Comput Chem* **25**, 1656-1676, doi:Doi 10.1002/Jcc.20090 (2004).

294 Vitalini, F., Mey, A. S. J. S., Noe, F. & Keller, B. G. Dynamic properties of force fields. *J Chem Phys* **142**, doi:Artn 084101 Doi 10.1063/1.4909549 (2015).

295 Guvench, O. & MacKerell, A. D., Jr. Comparison of protein force fields for molecular dynamics simulations. *Methods in molecular biology* **443**, 63-88, doi:10.1007/978-1-59745-177-2_4 (2008).

296 Shaw, D. E., Maragakis, P., Lindorff-Larsen, K., Piana, S., Dror, R. O., Eastwood, M. P., Bank, J. A., Jumper, J. M., Salmon, J. K., Shan, Y. B. & Wriggers, W. Atomic-Level Characterization of the Structural Dynamics of Proteins. *Science* **330**, 341-346, doi:Doi 10.1126/Science.1187409 (2010).

297 Gotz, A. W., Williamson, M. J., Xu, D., Poole, D., Le Grand, S. & Walker, R. C. Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 1. Generalized Born. *J Chem Theory Comput* **8**, 1542-1555, doi:Doi 10.1021/Ct200909j (2012).

298 Pierce, L. C., Salomon-Ferrer, R., Augusto, F. d. O. C., McCammon, J. A. & Walker, R. C. Routine Access to Millisecond Time Scale Events with Accelerated Molecular Dynamics. *Journal of chemical theory and computation* **8**, 2997-3002, doi:10.1021/ct300284c (2012).

299 Salomon-Ferrer, R., Gotz, A. W., Poole, D., Le Grand, S. & Walker, R. C. Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. *J Chem Theory Comput* **9**, 3878-3888, doi:Doi 10.1021/Ct400314y (2013).

300 Kohlhoff, K. J., Shukla, D., Lawrenz, M., Bowman, G. R., Konerding, D. E., Belov, D., Altman, R. B. & Pande, V. S. Cloud-based simulations on Google Exacycle reveal ligand modulation of GPCR activation pathways. *Nat Chem* **6**, 15-21, doi:Doi 10.1038/Nchem.1821 (2014).

301 Pande, V. S., Baker, I., Chapman, J., Elmer, S. P., Khaliq, S., Larson, S. M., Rhee, Y. M., Shirts, M. R., Snow, C. D., Sorin, E. J. & Zagrovic, B. Atomistic protein folding simulations on the submillisecond time scale using worldwide distributed computing. *Biopolymers* **68**, 91-109, doi:Doi 10.1002/Bip.10219 (2003).

302 Chodera, J. D. & Noe, F. Markov state models of biomolecular conformational dynamics. *Curr Opin Struc Biol* **25**, 135-144, doi:Doi 10.1016/J.Sbi.2014.04.002 (2014).

303 Prinz, J. H., Wu, H., Sarich, M., Keller, B., Senne, M., Held, M., Chodera, J. D., Schutte, C. & Noe, F. Markov models of molecular kinetics: Generation and validation. *J Chem Phys* **134**, doi:Artn 174105 Doi 10.1063/1.3565032 (2011).

304 Prinz, J. H., Keller, B. & Noe, F. Probing molecular kinetics with Markov models: metastable states, transition pathways and spectroscopic observables. *Phys Chem Chem Phys* **13**, 16912-16927, doi:Doi 10.1039/C1cp21258c (2011).

305 Sjoberg, T. J., Kornev, A. P. & Taylor, S. S. Dissecting the cAMP-inducible allosteric switch in protein kinase A RIalpha. *Protein Sci* **19**, 1213-1221, doi:10.1002/pro.400 (2010).

306 Malmstrom, R. D., Kornev, A. P., Taylor, S. S. & Amaro, R. E. Allostery through the Computational Microscope: cAMP Activation of a Canonical Signaling Domain. *Nature Communications* (2015).

307 Malmstrom, R. D., Lee, C. T., Van Wart, A. T. & Amaro, R. E. Application of Molecular-Dynamics Based Markov State Models to Functional Proteins. *J Chem Theory Comput* **10**, 2648-2657, doi:Doi 10.1021/Ct5002363 (2014).

308 Vanden-Eijnden, E. & Tal, F. A. Transition state theory: Variational formulation, dynamical corrections, and error estimates. *J Chem Phys* **123**, doi:Artn 184103 Doi 10.1063/1.2102898 (2005).

309 Bernardi, R. C., Melo, M. C. R. & Schulten, K. Enhanced sampling techniques in molecular dynamics simulations of biological systems. *Bba-Gen Subjects* **1850**, 872-877, doi:10.1016/j.bbagen.2014.10.019 (2015).

310 McGuffee, S. R. & Elcock, A. H. Diffusion, crowding & protein stability in a dynamic molecular model of the bacterial cytoplasm. *PLoS computational biology* **6**, e1000694, doi:10.1371/journal.pcbi.1000694 (2010).

311 Schoneberg, J. & Noe, F. ReaDDy - A Software for Particle-Based Reaction-Diffusion Dynamics in Crowded Cellular Environments. *Plos One* **8**, doi:ARTN e74261 DOI 10.1371/journal.pone.0074261 (2013).

312 Geyer, T. Many-particle Brownian and Langevin Dynamics Simulations with the Brownmove package. *Bmc Biophys* **4**, doi:Artn 7 Doi 10.1186/2046-1682-4-7 (2011).

313 Dlugosz, M., Zielinski, P. & Trylska, J. Software News and Updates Brownian Dynamics Simulations on CPU and GPU with BD_BOX. *J Comput Chem* **32**, 2734-2744, doi:Doi 10.1002/Jcc.21847 (2011).

314 Dickson, C. J., Madej, B. D., Skjevik, A. A., Betz, R. M., Teigen, K., Gould, I. R. & Walker, R. C. Lipid14: The Amber Lipid Force Field. *J Chem Theory Comput* **10**, 865-879, doi:Doi 10.1021/Ct4010307 (2014).

315 Klauda, J. B., Venable, R. M., Freites, J. A., O'Connor, J. W., Tobias, D. J., Mondragon-Ramirez, C., Vorobyov, I., MacKerell, A. D. & Pastor, R. W. Update of the CHARMM All-Atom Additive Force Field for Lipids: Validation on Six Lipid Types. *J Phys Chem B* **114**, 7830-7843, doi:Doi 10.1021/Jp101759q (2010).

316 Holst, M. Adaptive numerical treatment of elliptic systems on manifolds. *Adv Comput Math* **15**, 139-191, doi:Doi 10.1023/A:1014246117321 (2001).

317 Gu, J. & Bourne, P. E. *Structural bioinformatics*. 2nd edn, (Wiley-Blackwell, 2009).

318 Leach, A. R. *Molecular modelling : principles and applications*. 2nd edn, (Prentice Hall, 2001).

319 Allen, M. P. & Tildesley, D. J. *Computer simulation of liquids*. (Clarendon Press; Oxford University Press, 1987).

320 Elcock, A. H., Sept, D. & McCammon, J. A. Computer simulation of protein-protein interactions. *J Phys Chem B* **105**, 1504-1518, doi:Doi 10.1021/Jp003602d (2001).

321 Gabdoulline, R. R. & Wade, R. C. Effective charges for macromolecules in solvent. *J Phys Chem-Us* **100**, 3868-3878 (1996).

322 Gabdoulline, R. R. & Wade, R. C. Biomolecular diffusional association. *Current opinion in structural biology* **12**, 204-213 (2002).

323 Su, Y., Dostmann, W., Herberg, F., Durick, K., Xuong, N., Ten Eyck, L., Taylor, S. & Varughese, K. Regulatory subunit of protein kinase A: structure of deletion mutant with cAMP binding domains. *Science* **269**, 807-813 (1995).

324 Kim, C., Cheng, C. Y., Saldanha, S. A. & Taylor, S. S. PKA-I holoenzyme structure reveals a mechanism for cAMP-dependent activation. *Cell* **130**, 1032-1043, doi:Doi 10.1016/J.Cell.2007.07.018 (2007).

325 Votapka, L. W. & Amaro, R. E. Multiscale estimation of binding kinetics using molecular dynamics, brownian dynamics, and milestoning. *J Biomol Struct Dyn* **33**, 26-27, doi:10.1080/07391102.2015.1032587 (2015).

326 Mugnai, M. L. & Elber, R. Extracting the diffusion tensor from molecular dynamics simulation with Milestoning. *J Chem Phys* **142**, 014105, doi:10.1063/1.4904882 (2015).

327 Edeson, R. O., Yeo, G. F., Milne, R. K. & Madsen, B. W. Graphs, random sums, and sojourn time distributions, with application to ion-channel modeling. *Math Biosci* **102**, 75-104 (1990).

328 Lampert, A. & Korngreen, A. Markov modeling of ion channels: implications for understanding disease. *Prog Mol Biol Transl Sci* **123**, 1-21, doi:10.1016/B978-0-12-397897-4.00009-7 (2014).

329 Gurkiewicz, M., Korngreen, A., Waxman, S. G. & Lampert, A. Kinetic modeling of Nav1.7 provides insight into erythromelalgia-associated F1449V mutation. *J Neurophysiol* **105**, 1546-1557, doi:10.1152/jn.00703.2010 (2011).

330 Giugliano, M. Synthesis of generalized algorithms for the fast computation of synaptic conductances with Markov kinetic models in large network simulations. *Neural Comput* **12**, 903-931 (2000).

331 Qin, F., Auerbach, A. & Sachs, F. Estimating single-channel kinetic parameters from idealized patch-clamp data containing missed events. *Biophys J* **70**, 264-280 (1996).

332 Rudy, Y. & Silva, J. R. Computational biology in the study of cardiac ion channels and cell electrophysiology. *Q Rev Biophys* **39**, 57-116, doi:10.1017/S0033583506004227 (2006).

333 Clancy, C. E., Zhu, Z. I. & Rudy, Y. Pharmacogenetics and anti-arrhythmic drug therapy: a theoretical investigation (vol 292, pg 66, 2007). *Am J Physiol-Heart C* **292**, H1641-H1642, doi:DOI 10.1152/ajpheart.zh4-7410-corr.2007 (2007).

334 Yang, J. H. & Saucerman, J. J. Phospholemman is a negative feed-forward regulator of Ca2+ in beta-adrenergic signaling, accelerating beta-adrenergic inotropy. *J Mol Cell Cardiol* **52**, 1048-1055, doi:DOI 10.1016/j.yjmcc.2011.12.015 (2012).

335 Herberg, F. W., Taylor, S. S. & Dostmann, W. R. Active site mutations define the pathway for the cooperative activation of cAMP-dependent protein kinase. *Biochemistry* **35**, 2934-2942, doi:10.1021/bi951647c (1996).

336 Neitzel, J. J., Dostmann, W. R. & Taylor, S. S. Role of MgATP in the activation and reassociation of cAMP-dependent protein kinase I: consequences of replacing the essential arginine in cAMP binding site A. *Biochemistry* **30**, 733-739 (1991).

337 Khavrutskii, I. V., Grant, B., Taylor, S. S. & McCammon, J. A. A transition path ensemble study reveals a linchpin role for Mg(2+) during rate-limiting ADP release from protein kinase A. *Biochemistry* **48**, 11532-11545, doi:10.1021/bi901475g (2009).

338 Anderson, K. B. & Conder, J. A. Discussion of Multicyclic Hubbert Modeling as a Method for Forecasting Future Petroleum Production. *Energ Fuel* **25**, 1578-1584, doi:Doi 10.1021/Ef1012648 (2011).

339 Moraru, I. I., Schaff, J. C., Slepchenko, B. M., Blinov, M. L., Morgan, F., Lakshminarayana, A., Gao, F., Li, Y. & Loew, L. M. Virtual Cell modelling

and simulation software environment. *Iet Syst Biol* **2**, 352-362, doi:Doi 10.1049/Iet-Syb:20080102 (2008).

340 Marsden, A. L., Feinstein, J. A. & Taylor, C. A. A computational framework for derivative-free optimization of cardiovascular geometries. *Comput Method Appl M* **197**, 1890-1905, doi:Doi 10.1016/J.Cma.2007.12.009 (2008).

341 Christensen, A. E., Selheim, F., de Rooij, J., Dremier, S., Schwede, F., Dao, K. K., Martinez, A., Maenhaut, C., Bos, J. L., Genieser, H. G. & Doskeland, S. O. cAMP analog mapping of Epac1 and cAMP kinase - Discriminating analogs demonstrate that Epac and cAMP kinase act synergistically to promote PC-12 cell neurite extension. *Journal of Biological Chemistry* **278**, 35394-35402, doi:Doi 10.1074/Jbc.M302179200 (2003).

342 Scott, J. D. & Santana, L. F. A-Kinase Anchoring Proteins Getting to the Heart of the Matter. *Circulation* **121**, 1264-1271, doi:Doi 10.1161/Circulationaha.109.896357 (2010).

343 Kerr, R. A., Bartol, T. M., Kaminsky, B., Dittrich, M., Chang, J. C. J., Baden, S. B., Sejnowski, T. J. & Stiles, J. R. Fast Monte Carlo Simulation Methods for Biological Reaction-Diffusion Systems in Solution and on Surfaces. *Siam J Sci Comput* **30**, 3126-3149, doi:10.1137/070692017 (2008).

344 Hake, J., Kekenes-Huskey, P. M. & McCulloch, A. D. Computational modeling of subcellular transport and signaling. *Curr Opin Struc Biol* **25**, 92-97, doi:10.1016/j.sbi.2014.01.006 (2014).

345 Horvath, A., Bertherat, J., Groussin, L., Guillaud-Bataille, M., Tsang, K., Cazabat, L., Libe, R., Remmers, E., Rene-Corail, F., Faucz, F. R., Clauser, E., Calender, A., Bertagna, X., Carney, J. A. & Stratakis, C. A. Mutations and Polymorphisms in the Gene Encoding Regulatory Subunit Type 1-Alpha of Protein Kinase A (PRKAR1A): An Update. *Hum Mutat* **31**, 369-379, doi:Doi 10.1002/Humu.21178 (2010).

346 Cruickshank, J. M. beta blockers in hypertension. *Lancet* **376**, 415-415, doi:Doi 10.1016/S0140-6736(10)61217-2 (2010).