UNIVERSITY OF CALIFORNIA,
IRVINE

Enhanced Sampling Methods for the Computation of Conformational Kinetics in Macromolecules

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Chemistry

by

Gianmarc Grazioli

Dissertation Committee:
Professor Ioan Andricioaei, Chair
Professor Craig C. Martens
Professor Douglas J. Tobias

2016

DEDICATION

To my mom, my grandparents Emilio and Maria Grazioli, Lucien the fuzzy wonder, and to the love of my life, Kristina.

TABLE OF CONTENTS

LIST OF FIGURES

Page

ACKNOWLEDGMENTS

I would first like to thank my wonderful advisor, Prof. Ioan Andricioaei. Ioan, my time as a member of your research group has been nothing short of a transformative experience. When I began the PhD program in Theoretical Chemistry at UCI, I had zero programming experience, zero research experience in physical chemistry, biophysics, or chemical physics, and my math was very rusty, and you still gave me the opportunity to work on some very challenging research projects that required me to develop proficiency in all of these areas! Thank you for taking a chance on me, and for the best job I've ever had. I tried many career paths that didn't quite suit me before this one, but thanks to you, I now love what I do, and I look forward to continuing to improve for the rest of my career.

I would like to thank the members of my advancement committee (Profs. Nizkorodov, Dennin, Tobias, Mandelshtam, and Mukamel) for truly testing my mettle as a scientist, it was an excellent experience for me. I would also like to additionally thank committee members Prof. Vladimir Mandelshtam for his excellent course, Math Methods in Chemistry, Prof. Michael Dennin for his excellent Computational Methods course, and Prof. Shaul Mukamel for the opportunity to learn about Molecular Spectroscopy straight from the source.

I would like to thank Prof. Doug Tobias for chairing my advancement committee, being on my dissertation committee, and providing great advice throughout my time at UCI.

I would like to thank Prof. Craig Martens for two great classes, being on my dissertation committee, plenty of great conversations, and advice. Even while discussing non-scientific things with you, I learned a lot from your science related jokes and metaphors!

I would like to thank Prof. Rachel Martin for the great career advice and conversations, one of my favorite TAing assignments, and of course for bringing me in on the "Chemists Know" project!

I would like to thank Prof. Carter T. Butts for co-advising me on the Data Science Initiative project featured in chapter 5 of this dissertation, and some great conversations, it was a real pleasure working with you!

To all of my esteemed colleagues of the LTL, thank you for the stimulating conversation, camaraderie, and advice. Our meetings will truly be missed.

Thank you to Dr. De Gallow and everyone at the UCI Center for Engaged Instruction for an enriching experience as a Pedagogical Fellow. I will certainly draw on this experience in any future teaching roles.

Thank you to everyone in my cohort who has helped me along the way, especially my friends and problem set collaborators: Eric Wong, Paolo Reyes, Mya Le-Thai, Jerry Guo, and Jung-Gun Song. First year would have been brutal without you!

I would like to thank my fellow Andricioaei Research Group members whom I have had

the pleasure of working with: Jason Deckman, Ahmet Mentes, Nick Preketes, Mahua Roy, Daun Jeong, Maryna Taranova, Emel Ficici, Gavin Bascom, Anupam Chatterjee, and Jim McSally. Each one of you has helped me along the way in one way or another, thank you!

I would like to thank everyone at Villanova University who contributed to my education, especially my M.S. thesis advisor, Prof. Barry Selinsky, Prof. Andy Woldar, for your support throughout my PhD program, long after I was officially your student, and Prof. John Ullrich, whose undergraduate organic chemistry course reignited my love of science, and inspired me to pursue a career in chemistry.

Thank you to Mario, Carmen, and Krystal for your continued love, patience, and support throughout Kris's and my journeys, both educational and geographical, I love you all!

I would like to thank my mother, Emilia, for teaching me the value of education from an early age, sacrificing so that I could receive the best education, and giving me every opportunity to develop my creative abilities since the time of my earliest memories. I credit you for my best asset as a researcher, my creativity. Thanks Mom, love you!

Finally, I would like to thank Kristina, my amazing wife! It's impossible to find words that describe my gratitude for your love and support throughout this process. When I had the crazy idea to pursue a career in rock music, you encouraged me. When I quickly got fed up with trying to make a living that way, it was you who said, "Well, weren't you pretty good at science?" And when I got the crazy idea for us to leave two perfectly good jobs and move across the country so that the two of us could pursue PhDs in two different time zones, you actually went along with it! Now here we are, and I couldn't be happier with the life we've made. Thank you for being you and for bringing out the best in me, I love you!

I gratefully acknowledge funding received during my graduate study from the UCI Pedagogical Fellows program, the UCI Data Science Initiative, a University of California Regents Dissertation Fellowship, and teaching assistantships through the UCI Chemistry Department.

Appendix A is a reprint of "Rate turnover in mechano-catalytic coupling: A model and its microscopic origin" published by the American Institute of Physics in 2015 with permission of authors Ioan Andricioaei and Mahua Roy, who directly supervised or carried out the work along with myself.

CURRICULUM VITAE

Gianmarc Grazioli

EDUCATION

Doctor of Philosophy in Theoretical Physical Chemistry — **2016**
University of California, Irvine — *Irvine, CA*

Master of Science in Chemistry — **2007**
Villanova University — *Villanova, PA*

Bachelor of Science in Comprehensive Science — **2005**
Villanova University — *Villanova, PA*

Bachelor of Arts in Liberal Arts — **2002**
Villanova University — *Villanova, PA*

RESEARCH EXPERIENCE

Graduate Research Assistant — **2011–2016**
University of California, Irvine — *Irvine, California*

Graduate Research Assistant — **2005–2007**
Villanova University — *Villanova, PA*

HONORS AND AWARDS

Journal of Chemical Physics 2015 Editors Choice Award for publication entitled Rate turnover in mechano-catalytic coupling: A model and its microscopic origin — **2015**

University of California, Irvine Data Science Initiative Summer Fellowship — **2015**

Acceptance with Full Scholarship to the Telluride School on Theoretical Chemistry — **2015**

University of California Regents Dissertation Fellowship — **2015**

University of California, Irvine Pedagogical Fellowship — **2014**

INDUSTRY EXPERIENCE

Pharmaceutical Testing Lab Data Analyst **2007–2010**
Merck and Co. *West Point, PA*

Clinical Quality Technician **2003–2004**
Cardinal Health *Philadelphia, PA*

TEACHING EXPERIENCE

Teaching Assistant **2011–2016**
University of California, Irvine *Irvine, CA*

Part Time Lecturer **2010–2011**
University of Colorado, Boulder *Boulder, CO*

Teaching Assistant **2005–2007**
Villanova University *Villanova, PA*

High School Physics Teacher **2002–2003**
Bethlehem Catholic High School *Bethlehem, PA*

Mathematics Teaching Assistant, STOVS Program **Summer 2011**
Villanova University *Villanova, PA*

Adjunct Music Professor (Guitar) **2002–2009**
Villanova University *Villanova, PA*

Guitar Instructor **2003**
Music Learning Center *Rosemont, PA*

REFEREED JOURNAL PUBLICATIONS

Rate turnover in mechano-catalytical coupling: A model and its microscopic origin **2015**
Journal of Chemical Physics

Advancements in Milestoning I: Accelerated Milestoning via Wind Assisted Re- weighted Milestoning (WARM) **2016**
In preparation for Journal of Chemical Physics

Advancements in Milestoning II: Calculating Autocorrelation from Milestoning Data Using Stochastic Path Integrals in Milestone Space **2016**
In preparation for Journal of Chemical Physics

REFEREED CONFERENCE PUBLICATIONS

Advancements in Milestoning: 1) Computational speedup by re-weighting artificially accelerated trajectories, and 2) Venturing into the non-equilibrium with coarse-grained random walks in Milestone — **March 2016**
American Chemical Society National Conference

Calculating Watson-Crick to Hoogsteen Transition Kinetics in DNA with Langevin Dynamics and Fokker-Planck Diffusion in Reduced Configuration Space — **March 2016**
Biophysical Society Annual Meeting

A Smoluchowski Equation for Force-Modulated Chemistry in Single Molecule Pulling Experiments. — **Jan 2015**
Biophysical Society Annual Meeting

Structural Requirements for Time-Dependent and Time- Independent Inhibition of Prostaglandin Synthase I (COX I) — **Jan 2006**
American Chemical Society National Conference

ABSTRACT OF THE DISSERTATION

Enhanced Sampling Methods for the Computation of Conformational Kinetics in Macromolecules

By

Gianmarc Grazioli

Doctor of Philosophy in Chemistry

University of California, Irvine, 2016

Professor Ioan Andricioaei, Chair

Calculating the kinetics of conformational changes in macromolecules, such as proteins and nucleic acids, is still very much an open problem in theoretical chemistry and computational biophysics. If it were feasible to run large sets of molecular dynamics trajectories that begin in one configuration and terminate when reaching another configuration of interest, calculating kinetics from molecular dynamics simulations would be simple, but in practice, configuration spaces encompassing all possible configurations for even the simplest of macromolecules are far too vast for such a brute force approach. In fact, many problems related to searches of configuration spaces, such as protein structure prediction, are considered to be NP-hard. Two approaches to addressing this problem are to either develop methods for enhanced sampling of trajectories that confine the search to productive trajectories without loss of temporal information, or coarse-grained methodologies that recast the problem in reduced spaces that can be exhaustively searched. This thesis will begin with a description of work carried out in the vein of the second approach, where a Smoluchowski diffusion equation model was developed that accurately reproduces the rate vs. force relationship observed in the mechano-catalytic disulphide bond cleavage observed in thioredoxin-catalyzed reduction of disulphide bonds. Next, three different novel enhanced sampling methods developed in the vein of the first approach will be described, which can be employed either separately or in

conjunction with each other to autonomously define a set of energetically relevant subspaces in configuration space, accelerate trajectories between the interfaces dividing the subspaces while preserving the distribution of unassisted transition times between subspaces, and approximate time correlation functions from the kinetic data collected from the transitions between interfaces.

Chapter 1

Introduction

Ours is a truly exciting era in the history of science. We have the privilege of witnessing powerful digital implementations of ingenious numerical methods, some even predating the invention of the transistor, employed toward solving problems with staggering large degrees of complexity. This explosion in computational power has allowed for unprecedented computational exploration into the underlying physics behind the physical properties of important macromolecules such as proteins and nucleic acids. There are three main categories of applications of molecular dynamics to the study of macromolecules: configuration space sampling, obtaining equilibrium descriptions of systems, such as the calculation of thermodynamic properties, and the study of dynamical properties, which require not only adequate sampling in configuration space for accurate Boltzmann statistics, but also adequate sampling of the relevant time scales between configurations [54]. The research described in this work belongs in the third category. Two different types of approaches to addressing the challenge of calculating dynamical properties of macromolecules were explored. In one vein, a coarse-grained methodology was developed, where the mechano-catalytic property of thioredoxin in single molecule pulling experiments [4] was modeled as diffusion in a reduced space using a Smoluchowski formalism. The approach taken was similar in spirit to that of Liu and

Ou-Yang's model for catch slip bonds [69]. In the other vein, three different novel enhanced sampling methods were developed, which can be employed either separately or together to 1) autonomously define a set of energetically relevant subspaces in configuration space, 2) accelerate trajectories between the interfaces dividing the subspaces while preserving the distribution of unassisted transition times between subspaces, and 3) approximate time correlation functions from the kinetic data collected from the transitions between interfaces. Although these three methods are not exclusive to the Milestoning method, introduced by Faradjian and Elber in 2004 [33], the Milestoning method was chosen as the theoretical foundation and was also utilized in the accompanying proof of concept numerical demonstrations. Since detailed background information specific to each of the four distinct research projects is provided within each respective chapter, this chapter will be devoted to reviewing the more fundamental theoretical underpinnings in the field of non-equilibrium statistical mechanics which provide the unifying foundation for the collective work. Although multiple texts are available on the subject, this description will draw most heavily from *Nonequilibrium Statistical Mechanics* by Robert Zwanzig [132] and *Chemical Dynamics in Condensed Phases* by Abraham Nitzan [75].

1.1 Langevin Dynamics

The seed from which Langevin dynamics grew was planted in 1827, when the botanist Robert Brown first observed pollen grains in water moving along random trajectories under the microscope. With further contributions from Albert Einstein, Jean Perrin, and others, the theory of Brownian motion was developed, which provided a mathematical formalism for using statistics to approximate the complex interactions between a system of interest, like the pollen grain, and a "bath" comprised of numerous objects, in this case water molecules.

Due to the fact that this thesis is rooted in classical non-equilibrium statistical mechanics,

we begin our discussion of Langevin dynamics with Newton's famous equation for obtaining the total force on a classical object as a function of its trajectory in some configuration space x:

$$m\frac{\partial^2 x}{\partial t^2} = F_{Total}(t) \tag{1.1}$$

Now if our system is in the presence of a conservative field, i.e. the non-dissipative dynamics of our system are defined in such a way that the work needed to move it from one point in a configuration space to another is independent of the path taken, the force can be expressed as the negative gradient of the potential dictating the dynamics in our configuration space x:

$$m\frac{\partial^2 x}{\partial t^2} = -\nabla U(x) \tag{1.2}$$

An example of such a system would be a harmonic oscillator. Now if we want to consider a dissipative system, for example a damped harmonic oscillator, we need to include some sort of dissipative term. Mathematical models of dissipation, whether due to the viscosity of a liquid, or air resistance, or any other dissipative force, can quickly grow in complexity depending on the level of accuracy desired. The general trend, however, is that the faster an object moves through a dissipative medium, the stronger the dissipative force will resist. This trend is commonly observed in the automotive industry, where vehicles designed to travel at top speeds of maybe 50 miles per hour, like mail delivery trucks, would exhibit minimal performance gains if given a more aerodynamic shape, hence their boxy shape, while cars intended to travel at top speeds of 200 miles per hour must overcome tremendous

forces due to air resistance, and must be designed with great attention to aerodynamics. Given this overall trend in the magnitude of dissipative forces, let us define our dissipative resistive force term as being linearly dependent but opposite in sign to the velocity of our object by a coefficient of ζ:

$$m\frac{\partial^2 x}{\partial t^2} = -\nabla U(x) - \zeta\frac{\partial x}{\partial t} \tag{1.3}$$

At this point, we have a deterministic equation of motion, but the Langevin equation is a stochastic differential equation, where the total force must include a random term representing a force applied to the system by the bath. These random forces are colloquially referred to as kicks, and represented by the term $\delta\xi(t)$:

$$m\frac{\partial^2 x}{\partial t^2} = -\nabla U(x) - \zeta\frac{\partial x}{\partial t} + \delta\xi(t) \tag{1.4}$$

where a delta function is used to indicate that these random kicks occur as instantaneous pulses in time. Since the random force shows no directional bias, the mean value of the random force is zero. Additionally, since each random kick supplied by the random force is completely independent of all previous kicks, the time correlation function is a delta function multiplied by a constant factor dependent on the strength of the fluctuating force for any given lag time:

$$\langle\delta\xi(t)\rangle = 0, \quad \langle\delta\xi(t)\delta\xi(t')\rangle = 2B\delta(t-t') \tag{1.5}$$

Our Langevin equation is now complete in mathematical form, however; we must develop it further in order to ensure physical relevance. Given that both the random kick term and the dissipative force term are due to interactions between our system of interest and the bath, it seems apparent that there ought to be some relationship between the two terms. In fact, the two terms are related by what is known as the fluctuation dissipation theorem. In demonstrating this relationship, there is no need to consider force due to the potential $U(x)$, so let us consider diffusion of a free particle, for which $U(x)$ is a constant making its gradient term equal to zero:

$$m\frac{\partial^2 x}{\partial t^2} = -\zeta\frac{\partial x}{\partial t} + \delta\xi(t) \tag{1.6}$$

Since the demonstration will require us to invoke the equipartition theorem, which relates the temperature of a system to the mean squared velocity of the particles that make up that system, let us next recast our Langevin equation in terms of velocity instead of acceleration:

$$\frac{\partial v}{\partial t} = -\frac{\zeta}{m}v + \frac{1}{m}\delta\xi(t) \tag{1.7}$$

This first order linear inhomogeneous differential equation can then be solved using substitution to obtain the solution:

$$v(t) = e^{-\zeta t/m}v(0) + \int_0^t dt' e^{-\zeta(t-t')/m}\delta\xi(t')/m \tag{1.8}$$

By squaring both sides and integrating over time, the two cross terms go to zero as a result of the random kicks averaging to zero. The second squared term contains an integration over time of the square of the random force, which we know from equation 1.5 is simply a delta function multiplied by a factor of $2B$. After integration of the two squared terms, we obtain the expression for the mean squared velocity:

$$\langle v(t)^2 \rangle = e^{-2\zeta t/m} v(0)^2 + \frac{B}{\zeta m}(1 - e^{-2\zeta t/m}) \tag{1.9}$$

Notice that for long times, i.e. times long enough for equivalence between a time average and an ensemble average, known as ergodicity, to be reached, the exponential functions will approach zero, yielding a long time average of $\langle v(t)^2 \rangle = \frac{B}{\zeta m}$. Given that the equipartition theory states that the mean squared velocity of an ensemble of particles is equal to $\frac{k_B T}{m}$, the constant B representing the strength of our random force must be equal to $B = \zeta k_B T$, and we have established the fluctuation dissipation condition relating our random force to our dissipative force.

1.2 Fokker-Planck/Smoluchowski Diffusion Equations

The Langevin equation provides a description of individual trajectories moving through a configuration space subject to both deterministic and stochastic forces. It is possible to construct time dependent probability density functions, that are a function of both time and initial conditions by running numerous Langevin trajectories, for a system of interest and then constructing time dependent histograms using a stochastic path integral approach. A more direct approach to calculating these same time-dependent probability density functions is to solve the partial differential equations, called Fokker-Planck equations, that directly

describe the time evolution of these probability density functions. This is indeed a very powerful approach, but as usual, there are no free lunches, and so it should be noted that this approach can be very difficult to execute for systems with more than a few degrees of freedom. For this reason, applications of this approach are best when the degrees of freedom for the system can be projected onto some reduced space.

Given some normalized initial probability density function for a system in a configuration space, that spans all possible configurations, at time $t = 0$, $\rho(x, 0)$, we want to solve a differential equation that describes the time evolution of $\rho(x, t)$. Since the configuration space spans all possible configurations for our system, $\rho(x, t)$ must maintain its normalization condition for all time, i.e. the manifold describing probability density does not "leak." This implies a conservation law resembling that of an incompressible fluid, where if probability density is lost in one region, the exact same volume must be gained in some other region. In more mathematical terms, the change in probability density over time is exactly balanced by the divergence of flux:

$$\frac{\partial \rho}{\partial t} = -\frac{\partial}{\partial x} \cdot \left(\frac{\partial x}{\partial t} \rho \right) \tag{1.10}$$

It should be noted that if x is defined as a phase space, where the elements of the vector x include both all spatial coordinates of configuration spaces, as well as all momenta, and the appropriate substitutions of equivalent partial derivatives with respect to the Hamiltonian are made, this is equivalent to the Liouville equation. Using the conservation law of equation 1.10 as a constraint for the probability density, the Fokker-Planck equation describing an ensemble of diffusive systems can be derived from the Langevin equation used to describe individual trajectories. For diffusive systems like the ones described in this thesis, collision frequencies are very large compared to the frequencies describing the molecular motions of

the systems of interest, leading to dynamics where the effects of inertia are drowned out by the forces due to the potential and the random kicks from the random force [80]. In these cases, where the effects of inertia are negligible, it is a suitable approximation to set the mass equal to zero in our Langevin equation, and solve for velocity as a function of time, also known as Brownian Dynamics:

$$\frac{\partial x}{\partial t} = -\frac{1}{\zeta}\frac{\partial U}{\partial x} + \frac{1}{\zeta}\delta\xi(t) \tag{1.11}$$

Taking a cue from Abraham Nitzan [75], it is best to take on this derivation using an operator approach. In order to simplify the application of that approach, let us define a scaled velocity $\nu(t) \equiv \frac{1}{\zeta}v(t)$. Next this scaled velocity is substituted into the conservation law, equation 1.10:

$$\frac{\partial \rho}{\partial t} = \frac{\partial}{\partial x}\left(\frac{\partial U}{\partial x}\rho - \delta\xi(t)\rho\right) \tag{1.12}$$

If an operator $\hat{L}(t)$ is defined:

$$\hat{L}(t) \equiv \frac{\partial}{\partial x}\left(\frac{\partial U}{\partial x} - \delta\xi(t)\right) \tag{1.13}$$

where the time dependence of the operator is due to its stochastic term, then equation 1.12

can be recast in operator form as:

$$\frac{\partial \rho}{\partial t} = \hat{L}(t)\rho \tag{1.14}$$

Integrating equation 1.14 over a very small span of time from t to $t+\Delta t$ yields the expression:

$$\rho(x, t+\Delta t) = \rho(x,t) + \int_t^{t+\Delta t} dt_1 \hat{L}(t_1)\rho(x,t_1) \tag{1.15}$$

This result is an indication that time evolution of the probability density function $\rho(x,t)$ from time t to $t+\Delta t$ is equivalent to adding the integral over time of the operator $\hat{L}(t)$ operating on $\rho(x,t)$ to $\rho(x,t)$. Further, this implies that the time evolution shown in equation 1.15 is really just the first step of an expansion of iterative integration:

$$\rho(x,t+\Delta t) - \rho(x,t) = \left[\int_t^{t+\Delta t} dt_1 \hat{L}(t_1) + \int_t^{t+\Delta t} dt_1 \int_t^{t_1} dt_2 \hat{L}(t_1)\hat{L}(t_2) + ... \right] \rho(x,t) \tag{1.16}$$

The other important realization at this step is that since the Smoluchowski equation describes the time evolution for the probability density function in configuration space representing all possible manifestations of the overdamped Langevin equation, a stochastic differential equation, it is necessary that all operations be time averaged:

$$\rho(x,t+\Delta t) - \rho(x,t) = \left[\int_t^{t+\Delta t} dt_1 \langle \hat{L}(t_1) \rangle + \int_t^{t+\Delta t} dt_1 \int_t^{t_1} dt_2 \langle \hat{L}(t_1)\hat{L}(t_2) \rangle + ... \right] \rho(x,t) \tag{1.17}$$

Consider now the two terms in the operator $\hat{L}(t)$. The first term is deterministic, and the second term is stochastic. Since the first term of the expansion is first order in the operator, and given that the mean value of the random force $\xi(t)$ is zero, the effects of the stochastic term in the operator average out to zero for the first term of the expansion. This leaves only the time independent deterministic term, which can be factored out of the time integral to yield, what we'll refer to as the deterministic operator $\hat{A}$:

$$\hat{A} \equiv \frac{\partial}{\partial x}\frac{\partial U(x)}{\partial x}\Delta t \tag{1.18}$$

Moving on to the second order term of the expansion, we first note that the time independent deterministic term will get integrated over time twice, resulting in a factor of Δt^2. Given that Δt was defined as being infinitesimally small, any terms second order or higher in Δt will become vanishingly small. This leaves the effects of the stochastic term in the operator to dictate the value of the second term in the expansion. Since the stochastic term of the operator is the random force $\xi(t)$ multiplied by the differential operator $\frac{\partial}{\partial x}$, the second term in the expansion simplifies to:

$$\int_t^{t+\Delta t} dt_1 \int_t^{t_1} dt_2 \langle \xi(t_1)\xi(t_2)\rangle \frac{\partial^2}{\partial x^2} = \int_t^{t+\Delta t} dt_1 k_B T \frac{\partial^2}{\partial x^2} = k_B T \frac{\partial^2}{\partial x^2}\Delta t \tag{1.19}$$

Since all other terms in the expansion are of order two or higher in Δt, they can be truncated. Having carried out all the non-zero integrals within the square brackets in equation 1.17, that portion can be replaced with the integrands representing the deterministic and stochastic

terms of the operator:

$$\rho(x, t + \Delta t) - \rho(x, t) = \left[\frac{\partial}{\partial x}\frac{\partial U(x)}{\partial x} + k_B T \frac{\partial^2}{\partial x^2}\right]\rho(x, t)\Delta t$$

$$\frac{\rho(x, t + \Delta t) - \rho(x, t)}{\Delta t} = \left[\frac{\partial}{\partial x}\frac{\partial U(x)}{\partial x} + k_B T \frac{\partial^2}{\partial x^2}\right]\rho(x, t) \tag{1.20}$$

All that is left to do now is realize that this is an expression for a finite difference derivative for an infinitesimally small time step, i.e. a derivative, and remember to put the scaling factor of $\frac{1}{\zeta}$ back into our velocity expression, and we have fully demonstrated that the Smoluchowski equation is the natural consequence of noise averaged Langevin dynamics subject to the conservation law equation 1.10, which is really just a statement that normalization is maintained in time:

$$\frac{\partial \rho(x, t)}{\partial t} = \left[\frac{1}{\zeta}\frac{\partial}{\partial x}\frac{\partial U(x)}{\partial x} + \frac{k_B T}{\zeta}\frac{\partial^2}{\partial x^2}\right]\rho(x, t) \tag{1.21}$$

1.3 Rate Theories

1.3.1 Transition State Theory

We now turn our focus toward applying Langevin dynamics and Fokker-Planck diffusion equations toward calculating the conformational kinetics of macromolecules using various

rate theories, such as Transition State Theory (TST) [37] [121], Kramers Theory [60], and Transition Path Sampling [16].

Transition State Theory is a method for calculating reaction rates using transitions over energetic barriers between two subspaces in phase space (let's call them A and B), and is one of the earliest theoretical treatments for chemical kinetics. A necessary element in any formalism for describing such transitions is a rigorously defined interface that separates the two regions in phase space. The approach utilized in TST is to first determine an appropriate reaction coordinate along a single dimension in configuration space x, for which there should be two well-defined minima separated by a barrier, then assign a value of zero to the point along this coordinate corresponding to the top of the barrier, allowing all points in phase space for which $x > 0$ to be classified as being in state A and all points in phase space for which $x < 0$ to be classified as being in state B. This treatment allows for a step function $\Theta(x)$ to be defined, where $\Theta(x) = 0$ for state A and $\Theta(x) = 1$ for state B. By multiplying the time-dependent probability density function in phase space $f(p_1, x_1, \mathbf{X}; t)$ by this step function, only the portion of phase space within state B remains, which can then be integrated over all phase space variables to obtain the probability of finding the system in state B at time t:

$$P_B(t) = \int\int\int dx_1 dp_1 d\mathbf{X} \Theta(x_1) f(p_1, x_1, \mathbf{X}; t) \tag{1.22}$$

Since the goal is obtaining the rate of formation of B, the time derivative of $P_B(t)$ is the function of interest, which can be taken by operating on the probability density function

$f(p_1, x_1, \mathbf{X}; t)$ with the Liouville operator introduced in equation 1.10.

$$\frac{d}{dt}P_B(t) = -\int\int\int dx_1 dp_1 d\mathbf{X}\Theta(x_1)\hat{L}f(p_1, x_1, \mathbf{X}; t) \tag{1.23}$$

Much like the way the expectation value of an observable as a function of time can be calculated from the vantage point of averaging either time-dependent operators and static quantum states, or static operators operating and a wavefunction evolving in time, depending on whether the Heisenberg or Schrodinger picture is being employed, the fact that the Liouville operator is anti-self adjoint in phase space can be leveraged here to greatly simplify this integral [132]. Instead of operating on the time-dependent probability density function, we apply it to the step function:

$$\frac{d}{dt}P_B(t) = \int\int\int dx_1 dp_1 d\mathbf{X}(\hat{L}\Theta(x_1))f(p_1, x_1, \mathbf{X}; t) \tag{1.24}$$

Since the derivative of a step function is a delta function, we can write:

$$\hat{L}\Theta(x_1) = \frac{p_1}{m_1}\frac{\partial}{\partial x_1}\Theta(x_1) = \frac{p_1}{m_1}\delta(x_1) \tag{1.25}$$

Substituting this delta function expression into equation 1.24 serves to select only the slice of the probability density function in phase space for which $x_1 = 0$ upon integration over dx_1. Also, suppose we are interested in calculating the rate of our chemical species going from state B to state A (moving with a momentum in the negative direction, given the convention

of our step function), implying integration limits of $-\infty$ to 0 for dp_1, our expression becomes:

$$\left(\frac{d}{dt}P_B(t)\right)_{B\to A} = \int_{-\infty}^{0} dp_1 \int d\mathbf{X}\frac{p_1}{m_1}f(p_1, 0, \mathbf{X}; t) \tag{1.26}$$

Since the shape of the distribution in state B should have the same shape as the equilibrium distribution, the time dependence of $f(p_1, 0, \mathbf{X}; t)$ can be factored out as a ratio of the probability of the system being found in B as a function of time divided by its equilibrium value:

$$\left(\frac{d}{dt}P_B(t)\right)_{B\to A} = \int_{-\infty}^{0} dp_1 \int d\mathbf{X}\frac{p_1}{m_1}f(p_1, 0, \mathbf{X})\frac{P_B(t)}{P_B(eq)} \tag{1.27}$$

With the time dependence now confined to a factor of $P_B(t)$, the phase space integral can be carried out, yielding our desired rate equation:

$$\left(\frac{d}{dt}P_B(t)\right)_{B\to A} = -k_{AB}P_B(t) \tag{1.28}$$

Although TST is a venerable formalism that provided the foundation upon which other important theories of chemical kinetics could be built, it does have an Achilles heel. Notice that the multiplication of $\Theta(x_1)$ by $f(p_1, x_1, \mathbf{x}; t)$ in the initial statement of equation 1.24 implies separability between the function that determines which state the system is in and the function describing the density of states in phase space. Implicit within this detail of the formalism is a statement that the time scale on which the probability density function

$f(p_1, x_1, \mathbf{x}; t)$ equilibrates is very fast compared to the amount of time needed for a transition to take place so that the two events are completely uncorrelated. To make another comparison to quantum mechanics, this has a similar feel to the Born-Oppenheimer approximation, where the motion of electrons is so fast compared to the motion of nuclei, that the two motions can be approximated as being separable. Unfortunately, the difference in time scales for transitions between states and probability density function equilibration time for physically relevant classical systems is often not large enough for this separability approximation to be valid.

1.3.2 First Passage Times

Since probability distributions of the incubation times for transitions from one region of phase space or configuration space to another, known as first passage times, are central the Milestoning method, upon which much of this thesis is built, this section will delve into first passage times and their application to the Kramers theory for the rate of escape from potential wells by diffusive systems. First, let us establish a shorthand for the Smoluchowski operator shown in square brackets in equation 1.21.

$$\frac{\partial \rho(x,t)}{\partial t} = \mathcal{D}\rho(x,t) \tag{1.29}$$

which then allows for the operator solution:

$$\rho(x,t) = e^{t\mathcal{D}}\delta(x - x_0) \tag{1.30}$$

where the initial condition for the probability distribution in configuration space $x = x_0$ is represented be a delta function centered at x_0. If we are interested in the amount of time it takes for trajectories to leave a given region of volume V, we must first know the amount of probability density remaining there after a given time t. It is important to note that since we are addressing first passage times, there must be an implicit assumption of absorbing boundary conditions for V. From here, a function for the probability density within V after time t, given an initial configuration x_0 can be defined:

$$S(t, x_0) = \int_V dx \rho(t, x) \tag{1.31}$$

If we now consider $k(t, x_0)$, a probability distribution function of transition times out of volume V, we can write an expression for the change in $S(t, x_0)$, given an infinitesimal change in time dt:

$$S(t, x_0) - S(t + dt, x_0) = k(t, x_0)dt \tag{1.32}$$

which can easily be rearranged via finite difference to arrive at a differential equation for $k(t, x_0)$:

$$k(t, x_0) = -\frac{dS(t, x_0)}{dt} \tag{1.33}$$

Mean first passage $\tau(x_0)$ can then be obtained from the distribution in the usual way of

calculating an average from a distribution:

$$\tau(x_0) = \int_0^\infty tk(t, x_0)dt \tag{1.34}$$

This expression can then be rewritten in terms of the Smoluchowski operator $\mathcal{D}$ using integration by parts, equations 1.33, and 1.30, and the fact that the absorbing boundary conditions imply that $S(\infty, x_0) = 0$ (given infinite time, all trajectories will find their way out of V):

$$\tau(x_0) = -\int_0^\infty t\frac{dS(t, x_0)}{dt}dt$$

$$\tau(x_0) = -tS(t, x_0)\Big|_0^\infty + \int_0^\infty dtS(t, x_0)$$

$$\tau(x_0) = \int_0^\infty dtS(t, x_0)$$

$$\tau(x_0) = \int_0^\infty dt \int_V dx e^{t\mathcal{D}}\delta(x - x_0) \tag{1.35}$$

Interestingly, the fact that the adjoint of the Smoluchowski operator is defined, allows us

to commute the exponential of the Smoluchowski operator with the delta function, causing the Smoluchowski operator to operate on the factor to the right of the delta function, the number 1:

$$\tau(x_0) = \int_0^\infty dt \int_V dx e^{t\mathcal{D}} \delta(x - x_0) = \int_0^\infty dt \int_V dx \delta(x - x_0)(e^{t\mathcal{D}^\dagger} 1) \tag{1.36}$$

This strategy allows for a trivial integral in x:

$$\tau(x) = \int_0^\infty dt e^{t\mathcal{D}^\dagger} 1 \tag{1.37}$$

The differential equation for calculating mean first passage time of a diffusive system can be obtained from this expression by operating on both sides with the Smoluchowski operator, taking the well known integral of the form $\int dt a e^{at}$, and applying the absorbing boundary condition, which mandates that the function be equal to zero at time $t = \infty$:

$$\mathcal{D}^\dagger \tau(x) = \int_0^\infty dt e^{t\mathcal{D}^\dagger} 1 = \int_0^\infty dt \frac{d}{dt} e^{t\mathcal{D}^\dagger} 1 = -1 \tag{1.38}$$

Thusly, it is shown that mean first passage time for Smoluchowski diffusion can be expressed as a differential equation with boundary condition:

$$\mathcal{D}^\dagger \tau(x) = -1 \qquad \tau(x) = 0 \text{ on } \partial V \tag{1.39}$$

where mean first passage time on the surface element ∂V must be equal to zero because it takes no time to complete a journey that begins at the destination.

1.3.3 Milestoning

Included among the computational methods developed to address the challenge of calculating chemical kinetics thus far are transition state theory (TST) [37] [121], transition path sampling (TPS) [16], transition path theory (TPT) [71], and transition interface sampling (TiS) [113]. As previously shown, although TST has been successfully employed in the determination of kinetics for many systems with well-defined reactant and product states, interesting problems in biophysics and elsewhere are frequently encountered where the assumption implicit in TST that equilibration of stable states occurs on a much faster time scale than transition events to the point where the two can be considered uncorrelated cannot be made. In contrast, transition path sampling approaches require no intuition for reaction mechanisms or advance knowledge of transition state, although the requirement of a "dynamical bottleneck" does persist [16] [114]. In this category of methods is the Milestoning algorithm created by Ron Elber et al., which is a method for calculating kinetic properties, where the fundamental objects are the first passage time distributions $K_{AB}(\tau)$, as described in equation 1.33, between adjacent protein configuration milestone states (configurations A and B in this case), where the milestone states do not necessarily need to be meta-stable states as in transition state theory. The key feature of the Milestoning method is that the kinetics of configuration changes which occur over trajectories well outside the time scale of standard molecular dynamics simulations can be accessed by subdividing either configuration space or phase space into subspaces that are small enough for shorter trajectories to be run between the interfaces separating the subspaces. This results in a linear network of transition probabilities between milestones, which can then be solved for such quantities as first passage time between any pair of milestones, including those at the extreme ends of the

space, and the flux through a given milestone, s, as a function of time. Some of the computational gains from this treatment are that breaking up these long trajectory pathways into a network of shorter trajectories leads to increased sampling of the would-be under-sampled areas, and that gains in computational efficiency are possible due to the capacity to run these short trajectories in parallel [33].

The central quantity in milestoning is the flux through a given milestone [33]; it is prescribed by the probabilities

$$P_s(t) = \int_0^t Q_s(t') \left[1 - \int_0^{t-t'} K_s(\tau) d\tau \right] dt'$$

$$Q_s(t) = 2\delta(t) P_s(0) + \int_0^t Q_{s\pm1}(t'') K_{s\pm1}^{\mp}(t - t'') dt'', \tag{1.40}$$

where $P_s(t)$ is the probability of being at milestone s at time t, (or, more specifically, arriving at any time $t' < t$ and not leaving before time t [33]), $Q_s(t)$ is the probability of a transition to milestone s at time t and $K_s(\tau)$ is the probability of transitioning out of milestone s after an incubation time of τ. Thus $\int_0^{t-t'} K_s(\tau) d\tau$ is the probability of an exit from milestone s anytime between 0 and $t-t'$, which makes $1 - \int_0^{t-t'} K_s(\tau) d\tau$ the probability of there *not* being an exit from milestone s over that same time period. Since the probability of two independent simultaneous events happening concurrently is the product of the two events, the equation for $P_s(t)$ is simply integrating the concurrent probabilities of arriving at milestone s and not leaving over the time frame from time 0 to t. In dissecting the meaning of $Q_s(t)$, the first term, $2\delta(t)P_s(0)$, simply represents the probability that the system is already occupying milestone s at time $t = 0$, where the factor of 2 is present since the δ function is centered

at zero, meaning only half of its area would be counted without this factor. $Q_{s\pm1}(t'')$ is the probability that the system transitioned into one of the two milestones adjacent to s at an earlier time t''. $K^{\mp}_{s\pm1}(t-t'')$ is the probability of a transition from milestones $s\pm1$ into milestone s. Thus the second term of the second line of the milestoning equation is another concurrent probability: the probability of the system entering an adjacent milestone at an earlier time, and then transitioning into milestone s between time t and 0. It is important to note that all functions $P_s(t)$ and $Q_s(t)$ are calculated using the respective values of $K_s(\tau)$ between adjacent milestones, thus the set of $K_s(\tau)$ between all milestones of interest contains all the information needed to calculate kinetics using the milestoning method.

Chapter 2

A Smoluchowski Equation for Force-Modulated Chemistry in Single Molecule Pulling Experiments

2.1 Introduction

The material presented in this chapter is meant to first provide context for and then highlight my contribution to the research project described in detail in Appendix A, which is a reprint of an article published by Roy, Grazioli, and Andricioaei in the Journal of Chemical Physics under the title "Mechano-chemical coupling in force-catalyzed single-bond cleavage kinetics: Modeling and simulations of single-molecule pulling" [92].

Single-molecule manipulation techniques are increasingly often revealing important biomolecular conformational changes, one molecule at a time. Thereby, they enable one to identify intermediates and to characterize heterogeneity in conformational pathways, properties that

otherwise would be masked by the averaging inherent in usual bulk experiments. Typical techniques include pulling by atomic force microscopy (AFM) and by optical or magnetic tweezers to probe individual folding in biomolecules or binding in biomolecular complexes [55, 23, 35]. Other examples include applying forces and torques to study supercoiled DNA [68, 97] and DNA-protein [24, 72] or DNA-nanoparticle complexes [91], to unzip [27, 49, 118] or generate novel forms of DNA[119, 21], to reveal the details of viral packing [99, 20], or to probe the interaction of proteins with lipid membranes [5]. An area of related work concerns understanding how chemical steps, such as ATP hydrolysis, can lead to the generation of force and to the movement of biomolecular machines. Single molecule techniques have here been crucial in detecting and estimating mechano-chemical coupling *i.e.*, the coupling between movement and the chemistry of ATP hydrolysis in molecular motors [3, 93, 127, 44, 52, 76, 31]. Application of external forces induces conformational motion and motion couples to chemistry. It is therefore relevant to seek ways in which applied forces modulate chemistry. This is precisely what has been explored recently by Fernandez and coworkers [66, 43, 59, 123, 87, 65], who, in a novel experiment, have studied how external forces affect the the quintessential chemical act of catalysis. The technique used was single molecule force clamp spectroscopy (SMFCS), a method of precise constancy in the force application [22, 96, 79, 39], that has proved particularly useful previously in the characterization of the mechanical unfolding/refolding of proteins. They revealed a coupling between mechanically applied forces and the chemistry of bond cleavage in a *catalytic* reaction, i.e., the rate of force-catalyzed chemical reactions at the single molecule level. In particular, they studied the force dependence of the reduction of disulphide bonds in a protein substrate, titin when catalyzed by the enzyme thioredoxin [4] (reduction which occurs *in-vivo*), and when catalyzed by different small nucleophiles [123, 43, 66, 102]. Two opposing mechanical forces were applied via AFM to pull apart the C- and N-termini of the immunoglobulin-like domain number 27 (I27) of titin, which had an engineered disulphide bond between residues 32 and 75 (see Fig. A.1). The protein unfolded from the two termini up to the sequestered

disulphide bond, which, from being buried inside the folded protein, now became both exposed to the nucleophilic moiety and stretched by the same mechanical force that caused the unfolding of the intervening protein backbone "handles" $1-32$ and $75-89$ (Figure A.1). The disulphide bond was subsequently reduced (cleaved) by thioredoxin or, in their subset of experiments, by the small nucleophiles, and the cleavage resulted in further extension of titin.

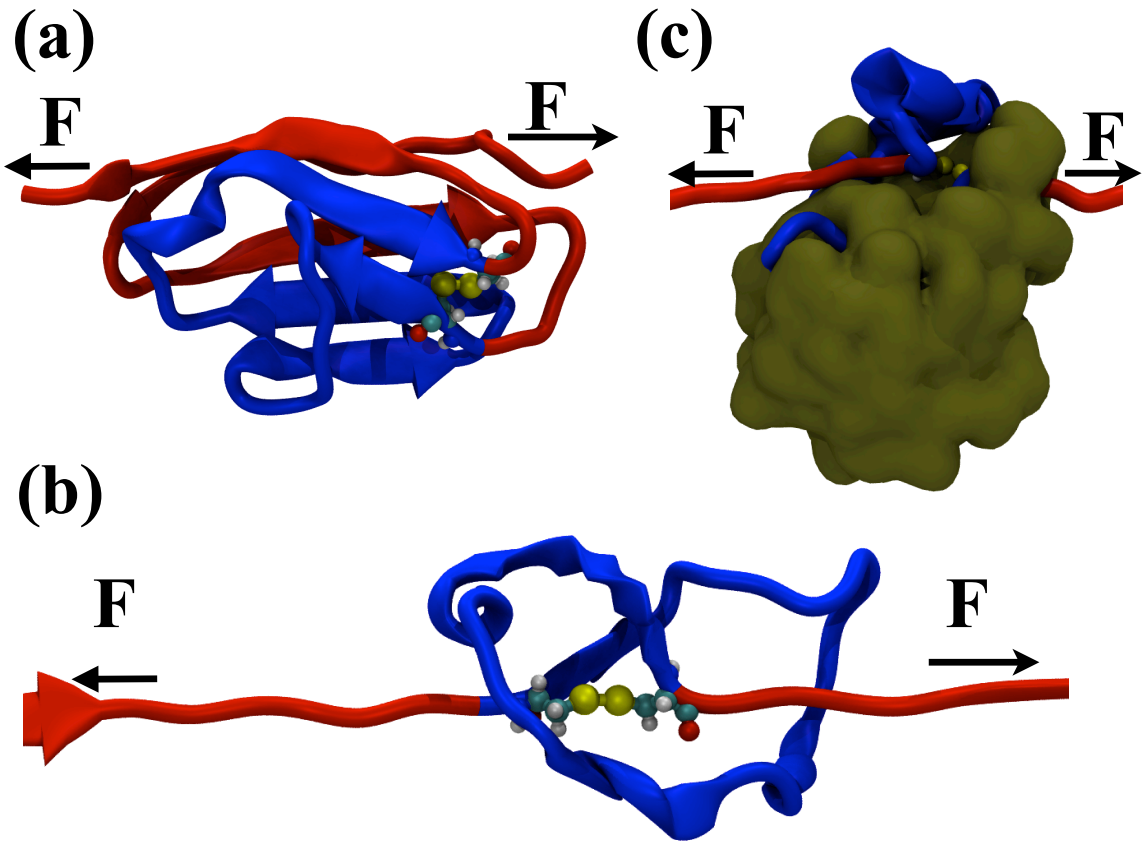

Figure 2.1: Different stages of external force application (a) Force applied on the native titin I27 with an engineered disulphide bond at Cys32-Cys75. (b) Force applied to exposed disulphide bond through the extended C- and N-termini "handle" (viewed from the opposite direction relative to panel (a)). (c) Force applied on titin "handles" while the exposed disulphide bond is in close proximity to the active site of enzyme thioredoxin, shown in green (PDB : 1XOB).

When the small nucleophiles were present, the disulphide reduction followed the kinetics of an $S_N{}^2$ reaction with a first order dependence of the reaction rate on the nucleophilic

concentration and an exponential dependence on the force as in Bell's model [12],

$$\text{rate} = A \exp(-(E_a - F\Delta x_r)/k_B T)[\text{Nu}], \tag{2.1}$$

where [Nu] is the concentration of the nucleophile and E_a is the activation energy barrier, lowered by the external force, F by an amount $F\Delta x_r$, with Δx_r the distance to the transition state along the reaction coordinate, identified as the elongation of the disulphide bond [66, 89, 59]. By and large, the rate of reduction by small nucleophiles was exponentially accelerated by the force on titin. However, when thioredoxin was the catalyst, disulphide bond reduction exhibited a biphasic force dependence [4, 122, 86] in the form of a turnover in the rate-force plots. Both eukaryotic and bacterial thioredoxins were studied. While all thioredoxins showed a negative force sensitivity at lower forces (the rate decreased with force), this was followed, at larger forces, by a force independent behavior for eukaryotic thioredoxins and an increase in the rate with the force for bacterial thioredoxin. The distinct chemical mechanisms underlying the catalytic activity of the two types of thioredoxin enzymes not seen in small nucleophiles was rationalized to be modulated by the highly conserved active site in the enzyme, defined by two "vicinal" cysteine residues at the 32 and 35 positions [48, 38], as well as the surface and depth of the substrate binding groove [87, 50, 9, 25]. Molecular dynamics simulation studies for thioredoxins of different origins attributed the biphasic kinetics in *E-coli* thioredoxin (bacterial) to the shallow binding groove controlling the chemistry of the reaction at lower forces [87] (substrate binding being the rate limiting in the absence of the force), according to a Michaelis-Menten mechanism. At higher forces, the reaction proceeds according to a simple ${S_N}^2$ mechanism, and the formation of the enzyme substrate complex is no longer the rate determining step. The force independent behavior in the case of eukaryotic thioredoxin can be explained on the basis of a single-electron transfer reaction [100] taking place irrespective of the orientation of the disulphide bond [90]. The experimental and molecular dynamics simulations of peptide bound enzymes also confirmed

that the eukaryotic thioredoxins have a deeper binding groove which can lock the substrate disulphide bond, preventing further conformational variability. Even though the reduction rate was found to be force accelerated, following a Bell model (or generalizations thereof) in the case of the small nucleophile, this was not the case for the biphasic turnover in the force-dependence of the thioredoxin catalyzed chemical rate. This points to the idea that the force also modulates the behavior of the protein environment surrounding the cleaved disulphide.

The force dependence observed in titin's disulphide cleavage reaction is reminiscent of an otherwise unrelated class of complexes with similarly biphasic rate-vs-force profiles. The prime feature of this class is a so-called "catch-slip" transition, seen when pulling apart certain adhesive supramolecular complexes, such as the binding to P-selectin and L-selectin of the P-selectin glycoprotein ligand-1(PSGL-1) [10, 94, 104] or the adhesion of the protein FimH to bacterial host cells [107]. The concept, introduced by Dembo *et al.* in 1988 [30], describes rates of dissociation of the ligand that, counterintuitively, first decrease with the pulling force, a range for which interactions are coined "catch bonds" (although no actual covalent bond exists). Subsequently, rates increase with force beyond a certain threshold, a force regime termed "slip bonds". Theories and phenomenological models explaining the dynamic transitions in catch-slip bonds have been developed, chiefly based on the existence of an energy landscape with two bound states or two pathways [106, 128, 83, 11, 130, 105, 36, 85].

A natural framework –perhaps the simplest– to rationalize the qualitative change in the reaction rate for disulphide cleavage with the force applied to the protein is a two-dimensional reaction-diffusion model. An earlier incarnation of a related approach is the venerable Agmon-Hopfield model [2] (see also Ref. [8]), which describes a first-order kinetic process (more precisely, CO binding to a protein) whose rate depends on the "protein coordinate," i.e., on a variable that diffuses in time [81]. The protein coordinate can be thought of, in ef-

fect, as a displaced normal mode, or some linear combination of normal modes, of the protein [131]. An essentially similar description, with the chemistry being this time electron transfer, is the Sumi-Markus model [100]. In any case, in such models, the reaction coordinate, r, is the one along which chemistry occurs and it is coupled to an orthogonal coordinate, the conceptual protein coordinate, x, which evolves according to a Smoluchowski-based reaction-diffusion equation,

$$\frac{\partial \rho(x,t)}{\partial t} = D\frac{\partial^2 \rho(x,t)}{\partial x^2} + \frac{D}{k_B T}\frac{\partial}{\partial x}(\rho(x,t)\frac{\partial V}{\partial x}) - k(x)\rho(x,t), \tag{2.2}$$

where $\rho(x,t)$ is the probability density of x, D is the diffusion coefficient and the last term relates to the rate of the reaction.

Herein we pursue a description of the force-modulated kinetics of disulphide cleavage using a framework inspired by the Agmon-Hopfield or Sumi-Markus models embodied by Eq. (A.2). Our approach is, in spirit, similar to the treatment of catch-slip bond transitions proposed by Liu and Ou-Yang [69], who assume that the distribution of protein conformations involved in the adhesive complex is modulated by the external force, which also couples to the catch-slip detachment coordinate. A quasiharmonic analysis of the substrate protein titin was carried out by Dr. Mahua Roy in order to evaluate candidate modes that can describe the collective motion of the protein coordinate. Our study highlights the importance of involving force-dependent protein modes in theoretical descriptions of mechano-chemical coupling. A novel treatment of the applied force in force-modulated enzyme-catalyzed disulphide bond reduction experiments is introduced, where the force is represented as a vector with components in the reaction coordinate and protein coordinate, and a Smoluchowski-based formalism for reaction-diffusion in this two dimensional space, which can be routinely solved using numerical methods, is presented.

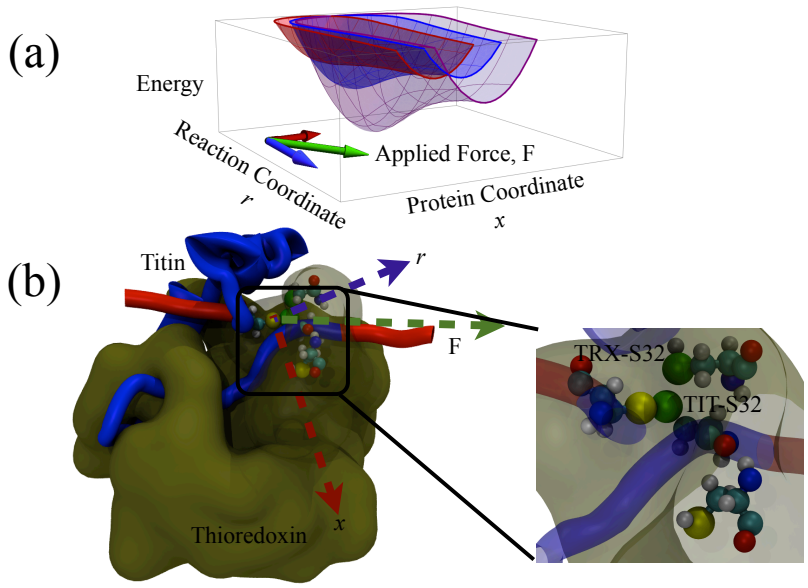

Figure 2.2: (a) Sketch of a potential energy surface distorting under an applied force (green arrow) with vector components along the 'protein coordinate' (red arrow) and 'reaction coordinate' (blue arrow). The red surface corresponds to an applied force of smallest magnitude, with blue being larger, and purple being the largest. The small increase in applied force magnitude going from the red surface to the blue surface causes an increase in well depth for the reaction coordinate, but little change in energy for the non-bonded state along the reaction coordinate, while the highest magnitude applied force shows a pronounced decrease in energy for the non-bonded state (compare the positions of the parabolic cross-sections of the surfaces). This is analogous to the turnover behavior observed in force-modulated disulphide bond reduction, where smaller magnitude forces favor the bound state, while higher magnitude forces increase the rate of bond breakage, i.e., favor the unbound state. (b) Cartoon representation of thioredoxin in complex with the substrate protein (titin) at the transition state as modeled during atomistic simulations carried out by M. Roy (see Appendix A). The arrows represent the same orthogonal coordinates as in part (a), and employ the same color scheme.

2.2 Results

As described above, a diversity of mechanisms for the thioredoxin-catylized disulphide bond cleavage of titin, as it was subjected to pulling forces, were experimentally observed. There were two distinct rate-force dependencies observed, depending on the variety of thioredoxin present [87]. The two types of dependences corresponded to the two distinct families of thioredoxins studied, i.e., eukaryotic vs. bacterial. A prime role in this dichotomy was

played by the binding groove, which was different in the two families, and which controlled the force application on the active site and hence on catalysis. However, a quantitative description of the rate remained to be obtained. Here, we investigate a model in which the diversity in the rates of disulphide cleavage is shown to result from a different force response of the protein coordinate, which in turn affects the reaction coordinate by modulating its energy barrier and transition state position.

To correlate the collective motions the enzyme-substrate complex to the biphasic kinetics of the enzymatic disulphide reduction reaction, we propose to use a force modulated diffusion model - a "bounded diffusion" orthogonal to the reaction coordinate, bound by a force-modulated harmonic potential. The conformational variations corresponding to low frequency vibrations give rise to a fluctuating intrinsic energy barrier, which being modulated by force is revealed as a biphasic behavior in the rate. The particular force-modulated reaction-diffusion equation utilized here is the generalization of the Agmon-Hopfield model [2] by Liu and Ou-Yang [69], which is in effect, an expanded form of the Smoluchowski dynamics in an external field [124].

$$\frac{\partial \rho(x,t)}{\partial t} = D\frac{\partial^2 \rho(x,t)}{\partial x^2} + D\beta\frac{\partial}{\partial x}\rho(x,t)\frac{\partial V\left(x, F_{\perp}\right)}{\partial x} - k_{\text{off}}(x, F_{\parallel})\rho(x,t) \tag{2.3}$$

where, $\rho(x,t)$ is the probability density of finding the value x at a time t, and the two components of the pulling force $\mathbf{F}$ along the protein and reaction coordinate are, respectively, given by

$$F_{\perp} = \mathbf{F} \cdot \hat{x} = F \sin\theta \tag{2.4}$$

$$F_{\parallel} = \mathbf{F} \cdot \hat{r} = F \cos\theta, \tag{2.5}$$

with θ the angle between the applied force and the reaction coordinate, r; D is the diffusion coefficient and β the inverse temperature. $V(x, F_\perp)$ is the force-modulated potential acting on the system (relatedly, a reduced two-dimensional version of a multi-dimensional analytical free energy function was proposed by Suzuki and Dudko [101]), which is a function of both the position along the protein coordinate, x, and the perpendicular component of the applied force. The potential is expressed as:

$$V(x, F_\perp) = V_0 + \frac{1}{2}\kappa(x - x_0)^2 - F_\perp x, \tag{2.6}$$

with V_0 the minimum value of the potential of force constant κ, and x_0 the location of the minimum along the protein coordinate, x. The term $-F_\perp x$ models the amount by which the component of the force along the protein coordinate x tilts the energy along x. Herein we define the *protein coordinate* x as a conformational coordinate along which the motion of the system is orthogonal to the reaction coordinate r; x can be thought of as a linear combination of protein "breathing" modes, a physical correlation to a second degree of freedom affecting kinetic turnover [61]. The final, sink term in the Smoluchowski equation, Eq. (A.3), consumes probability density directly proportionally to the reaction rate coefficient, k_{off}. The rate coefficient is itself a function of the protein coordinate (via the $x-$dependent energy barrier height), as well as of the parallel component of the applied force:

$$k_{\text{off}}(x, F_\parallel) = k_0 \exp[-\beta(\Delta V^\ddagger(x) - F_\parallel r^\ddagger)] \tag{2.7}$$

The component of the force parallel to the reaction coordinate r tilts the energy surface by the amount $-F_\parallel r^\ddagger$, with $r^\ddagger$ the distance to the barrier, namely the separation between the bound state and the energy barrier for disulphide cleavage. Following Liu and Ou-Yang [69], the shape of the reaction energy barrier height $\Delta V^\ddagger$ as a function of the protein coordinate x was modeled as a piecewise function, initially with a positive slope until an equilibrium

distance, after which it assumed a zero slope. The Smoluchowski equation was then solved numerically for $\rho(x,t)$ using the partial differential equation solver in *Mathematica* [125], with Dirichlet boundary conditions $\rho(x_{\max},t)=0$ and $\rho(x_{\min},t)=0$, and the initial condition:

$$\rho(x,0)=\frac{1}{\sqrt{\pi}}e^{-(x-1)^2}. \tag{2.8}$$

Although Neumann boundary conditions (which cancel the flux at the boundaries) are also possible, the simpler to implement Dirichlet boundary conditions were employed here without loss of precision by placing them at values of $x_{\min}$ and $x_{\max}$ which corresponded to states energetically inaccessible to the system at room temperature (we chose $x_{\min}=-22$ and $x_{\max}=22$). Thus the potential itself keeps the system bounded in the x direction, and the boundary conditions merely serve to define zero-valued endpoints for the numerical method. The ultimate goal of generating the ρ(x,t) surfaces is to integrate them over all space and time, in order to generate τ, the disulphide bond lifetime:

$$\tau=\int_0^\infty\int_{-\infty}^\infty \rho(x,t)dxdt. \tag{2.9}$$

The sink term in the Smoluchowski equation ensures that the probability density decays to zero at long times. In the numerical implementation, integrating the surface over infinite time was made tractable by setting the upper limit of the time integral to a value larger than the time t at which $\rho(x,t)$ has decayed more than a cutoff of 10^{-13} of the initial-time value integrated over all x. Since $\rho(x,t)$ decays to zero long before reaching the boundaries, integrating between the boundary conditions is equivalent to integrating over all space. The above process of numerically solving for $\rho(x,t)$ and then integrating it with Eq. (A.9) was repeated for increasing values of force from 0 to 600 pN, and the resulting values for lifetime τ were inverted to find the numerical values of the reaction rates. These reaction rates could then be plotted as a function of the applied pulling force and was fitted to the experimental data of Perez-Jimenez *et al.* [4] by collectively varying the distance to the reaction barrier

$r^{\dagger}$, the reaction attempt frequency k_0, and the force constant for the effective harmonic potential of the protein coordinate κ. Goodness of fit was monitored using a cost function which summed the squared differences between the experimentally measured rates and the calculated rates for each given force value.

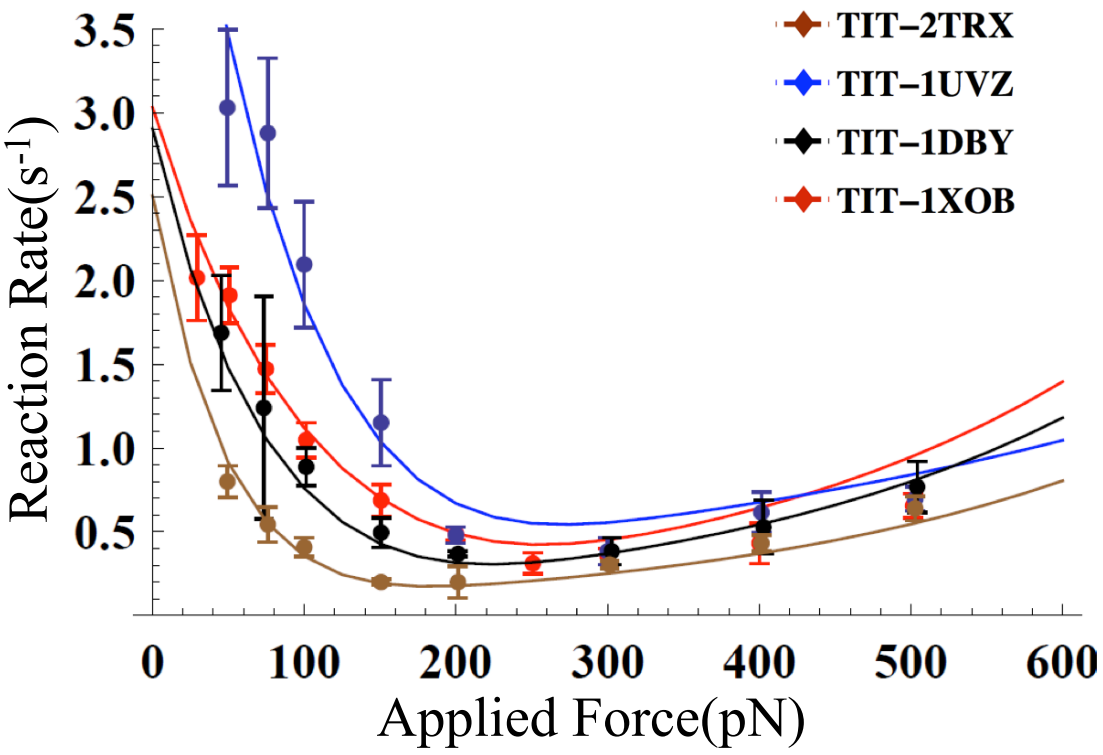

Figure 2.3: Curves generated by numerically solving and integrating the Agmon-Hopfield-Smoluchowski equation for increasing values of applied force and fit to experimentally measured disulphide bond reduction rates from Perez-Jimenez et al. (fit parameters given in Table A.1). Color coding scheme same as in Fig. A.3.

Varying the force constant κ of the force-modulated potential $V(x, F_{\perp})$ varied the rate, reinforcing the importance of the "softness" of the underlying protein coordinate. The matching protein coordinates were subsequently identified atomistically from the the low frequency modes of the quasi-harmonic analysis. *i.e.*, when the distribution of the protein coordinate was bound by a harmonic potential of different force constants κ for each enzyme-substrate complex. The variability in the low frequency quasi-harmonic modes in the inset to

Fig. A.3 originates from the different protein environment engulfing the substrate disulfide in each binding groove pictured in Fig. A.2(b) and results in the variety of curves for the rates in Fig. A.4.

The two key parameters that strongly control the shape of the force-rate curves in Fig. A.4 are κ, the force constant defining the force modulated harmonic potential along the protein coordinate and $r^{\dagger}$, the distance along the reaction coordinate from the bottom of the reactant well to the transition state. The values for κ ranged from 8 pN nm^{-1} (2TRX) to 24 pN nm^{-1} (1XOB), and $r^{\dagger}$ ranged between 0.008 nm (1UVZ) and 0.02 nm (2TRX). Adding to the information obtained through low frequency normal modes, we used molecular dynamics simulations to capture the conformational transitions within the substrate in the presence of force [42, 122] that occur in close proximity to the enzyme thioredoxin, proximity being a measure of r, the distance to the transition state. The force constant, κ, is the key parameter in establishing a connection between the Smoluchowski model for force-modulated chemistry and the atomistic simulations presented herein. Quasi-harmonic mode analysis of the protein/substrate system resulted in a matching distribution of low frequencies from less than 1 to around 30 cm^{-1}. Upon solving for κ from $\omega = \sqrt{\kappa/m}$, where ω is the frequency and m the reduced mass of the oscillator, a value of κ can be identified. This relates the force constant of the Smoluchowski model to a particular region of the spectrum of quaisharmonic modes. For example, for a total mass of the atomistic system in the simulation of roughly 15,000 amu, the larger reduced mass is $7500 \times 7500/15000 = 3570$ amu, which would correspond to a large scale oscillation of two portions of equal mass. This establishes the upper limit for reduced mass, with lesser values possible, (all the way down to 1 amu, corresponding to a single hydrogen atom oscillating against the rest of the system). As an example, if we assume that the normal mode which best corresponds to our protein coordinate is a low frequency mode approximated as motion involving about 5% of the full structure oscillating from the remainder of the structure, then we can use the previously stated relationships to establish a range for κ of around 10 pN/nm corresponding to frequency = 0.5 cm^{-1} to around

27 pN/nm for frequency = 0.8 cm^{-1}. In comparison, for the Smoluchowski equation used to generate the rate vs. force curves in Fig. A.4, the κ parameter was varied between 8 and 24 pN/nm to produce the fits. Although this matching strategy does not generate a unique mode identification, it does provide a reasonable picture of the protein coordinate, and it invites further exploration into ascribing physical meaning to motion along it using linear combinations of low-frequency normal modes.

	κ (pN/nm)	$r^{\dagger}$ (nm)	θ (rad)	k_0 (s^{-1})
Blue	9.9798	0.008	0.1	6.0×10^6
Red	23.9798	0.01215	0.166	2.8×10^6
Black	11	0.016	0.1	2.0×10^6
Brown	8	0.02	0.1	1.1×10^6

Table 2.1: Parameter values from the Smoluchowski equation used to fit the curves representing disulphide bond cleavage rate as a function of the applied force to the experimentally measured values for different forms of Thioredoxin, as color coded in Fig. A.4.

2.3 Concluding Discussion

Recent single molecule experiments on thioredoxin-catalyzed thiol-disulphide exchange unveiled a nontrivial rate-force dependence. The rate initially decreased for forces F below 200 pN and increased at higher forces. In principle, the initial decrease could formally be considered as a consequence of an effective distance-to-barrier increase with force, modeled by Δx_r and F having opposite signs if the Bell model (Eq.(A.1)) is imposed [83]. This is would be then followed by a subsequent increase at larger forces, caused by a coupling of the force with the elongation in bond length (a regular, positive Δx_r). Such a force-reaction coupling modeled à la Bell is necessarily one-dimensional. In the case of complex macromolecular systems, as are protein-enzyme complexes, the force applied to the substrate protein is more likely to act along directions other than the reaction coordinate. Hence the factors dictating the distance to the transition state and subsequent kinetics cannot be identified by a single bond elongation, Δx_r, but by a combination of several parameters. Here we offered evidence that

such parameters arise naturally from internal protein coordinates, that vary or are modulated by force as the system progresses from the reactant to the transition state. We modeled an internal protein coordinate as a linear combination of low-frequency quasi-harmonic modes which differ in different active site environments, as represented by complexes of the same substrate with different enzymes. Accordingly, rates of different complexes exhibited a slight difference that was in agreement with experiments and the measured the biphasic behavior observed in all bacterial complexes. We successfully reproduced the biphasic force dependency of rates by simultaneously propagating the protein coordinate and reaction coordinate along the two components of a force modulated potential by solving a generalized version of a reaction-diffusion equation.

To conclude, we provided microscopic evidence of the protein conformational coordinate through our simulations and quahiharmonic mode analysis which successfully validated our model of force modulated diffusion of protein and reaction coordinate along two perpendicular dimensions. Similar descriptions are also relevant for the force and torque effects on the activity of enzymes on nucleic acid substrates during genetic transactions [6]. We expect our study to be important as more experimental examples of mechano-chemical coupling, new sono-chemical coupling [26, 47] or coupling to electrical fields [41] become available.

Chapter 3

Advancements in the Milestoning Technique: I. Enhanced Sampling via "Wind" Assisted Re-weighted Milestoning (WARM)

3.1 Introduction

The task of calculating kinetic properties from molecular dynamics simulations is a complex problem of considerable interest [29] [28]. In contrast to computational methods designed for equilibrium calculations, in which the basic observables are thermodynamic averages over conformational points (structures) generated over an invariant measure without the need to obey exact dynamical equations, studies of kinetics require physically correct time-ordered trajectories to obtain time-correlation functions as the basic objects [57]. Since each time-correlation function describes the relaxation under investigation as an average over

all relevant trajectories, adequate sampling for accurate calculation of long-time kinetics can quickly become computationally intractable via direct simulation [14]. This is because a direct, brute force method of this type would require sufficiently long simulation times such that the system would be able to transition between states of interest enough times that a statistically significant distribution of first passage times could be generated. Several computational methods have been developed to address the challenge of calculating chemical kinetics, starting with the venerable transition state theory (TST) [37] [121], and continuing with more recent developments, such as, transition path sampling (TPS) [16], transition path theory (TPT) [71], and transition interface sampling (TiS) [113]. Although transition state theory has been successfully used in the determination of the kinetics for many systems with well-defined reactant and product states, for which the "dynamical bottleneck" can be identified [110], there are many interesting problems in biophysics, and elsewhere, for which these assumptions do not hold. In contrast, transition path sampling approaches require no intuition for reaction mechanisms or advance knowledge of transition state, although the requirement of a "dynamical bottleneck" does persist [16] [114]. In the same category of methods is the milestoning algorithm created by Ron Elber et al., which is a method for calculating kinetic properties, where the fundamental objects are the first passage time distributions $K_{AB}(\tau)$ between adjacent protein configuration milestone states (configurations A and B in this case), where the milestone states do not necessarily need to be meta-stable states as in transition state theory. The key feature of the milestoning method is that long trajectory pathways for large scale configuration changes can be broken up into shorter trajectories for which a linear network of transition probabilities between milestones can be devised. The aforementioned linear networks of transition probabilities can then be solved for such quantities as first passage time between any pair of milestones, including those at the extreme ends of the space, and the flux through a given milestone, s, as a function of time, written as $P_s(t)$ (equation 3.1). Some of the key gains from this treatment are that breaking up these long trajectory pathways into a network of shorter trajectories leads to

increased sampling of the would-be under-sampled areas, and that gains in computational efficiency are possible due to the capacity to run these short trajectories in parallel [33]. In practice, previous milestoning calculations have been limited to calculating the constant flux values representative of the system at equilibrium, which can be thought of as the long time flux values $\lim_{t\to\infty} P_s(t)$. A method for calculating the time-dependent flux through a given milestone $P_s(t)$ can be found in chapter 4. The aim of the technique we present in this paper is to increase the computational speed of the milestoning method via the addition of an artificial constant force ($\mathcal{F}_{wind}$) along the vector pointing from the initial state to the final state for each pair of milestones in the simulation, causing the system to arrive at the destination configuration in far fewer time steps than if it were left to Brownian dynamics alone. The key idea which makes this possible is the use of a re-weighting function we have introduced previously [77] [126] [51] [78], which generates a re-weighting coefficient for each trajectory, thus allowing the true distribution of first passage times to be recovered from the artificially accelerated trajectories. Preliminary calculations conducted on model systems, described in the Numerical Demonstration section, have demonstrated a computation time speedup by a factor of nearly 40 using this method.

3.2 Theory

The central quantity in milestoning is the flux through a given milestone [33]; it is prescribed by the probabilities

$$P_s(t) = \int_0^t Q_s(t') \left[1 - \int_0^{t-t'} K_s(\tau) d\tau \right] dt'$$

$$Q_s(t) = 2\delta(t)P_s(0) + \int_0^t Q_{s\pm1}(t'')K^{\mp}_{s\pm1}(t-t'')dt'', \tag{3.1}$$

where $P_s(t)$ is the probability of being at milestone s at time t, (or, more specifically, arriving at any time $t' < t$ and not leaving before time t [33]), $Q_s(t)$ is the probability of a transition to milestone s at time t and $K_s(\tau)$ is the probability of transitioning out of milestone s after an incubation time of τ. Thus $\int_0^{t-t'} K_s(\tau)d\tau$ is the probability of an exit from milestone s anytime between 0 and $t-t'$, which makes $1-\int_0^{t-t'} K_s(\tau)d\tau$ the probability of there *not* being an exit from milestone s over that same time period. Since the probability of two independent simultaneous events happening concurrently is the product of the two events, the equation for $P_s(t)$ is simply integrating the concurrent probabilities of arriving at milestone s and not leaving over the time frame from time 0 to t. In dissecting the meaning of $Q_s(t)$, the first term, $2\delta(t)P_s(0)$, simply represents the probability that the system is already occupying milestone s at time $t = 0$, where the factor of 2 is present since the δ function is centered at zero, meaning only half of its area would be counted without this factor. $Q_{s\pm1}(t'')$ is the probability that the system transitioned into one of the two milestones adjacent to s at an earlier time t''. $K^{\mp}_{s\pm1}(t-t'')$ is the probability of a transition from milestones $s \pm 1$ into milestone s. Thus the second term of the second line of Eq. (3.1) is another concurrent probability: the probability of the system entering an adjacent milestone at an earlier time, and then transitioning into milestone s between time t and 0. It is important to note that all functions $P_s(t)$ and $Q_s(t)$ are calculated using the respective values of $K_s(\tau)$ between adjacent milestones, thus the set of $K_s(\tau)$ between all milestones of interest contains all the information needed to calculate kinetics using the milestoning method.

The essential connection to make in regard to combining the milestoning method with re-weighting of artificially accelerated trajectories is that a K function between two milestones located at $x = A$ and $x = B$, $K_{AB}(\tau)$, is nothing more than a probability distribution as a

function of lifetime describing the conditional probability that a system found in state A at time $t = 0$ will be found, for the first time, in state B at time $t = \tau$:

$$K_{AB}(\tau) = P(x_B, \tau | x_A, 0) \tag{3.2}$$

Given this relationship, we can now begin to make the connection between milestoning and re-weighting of artificially accelerated trajectories. Assuming Langevin dynamics with the addition of a *wind* force:

$$m\ddot{x} = -\gamma m\dot{x} - \nabla V(x) + \xi(t) + \mathcal{F}_{wind} \tag{3.3}$$

where γ is the friction coefficient, $V(x)$ is the potential, $\xi(t)$ is the random force, and $\mathcal{F}_{wind}$ is a constant force applied in the direction of the goal milestone for each run; conditional probabilities reflecting first passage transitions from milestone A to B can be expressed as:

$$P(x_B, \tau | x_A, 0) = \int D\xi W[\xi] \delta(x(\tau) - x_B) \tag{3.4}$$

In this equation, $W[\xi(t)]$ is the probability distribution representing the joint probabilities of all possible series of random kicks, so multiplying by the delta function $\delta(x(\tau) - x_B)$ and integrating selects for only the portion of the distribution which represents a series of random kicks which results in a transition from state A to state B given an incubation time τ. It then follows suit that the integral in this equation is simply the expectation value for the probability of a transition from A to B for each incubation time point τ, which, again, is

equivalent to $K_{AB}(\tau)$. Because of fluctuation dissipation, $\langle \xi(t)\xi(t') \rangle = 2k_B T m\gamma$, the random force in Langevin dynamics is a Gaussian distribution with variance $w \equiv 2k_B T m\gamma$. Thus it can be show that:

$$W[\xi(t)] = \exp(-\frac{1}{2w}\int_0^t \xi(t')^2 dt') \tag{3.5}$$

With $W[\xi(t)]$, our weighting function for joint probabilities of random kick sequences in terms of our random force $\xi(t)$ in hand, we can now write the noise history $\xi(t)$ in terms of the trajectory $x(t)$ it generates:

$$\xi(t)^2 = (m\ddot{x} + \gamma m\dot{x} + \nabla V(x) - \mathcal{F}_{wind})^2 \tag{3.6}$$

We are ultimately interested in measuring conditional probability distributions in configuration space, $x(t)$, not random force space, $\xi(t)$, but since the Jacobian is built into the measure, x, we can define $S[x(t)]$ thusly:

$$S[x(t)] \equiv \xi(t)^2 = (m\ddot{x} + \gamma m\dot{x} + \nabla V(x) - \mathcal{F}_{wind})^2 \tag{3.7}$$

Then, using the Ito formalism for stochastic calculus, we can express our conditional prob-

ability using the Wiener formalism of path integrals [58] as:

$$P(x_B, \tau | x_A, 0) = \int_{(x_A,0)}^{(x_B,\tau)} Dx \exp\left(-\frac{S[x(t)]}{2w}\right) \tag{3.8}$$

In this form, it is clear that the exponential function represents the weighting function for the trajectory $x(t)$:

$$W[x(t)] \equiv \exp\left(-\frac{S[x(t)]}{2w}\right) \tag{3.9}$$

With the weight of each trajectory, $W[x(t)]$, now formally defined, we can define the re-weighting factor for obtaining the true weight of an artificially accelerated trajectory as:

$$\frac{W[x(t)]}{W_f[x(t)]} = \exp\left(-\frac{S[x(t)] - S_f[x(t)]}{2w}\right) \tag{3.10}$$

where the f subscript indicates a function generated under the influence of the artificial $\mathcal{F}_{wind}$ force. In practice, once a trajectory $x(t)$ is generated (in the presence of the wind force), the actions are calculated in discrete numerical form using:

$$S_f[x(t)] \approx \sum_i \left(m\frac{\Delta\nu_i}{\Delta t} + m\gamma\frac{\Delta x_i}{\Delta t} + \nabla V_i - \mathcal{F}_{wind}\right)^2 \Delta t$$

$$S[x(t)] \approx \sum_i \left(m\frac{\Delta\nu_i}{\Delta t} + m\gamma\frac{\Delta x_i}{\Delta t} + \nabla V_i \right)^2 \Delta t \tag{3.11}$$

The re-weighting factor is calculated from Eq. (10) and stored in an array. When post-processing to compute the $K_{AB}(\tau)$ distribution for a particular pair of milestones A and B by histogramming trajectories by lifetime τ, instead of adding 1 to a particular bin each time the lifetime of a particular trajectory falls within the bounds of that bin, the weight $W[x(t)]$ corresponding to that trajectory is instead added. It is clear from equation 10 that as $S_f[x(t)]$ for the artificially accelerated trajectory approaches $S[x(t)]$, the weight of the trajectory approaches unity, thus the method reduces to an unweighted histogram for $\mathcal{F}_{wind} = 0$ as it should.

3.3 Numerical Demonstration

Model System in One Dimension

A simple two well potential (see inset of figure 7) of equation $y = (x-1)^2(x+1)^2$ was chosen to be the model system upon which the wind-assisted milestoning methodology could be developed. In running wind-assisted milestoning, the potential to which the particle is being subjected is first divided into any number of milestones, in this case, 7 milestones, thus 6 separate spaces. Next, numerous Langevin trajectories are run both from left to right and right to left between each pair of adjacent milestones. The number of time steps required to go from the starting milestone to the destination milestone for each trial of each pair and the weight of each trajectory is then stored in an array as mentioned in the theory section. As shown in figures 1 and 2, this method has brought about a more than tenfold speedup in

computation time with very little sacrifice in terms of accuracy.

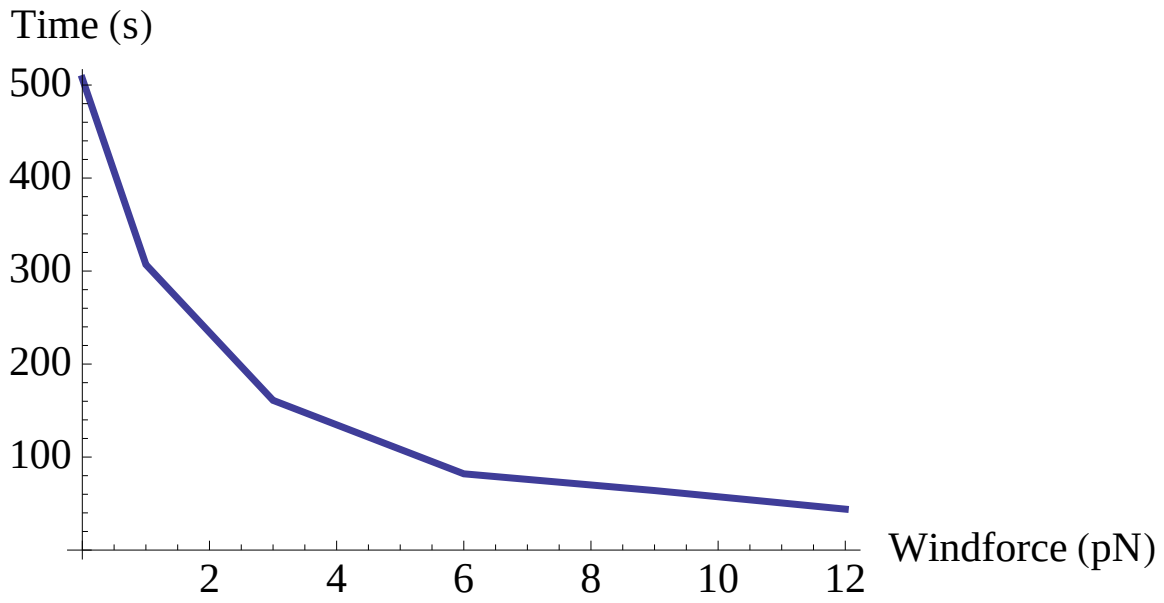

Figure 3.1: Shown here is calculation time as a function of the $\mathcal{F}_{wind}$ force for all $K(\tau)$ distributions in both directions for six subspaces ranging from -2 to 2 on the bistable harmonic potential. The calculation took 507 seconds for $\mathcal{F}_{wind} = 0pN$ and just 44 seconds for $\mathcal{F}_{wind} = 12pN$

.

Model System in One Dimension with Distortion

Thus far, we have approached the WARM method from the standpoint of speeding up the calculation by pushing $\mathcal{F}_{wind}$ until the $K(\tau)$ functions begin to distort. Here we will explore the possibility that even slightly distorted $K(\tau)$ functions can yield useful information, allowing for even greater computational speedup. The flux value for a given milestone s, $P_s(t)$, should approach the probability predicted by the Boltzmann distribution generated from configurational partition function as time approaches infinity. Given a discrete space in x, subject to our 1D potential $y = (x-1)^2(x+1)^2$, the Boltzmann distribution function

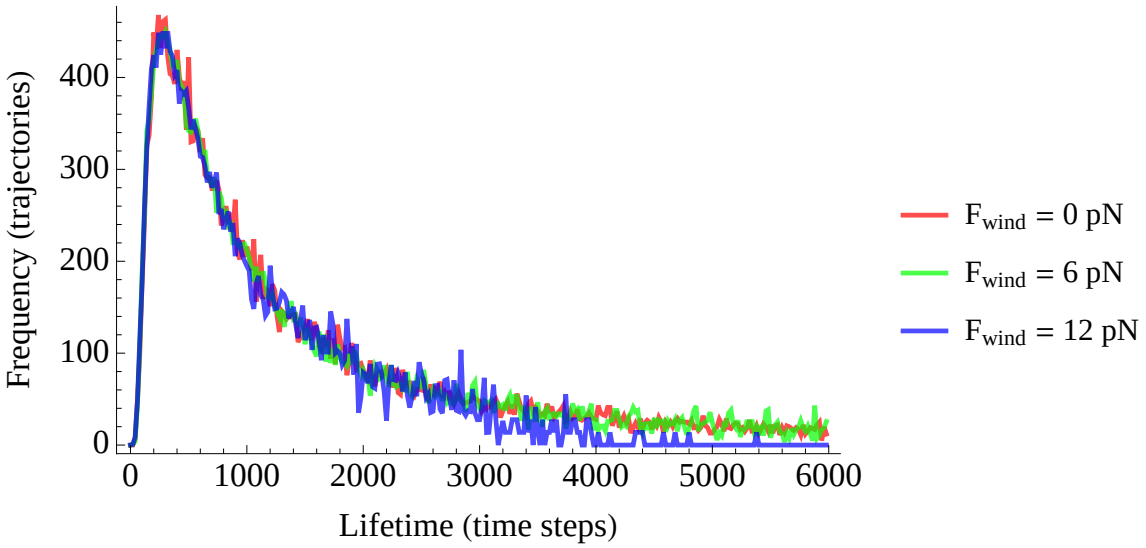

Figure 3.2: Shown in this figure is the transition probability distribution $K_{23}(\tau)$, i.e. the transition probability from milestone 2 to milestone 3 as a function of lifetime, calculated using $\mathcal{F}_{wind}$ forces ranging from 0 to 12 pN. The plots indicate that the rapid decrease in computation time due to the added $\mathcal{F}_{wind}$ force has almost no effect on accuracy.

can be obtained in the usual way, shown in equation 3.12, below:

$$\lim_{t\to\infty} P_s(t) = \frac{e^{-\beta U(x_s)}}{\sum_{n=1}^{N_S} e^{-\beta U(x_n)}} \tag{3.12}$$

where N_s is the total number of milestone configurations, and x_n signifies the spatial position of each milestone. This discrete space approximation for the equilibrium flux values is utilized below as a test for accuracy in figure 3.3 (dashed lines). The numerical demonstration in this section consists of dividing the space for the bistable 1D potential between $x = -2$ and $x = 2$ into 11 subspaces bounded by 12 milestones. First hitting trajectories were run between each pair of adjacent milestones, and then each pair of $K(\tau)$ functions describing a transition away from each milestone were normalized (e.g. for milestone 3, all trajectories must terminate at either milestone 2 or milestone 4, therefore $\int_0^\infty K_{32}(\tau)d\tau + \int_0^\infty K_{34}(\tau)d\tau = 1$). The normalized $K(\tau)$ functions are then integrated over all time, and these values are placed

in a matrix, $\mathbf{K}$, of equilibrium transition probabilities. The equilibrium flux values for the vector representing the set of milestones, $\mathbf{P}$, is then found by numerically solving for the eigenvector: $\mathbf{P}\cdot\mathbf{K} = \mathbf{P}$ [32]. By using this method to determine equilibrium flux values, it is demonstrated in figures 3.3 and 3.4, that even when $\mathcal{F}_{wind}$ is set to a value strong enough to distort the $K(\tau)$ functions, accurate equilibrium flux values can still be calculated.

Model Systems in Two Dimensions

Two additional test systems for the WARM technique were implemented for further validating the method in two dimensions. Both systems have double well shapes, however for one well, the barrier to transition from one well to the other is primarily energetic, while the other is primarily entropic (see figure 3 below). The potential with the energetic barrier is a generalization of the 1D potential from the previous section to two dimensions, and the potential with the entropic barrier is the same potential implemented by Elber and Faradjian in their original paper on milestoning [33]. As can be seen in the data below, the WARM method successfully re-weighted first passage time distributions ($K(\tau)$) generated using artificially accelerated trajectories to yield the true first passage time distributions which would have resulted from trajectories in the absence of the *wind* force. In both cases, the method achieved more than 60% faster computation times with very little sacrifice in terms of accuracy.

Model System in Eleven Dimensions

In order to demonstrate that the WARM method possesses no inherent limitations due to scaling, the method was applied to an 11 dimensional hyperspace. For this model, the $11D$

potential was defined as:

$$V(x_1, x_2, ..., x_{11}) = (x_1 - 1)^2(x_1 + 1)^2 - \frac{1}{2}\sum_{n=2}^{11} x_n^2 x_1^2 + \sum_{n=2}^{11} x_n^4 \tag{3.13}$$

where the first term is the same bistable potential in x_1 used in the first one dimensional example, the second term couples motion in the 10 dimensions orthogonal to barrier height in x_1, and the third term simply confines the system to a reasonably sized configurational space using a quartic potential. In order to develop some intuition for this potential, see figure 3.10, then just imagine that there are nine other dimensions which have the same effect as y on barrier height in x_1.

Accordingly, the milestones must be defined as hyperplanes, given the general definition of a hyperplane:

$$a_1x_1 + a_2x_2 + a_3x_3...a_nx_n = b \tag{3.14}$$

To keep things simple, we set a_2 through a_{11} equal to zero, and $a_1 = 1$, allowing us to define two hyperplanes as $x_1 = -1$ and $x_1 = 1$. In this scenario, the features of interest are the transitions between the wells at $x_1 = -1$ and $x_1 = 1$, thus the 11D $\mathcal{F}_{wind}$ is applied with zero components in all dimensions except for x_1 where it is used to push the system over the barrier between wells. The WARM method was successfully applied to this 11 dimensional potential, and a speedup by a factor of 4.5 was observed (figures 3.11 and 3.12).

Wind Force as a Vector Field

In all of the preceding examples, $\mathcal{F}_{wind}$ was applied to the system as a constant force applied in a straight line, perpendicular to the parallel milestone hyperplanes, but $\mathcal{F}_{wind}$ can be defined any way we choose. This section demonstrates a method whereby the directionality of $\mathcal{F}_{wind}$ is defined by a vector field which allows $\mathcal{F}_{wind}$ to blow in a curved path between two nearly orthogonal milestones (see figures 3.13 and 3.14), i.e. our wind has become a tornado! In order to define this vector field, the point of intersection between the two planes was determined, then a function was created which finds the straight line connecting the current position to this point of intersection, and then defines $\mathcal{F}_{wind}$ at that point to be a vector both orthogonal to that straight line and pointing in a clockwise direction. When first passage times were calculated going from the milestone shown in red toward the milestone shown in green (figure 3.13), the vector field is simply multiplied by -1 to cause our tornado to spin counterclockwise. Using a wind force defined in this manner, we obtain an efficient directionality for $\mathcal{F}_{wind}$ which biases the system toward both leaving its initial milestone in the right direction and approaching its destination milestone, regardless of the positioning of the milestones in configuration space. Another advantage of this scheme is that our curved vector field of $\mathcal{F}_{wind}$ can be defined without any knowledge of the system itself, we only need to know the positions of the milestones, which are always known in milestoning calculations. Figures 3.13 through 3.18 illustrate the application and results of this method using two different 2D potentials, the Muller-Brown potential [74], and a simpler Muller-inspired potential with two Gaussian wells we'll call our Gaussian potential. The Gaussian potential is defined as:

$$V(x,y) = -exp[-(2(x-.8)^2 + y^2)] \tag{3.15}$$
$$-1.3exp[-((x+1)^2 + (y-1.5)^2)]$$
$$+.2x^2 + .2y^2$$

and the Muller potential is defined as:

$$V(x,y) = h\sum_{k=1}^{4} exp[a_k(x-x_k^0)^2 \tag{3.16}$$
$$+b_k(x-x_k^0)(y-y_k^0) + c_k(y-y_k^0)^2]$$

where:

$$A = (-200, -100, -170, 15), a = (-1, -1, -6.5, .7)$$
$$b = (0, 0, 11, .6), c = (-10, -10, -6.5, .7)$$
$$x^0 = (1, 0, -.5, -1), y^0 = (0, .5, 1.5, 1)$$
$$h = .005$$

A speedup factor of 4 was observed in both the Muller potential and the Gaussian potential, although the $K(\tau)$ functions in the Gaussian potential example displayed less distortion than those produced in the calculations performed using the Muller potential.

3.4 Concluding Discussion

We have presented and tested a method for accelerating milestoning calculations, whereby the true probability density functions for first passage transition time between milestones, $K_{AB}(\tau)$, are recovered from artificially accelerated trajectories via the re-weighting method described in the Theory section. These $K_{AB}(\tau)$ functions are central to milestoning calculations, thus the WARM method presented herein shows potential for broad application.

Our method has been shown to be effective on one and two dimensional potentials with both energetic and entropic barriers, as well as an 11 dimensional hyperspace, implying that the method should have no scaling limitations, thus the next step will be to test the method on chemical systems. The simplest application would be to apply a single force vector to a single atom which pushes the system toward a configurational change of interest. In this case, the re-weighting factors $S[x(t)]$ and $S_f[x(t)]$ could be calculated by summing the force components in the x, y, and z directions both with and without the components of the applied force, respectively.

The main limitation of the WARM method, regardless of the number dimensions are present, is obtaining good re-weighting in the longer τ range. This is simply a matter of under-sampling. If $\mathcal{F}_{wind}$ is pushing the system to the next milestone so quickly that longer values of τ, relevant to the true $K_{AB}(\tau)$ distribution, are not being sampled, then there just isn't enough density present (or even none at all) to re-weight. This is why the accuracy in the low tau regime is often still quite good when too high of an $\mathcal{F}_{wind}$ has caused the latter portion of the distribution to turn to noise. Thus far, the limitations on the WARM method appear to be solely dependent on whether or not we've pushed the force so hard that trajectories in the longer τ region of the true $K_{AB}(\tau)$ are even being sampled. For this reason, systems whose true $K_{AB}(\tau)$ distribution functions possess fat tails place the greatest limitations on the degree of computational speedup achievable by WARM. This issue can be addressed in

a couple different ways. One approach is to simply define more milestones in the space, the other is to combine the WARM method with some sort of artificial heating method, both modifications which will yield $K_{AB}(\tau)$ functions which decay more rapidly after their peak.

It should be noted that our application of the WARM method to both the high dimensional model, and our vector field-based $\mathcal{F}_{wind}$ implementation of demonstrate that this technique can be applied to systems too complex to intuit the placement of the artificial forces. Given an initial and a final milestone configuration, one could determine the vectors pointing from each atom's initial position to it's final position. Artificial forces, $\mathcal{F}_{wind}$, could then be placed upon all atoms in the system pointing in the direction of these vectors and with a magnitude proportional to the length of the vectors. A zero cutoff could also be added so as not to waste computational resources on applying and accounting for $\mathcal{F}_{wind}$ forces atoms which are beginning at a position fairly close to their destination. We believe that, upon implementation into a molecular dynamics package such as MOIL [34], the WARM method has the potential to be a useful tool for the determination of the kinetic properties of macromolecules.

3.5 Acknowledgments

IA acknowledges funds from an NSF CAREER award (CHE-0548047).

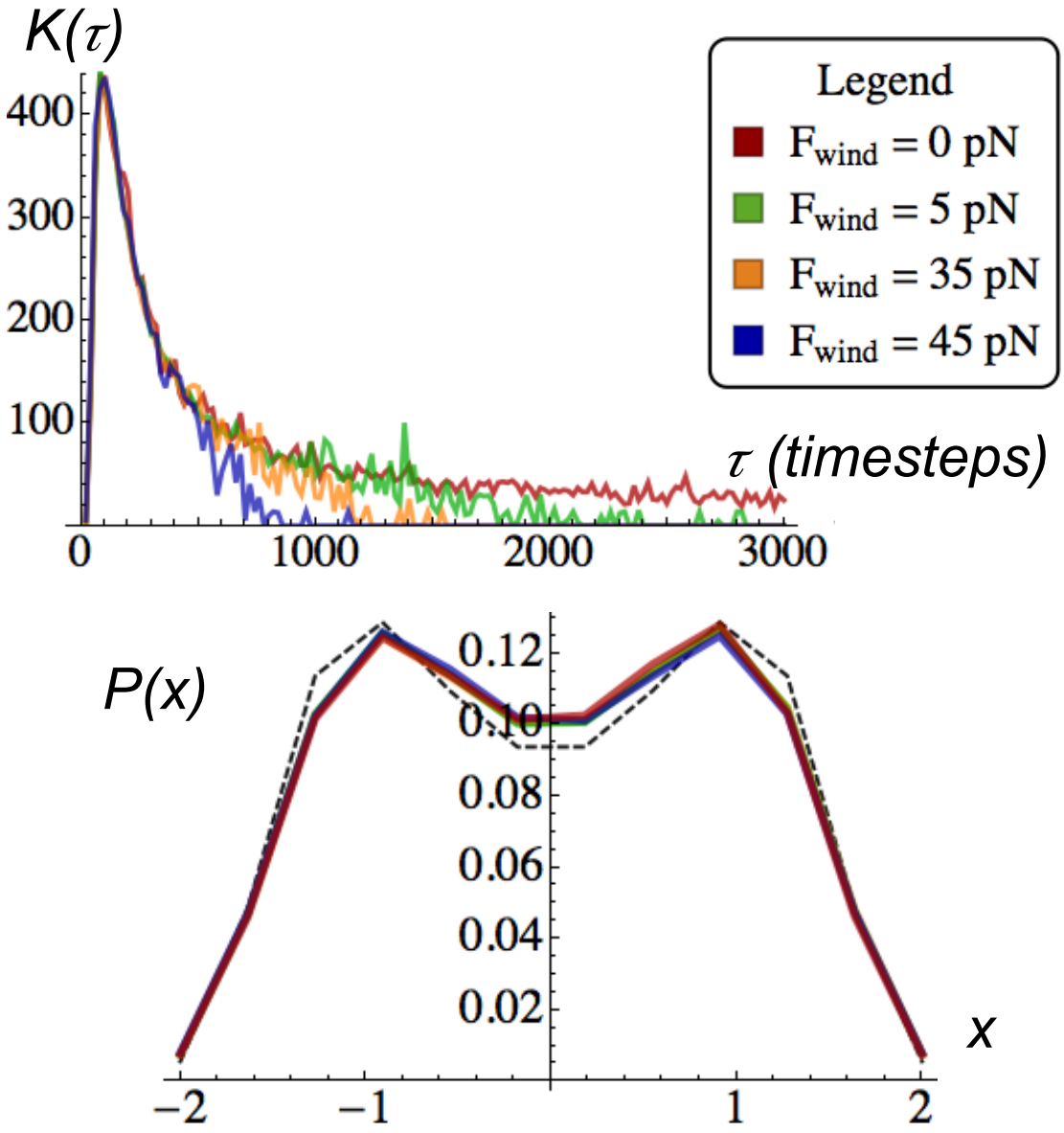

Figure 3.3: The plot at the top of this figure shows plots for one of the transition probability distributions $K(\tau)$ for the bistable 1D potential with different values of $\mathcal{F}_{wind}$ implemented. Note that although the distributions distort considerably for higher values of τ when the system is pushed with high magnitude $\mathcal{F}_{wind}$, the equilibrium flux values in the plot below remain fairly constant. The color scheme legend applies to both plots.

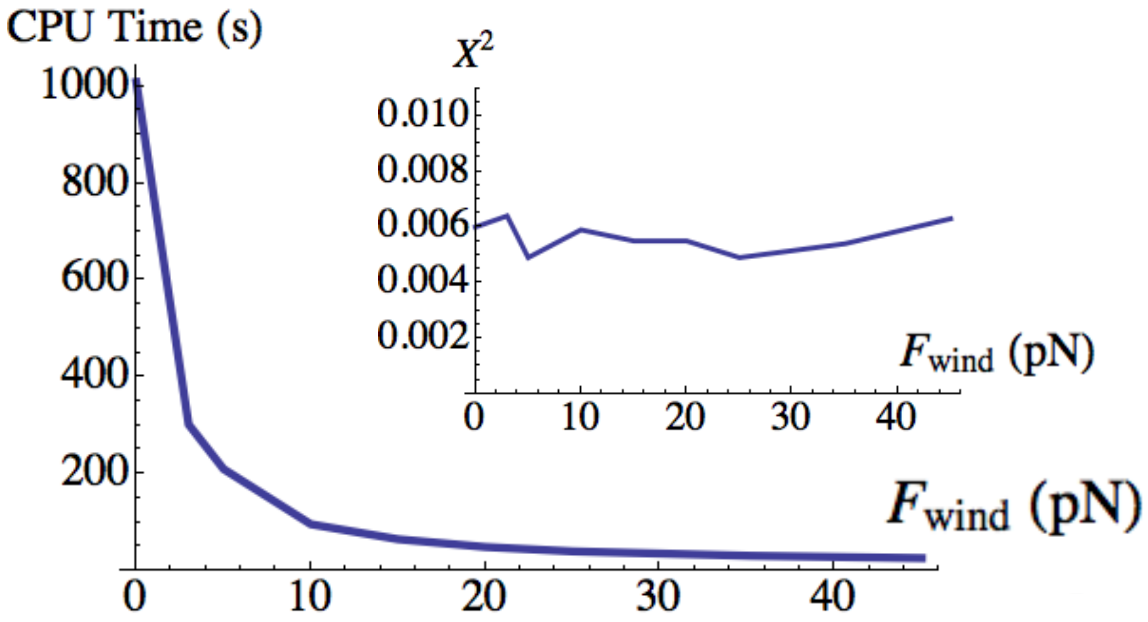

Figure 3.4: Here we show effects of applying higher magnitude $\mathcal{F}_{wind}$ which are strong enough to significantly distort the $K(\tau)$ functions. This figure facilitates a direct comparison of gain in computational speed with the accuracy of the equilibrium flux values (measured as X^2). Note that while there is no appreciable change in accuracy, calculation time drops from 1109 s to 26 s, a speedup by a factor of nearly 40.

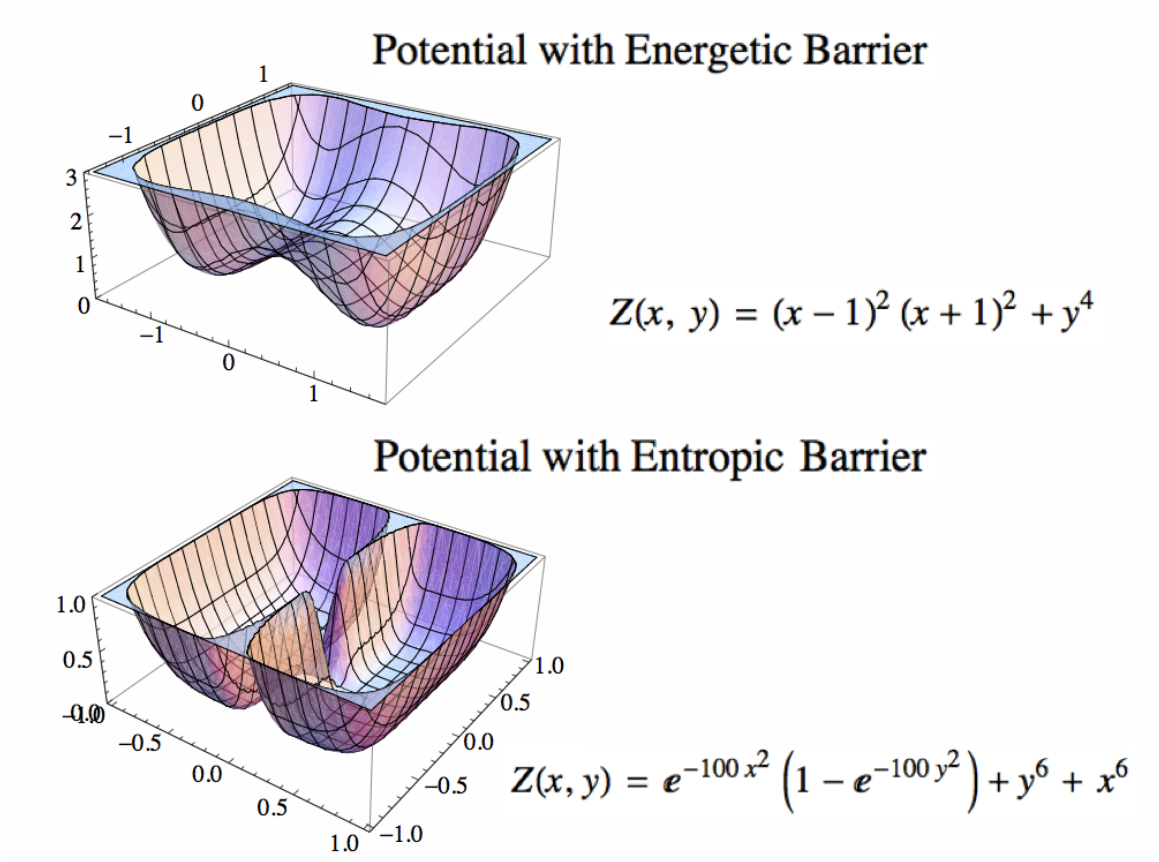

Figure 3.5: Show here are the potentials used in the 2D WARM calculations. In the first case, the primary barrier to crossing from one well to the other is the height of the barrier relative to the strength of the "kicks" from the random force in the Langevin equation. In the second potential [33], the barrier to crossing between wells is entropic, in that a trajectory which results in a transition between wells must find its way through the gap at the center, i.e. the likelihood of a transition is not limited by any sort of uphill battle, but instead by decreased degeneracy in the number of possible trajectories which result in a transition.

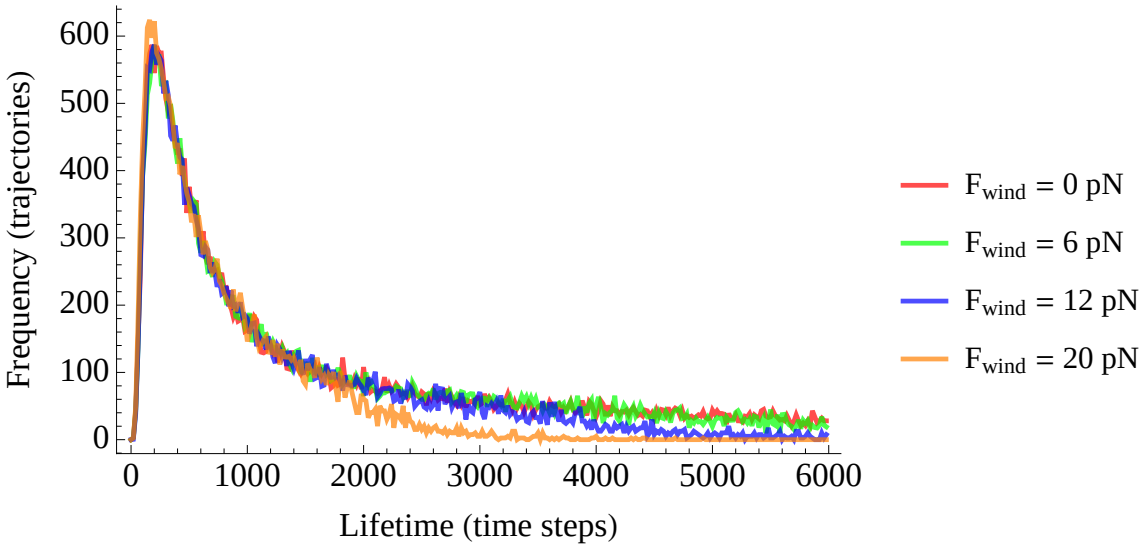

Figure 3.6: Shown in this figure is the transition probability distribution $K_{12}(\tau)$, i.e. the transition probability from milestone 1 (the line $x = -1$) to milestone 2 (the line $x = 0$) on the the 2D potential with the energetic barrier as a function of lifetime, calculated using $\mathcal{F}_{wind}$ forces ranging from 0 to 12 pN. The plots indicate that the rapid decrease in computation time due to the added *wind* force has almost no effect on accuracy.

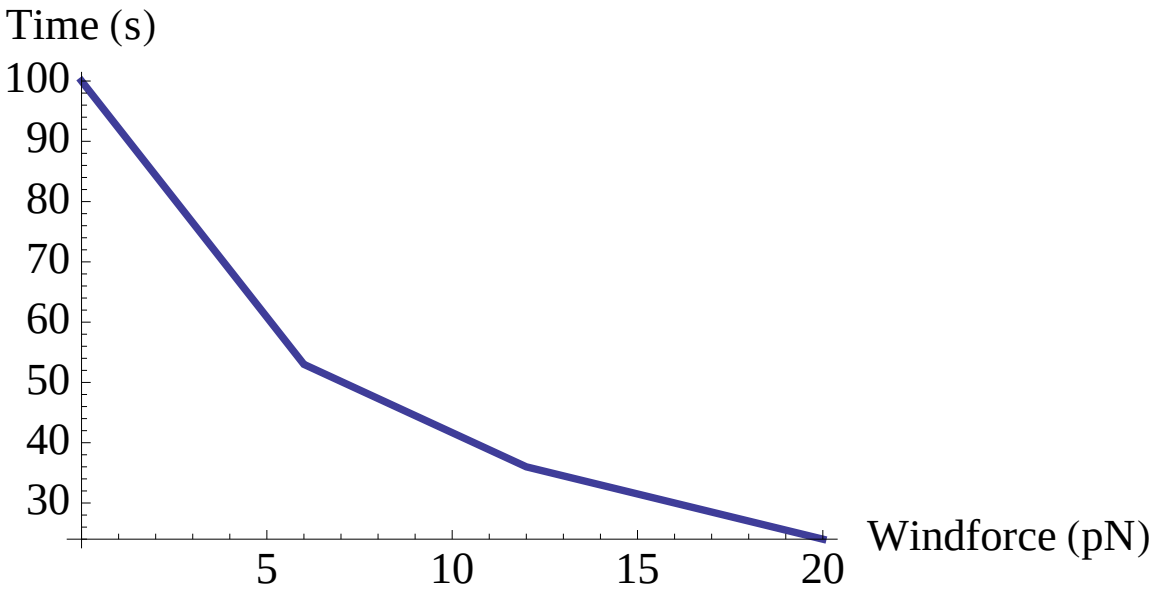

Figure 3.7: Shown here is calculation time as a function of the $\mathcal{F}_{wind}$ force for all $K(\tau)$ distributions in both directions for two subspaces ranging from -1 to 1 on the x axis of the 2D potential with the energetic barrier. All trajectories were run using $\beta = .123$. The highest value of $\mathcal{F}_{wind}$ yielded a faster computation time by a factor of 4.17 than the unassisted calculation with very little distortion to the $K(\tau)$ function.

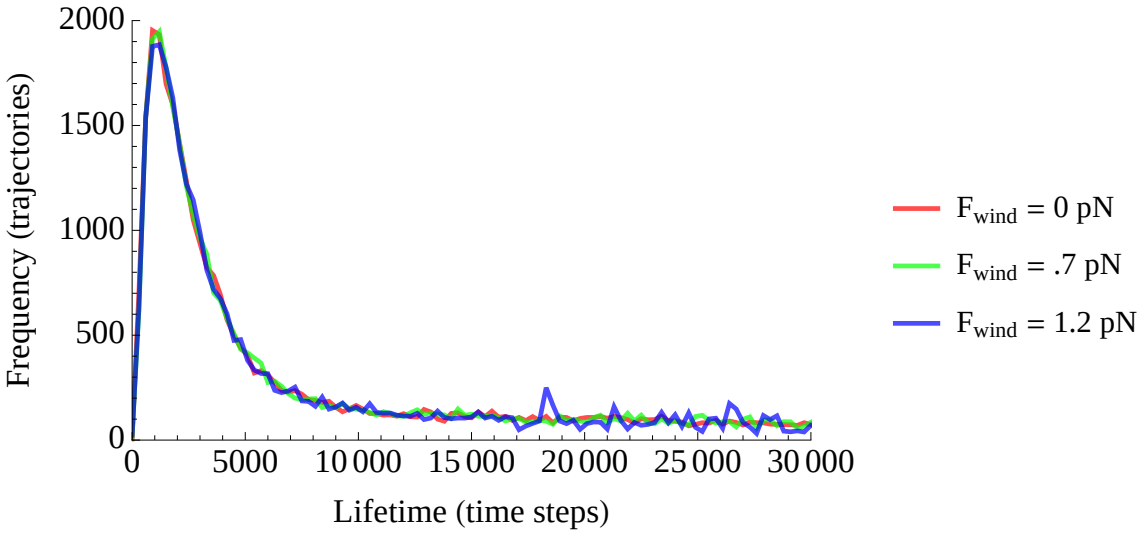

Figure 3.8: Shown in this figure is the transition probability distribution $K_{12}(\tau)$, i.e. the transition probability from milestone 1 (the line $x = -.5$) to milestone 2 (the line $x = 0$) on the the 2D potential with the entropic barrier as a function of lifetime, calculated using $\mathcal{F}_{wind}$ forces ranging from 0 to 1 pN. The plots indicate that the rapid decrease in computation time due to the added $\mathcal{F}_{wind}$ force has almost no effect on accuracy.

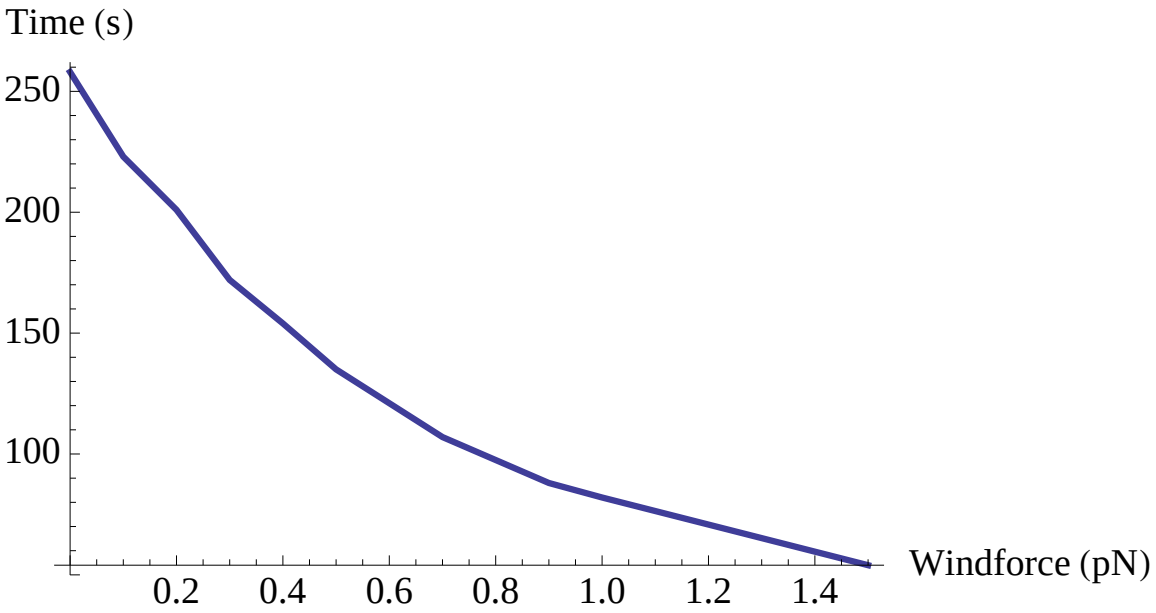

Figure 3.9: Shown here is calculation time as a function of the $\mathcal{F}_{wind}$ force for all $K(\tau)$ distributions in both directions for two subspaces ranging from -.5 to .5 on the x axis of the 2D potential with the entropic barrier. All trajectories were run using $\beta = 3.0$ so as to ensure that transitions over the barrier instead of through the small gap were highly unlikely. The highest value of $\mathcal{F}_{wind}$ yielded a faster computation time by a factor of 4.78 than the unassisted calculation with almost no distortion to the $K(\tau)$ function.

.

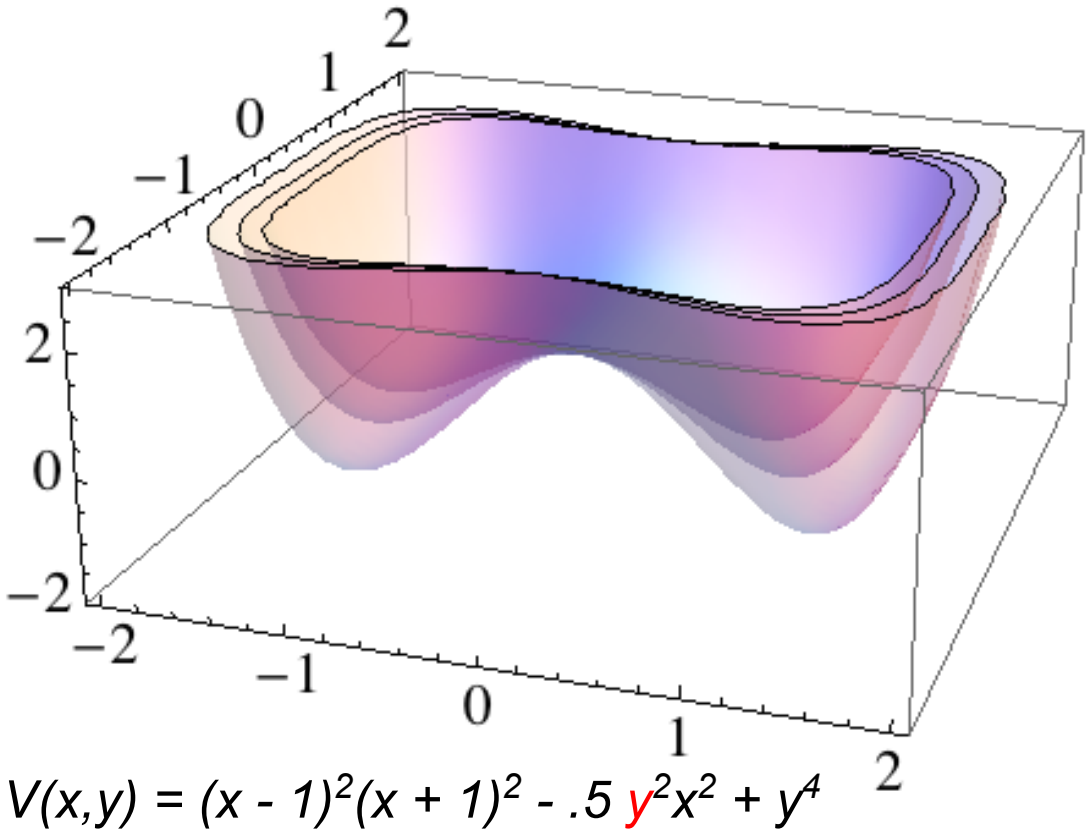

Figure 3.10: Show here is a 2D representation of the 11D coupled potential. The y in the second term (red) has been left as a parameter in this plot. The surfaces shown are for values for the parametric y of $0, \pm 1$, and ± 1.5, where the deepest well corresponds to $y = 1.5$ and the shallowest corresponds to parametric $y = 0$. Just as the well becomes deeper, the further from the system wanders from the origin in the y direction in this 2D model, the 11D system also encounters deeper wells in the x_n dimensions the further it wanders from the origin in each x_n dimension.

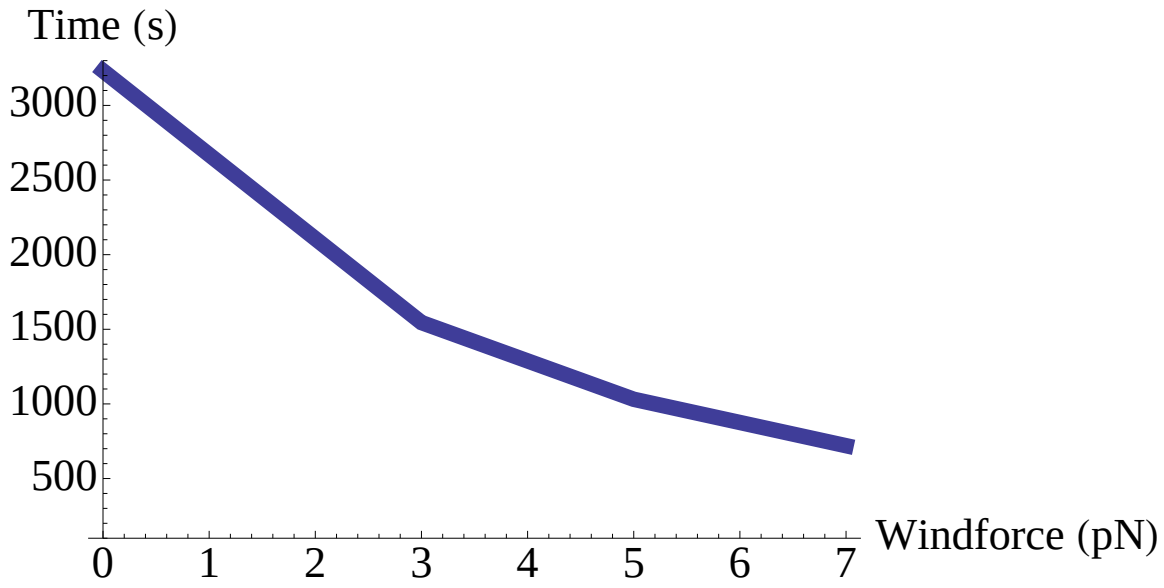

Figure 3.11: This plot shows CPU time as a function of the magnitude of the $\mathcal{F}_{wind}$ in 11D. The maximum speedup measured was a factor of 4.5.

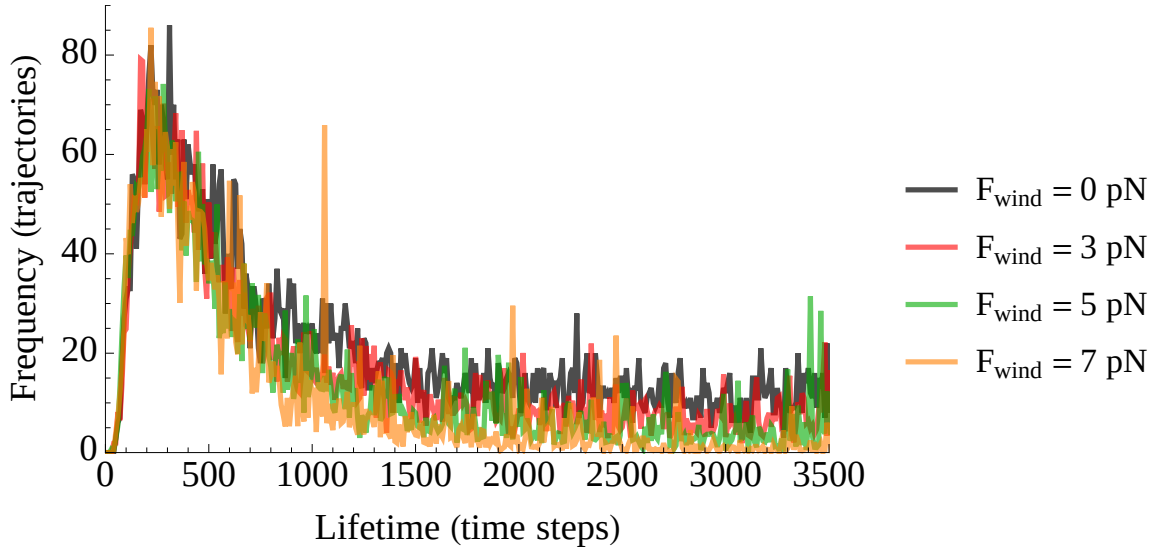

Figure 3.12: Shown in this figure are the $K(\tau)$ functions generated for each data point in the CPU time vs. $\mathcal{F}_{wind}$ plot for the 11D system.

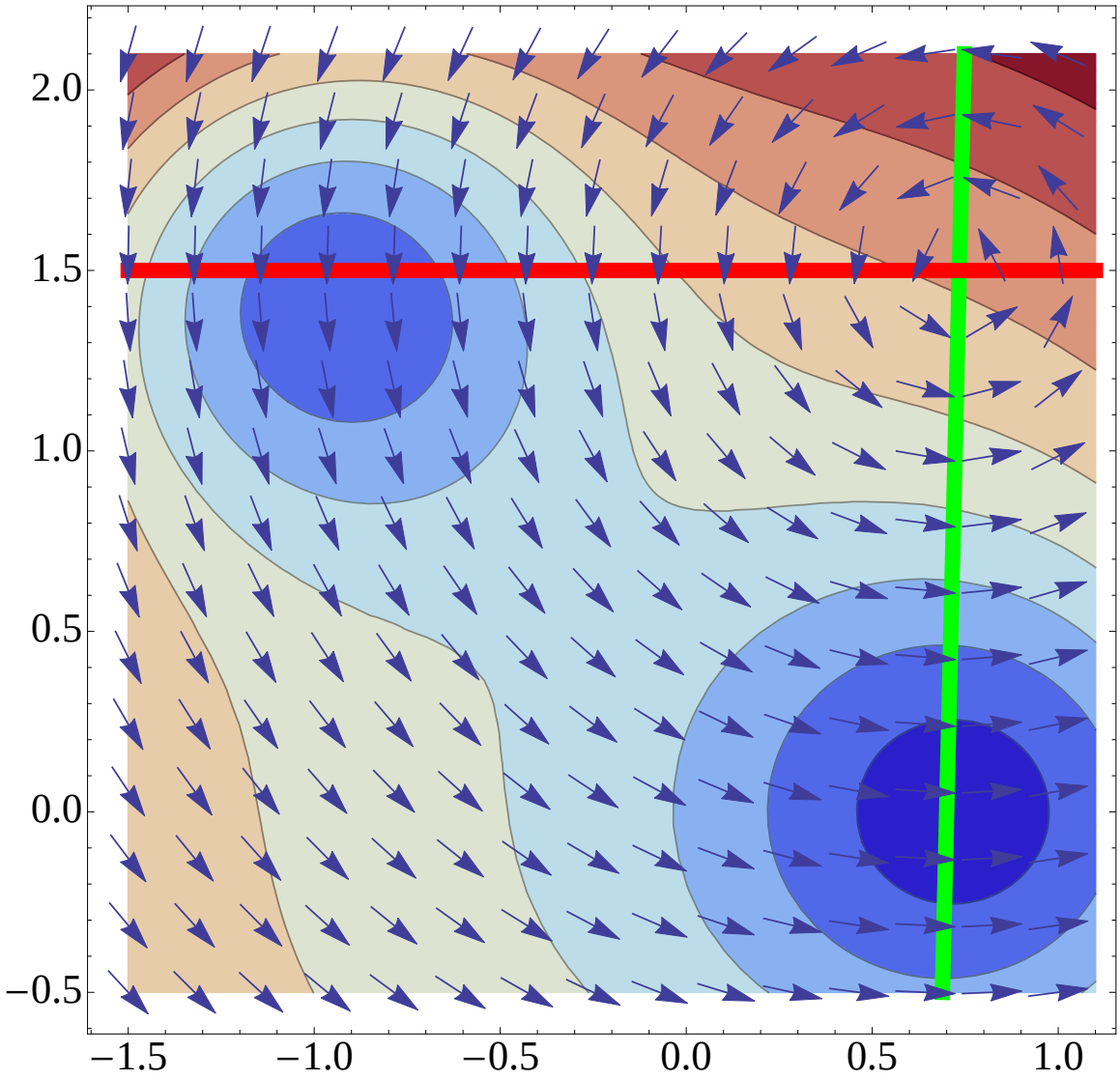

Figure 3.13: Shown here is a representation of the vector field approach to applying $\mathcal{F}_{wind}$ to push milestoning trajectories between two nearly orthogonal planes, subject to our Gaussian potential. The green milestone is defined as the plane for which $\frac{y}{44} - x = -.7$ and the red milestone is defined as the plane for which $y = 1.5$. The vector wind is configured to show the $\mathcal{F}_{wind}$ scheme for accelerating trajectories going from red to green.

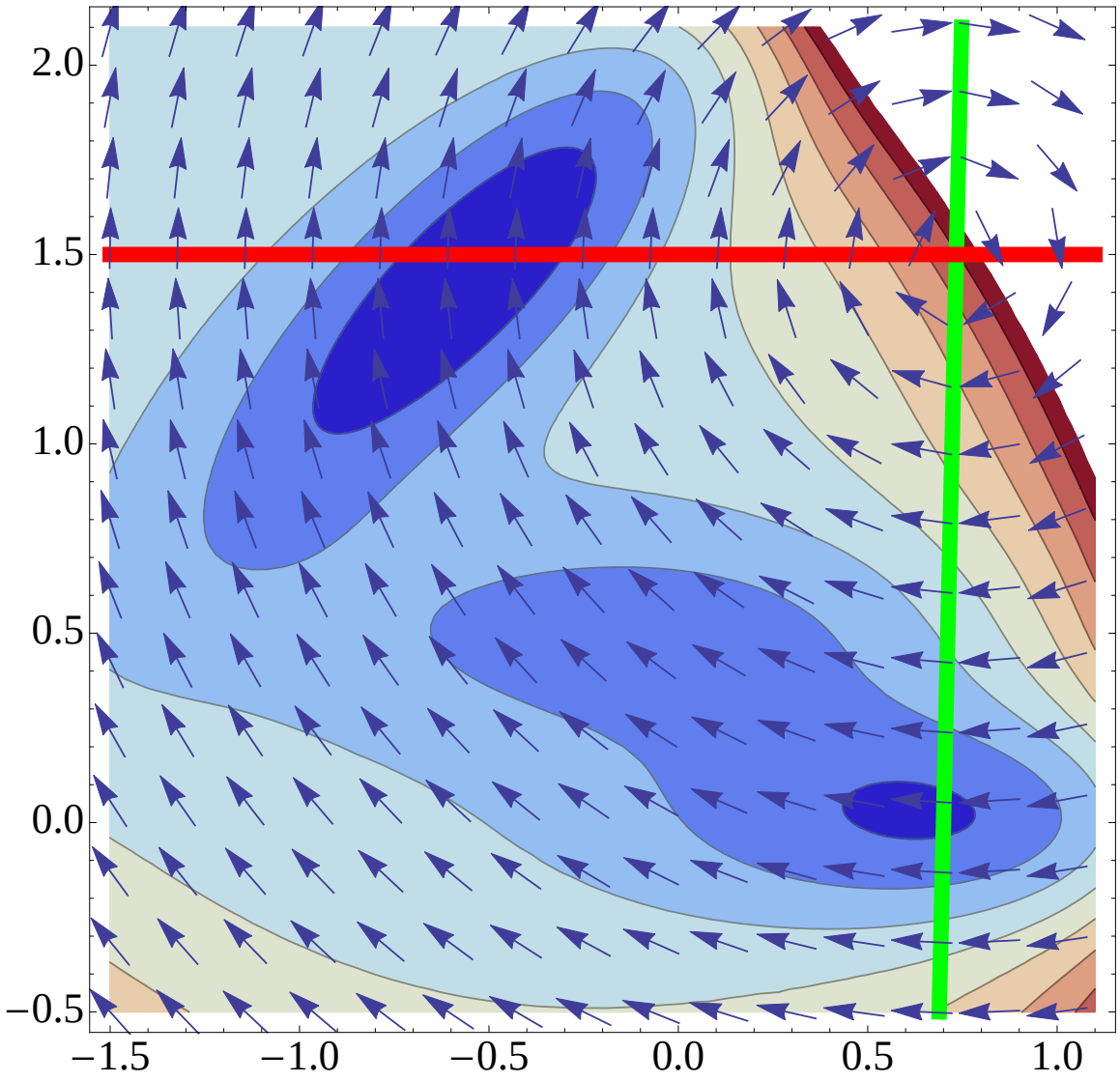

Figure 3.14: This plot shows the same milestone placement and $\mathcal{F}_{wind}$ scheme as the Gaussian potential example applied to the Muller potential and with a directionality for accelerating trajectories from the green milestone to the red one.

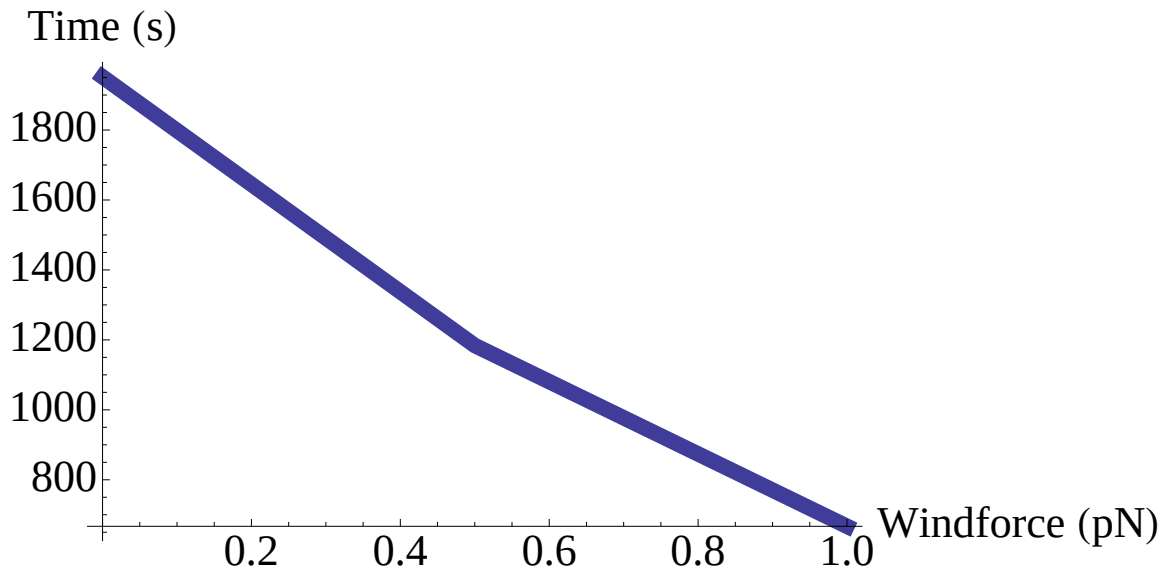

Figure 3.15: This plot shows CPU time as a function of $\mathcal{F}_{wind}$ magnitude for the Gaussian potential.

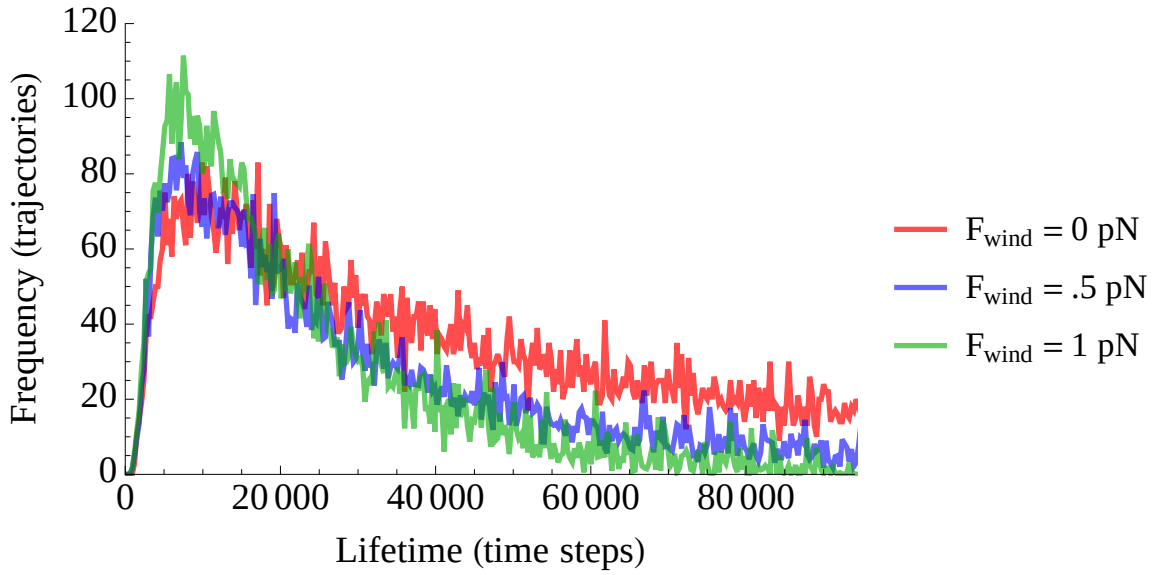

Figure 3.16: This plot shows the $K(\tau)$ functions corresponding to different magnitudes of $\mathcal{F}_{wind}$ as applied to the Gaussian potential.

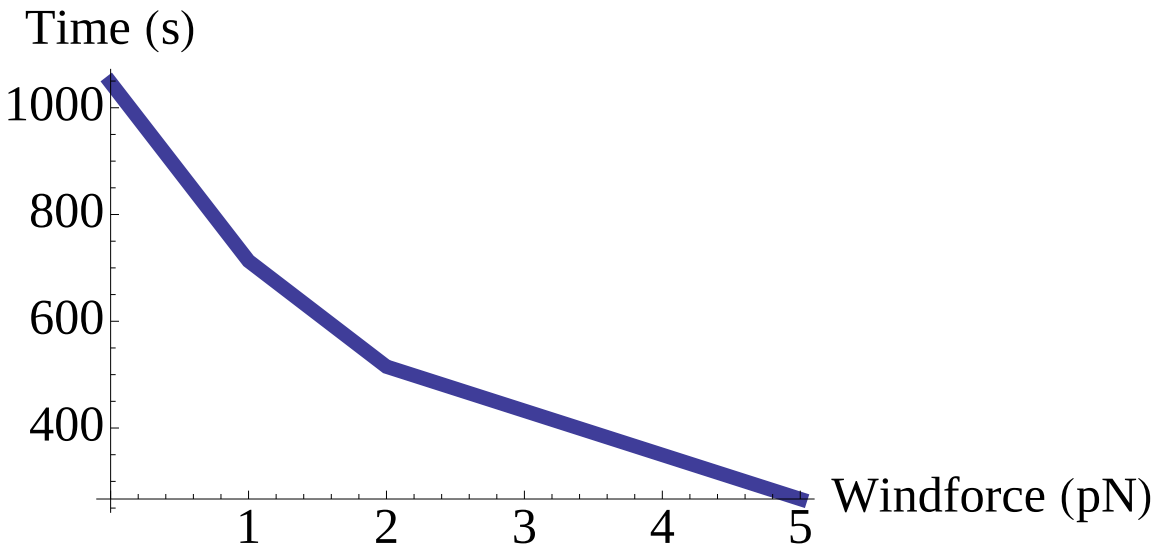

Figure 3.17: This plot shows CPU time as a function of $\mathcal{F}_{wind}$ magnitude for the Muller potential.

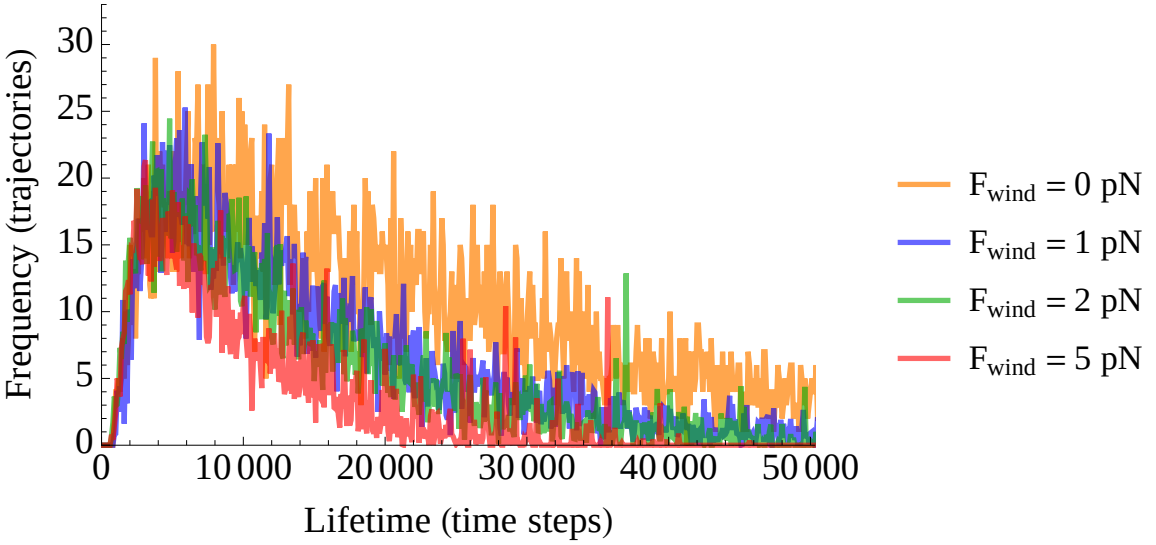

Figure 3.18: This plot shows the $K(\tau)$ functions corresponding to different magnitudes of $\mathcal{F}_{wind}$ as applied to the Muller potential.

Chapter 4

Advancements in Milestoning II: Calculating Autocorrelation from Milestoning Data Using Stochastic Path Integrals in Milestone Space

4.1 Introduction

The calculation of time correlation functions from time series measurements made along molecular dynamics trajectories plays the same central role in kinetics as calculating partition functions from sets of molecular configurations and their respective energies in the realm of thermodynamics. To put the magnitude of this task into perspective, consider a simple system where 100 different configurations are possible, and a transition between any pair of these configurations is possible. In this simple system, there are over 1.7×10^{13} different 10 step trajectories possible (100 choose 10) without even considering the fact that the same

series of 10 configurations can occur with different transition times which makes the number of possible trajectories proliferate even further! All important experimental properties can be calculated from time correlation functions measured from molecular dynamics simulations, but these effects are typically only measurable on timescales which are out of reach for brute force molecular dynamics. An example would be calculating RDCs (Residual Dipole Couplings) from NMR experiments from bond vector time correlation functions. The challenge of and demand for calculating kinetic properties from molecular dynamics simulations have caused it to become a major growth area in chemical physics [29] [28], leading to the development of several methods, spanning from early treatments using transition state theory (TST) [37] [121], to more recently, transition path sampling (TPS) [16], transition path theory (TPT) [71], and transition interface sampling (TiS) [113]. A common strategy in measuring kinetics in molecular dynamics simulations is the measurement of fluxes of trajectories through hyperplanes in phase space or configuration space [33] [112] . More recently, the use of the hyperplanes in the Milestoning method has been generalized to subdividing phase space into Voronoi cells, where the milestones exist as the interfaces between cells [115]. Thus far, Milestoning has been used to calculate many useful properties, such as equilibrium flux values through the set of milestones, rate constants [120], and other equilibrium properties such as mean first passage times between states [13], but the method has never before been used to calculate non-equilibrium dynamical objects such as time correlation functions. In our first paper, Advancements in Milestoning I, we introduced a methodology for rapid calculation of transition time density functions between milestone hyperplanes, the central objects of milestoning calculations, by artificially pushing the system toward the target milestone and then re-weighting the distribution to recover the true transition time distribution [46]. In this paper, we venture into this realm by introducing a method for calculating time correlation functions from milestoning data. In order to calculate autocorrelation from milestoning data, not only must we know the equilibrium flux values through each interface, we must also know the flux through each interface as a function of time and initial

configuration. For this reason, it was necessary that we also introduce our stochastic path integral approach to calculating the time-dependent fluxes, in addition to the methodology for calculating time correlation functions from these time-dependent fluxes.

4.2 Theory

Milestoning Theory

A more in-depth overview of milestoning theory can be found in our first paper [46], or in [120], but let us review a few of the key premises upon which our method for calculating time correlation functions hinge. The quantity of most fundamental importance in milestoning is the flux through a given milestone, for which the equation is [33]:

$$P_s(t) = \int_0^t Q_s(t') \left[1 - \int_0^{t-t'} K_s(\tau) d\tau \right] dt',$$

$$Q_s(t) = 2\delta(t) P_s(0) + \int_0^t Q_{s\pm1}(t'') K_{s\pm1}^{\mp}(t - t'') dt'' \tag{4.1}$$

where $P_s(t)$ is the probability of being at milestone s at time t, (or, more specifically, arriving at time t' and not leaving before time t [33]), and $Q_s(t)$ is the probability of a transition to milestone s at time t. $K_s(\tau)$ indicates the probability of transitioning out of milestone s given an incubation time of τ, thus $\int_0^{t-t'} K_s(\tau) d\tau$ is the probability of an exit from milestone s anytime between 0 and $t - t'$, which makes $1 - \int_0^{t-t'} K_s(\tau) d\tau$ the probability of there *not*

being an exit from milestone s over that same time period. Since the probability of two independent events happening concurrently is the product of the two events, the equation for $P_s(t)$ is simply integrating the concurrent probabilities of arriving at milestone s and not leaving over the time frame from time 0 to t. Turning our attention towards the meaning of the first term, $Q_s(t)$, $2\delta(t)P_s(0)$, simply represents the probability that the system is already occupying milestone s at time $t = 0$, where the factor of 2 is present since the δ-function is centered at zero, meaning only half of its area would be counted without this factor. $Q_{s\pm1}(t'')$ is the probability that the system transitioned into one of the two milestones adjacent to s at an earlier time t''. $K^{\mp}_{s\pm1}(t-t'')$ is the probability of a transition from milestones $s\pm1$ into milestone s. Thus the second term of the second line of equation 14 is another concurrent probability: the probability of the system entering an adjacent milestone at an earlier time, and then transitioning into milestone s between time t and 0. It is important to note that all functions $P_s(t)$ and $Q_s(t)$ are calculated using the respective values of $K_s(\tau)$ between adjacent milestones, thus the set of $K_s(\tau)$ between all milestones of interest contains all the information needed to calculate kinetics using the milestoning method. It is also important to note that a K function between two milestones $x = A$ and $x = B$, $K_{AB}(\tau)$, is simply a probability distribution representing the lifetime for the system remaining in state A before transitioning to state B.

Time Correlation from Milestoning Data

This approach aims to glean the time correlation function $C(t)$ of an observable from Milestoning data. The key insight into this method is the approximation of the continuous configuration space, which we define as x, as a discrete space of milestone configurations. Although the formalism presented below requires that the equilibrium distribution of configurations occupied, $f(x)$, is known, any successful Milestoning simulation yields the equilibrium flux through the set of milestones, and so this set of fluxes will serve as the equilibrium

distribution of configurations in our discrete space. For the sake of clarity of notation, we will be limiting our derivation to observables which are a function of configuration x, but it should be noted that all developments presented herein can be easily generalized to observables which are a function of both position and velocity by considering our variable x as a phase space coordinate. We begin with the usual definition for a time correlation function for time-ordered measurements of an observable that is a function of configuration, $A(x;t)$, arising from the equilibrium distribution of configurations, $f(x)$:

$$C(t) = \langle A(x,0)A(x,t)\rangle = \int A(x_0,0)A(x,t)f(x)dx \tag{4.2}$$

where time t is the lag time between measurements. For time $t = 0$ the time correlation function has the lower limit $C(0) = \int A(x_0,0)A(x_0,0)f(x)dx = \langle A^2\rangle$, the variance. On the opposite extreme, given an infinite relaxation time, the mean value of x at time t will be equivalent to the mean at equilibrium, $\lim_{t\to\infty}\langle A(x,t)\rangle = \int A(x)f(x)dx$, which implies: $\lim_{t\to\infty} C(t) = \int A(x)\left(\int A(x)f(x)dx\right)f(x)dx = \int A(x)f(x)dx \int A(x)f(x)dx = \langle A\rangle^2$

So far, we have only discussed equilibrium probability distributions in configuration space, which we defined as $f(x)$, but let us now consider a time-dependent probability density function of configuration, which is a function of initial configuration $x(0)$. Keep in mind that time-dependent probability density functions such as these are the solutions to Fokker-Planck equations. Let us define this probability density function as $g(x,t;x_0,0)$, and express its mean value as a function of time and initial configuration, $\langle x(t,x_0)\rangle$, in the following manner:

$$\langle x(t,x_0)\rangle = \int x g(x,t;x_0,0)dx \tag{4.3}$$

Following suit, the expectation value of our observable A as a function of time can be written as:

$$\langle A(x,t;x_0,0)\rangle = \int A(x)g(x,t;x_0,0)dx \tag{4.4}$$

We can now substitute $\langle A(x,t;x_0,0)\rangle$ for $A(x,t)$ in the definition of a time correlation function:

$$C(t) = \int A(x)\left(\int A(x)g(x,t;x_0,0)dx\right)f(x)dx \tag{4.5}$$

As stated earlier in this section, our aim is to coarse grain the continuous configuration space of x into a discrete space of milestone configurations, from which we can calculate a time correlation function. Our first step in constructing this model will be to approximate the outermost integral in x with a sum over a discrete set of configurations $\{x_i\}$ multiplied by the equilibrium probability of finding the system in the configuration i. If we define the probability of the system being in configuration x_i at time t given an initial configuration x_0 as $P_i(t;x_0)$, then given that our system will reach equilibrium at infinite time regardless of initial configuration, the equilibrium probability can be expressed as $P_i(\infty)$. Thus we arrive at our first discrete approximation of time correlation:

$$C(t) \approx \sum_i A(x_i)P_i(\infty)\left(\int A(x)g(x,x(0),t)dx\right) \tag{4.6}$$

Our next task is to approximate the remaining integral in the equation with a sum over milestone states. Equation 4.4 gives us an expression for the mean value of $A(x)$ in a

continuous space, given an amount of time elapsed t and an initial configuration x_0. Now consider the case where x can only occupy discrete values from the set $\{x_s\}$. In this case, the integral in equation 4.4 is replaced by a sum in a weighted average expression where each discrete value of x_i multiplied by its statistical weight as a function of time:

$$\int A(x)g(x, x(0), t)dx \approx \sum_s A(x_s)P_s(t|x_0) \tag{4.7}$$

Next, we substitute this weighted sum approximation into equation 4.6:

$$C(t) = \sum_i \left(A(x_i)P_i(\infty) \sum_s A(x_s)P_s(t|x_0) \right) \tag{4.8}$$

Note that we have now arrived at a complete expression for a discrete approximation of time correlation, with the assumption that $P_s(t|x_0)$ and $P_i(\infty)$ can be obtained from milestoning calculations. Since the set of equilibrium fluxes, $P_i(\infty)$, have been calculated from milestoning simulations since the beginning, and we will introduce a novel method for calculating $P_s(t|x_0)$ from milestoning simulations in the Random Walk / Path Integral Methodology subsection later in the article, we are able to demonstrate that time correlation can indeed be calculated from Milestoning simulations.

4.3 Analytical Solution for 1D Harmonic Oscillator

In this section, we demonstrate the effectiveness of equation 4.8 in approximating the time correlation function for diffusion in a harmonic potential, for which there is an analytical solution. Our potential is defined as $V(x) = \frac{1}{2}kx^2$, and it's equilibrium distribution in x is the

Boltzmann distribution, $f(x) = e^{-\beta V(x)}$. The closed form expression for the time-dependent probability distribution for diffusion in a harmonic well is [1]:

$$p(x,t|x_0,0) = \frac{1}{\sqrt{2\pi k_B T S(t)/k}} \exp\left[-\frac{\left(x - x_0 e^{-2t/\bar{\tau}}\right)^2}{2k_B T S(t)/k}\right] \tag{4.9}$$

where $S(t) = 1 - e^{-4t/\bar{\tau}}$ and $\bar{\tau} = 2k_B T/kD$.

Given this analytical expression for $p(x,t,|x_0,0)$, we can obtain an analytical expression for $C(t)$ by substituting $p(x,t,|x_0,0)$ into equation 4.5 for $g(x,x_i(0),t)$ and integrating. This yields the exact time correlation function $C(t)$ for diffusion in a harmonic potential:

$$C(t) = \frac{2\sqrt{\pi}e^{-\frac{2t}{\bar{\tau}}}}{\left(\frac{k}{k_B T}\right)^{3/2}\sqrt{\frac{k_B T\left(1-e^{-\frac{4t}{\bar{\tau}}}\right)}{k}}\sqrt{\frac{k\left(\coth\left(\frac{2t}{\bar{\tau}}\right)+1\right)}{k_B T}}} \tag{4.10}$$

Alternatively, we can apply equation 4.8, and obtain a general closed form expression for approximating $C(t)$ by summing over a discrete configuration space of N milestones rather than integrating over a continuous one:

$$C(t) = \frac{1}{\sqrt{\frac{2\pi k_B T\left(1-e^{-\frac{4t}{\bar{\tau}}}\right)}{k}}} \sum_{i=1}^{N} x_i P_i(\infty)\Delta x \sum_{j=1}^{N}\left(x_j Q_{ji}(t)\Delta x + x_i Q_{ii}(t)\Delta x\right) \tag{4.11}$$

where

$$Q_{ji}(t) = \exp\left(-\frac{k\left(\coth\left(\frac{2t}{\bar{\tau}}\right)-1\right)\left(x_i - x_j e^{\frac{2t}{\bar{\tau}}}\right)^2}{4k_BT}\right)$$

$$Q_{ii}(t) = \exp\left(-\frac{x_i^2 k \tanh\left(\frac{t}{\bar{\tau}}\right)}{2k_BT}\right) \quad (4.12)$$

and Δx is the distance between the evenly spaced milestones. $Q_{ji}(t)$ represents the discrete time-dependent probability density as a function of time that our system is in configuration x_i at time t, given that the system was in state x_j at time $t = 0$. Likewise, $Q_{ii}(t)$ is the discrete probability density as a function of time that our system is still in configuration x_i at time t if it started in configuration x_i at time $t = 0$. Thinking in terms of the assumption of Markov statistics for transitions between milestones inherent to the Milestoning method, it makes sense that these probabilities are added given that we are interested in the outcome of finding our system in configuration x_i whether it was already there, or it arrived there from another configuration.

The most straightforward and intuitive way to compare equations 4.10 and 4.11 is to plot them. In figure 4.1, we can compare the exact time correlation function for diffusion in a harmonic potential (with parameters $\beta = .35, k = 5,$ and $D = .2857$) with the approximate $C(t)$ generated using equation 4.11. Discretizing the space to three milestones is clearly too coarse of an approximation, but the gain in accuracy in going from 6 to 9 milestones is quite modest. As one might expect, the discrete approximation of the time correlation function is most accurate for long times and least accurate for $C(0)$. It turns out that this sacrifice in accuracy is a meager one because $C(0)$ is always available from Milestoning data because it is equivalent to the sum approximation of the variance in configuration space at equilibrium, $\sum_{i=1}^{N} x_i^2 P_i(\infty)$. This will be leveraged to our advantage in the following section.

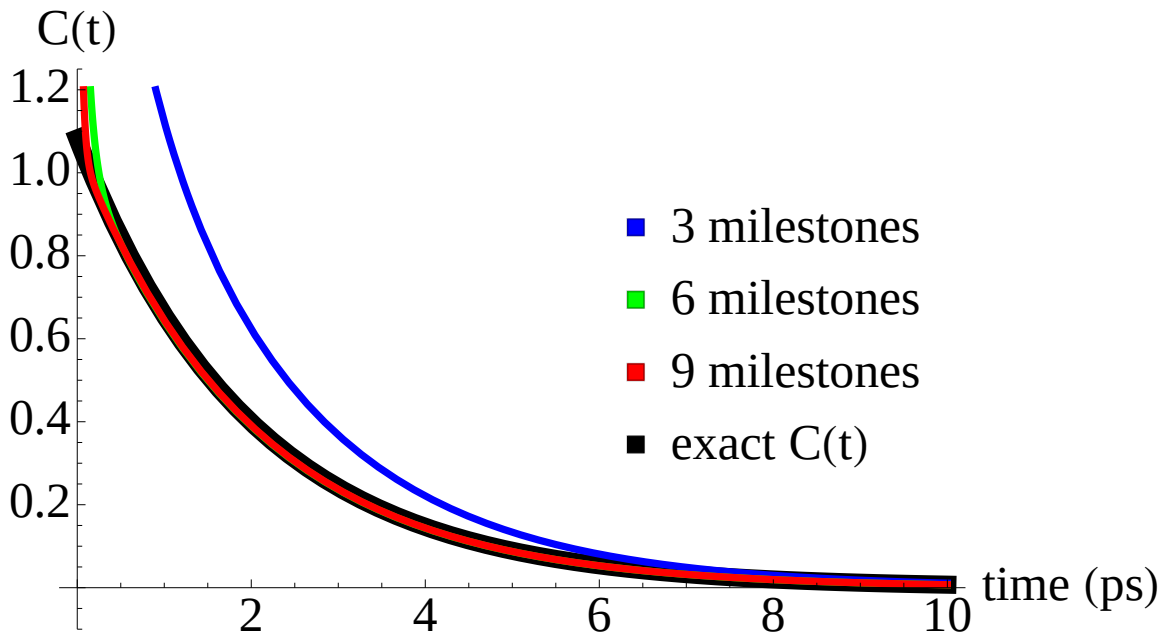

Figure 4.1: This figure shows the approximate time correlation functions calculated using equation 4.8 for 3, 6, and 9 milestones overlaid on top of the exact analytical function $C(t)$.

4.4 Numerical Demonstration

1D Fokker-Planck Diffusion on a Bistable Potential

In order to further validate the approach of calculating time correlation functions using the nested sum in equation 4.8 in a discrete configuration space to approximate integrating equation 4.2 in continuous conformation space, the method was applied to a simple two well potential of equation $y = (x-1)^2(x+1)^2$, where the time evolution of the probability density function in configuration space was calculated using a Fokker-Planck formalism:

$$\frac{\partial \rho(x,t)}{\partial t} = D\frac{\partial^2 \rho(x,t)}{\partial x^2} + \frac{D}{k_B T}\frac{\partial}{\partial x}(\rho(x,t)\frac{\partial V}{\partial x}) \tag{4.13}$$

By repeatedly solving equation 4.13 numerically with the using the *Mathematica* software package [125], using a normalized Gaussian distribution centered at the various $x_i(0)$ values as the initial condition, the manifolds $g(x, x_i(0), t)$ were obtained for each of the 10 milestone configurations x_i in the set $\{-2, -1.6, ..., 1.6, 2\}$. These manifolds were then used to find $C(t)$ using both the intermediate method described by equation 4.6 (shown as red circles in figure 4.2) as well as our fully developed discrete method described by equation 4.8 (shown as blue circles in figure 4.2). In the case of the equation 4.6, the integral $\int xg(x, x(0), t)dx$ was numerically integrated directly, while in the case of equation 4.8, the manifold $g(x, x(0), t)$ was used to obtain values of $P_i(x(0), t)$ by multiplying $g(x, x(0), t)\Delta x$, similar to the transformation from equation 4.6 to equation 4.8, but in reverse. The results are shown superimposed over a plot of the time correlation function for the system obtained in the traditional manner by running 10^9 steps of langevin dynamics and then calculating the time correlation function over this one long trajectory using the equation:

$$C(t) = \frac{1}{n-t} \sum_{i=1}^{n-t} x_i x_{t+i} \tag{4.14}$$

We would like to point out that, as we alluded to in the previous section, the data point for $C(0)$ is the only portion of the time correlation function approximated using equation 4.8 with any appreciable error. In practice, the data point for $C(0)$ can always be replaced with the value obtained from the sum $C(0) = \sum_i x_i^2 P_i(\infty)$ (shown as the green ring in figure 4.2), due to the fact that the set of equilibrium probabilities, $P_i(\infty)$ are always known from Milestoning simulations.

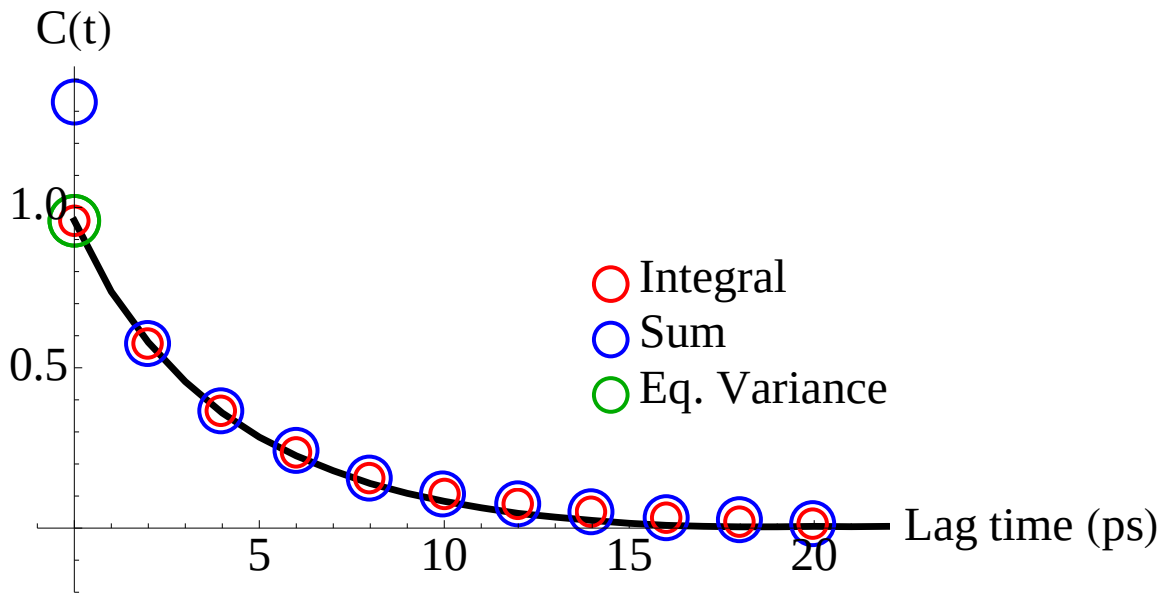

Figure 4.2: This plot demonstrates a successful implementation of our method for approximating time correlation functions in continuous space by summing over time dependent joint probabilities of transitions between discrete states, as obtained in Milestoning simulations. The red rings mark the data points from implementing equation 4.6, the blue data points indicate the positions where the full nested sum approximation of equation 4.8 was implemented, and the green ring is the data point for $C(0)$ calculated from equilibrium probabilities which is used to replace the value of $C(0)$ generated using equation 4.8. The data is shown superimposed over the time correlation function $C(t)$, represented by a solid black line, calculated using the traditional method of equation 4.14.

Random Walk / Path Integral Methodology

In order to make use of the formalism for obtaining autocorrelation in a discrete configuration space, as introduced in the Theory section, we require an expression for $P_s(t|x_i(0))$, i.e. the probability that our system is in configuration s at time t, given that it was in configuration i at time $t = 0$. Since previous implementations of the milestoning method have been "based on iterative determination of stationary flux vectors at milestones" [13], and not the determination of non-equilibrium time dependent fluxes given some initial configuration, it was necessary to devise a methodology for obtaining the function $P_s(t|x_r(0))$ from milestoning data. In the case of diffusive systems which can be described using a Fokker-Planck formalism (eq. 4.13), the Fokker-Planck equation can be solved for a manifold $\rho(x, t)$ which

represents a probability density of configurations evolving in time, where the distribution at time $t = 0$ is the distribution dictated by the initial condition and the distribution as $t \to \infty$ is equivalent to the equilibrium distribution in x. While this Fokker-Planck description can be directly solved for the time evolution of a probability density function of configurations (when tractable, as in figure 4.2), it is also possible to obtain the manifold $\rho(x, t)$ via a path integral approach using a large ensemble of trajectories generated using stochastic models such as Langevin dynamics. This equivalence was the inspiration behind the random walk / path integral method introduced in this section. There are some differences however, for example, instead of Langevin trajectories, we use random walks along the given set of milestones. Very long random walks, orders of magnitude longer than time scales accessible to molecular dynamics, can be quickly generated with minimal computational cost by taking advantage of two data sets which are already known in any milestoning calculation: the transition matrix $\mathbf{K}$ (essentially a Markov matrix) and the set of all $K_{AB}(\tau)$ functions, which are the probability density functions of transition times between milestone A and milestone B. The $K_{AB}(\tau)$ functions are obtained by histogramming transition times between milestones, and each element K_{ij} of the matrix $\mathbf{K}$ is obtained by integrating the distributions of transition times, $k_{ij}(\tau)$, over all time τ and then normalizing each row to impose the constraint that the system at state i has probability 1 of transitioning to one of the states to which it is coupled (j). Since the matrix $\mathbf{K}$ gives the equilibrium transition probabilities between milestones, and the k_{ij} functions are probability density functions for the transition time between connected milestones, these two pieces of information can be used to construct time-dependent random walks along a set of milestones. Each step taken from some current configuration i is chosen by selecting between each possible coupled state j, weighted by the transition probabilities from $\mathbf{K}$, next, the amount of time each selected transition from state i to j took is selected randomly from the distribution defined by $k_{ij}(\tau)$. In this manner, trajectories of arbitrary length in this discrete space can be very quickly generated in only the amount of CPU time necessary to select $2N$ random numbers, where N is the desired

number of steps in the random walk. Once a large set of these random walks is generated, they can be used to calculate discrete versions of the same $\rho(x,t)$ manifolds which would be obtained as the solutions to the Fokker-Planck equation (see figure 4.3). To elaborate on this, consider a single random walk along the milestone configurations. If, at each time step, we histogram the frequency with which our system has visited each milestone configuration up to that point in time into a normalized distribution, then we have constructed a discrete manifold in configuration space x and time t which represents the time evolution of the probability distribution of finding our system in a particular configuration for this particular realization of a random walk in our discrete configuration space. From here, it only remains to average the set of probability distributions generated from numerous manifestations of the random walk. An alternative approach to calculating time correlation functions from these random walks would be to "connect the dots" along the random walk using an interpolation method, and then use the traditional approach to numerically calculating time correlation, shown in equation 4.14, from the resulting continuous function, as shown in figure 4.5.

4.5 Application to Calculating Long-Time RDCs in Atomistic Simulations

Application of Discrete Space Time Correlation Methodology to the Alanine Dipeptide Bond Vector

In this section, we describe an application of our methodology to a molecular system. Shown in figure 4.6 is the molecular structure of our system, alanine dipeptide. After constraining the nitrogen and carbon atoms labeled in yellow to remain fixed at their initial positions, Langevin dynamics at $T = 300K$ was run for 4×10^7 time steps with a time step size of 0.001 ps for a total of 40 nanoseconds using the CHARMM molecular dynamics software package.

As the molecular dynamics simulation ran, the orientation of the bond vector extending from the center of the labeled nitrogen atom to the center of the hydrogen atom indicated by the purple arrow in figure 4.6 was recorded. Although this bond vector possesses three spatial degrees of freedom, it's orientation could be well approximated by a single rotational degree of freedom, as shown in figure 4.7. By counting the number of time steps between transitions from one milestone state to the next (shown graphically as the four colored planes in figure 4.7) over the course of the 40 nanosecond trajectory, probability distribution functions for the transition times between neighboring pairs were constructed as histograms to obtain the set of $k_{ij}(\tau)$ functions for each pair of neighboring milestone states. These $k_{ij}(\tau)$ functions were then used as the basis for the random walk / path integral approach described in the previous section. Thusly, the $P_s(t|x_0)$ functions necessary to calculate the time correlation function using equation 4.16 were calculated by averaging 75,000 different time-dependent probability distribution functions which each resulted from some particular manifestation of the random walk. The time correlation functions of interest for this system are those which can be calculated using the Lipari-Szabo formalism [67], as implemented by Xing and Andricioaei [126], using the equation:

$$C(t) = \langle L_2(\mathbf{u}(0)\mathbf{u}(t))\rangle \tag{4.15}$$

where $L_2(\mathbf{u}(0)\mathbf{u}(t))$ refers to plugging the scalar resulting from the dot product of time series measurements of the bond vector $\mathbf{u}$ into the second order Legendre polynomial. This motif of measuring the autocorrelation of this value is then applied to equation 4.8 to yield the

discrete space time correlation function relationship:

$$C(t) = \sum_i L_2 \left[\sum_s (\mathbf{u}_i(0) \cdot \mathbf{u}_s) P_s(t|\mathbf{u}_i(0)) \right] P_i(\infty) \tag{4.16}$$

where the vectors $\mathbf{u}_i$ represent the different possible values for the bond vector, given the coarse graining of the bond vector into a discrete space. The oscillatory and slower decay in correlation for the 4 milestone case is an effect of coarse graining the space. This is due to a loss in entropy in going from the continuous space to the discrete one, i.e. if only four possibilities exist for the position of the bond vector, the probability of pointing in the same direction as that of a previous time step increases compared to a system where 8 or more configurations are possible.

Notably, the oscillatory and slower decay in correlation for the 4 milestone case is an effect of coarse graining the space (the oscillations are reproduceable). This is due to a loss in entropy in going from the continuous space to the discrete one, i.e. if only four possibilities exist for the position of the bond vector, the probability of pointing in the same direction as that of a previous time step increases compared to a system where 8 or more configurations are possible.

4.6 Concluding Discussion

We have demonstrated for the first time that time correlation functions for continuous processes can be approximated using equation 4.8 to coarse grain the configuration space to a discrete one. Additionally, we have introduced a novel method for extending milestoning into non-equilibrium regimes by numerically calculating the time-dependent fluxes $P_s(t|x_i(0))$. The method consists of constructing random walks in the discrete configuration space, de-

fined by a set of milestone configurations, from transition time probability density functions $k_{ij}(\tau)$ obtained using the milestoning method, followed by calculating time-dependent histograms of milestone states occupied using the stochastic path integral method described in the Random Walk / Path Integral Methodology section.

The time correlation function for the harmonic oscillator calculated analytically using our discretization method showed excellent agreement with the true time correlation function $C(t)$, also obtained analytically, for a harmonic oscillator. There was also an excellent agreement between the $C(t)$ calculated for a discrete configuration space for a bistable potential and the true autocorrelation function, where $P_s(t|x_i(0))$ was obtained by numerically solving a Fokker-Planck equation. We also obtained a promising result from applying the discretization method of equation 4.16 in conjunction with the stochastic path integral method to an atomistic system. The autocorrelation function $C(t)$ for the bond vector calculated using the methods introduced herein showed a nice agreement with the true $C(t)$ calculated using equation 4.15. The limitations to the methods we have introduced appear to be limited to the challenges inherent to implementation of the milestoning method. A key advantage of our method is that the random walks between discrete configurations can be constructed at trivial computational cost, allowing for us to make predictions well into time regimes inaccessible to molecular dynamics simulations. We would like to note that, although the calculations described in this article were performed on systems where the observable of interest was constant along each milestone hyperplane, the method can easily be generalized for systems where the observable varies along each milestone hyperplane. In order to account for such observables, one must simply construct equilibrium probability distributions of the observable on each hyperplane, then select from this distribution at each time step of the random walk along the milestones. In other words, at each step, the algorithm must first choose the next step to take using the transition matrix, then select the transition time from the appropriate transition time distribution function, then select the value of the observable from the probability distribution describing the observable along that hyperplane. We

feel that the methods introduced in this paper have the potential to allow for the calculation of experimental observables from molecular dynamics simulations that are currently unattainable by brute force long time simulations. The method presented herein could also be further enhanced by combining it with the enhanced sampling methodology introduced in the companion article to this paper, also found within this publication [46].

4.7 Acknowledgments

IA acknowledges funds from an NSF CAREER award (CHE-0548047).

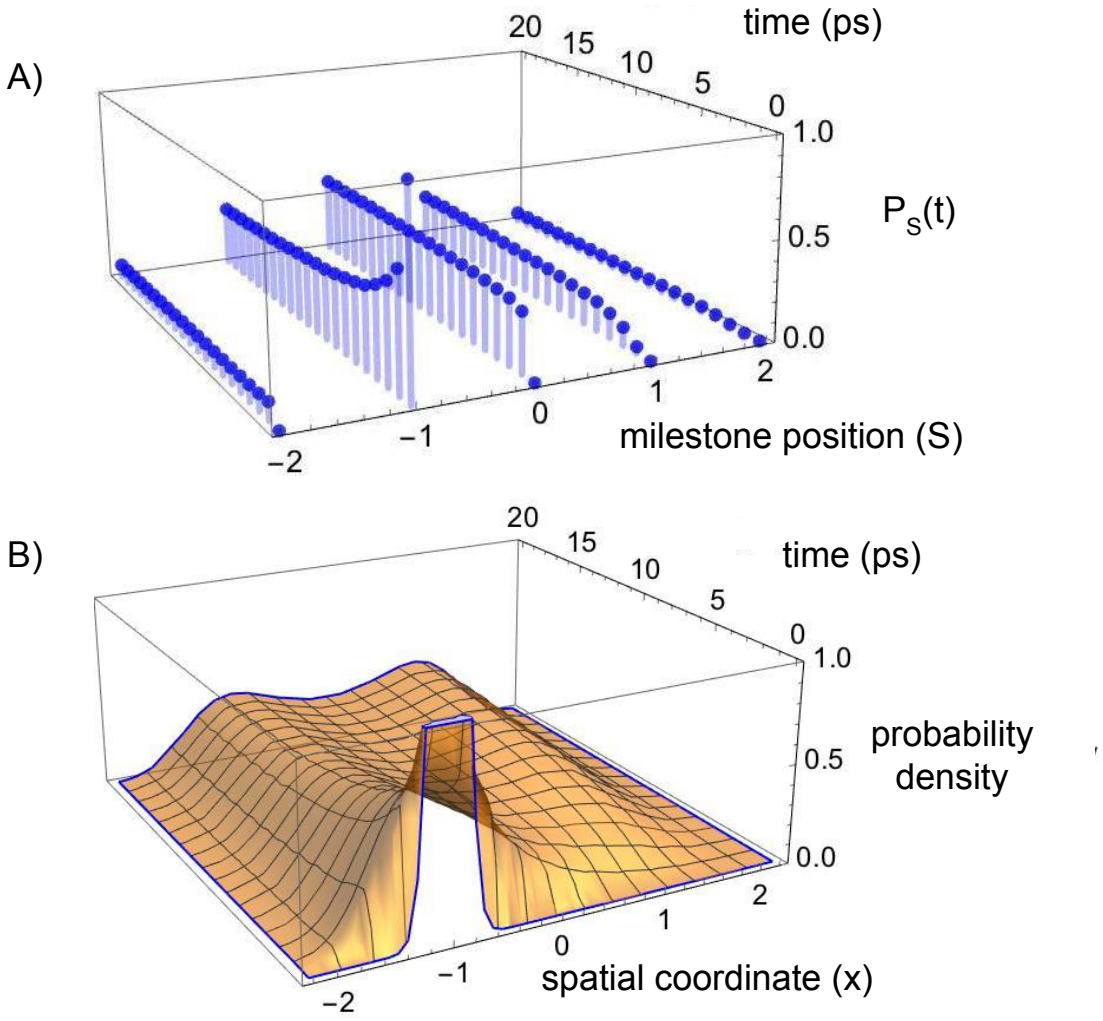

Figure 4.3: This figure shows a graphical comparison between the time evolution of a discrete probability distribution for a set of 5 milestone configurations subjected to the two well 1D potential found in the Numerical Demonstration section using our random walk / path integral methodology (part A), and the manifold representing the time evolution of a continuous probability density function of configurations for the same two well system subjected to Fokker-Planck diffusion (part B). Part A is the set of probabilities as a function of time for the system being found at each milestone configuration, given that the system was in configuration $x = -1$ at time $t = 0$, and part B shows Fokker-Planck diffusion on the same two well system. Note that the random walk in part A began at the milestone located at $x = -12$, thus we see a decay from $\{P_1(0) = 0, P_2(0) = 1, P_3(0) = 0, P_4(0) = 0, P_5(0)\}$ to the equilibrium distribution, the same way our initial continuous distribution, a normalized Gaussian centered at -1, decays to the equilibrium probability distribution predicted by the Bolzmann distribution for the two well potential, and both evolve in time on about the same time scale.

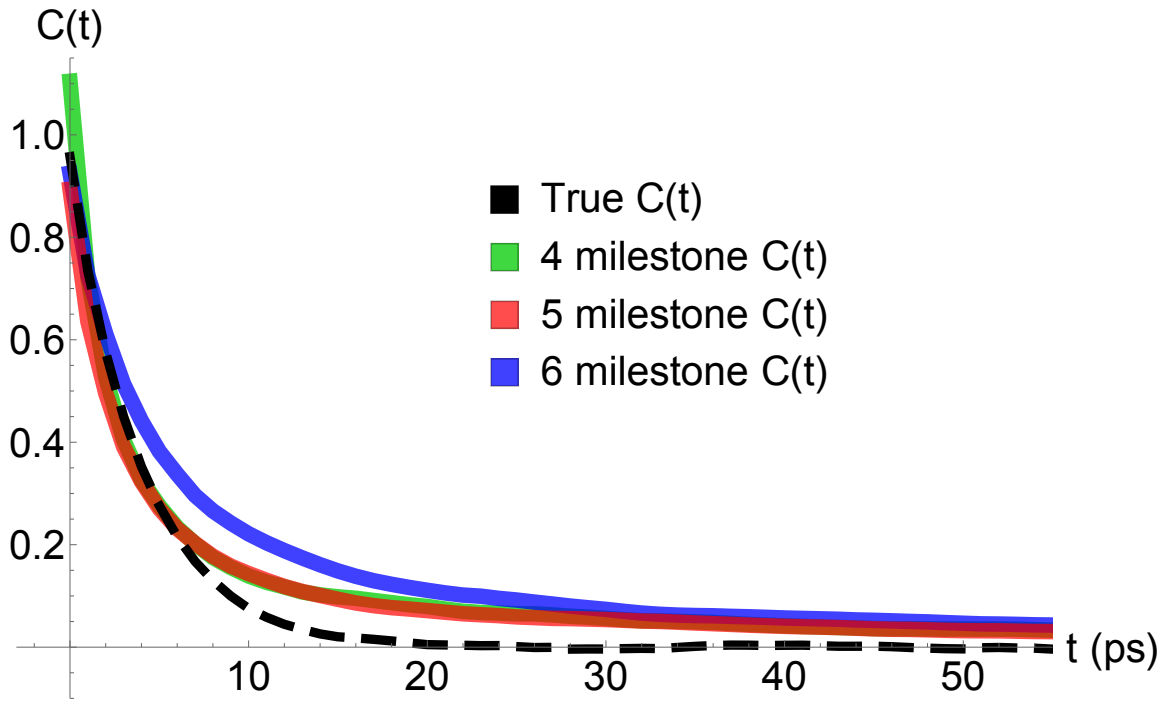

Figure 4.4: Shown here are time correlation functions calculated using equation 4.8, where the conditional probability as a functions of time, $P_s(t|x(0))$, are calculated using our random walk / path integral methodology, represented graphically in figure 4.3A.

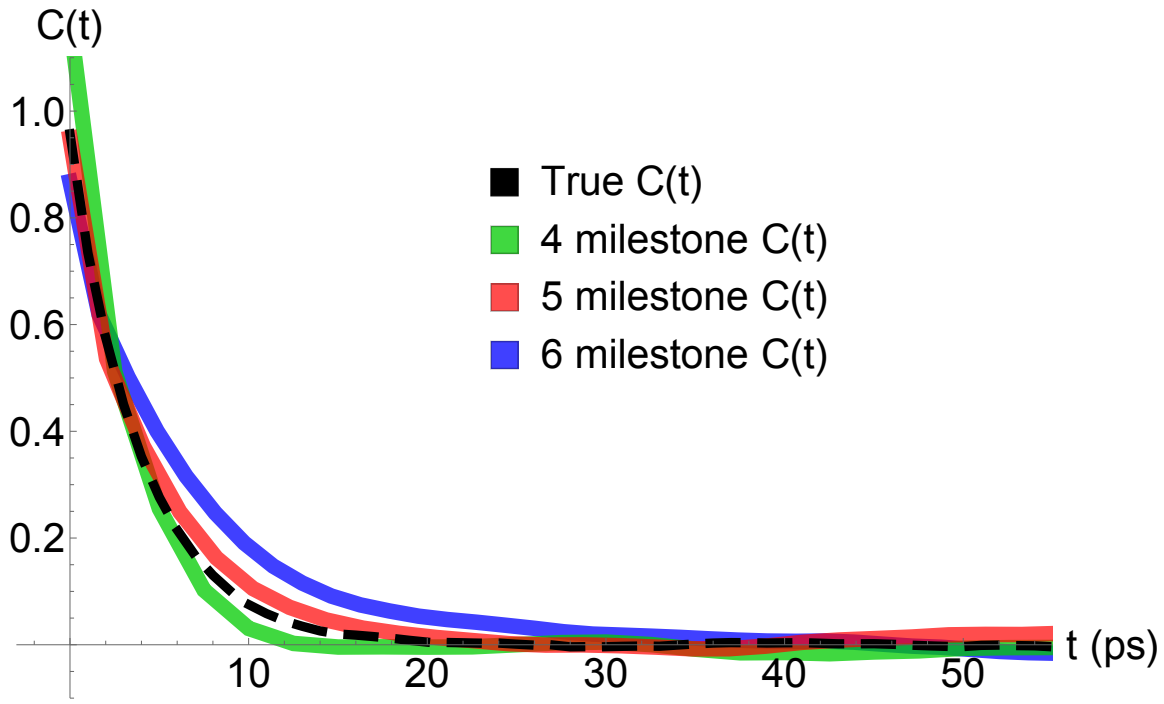

Figure 4.5: Shown here are time correlation functions which were calculated by first generating one long random walk using the method introduced in this article, then linking each point in the trajectory using linear interpolation, and finally using equation 4.14 to calculate $C(t)$.

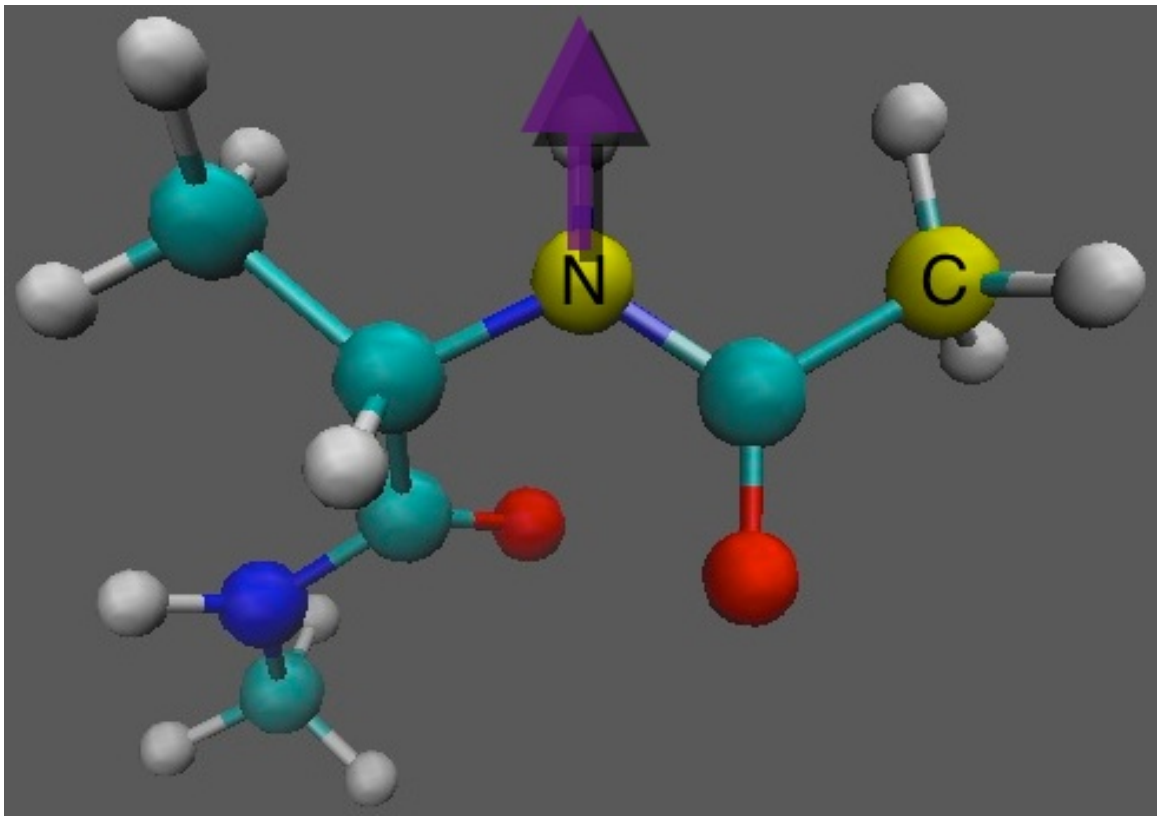

Figure 4.6: Shown in this figure is the alanine dipeptide molecule used as our model system. The two atoms shown in yellow were held fixed in space while the rest of the molecule was subjected to Langevin dynamics. The purple arrow gives the orientation of the bond vector which served as the measurable in our time correlation function calculations.

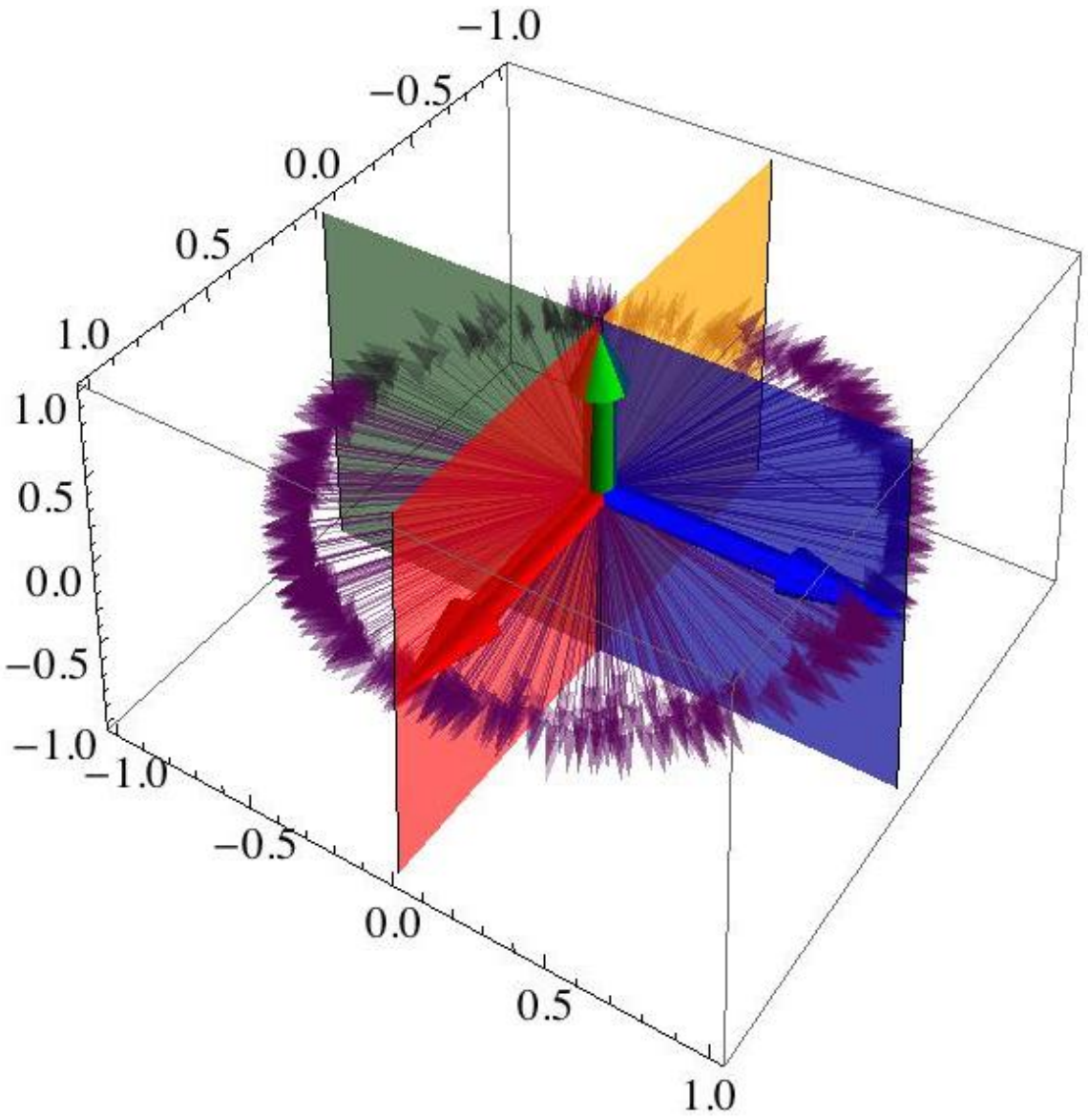

Figure 4.7: Shown here is a graphical representation of the four milestone configuration for measuring the time correlation function of the alanine dipeptide bond vector. Although the bond vector, shown as many thin, purple arrows, posses three degrees of freedom as it fluctuates in time, we are able to choose a frame of reference where the bulk of the motion is taking place as a rotation about the z-axis, shown as a thick green arrow. Using the four milestones, shown as the red, green, yellow, and blue planes, we can calculate transition time probability distributions between each pair of adjacent milestones.

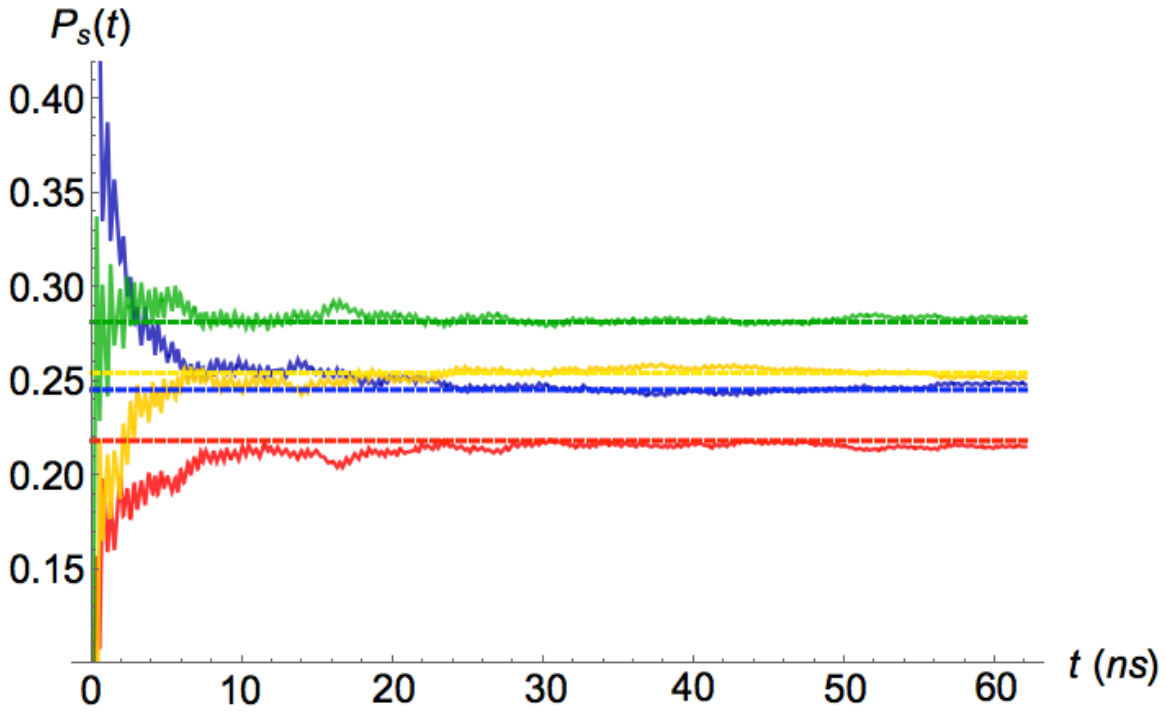

Figure 4.8: This plot gives the probability of finding our system in each of the four milestone configurations as a function time, given that we began the simulation with our system in the configuration shown as the blue plane, using the same color scheme as in figure 4.7. The probability of being found in the blue milestone is equal to 1 at time $t = 0$ of course, but the plot range stops shy of $P_s(t) = 1$ in order to provide a more detailed view. Note that the probability of the system being in any of the other three milestone configurations is equal to zero at time $t = 0$, as expected. These functions were calculated using the methodology described in the Random Walk / Path Integral Methodology section. These functions contributed to the calculation of $C(t)$ shown in figure 4.9. Note that the probabilities converge to their equilibrium values on roughly the same timescale that $C(t)$ converges to its long time value.

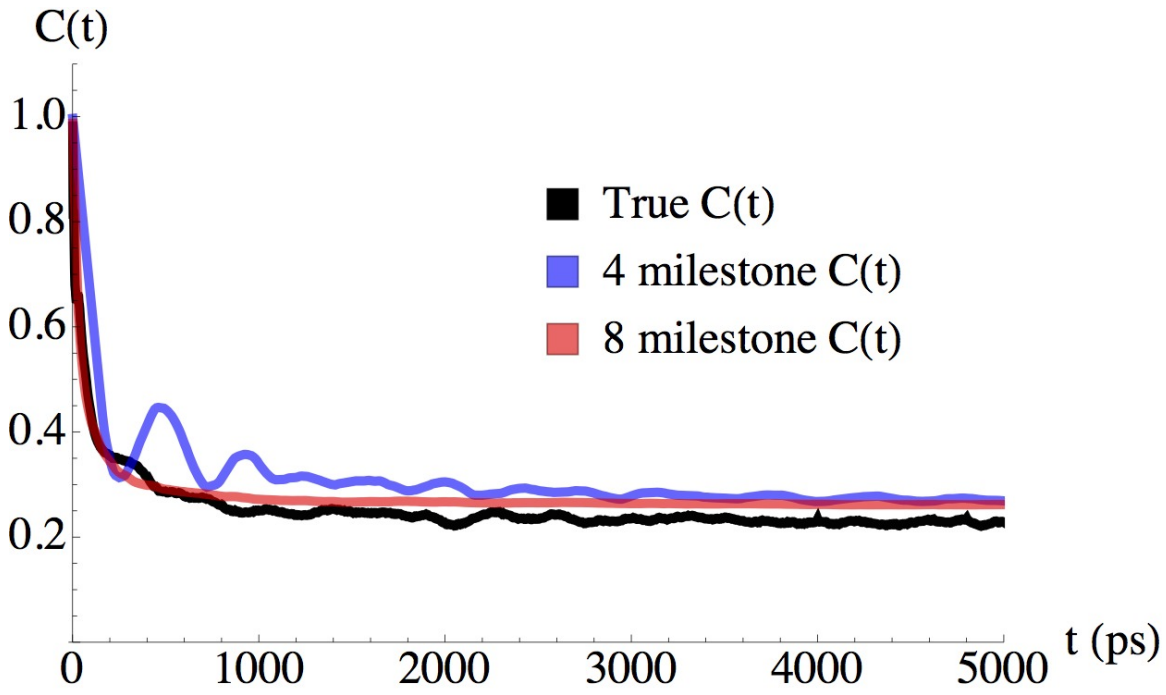

Figure 4.9: This figure shows the approximate time correlation functions calculated using equation 4.8 superimposed over the true time correlation function, calculated using equation 4.14. The 4 milestone $C(t)$ function was calculated with the milestones placed 90 degrees apart as illustrated in figure 4.7, while the 8 milestone configuration was the same motif, only with 8 planes placed 45 degrees apart.

Chapter 5

An Algorithm for Automated Definition of Hyperplane Interfaces for Measuring Conformational Kinetics of Macromolecules Using Machine Learning

5.1 Introduction

The Milestoning method , developed by Ron Elber et al. is an algorithm whereby the kinetics of configurational changes in complex macromolecules can be calculated from molecular dynamics simulations [13]. In this paper, we present an algorithm for a fully automated method for both finding the physically important regions of a molecule's configuration space, and defining the milestone hyperplanes which best serve to bound these regions. It is often the

case that only one configuration of a biomolecule of interest, typically obtained from x-ray crystallography experiments, is known. This known configuration can be used as a starting point for molecular dynamics simulations in order to determine other possible configurations of the molecule, but the molecular motions from these simulations are often too complex for a human to recognize the important configurational changes, let alone define the milestone hyperplanes in the space that best capture the transitions. Motivated by this, we introduce an algorithm that (1) efficiently explores the configuration space of a macromolecule, and (2) automatically partitions the space in a fashion suitable for use with milestoning. In accord with these two goals, our algorithm consists of two subroutines: first a search of the configuration space, where mutually repulsive clones (MRCs) of the system explore the space; and second a milestone designation step, where the superior pattern recognition capabilities of machine learning are harnessed to first define clusters of computer-generated configurations and then define hyperplane interfaces between these clusters to be used in identifying milestones. Our approach scales well to high-dimensional spaces, and accommodates low-energy regions of irregular (and even non-convex) shape in the configuration space.

The remainder of the paper is organized as follows: we first give a brief introduction to the Milestoning method, next describing our algorithm for exploration of the configuration space describing the molecular motions of our system and its subsequent partitioning into subspaces with milestone interfaces. We then illustrate our method on a two dimensional system, concluding with an application of our method to a simple molecular model.

5.2 Milestoning Theory

A more in-depth overview of milestoning theory can be found in [46] or [120], but we briefly review a few of the key premises upon which our method for calculating time correlation functions hinges. The quantity of most fundamental importance in milestoning is the flux

through a given milestone, for which the equation is [33]:

$$P_s(t) = \int_0^t Q_s(t') \left[1 - \int_0^{t-t'} K_s(\tau) d\tau\right] dt',$$

$$Q_s(t) = 2\delta(t)P_s(0) + \int_0^t Q_{s\pm1}(t'')K_{s\pm1}^{\mp}(t-t'')dt'' \tag{5.1}$$

where $P_s(t)$ is the probability of being at milestone s at time t, (or, more specifically, arriving at time t' and not leaving before time t [33]), and $Q_s(t)$ is the probability of a transition to milestone s at time t. $K_s(\tau)$ indicates the probability of transitioning out of milestone s given an incubation time of τ, thus $\int_0^{t-t'} K_s(\tau)d\tau$ is the probability of an exit from milestone s anytime between 0 and $t-t'$, which makes $1 - \int_0^{t-t'} K_s(\tau)d\tau$ the probability of there *not* being an exit from milestone s over that same time period. Since the probability of two independent events happening concurrently is the product of the two events, the equation for $P_s(t)$ is simply integrating the concurrent hazards of arriving at milestone s and not leaving over the time frame from time 0 to t. Turning our attention towards the meaning of the first term, $Q_s(t)$, $2\delta(t)P_s(0)$, simply represents the probability that the system is already occupying milestone s at time $t = 0$, where the factor of 2 is present since the δ-function is centered at zero, meaning only half of its area would be counted without this factor. $Q_{s\pm1}(t'')$ is the probability that the system transitioned into one of the two milestones adjacent to s at an earlier time t''. $K_{s\pm1}^{\mp}(t-t'')$ is the probability of a transition from milestones $s \pm 1$ into milestone s. Thus the second term of the second line of equation 14 is another concurrent probability: the probability of the system entering an adjacent milestone at an earlier time, and then transitioning into milestone s between time t and 0. It is important to note that all functions $P_s(t)$ and $Q_s(t)$ are calculated using the respective values of $K_s(\tau)$ between

adjacent milestones, thus the set of $K_s(\tau)$ between all milestones of interest contains all the information needed to calculate kinetics using the milestoning method. It is also important to note that a K function between two milestones $x = A$ and $x = B$, $K_{AB}(\tau)$, is simply a probability distribution representing the lifetime for the system remaining in state A before transitioning to state B.

5.3 Algorithm

The efficiency of the above calculation depends upon the division of configuration space into a set of subspaces that jointly cover the region of physical interest, while also allowing relatively inexpensive evaluation of the transition times between each pair of subspaces which share an interface. This is made difficult by the high dimensionality of the configuration space, and the potentially complex boundaries surrounding optimal configuration sets. We address this challenge via a scalable two-stage algorithm, as described below.

Configuration Space Exploration

A common pitfall in molecular dynamics simulations is the problem of broken ergodicity due to incomplete sampling of the configuration space. This can occur when the system becomes trapped in one particular region of configuration space. In calculating kinetics from molecular dynamics, trapping can lead to inaccuracies if physically relevant portions of configuration space are separated by barriers high enough that transitions over them are rare on the time scale of the molecular dynamics simulations. Since molecular dynamics simulations are, at their very best, on the order of a few milliseconds—while biological processes are often on the order of seconds, minutes, or even days—enhanced sampling of configuration space is a necessity for calculating kinetics from molecular dynamics simulations of many

systems. Many common solutions to the problem of trapping (e.g., tempering and annealing methods) are based on artificially raising the temperature in the simulation, leading to an effective lowering of transition barriers and allowing a more complete sampling of the space. Although methods such as this do allow a more complete sampling of configuration space, the artificially low barriers can lead to unphysical transitions between minima, a blurring of the energetic landscape, and also wasted computation time spent both exploring unphysical transition states (see figure 5.1, top right) and oversampling low energy regions. Although, many successful re-weighting methods, such as Umbrella Sampling, Simulated Annealing, and high temperature molecular dynamics [109] [56] [19] have been used to recover accurate low temperature thermodynamics from artificially elevated temperature simulations, these methods are not ideal for our approach to calculating kinetics using Milestoning. Motivated by a desire to efficiently sample all minima in an energetic landscape that renders them easily discernible to a clustering algorithm, we introduce the *mutually repulsive clone* (MRC) approach to configuration space exploration. In this sampling scheme, multiple copies of the system that experience an artificial repulsive force between them are allowed to explore an otherwise normal energetic landscape. The results of applying this method to a two dimensional model potential can be seen in figure 5.1.

The application of this method to molecular systems is a straightforward scaling of dimensions. Instead of our mutually repulsive points in a two dimensional configuration space, molecular simulations would be performed as mutually repulsive points in, at most, a 3N dimensional space, i.e. each atom of each clone would be repelled by its respective atom in the other clones. In order to save on computational expense, the clones can be made to interact via some subset of the total set of their atoms, for example allowing only the alpha carbons of a protein to interact between clones. Using a pairwise interaction approach to calculating the repulsion between clones scales as n^2, where n is the number of clones. In order to further cut down on computational expense, and also to facilitate parallelization, we propose the following method whereby a k-ary tree-like motif (see figure 5.3) of repulsive

stationary vertex structures (VS's) is generated from iterative simulations of sets of mutually repulsive clones (for this example assume that our computing resources can comfortably accommodate 5 MRCs per node):

1. Read initial structure
2. Store initial structure as a VS
3. Generate a family of five MRCs from VS
4. Run minimization with repulsion between MRCs and VS
5. Run MD loop below on all MRCs with mutual repulsion and repulsion from VS

   MD loop for each MRC simulation:

   while (stepCount < stepMax

   and RMSDvar > rmsdCutoff)

   (a) stepCount = stepCount + 1

   (b) Save configuration to cPoints array every n steps

   (c) Save RMSD from VS to rmsdArray every n steps, where rmsdArray is the array that holds the 15 most recent RMSD measurements

   (d) The variance of the 15 RMSD values is measured as soon as 15 values have been stored, and also each time a new RMSD value is added to rmsdArray thereafter. If this variance, RMSDvar, goes below rmsdCutoff, exit this loop in accordance with the while statement.
6. Save final configurations from each individual MRC simulation as new VS's in VSarray. Additionally, these final static configurations take the place of the dynamic MRC for any siblings within its family that are still running

7. Each time an MD simulation of a particular MRC terminates, its resulting VS is saved to the VS array, and then sent to a new node as a seed structure, where steps 3 through 6 are carried out with repulsion both between MRCs and away from all VSs in VSarray (zero cutoff can be used to avoid calculating repulsion between MRCs and very distant VS points in configuration space).

To recapitulate, our search algorithm repeatedly spawns families of MRCs on separate processors, each family exploring configuration space in parallel, and each individual MRC MD simulation outputting structures, i.e. points in configuration space, at a given frequency. MD simulations of individual MRCs are terminated either when their RMSD relative to their seed structure begins to converge (the RMSD distribution begins to narrow), indicating that they have reached a "dead end" in configuration space, or they reach an assigned maximum number of steps, whichever comes first. There are several advantages to this algorithm. One key gain in using this k-ary tree-like motif of stationary vertex structures is that it allows for a full parallelization of our MRC search algorithm. If we were to simply keep adding MRCs, then the number of MRC repulsion calculations would scale exponentially on the same processor, which is obviously problematic. By using an array of static vertex structures as nodes in a configuration space graph, we are able to include repulsion from previously visited portions of configuration space in simulations run completely independently from the simulations used to map out these previously visited regions. This prevents "backtracking" into regions of configuration space which have already been explored, in a similar fashion to the metadynamics pioneered by Parinello et al [] . In cases where our system of interest is too large to simulate 5 copies of simultaneously interacting MRCs on the same processor, the method can be run using only VS repulsion, allowing all clones to be run in parallel on separate processors (figure 5.2).

Automated Construction of Subspaces

The main purpose of our configuration exploration method introduced in the previous section was to optimize the sampling in configuration space for efficient use of computational resources, while maintaining the integrity of the energetic landscape. More specifically, the goal is for the configuration search to yield a set of points in configuration space which can easily be separated into clusters centered about minima in the energetic landscape. Once such sets of points in configuration space are obtained, they become the training set for a fully automated method for subdividing configuration space into a set of subspaces where each subspace contains a single energetic minimum and transitions through the milestone interfaces between subspaces correspond to transitions between the minima. The first step in the process is to separate the configuration space points into groups using a Euclidian distance clustering algorithm as shown in figure 5.4. The points can be clustered using only spatial dimensions as in figure 5.4, or other parameters, such as the energy do to the potential, can function as extra dimensions in configuration space as shown in figure 5.8. With each configuration space point in our set now assigned to a particular cluster, we now have a training set which can be fed into a Support Vector Machine (SVM) kernel-based machine learning algorithm that will return the boundaries in configuration space which best serve to partition our clustered configuration space points. These boundaries are known as support vectors, and they function as the milestones in our proposed automated milestoning approach. Following suit, an SVM classifier function is then used to determine when a transition between subspaces has occurred during the molecular dynamics simulations used to measure transition times between subspaces. Results indicating proof of concept for this approach can be found in figures 5.7 and 5.8.

5.4 Application to Atomistic System

So far, we have demonstrated our algorithm by applying it to a two dimensional model system, with the understanding that the method will easily generalize to higher dimensional configuration spaces. Now let us address a specific motif for defining a configuration space for fully automated machine learning-enhanced milestoning simulations. For some molecular systems, there exists an obvious reaction coordinate for the kinetics of a configuration change of interest. For example, if we are interested in the kinetics of transitions between the Watson-Crick and Hoogsteen conformations, we can use the dihedral angle describing the rotation of an Adenine residue as our reaction coordinate cite. Now consider a highly flexible structure, like an intrinsically disordered protein. If we are interested in developing a kinetic model for the numerous configurations it can occupy, we cannot expect to gain much insight from measuring some predetermined reaction coordinate intuited by the person carrying out the simulations. This is exactly the sort of system where a machine learning-based methodology can outperform Milestoning techniques that require definition of milestone hyperplanes based on user input. In this case, one approach could be to define our configuration space as the set of pairwise distances between all alpha carbons, or perhaps every $5th$ alpha carbon along the backbone if the system is too large. With the dimensions of our configuration space defined in this manner, we could then feed an initial structure into our configuration space exploration algorithm, then feed the set of configuration space points from the exploration step into clustering/SVM software to yield a set of hyperplanes to be used in a Milestoning simulation. From such a procedure, one could obtain the transition kinetics between energetically favored configurations of a system for which the user lacks any intuition. In order to provide a visualizable example of this type of configuration space, we have applied it to alanine dipeptide. For this simple molecular system, a three dimensional configuration space, composed of the distances between the carbon atoms shown in blue, green, and red, was devised (figure 5.9). The molecular dynamics governing the configurational fluctuations

of this system were then simulated, and these pairwise distances were outputted. Shown bottom right is the result of using a clustering algorithm to group the points into distinct sets, the interfaces between which could then be used in Milestoning simulations.

5.5 Concluding Discussion

This project has served to lay the groundwork and proof of concept for a fully automated machine learning-based method for calculating configurational kinetics from molecular dynamics simulations. The key advantage provided by this method is that it can be used to explore the conformational dynamics of a macromolecule for which the user has little to no intuition. Although, in theory, Milestoning calculations are insensitive to the placement of the milestones [73], in practice, the efficiency of transition sampling can be improved by initiating trajectories along maxima or saddle points between energetic minima. Additionally, and perhaps more importantly, the ability to calculate fluxes through hyperplanes in configuration space is most useful for gaining insight for a system of interest when those hyperplanes are placed through topologically significant regions of the potential energy surface, such as energetic maxima. Another benefit of partitioning configuration space into subspaces centered about individual energetic minima is that all transitions between milestones reflect a transition between individual barriers, whereas a less careful placement of milestone interfaces may contain multiple, unobserved transition barriers between them, or even none at all. Future work will be directed toward full implementation of this method for all atom systems.

5.6 Acknowledgments

IA acknowledges funds from an NSF CAREER award (CHE-0548047). CTB was supported by NSF award DMS-1361425. GG was supported by a UC Irvine Data Science Initiative summer research grant.

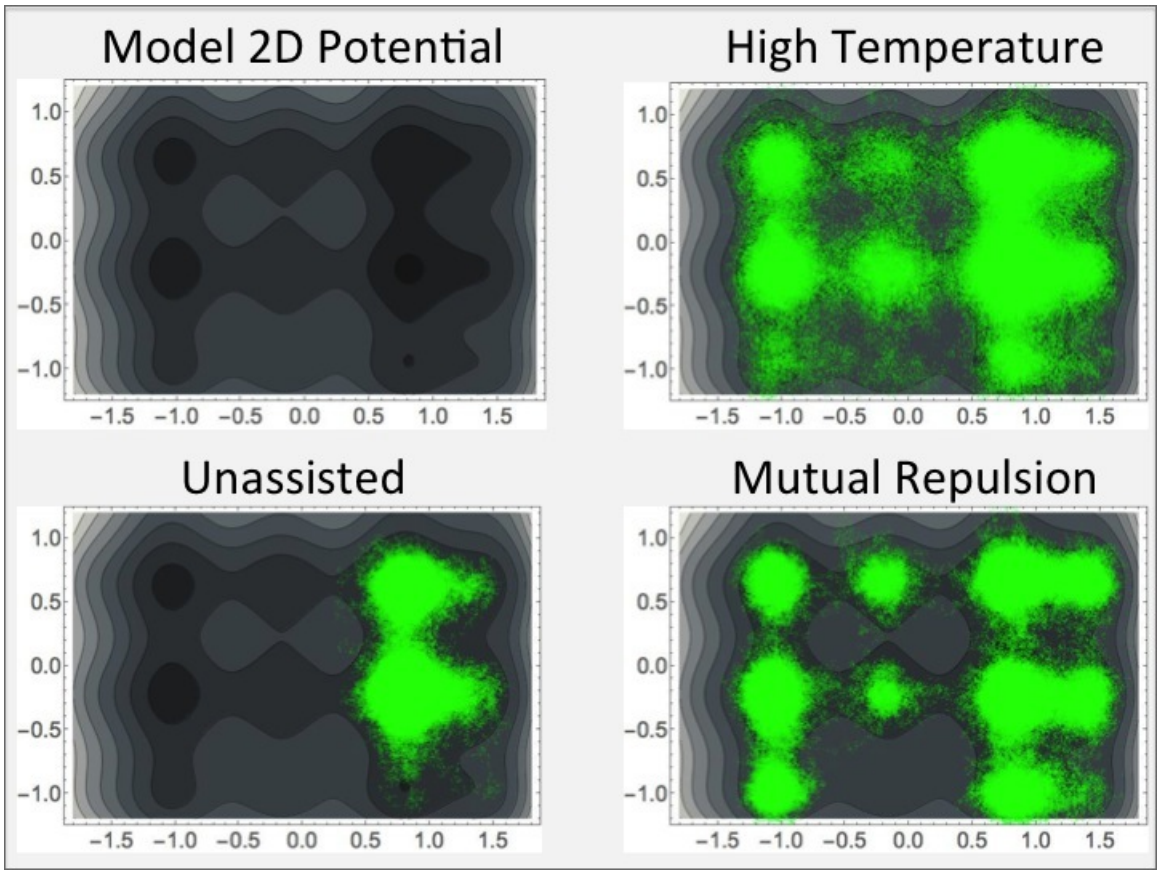

Figure 5.1: Shown here is a comparison between our mutually repulsive clone search, Langevin dynamics alone (Unassisted), and running Langevin dynamics at an artificially high temperature. In the top left is a relief map showing the shape of our potential. Each plot shows the density of points visited in configuration space in green, given that all methods began in the local minimum centered at approximately $\{.9, -.25\}$. The mutually repulsive clone exploration method outperformed both unassisted sampling, which remained trapped in two local minima, and the sampling performed at an artificially high temperature, which did manage to sample the entire configuration space by suffered from a blurring of the features in the energetic landscape.

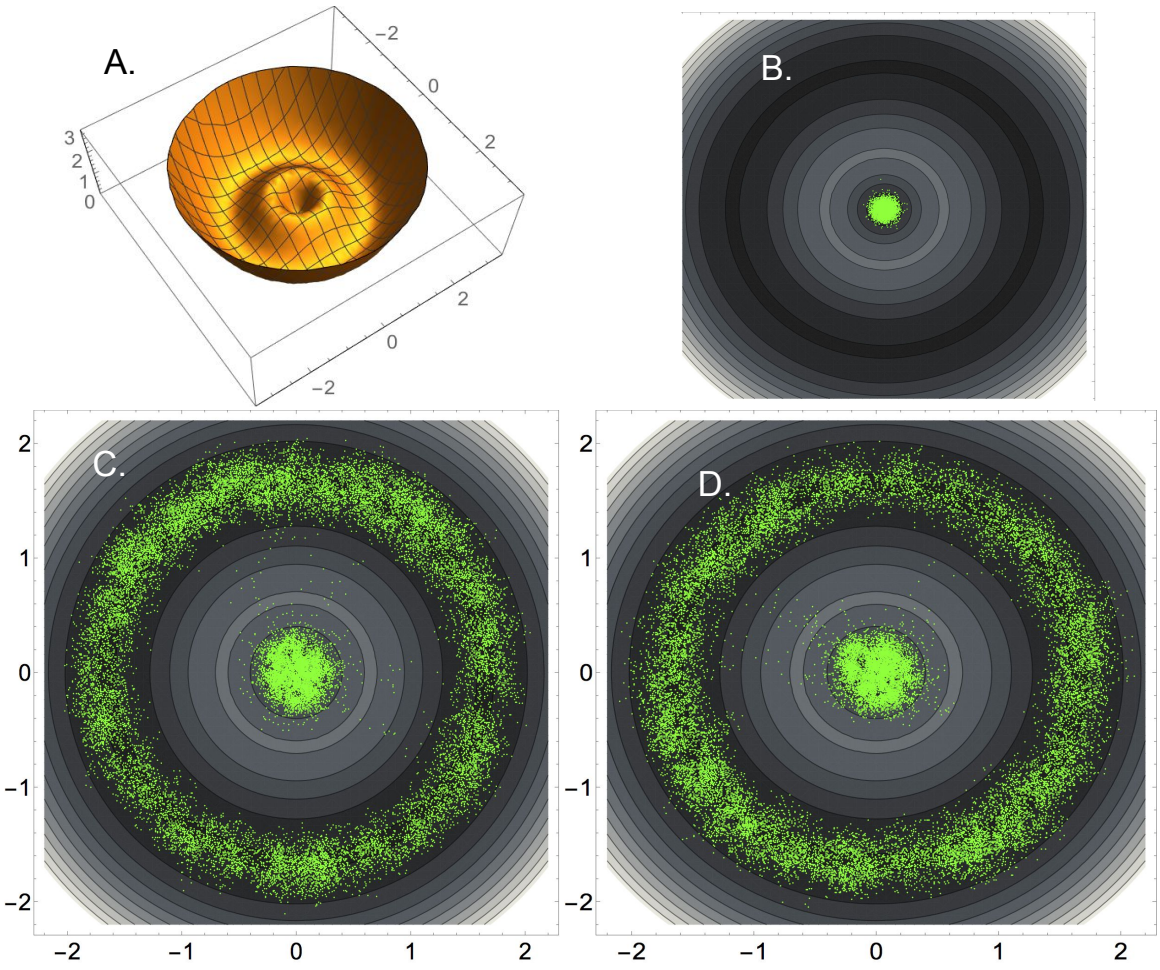

Figure 5.2: Figure A is a plot of the potential energy surface $f(x,y) = -\frac{3.5}{x^2+y^2+0.5} + \frac{30}{x^2+y^2+2} + x^2+y^2-7.5$, which features two concentric stable regions, a motif that could pose problems for a hyperplane-based milestoning methodology. The plot labeled B shows the configurations space points visited during a Langevin dynamics simulation of 2 million time steps for six non-interacting copies of our system beginning at the point $\{0,\ 0\}$. Note that under the conditions these simulations were run, without a bias, our system is trapped in the central minimum. Plot C shows tthe results of another 2 million step Langevin simulations where the current locations of 6 MRCs were saved as repulsive nodes every $40,000$ steps, and all other conditions were the same as in plot B. In plot D, we show the results of a simulation where repulsive nodes, or VSs, were used, but there was no repulsion between the clones. Note that, in this case, leaving out the mutual repulsion between active simulations of MRCs was in no way detrimental to the sampling of configuration space. Again, the advantage of calculating repulsion from VS configuration space points only is that all simulations can be run in parallel on separate processors.

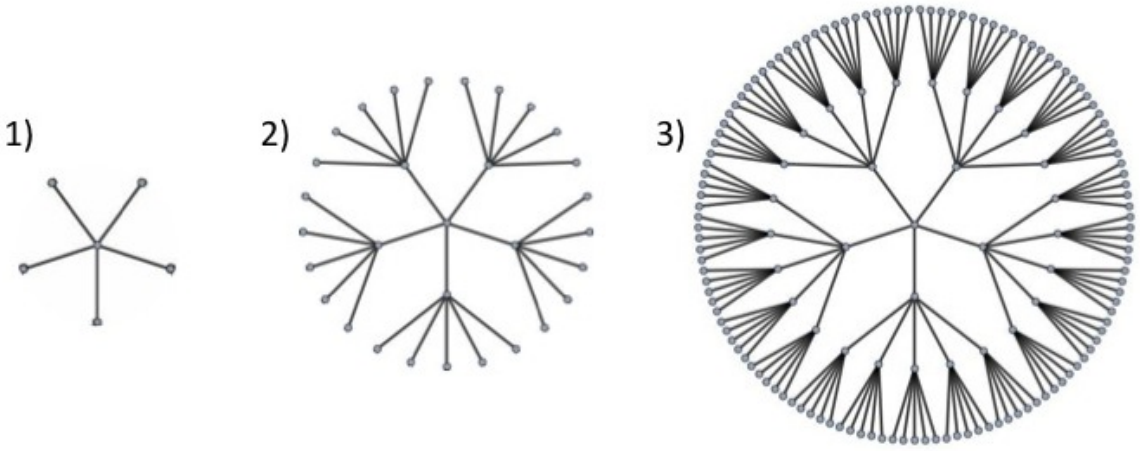

Figure 5.3: Shown here are the first, second, and third branching iterations of a k-ary tree of degree 5. Note how a free expansion of mutually repulsive clones occupying a two dimensional space results in the formation of a cell-like structure. If such an expansion occurred subject to both mutual repulsion and an external potential like in figure 5.1, the result would be a similar cell with boundaries deformed to fit the shape of the basin.

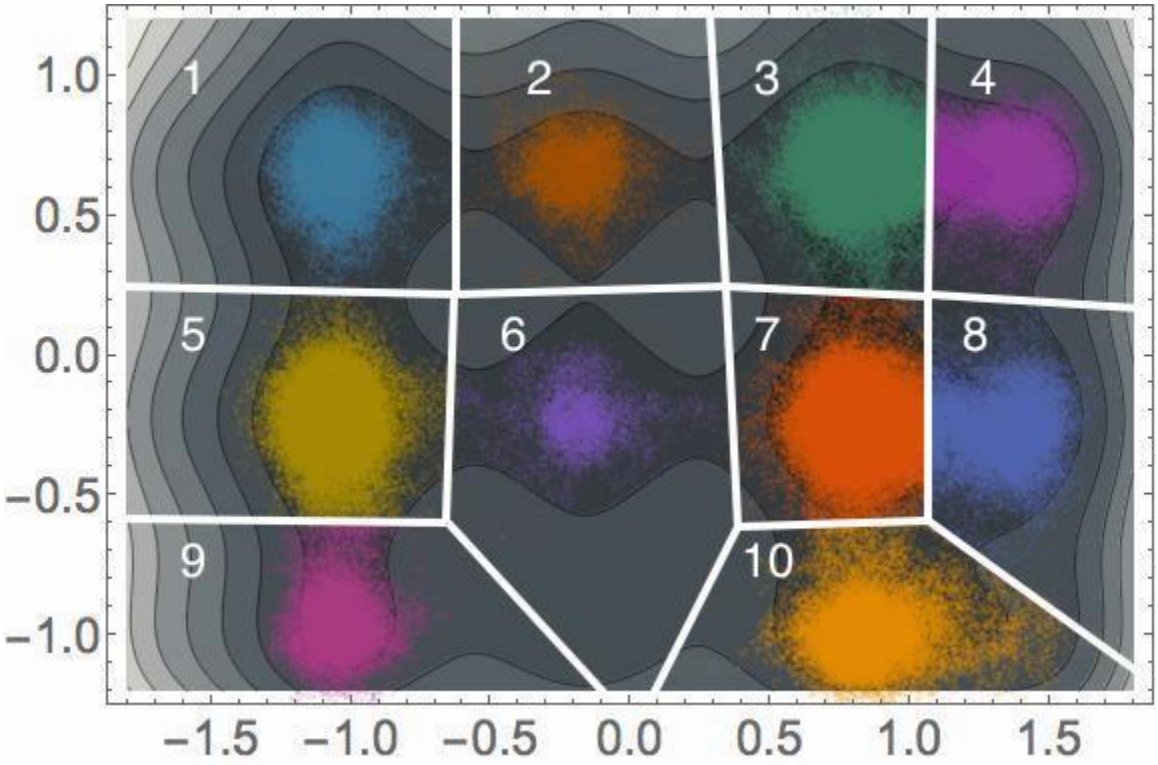

Figure 5.4: Shown here are the results of running a Euclidian distance clustering algorithm on the configuration space points generated using our mutually repulsive clone data (figure 5.1, lower right). Note that the clustering algorithm was able to define boundaries so that each subspace contains exactly one of the local minima.

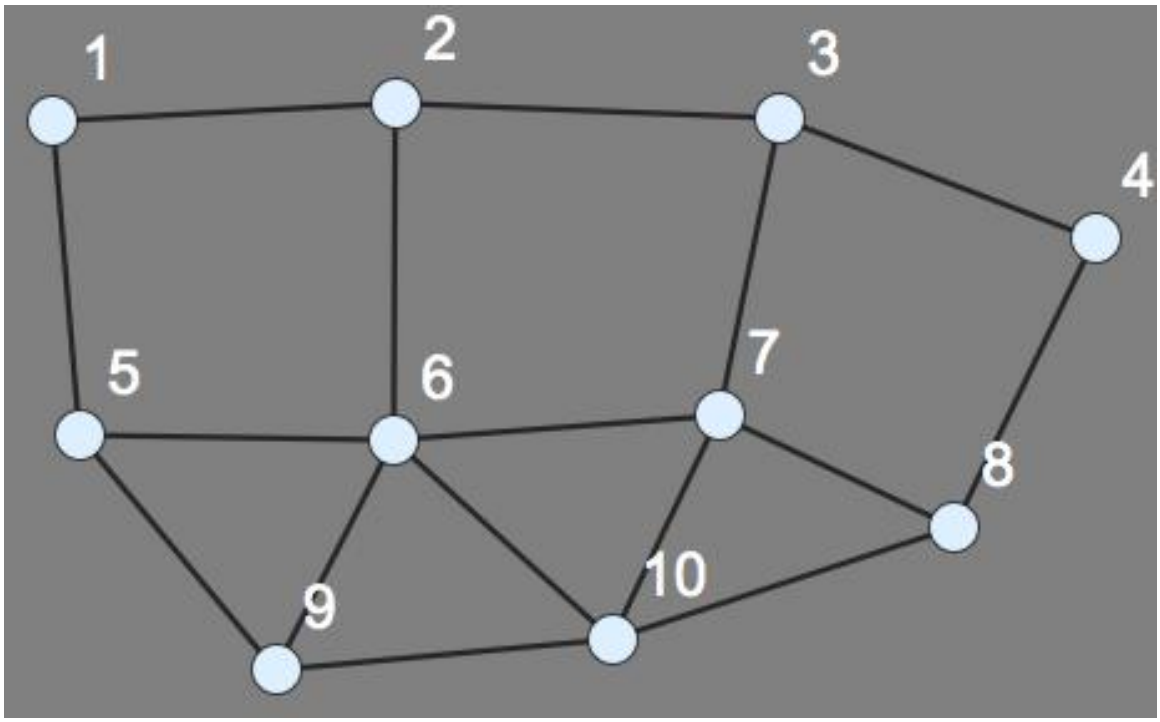

Figure 5.5: The subspaces in configuration space generated by our algorithm are best represented as a weighted directed graph. Each subspace is a node in the graph, and a pair of nodes share an edge if and only if they share a milestone hyperplane interface. The weight of the i, j edge represents the rate of flow between subspaces.

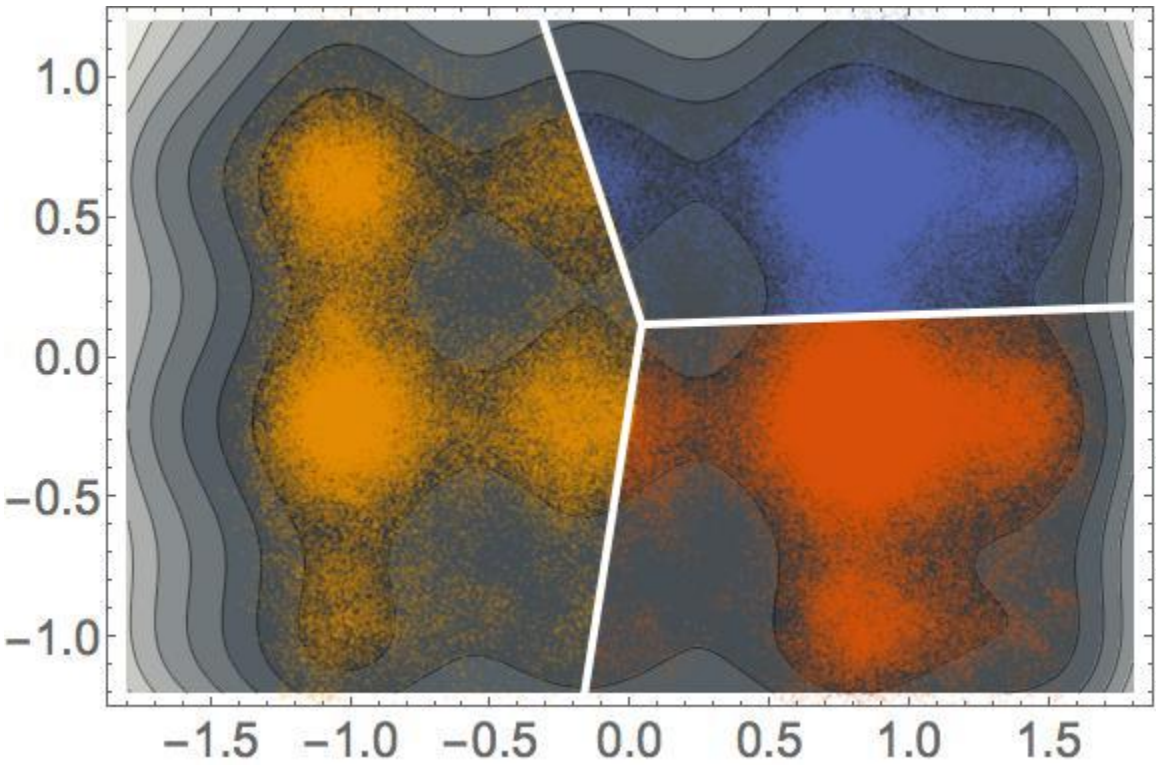

Figure 5.6: Shown here are the results of running a Euclidian distance clustering algorithm on the configuration space points generated using an artificially high temperature to allow sampling over high barriers (figure 5.1, upper right). Note that the blurring of the energetic landscape caused by running Langevin dynamics at an artificially high temperature has led to the same clustering method, run with the same parameters, to identify only three distinct regions. By lumping together regions with multiple local minima into the same subspaces in this manner, we would obtain a simplistic connected graph of three states representing the kinetics instead of the much richer representation shown in figure 5.5. Although a more stringent clustering method could be applied to this particular data set to yield better classification into clusters, the motivation behind this example is to demonstrate that sampling techniques that maintain the integrity of the energetic landscape, like our mutually repulsive clone method, can yield better results by providing the clustering algorithms with more physically relevant data sets.

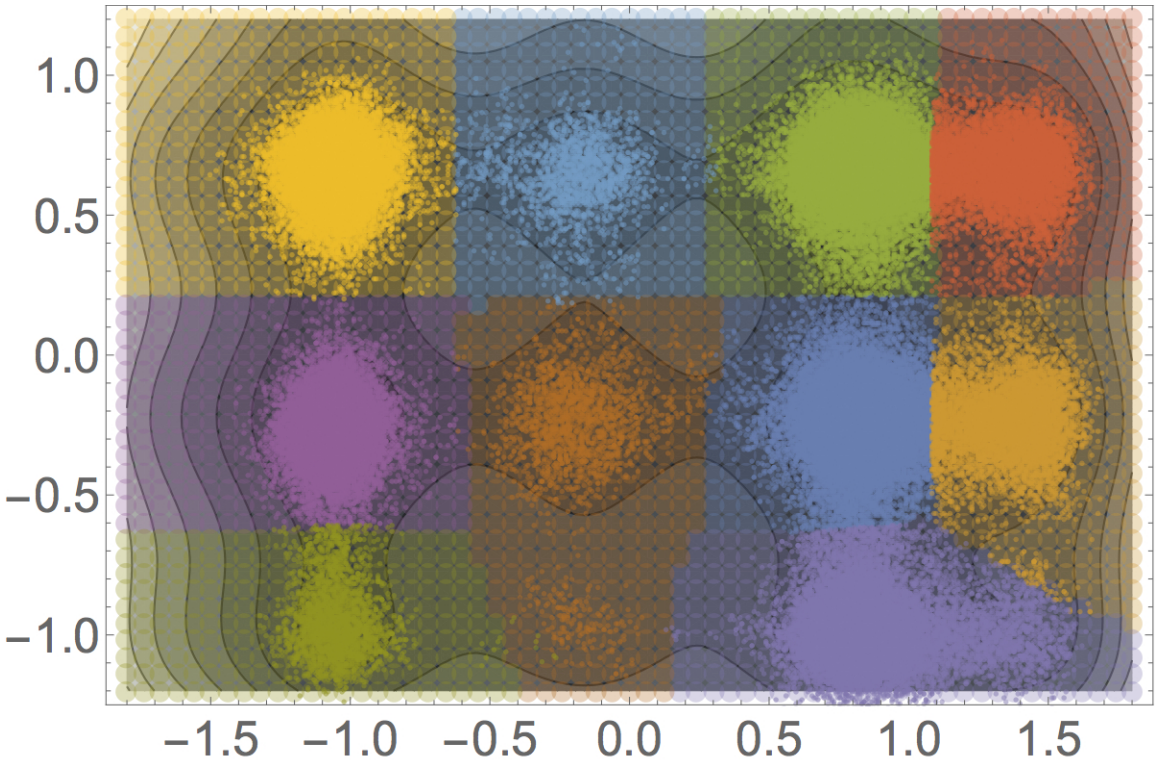

Figure 5.7: This figure displays the results of first applying a clustering algorithm to the data from figure 5.4, and then using the clustering data as a training set for a Support Vector Machine methodology for dividing the configuration space into a set of subspaces. The more opaque points display the configuration space points visited in the simulation, and the colored shading indicates the partitioning of configuration space into subspaces suitable for Milestoning. These results demonstrate our fully automated methodology for subdividing a configuration space in such a way that measuring transition kinetics between subspaces corresponds to transitions between local minima in the potential energy surface.

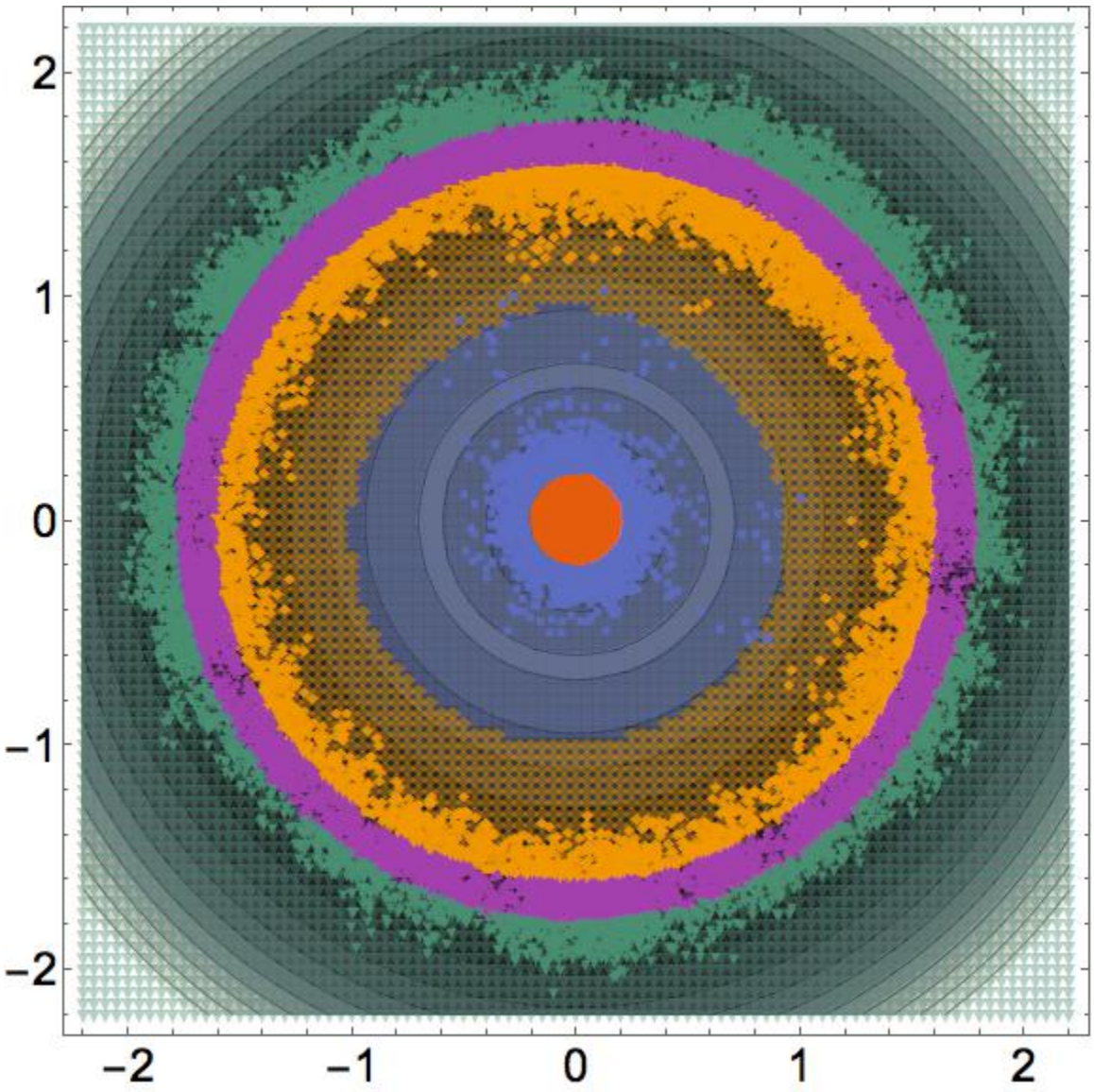

Figure 5.8: This plot shows a similar approach to the one in figure 5.7 applied to our concentric minima potential. In this example, the configuration space for the clustering step was defined in two dimensions, RMSD from the initial configuration and energy due to the potential energy surface, two quantities easily calculated from molecular simulations. Using the classifications of our data points generated by the clustering algorithm as a training set, the Support Vector Machine has divided the purely spatial configuration space into 5 distinct subspaces, one for the central minimum (dark orange), one for the transition state (blue), one for the outer minimum (magenta), as well as two intermediate states going away from the outer minimum (green and light orange). Note that using the kernel-based SVM has allowed for not only curved but concentric interfaces. Further, our method has autonomously defined 5 meaningful subspaces bounded by just four interfaces. In the case of ordinary hyperplanes, it would take four hyperplanes just to bound one of these regions by circumscribing it into a quadrilateral. Again, the coloring scheme applies to both the classification of the data points, shown shown using opaque points, and the subspaces, shown as lighter shading of the same color.

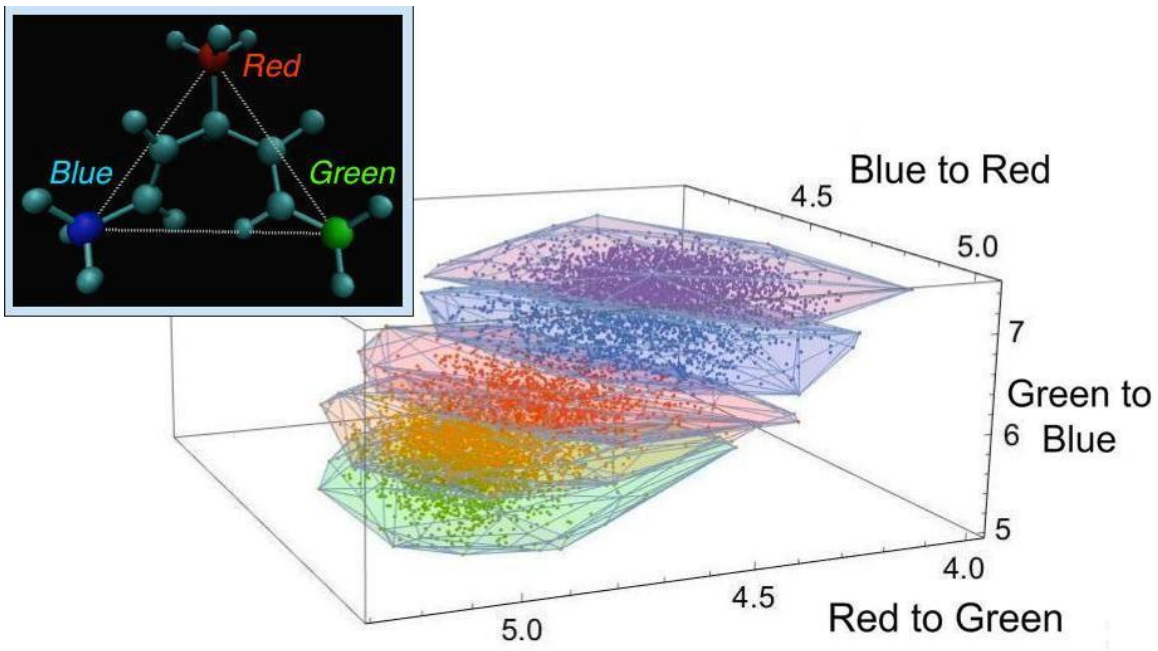

Figure 5.9: Here we have a demonstration of automated partitioning of configuration space into convex hulls. In this particular case, we have defined a three dimensional configuration space comprised of the pairwise interatomic distances between three labeled carbons shown as red, green, and blue (top left). In the interest of keeping our demonstration visualizable, only three dimensions were defined in the configuration space; however, the method can be easily generalized to any number of dimensions, for example the set of pairwise distances between all alpha carbons in a protein.

Chapter 6

Conclusions and Future Research

As important as the calculation of thermodynamic properties of macromolecules is to understanding the chemistry of life at the molecular level, in no way can we describe the state of any living thing as an equilibrium state. Every life process is a time-dependent process, and for this reason, no description of any biochemical system is complete without addressing kinetics. The advancement of kinetic calculations in the field of molecular dynamics has not trailed those of thermodynamic due to lack of interest, however; the bottleneck has been largely due to the scaling problem fundamental to transitioning from thermodynamic calculations to kinetic calculations. More concretely, the thermodynamic properties of macromolecular configurations are centered around the set of all possible configurations, which together with their energy-dependent statistical weights can be used to calculate the partition function, while kinetic information is encoded in the set of all possible trajectories through configuration space, from which time correlation functions can be derived, allowing for the calculation of chemical spectra from molecular dynamics simulations. The study of chemical kinetics then requires both the extensive exploration of configuration space inherent to thermodynamic calculations as well as extensive sampling of the even larger set of possible time-dependent transitions between these configurations. As daunting as this may be, access

to dynamical representations of the configurations of macromolecules are fundamental to one of the most broadly impactful applications of molecular dynamics: the synergistic exchange between theorists carrying out molecular dynamics simulations and experimentalists making spectroscopic measurements.

I feel that our work in developing Smoluchowski models for the mechano-catalytic behavior observed in thioredoxin has good potential for further development in providing microscopic insights into the results of single molecule pulling experiments. In this realm of scientific inquiry, the tether between the experimental system and simulation is the reaction rate, an incredibly expensive quantity to calculate from atomistic simulations, due to the usual bottleneck of sampling reactive trajectories. The fact that simulating mechano-catalytic behavior such as this requires that many data points be calculated for rate as a function of applied force, means that one of the few tractable approaches toward calculating such relationships are via coarse grain models, and hence the utility of our approach. The normal modes serve as a linear space of fundamental motions at atomic resolution that can be used to construct any motion of interest as a linear combination of normal modes, which in turn can be used to define the diffusion coordinate for the Smoluchowski equation. I feel that the step in this process most in need of further development is the point where the motion of interest is fit to a linear combination of normal modes. For this reason, I would propose future work be done to further cultivate this approach using simpler model systems. For example a coarse grain approach could be taken to build a model a system similar to the titin substrate with an engineered disulphide bond used in the experiments by Jimenez et al [4]. The model would consist of a Go model type approach [111], where there is a string of beads, bound along the length of the strand with harmonic bonds, some weak harmonic bonds between some non-neighboring beads to introduce some secondary structure, as well as a Lennard-Jones potential binding one pair of non-neighboring beads to model a disulphide bond. Additionally, pulling forces could be introduced at the ends of the string of beads. The linear space of motions would first be created by performing

normal mode analysis in the usual way. Next, longtime Langevin dynamics could then be run on this simple system at high enough temperatures that breakage of the Lennard-Jones bond would occur with some regularity over some distribution of incubation times. The trajectories encompassing the configurations just before, during, and after breakage of the Lennard-Jones bond would then be stored as reactive trajectories. From here, a similar approach to calculating the coefficients for basis functions representing a particular wave function in quantum calculations could be employed, whereby the vectors representing the reactive motions would undergo dot product multiplication with the vectors representing the characteristic motion of each normal mode. In this manner, individual reactive trajectories could be expressed as projections onto a linear space of normal modes. From here, the order parameter describing motion along the characteristic linear combination of normal modes as a function of energy in the Go type model could serve as the potential in the Smoluchowski diffusion model (a harmonic potential could also be fit to the Go type energy potential in each case). The rate determined by the Smoluchowski calculation would then serve as a test for each reactive trajectory, where rates calculated close to the observed rate indicate physically plausible trajectories which may be able to help elucidate mechanistic details of the reaction.

I feel that the next step for the three enhanced sampling methods for calculating the kinetics of conformational changes in macromolecules introduced in this thesis is implementation into a molecular dynamics software package. The WARM method is broadly applicable toward reducing computation time in any Milestoning simulation, and thus would be the best first candidate for implementation into a molecular dynamics software package. I also feel that the machine learning-guided approach to defining milestone interfaces introduced in this work is ready for implementation, which will provide the decisive test for its utility. The method for approximating time correlation functions from Milestoning simulations introduced in this work has promising potential for development into software for calculating spectroscopic data from Milestoning simulations, and I feel that the proof of concept calculations herein

provide strong evidence for the utility of this approximation, but again, implementation into molecular dynamics software will most effectively reveal the strengths and limitations of the approximation. In the long term, I would like to implement all three methods as a suite of enhanced Milestoning methods for calculating spectra of complex macromolecules from molecular dynamics simulations. One area of personal interest that I would very much like to explore using these three tools in conjunction is the study of intrinsically disordered proteins. Although these proteins typically lack a well-defined minimum energy structure that they will consistently occupy in the crystalline state [108], their structure can be understood in a dynamical sense, and I feel that these three enhanced sampling methods could be ideal for elucidating such dynamical models. The approach would begin by applying the search, cluster, and subspace defining steps of the machine learning-guided method to define the set milestone interfaces that best subdivide the configuration space. Next, the WARM method would be applied to collect the transition time distributions necessary for Milestoning. Finally, the method for obtaining discrete approximations of time correlation functions would be applied to calculate the relevant time correlation functions necessary for ultimately calculating particular spectroscopic measurements of interest. Given a set of trajectories which correspond to accurate spectra, one could then go back to the trajectories and mine them for the atomistic behaviors corresponding to experimentally observed spectra. Optimistically, if the implementation is sufficiently robust, my hope is that I can one day combine the three methods in this fashion to form a software package that serves as a "computational microscope" [62] for the study of macromolecules exhibiting complex dynamics that requires little to no prior intuition about the chemical system from the user.

Bibliography

[1] Phys 498 lecture notes, university of illinois urbana-champaign. `http://www.ks.uiuc.edu/Services/Class/PHYS498/LectureNotes/chp4.pdf`. Accessed: 2015-08-24.

[2] N. Agmon and J. Hopfiled. CO Binding to Heme-Proteins - a Model for Barrier Height Distributions and Slow Conformational-Changes. *J. Chem. Phys.*, 79(4):2042–2053, 1983.

[3] R. Ait-Haddou and W. Herzog. Force and motion generation of myosin motors: muscle contraction. *Journal of Electromyography and Kinesiology*, 12(6):435 – 445, 2002.

[4] J. Alegre-Cebollada, R. Perez-Jimenez, P. Kosuri, and J. M. Fernandez. Single-molecule force spectroscopy approach to enzyme catalysis. *Journal of Biological Chemistry*, 285(25):18961–18966, 2010.

[5] G. Andre, R. Brasseur, and Y. F. Dufrene. Probing the interaction forces between hydrophobic peptides and supported lipid bilayers using AFM. *Journal of Molecular Recognition*, 20(6):538–545, 2007.

[6] I. Andricioaei, A. Goel, D. Herschbach, and M. Karplus. Dependence of DNA polymerase replication rate on external forces: a model based on molecular dynamics simulations. *Biophysical journal*, 87(3):1478–1497, 2004.

[7] I. Andricioaei and M. Karplus. On the calculation of entropy from covariance matrices of the atomic fluctuations. *Journal of Chemical Physics*, 115:6289–6292, 2001.

[8] A. Ansari. Mean first passage time solution of the smoluchowski equation: Application to relaxation dynamics in myoglobin. *The Journal of Chemical Physics*, 112(5):2516–2522, 2000.

[9] E. S. J. Arner and A. Holmgren. Physiological functions of thioredoxin and thioredoxin reductase. *European Journal of Biochemistry*, 267(20):6102–6109, 2000.

[10] B. T. AU Marshall, M. Long, J. W. Piper, T. Yago, R. P. McEver, and C. Zhu. Direct observation of catch bonds involving cell-adhesion molecules. *Nature*, 423(6936):190 – 193, 2003.

[11] V. Barsegov and D. Thirumalai. Dynamics of unbinding of cell adhesion molecules: transition from catch to slip bonds. *Proceedings of the National Academy of Sciences of the United States of America*, 102(6):1835–1839, Feb. 2005.

[12] G. I. Bell. Models for the specific adhesion of cells to cells. *Science*, 200(4342):618–627, 1978.

[13] J. M. Bello-Rivas and R. Elber. Exact milestoning. *The Journal of Chemical Physics*, 142(9):–, 2015.

[14] H. Berendsen. Protein Folding - A Glimpse of the Holy Grail? *Science*, 282(5389):642–643, Oct. 1998.

[15] H. J. Berendsen and S. Hayward. Collective protein dynamics in relation to function. *Current Opinion in Structural Biology*, 10(2):165 – 169, 2000.

[16] P. Bolhuis, D. Chandler, C. Dellago, and P. Geissler. Transition path sampling: Throwing ropes over rough mountain passes, in the dark. *Annu. Rev. Phys. Chem.*, 53:291–318, 2002.

[17] D. B. Boyd. Conformational dependence of the electronic energy levels in disulfides. *Journal of the American Chemical Society*, 94(25):8799–8804, 1972.

[18] B. R. Brooks, D. Janezic, and M. Karplus. Harmonic-analysis of large systems .1. Methodology. *Journal of Computational Chemistry*, 16:1522–1542, 1995.

[19] R. E. Bruccoleri and M. Karplus. Conformational sampling using high-temperature molecular dynamics. *Biopolymers*, 29(14):1847–1862, 1990.

[20] C. Bustamante and J. Moffitt. *Viral DNA4 Packaging: One Step at a Time*, volume 96 of *Springer Series in Chemical Physics*. Springer Berlin Heidelberg, 2010.

[21] C. Bustamante, S. B. Smith, J. Liphardt, and D. Smith. Single-molecule studies of DNA mechanics. *Current Opinion in Structural Biology*, 10(3):279 – 285, 2000.

[22] Y. Cao and H. Li. Single-molecule force-clamp spectroscopy: Dwell time analysis and practical considerations. *Langmuir*, 27(4):1440–1447, 2011.

[23] M. Carrion-Vazquez, A. F. Oberhauser, T. E. Fisher, P. E. Marszalek, H. Li, and J. M. Fernandez. Mechanical design of proteins studied by single-molecule force spectroscopy and protein engineering. *Progress in Biophysics and Molecular Biology*, 74(12):63 – 91, 2000.

[24] C. Cecconi, E. Shank, F. Dahlquist, S. Marqusee, and C. Bustamante. Protein-DNA chimeras for single molecule mechanical folding studies with the optical tweezers. *European Biophysics Journal*, 37(6):729–738, 2008.

[25] P. T. Chivers and R. T. Raines. General acid/base catalysis in the active site of escherichia coli thioredoxin. *Biochemistry*, 36(50):15810–15816, 1997.

[26] G. Cravotto and P. Cintas. Forcing and controlling chemical reactions with ultrasound. *Angewandte Chemie International Edition*, 46(29):5476–5478, 2007.

[27] C. Danilowicz, V. W. Coljee, C. Bouzigues, D. K. Lubensky, D. R. Nelson, and M. Prentiss. DNA unzipped under a constant force exhibits multiple metastable intermediates. *Proceedings of the National Academy of Sciences*, 100(4):1694–1699, 2003.

[28] C. Dellago, P. G. Bolhuis, and D. Chandler. On the calculation of reaction rate constants in the transition path ensemble. *The Journal of Chemical Physics*, 110(14):6617–6625, 1999.

[29] C. Dellago, P. G. Bolhuis, F. S. Csajka, and D. Chandler. Transition path sampling and the calculation of rate constants. *The Journal of Chemical Physics*, 108(5):1964–1977, 1998.

[30] M. Dembo, D. C. Torney, K. Saxman, and D. Hammer. The reaction-limited kinetics of membrane-to-surface adhesion and detachment. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 234(1274):55–83, 1988.

[31] Y. F. Dufrene and D. J. Muller. Force generation: ATP-powered proteasomes pull the rope. *Current Biology*, 21(11):R427 – R430, 2011.

[32] R. Elber. Milestoning Theory and a Simple Example. `http://clsb.ices.utexas.edu/web/Milestoning%20theory%20and%20a%20simple%20example.pdf`. [Online; accessed 30-Sep-2008].

[33] R. Elber and A. Faradjian. Computing time scales from reaction coordinates by milestoning. *Journal of Chemical Physics*, 120(23):10880–9, Mar. 2004.

[34] R. Elber, A. Roitberg, C. Simmerling, R. Goldstein, H. Li, G. Verkhivker, C. Keasar, J. Zhang, and A. Ulitsky. Moil: A program for simulations of macromolecules. *Computer Physics Communications*, 91(13):159 – 189, 1995.

[35] E. Evans. Probing the relation between forcelifetimeand chemistry in single molecular bonds. *Annual Review of Biophysics and Biomolecular Structure*, 30(1):105–128, 2001.

[36] E. Evans, A. Leung, V. Heinrich, and C. Zhu. Mechanical switching and coupling between two dissociation pathways in a p-selectin adhesion bond. *Proceedings of the National Academy of Sciences of the United States of America*, 101(31):11281–11286, 2004.

[37] H. Eyring. The activated complex in chemical reactions. *The Journal of Chemical Physics*, 3(2):107–115, 1935.

[38] P. A. Fernandes and M. J. Ramos. Theoretical insights into the mechanism for thiol/disulfide exchange. *Chemistry A European Journal*, 10(1):257–266, 2004.

[39] J. M. Fernandez and H. Li. Force-clamp spectroscopy monitors the folding trajectory of a single protein. *Science*, 303(5664):1674–1678, 2004.

[40] M. J. Field, P. A. Bash, and M. Karplus. A combined quantum mechanical and molecular mechanical potential for molecular dynamics simulations. *Journal of Computational Chemistry*, 11(6):700–733, 1990.

[41] S. D. Fried, S. Bagchi, and S. G. Boxer. Extreme electric fields power catalysis in the active site of ketosteroid isomerase. *Science*, 346(6216):1510–1514, 2014.

[42] S. Garcia-Manyes, T.-L. Kuo, and J. M. Fernndez. Contrasting the individual reactive pathways in protein unfolding and disulfide bond reduction observed within a single protein. *Journal of the American Chemical Society*, 133(9):3104–3113, 2011.

[43] S. Garcia-Manyes, J. Liang, R. Szoszkiewicz, T.-L. Kuo, and J. M. Fernandez. Force-activated reactivity switch in a bimolecular chemical reaction. *Nat Chem*, 1(3):236 – 242, 2009.

[44] M. Garcia-Viloca, J. Gao, M. Karplus, and D. G. Truhlar. How enzymes work: Analysis by modern rate theory and computer simulations. *Science*, 303(5655):186–195, 2004.

[45] M. Grandbois, M. Beyer, M. Rief, H. Clausen-Schaumann, and H. E. Gaub. How strong is a covalent bond? *Science*, 283(5408):1727–1730, 1999.

[46] G. Grazioli and I. Andricioaei. Advancements in milestoning I: Accelerated milestoning via wind assisted re-weighted milestoning (WARM). *arXiv*, 2015.

[47] C. R. Hickenboth, J. S. Moore, S. R. White, N. R. Sottos, J. Baudry, and S. R. Wilson. Biasing reaction pathways with mechanical force. *Nature*, 446(7134):423–427, 2007.

[48] A. Holmgren. Thioredoxin structure and mechanism: conformational changes on oxidation of the active-site sulfhydryls to a disulfide. *Structure*, 3(3):239 – 243, 1995.

[49] C. Hyeon and D. Thirumalai. Mechanical unfolding of RNA: From hairpins to structures with internal multiloops. *Biophysical Journal*, 92(3):731 – 743, 2007.

[50] M.-F. Jeng, A. Campbell, T. Begley, A. Holmgren, D. A. Case, P. E. Wright, and H. Dyson. High-resolution solution structures of oxidized and reduced escherichia coli thioredoxin. *Structure*, 2(9):853 – 868, 1994.

[51] D. Jeong and I. Andricioaei. Reconstructing equilibrium entropy and enthalpy profiles from non-equilibrium pulling. *The Journal of Chemical Physics*, 138(11):–, 2013.

[52] M. Karplus and Y. Q. Gao. Biomolecular motors: the F1-ATPase paradigm. *Current Opinion in Structural Biology*, 14(2):250 – 259, 2004.

[53] M. Karplus and J. Kushik. Method for estimating the configurational entropy of macromolecules. *Macrmolecules*, 14, 1981.

[54] M. Karplus and J. A. McCammon. Molecular dynamics simulations of biomolecules. *Nature Structural & Molecular Biology*, 9(9):646–652, 2002.

[55] M. S. Z. Kellermayer, S. B. Smith, H. L. Granzier, and C. Bustamante. Folding-unfolding transitions in single titin molecules characterized with laser tweezers. *Science*, 276(5315):1112–1116, 1997.

[56] S. Kirkpatrick, M. P. Vecchi, et al. Optimization by simmulated annealing. *science*, 220(4598):671–680, 1983.

[57] S. Kirmizialtin and R. Elber. Revisiting and computing reaction coordinates with directional milestoning. *The Journal of Physical Chemistry A*, 115(23):6137–6148, 2011. PMID: 21500798.

[58] H. Kleinert. *Path Integrals in Quantum Mechanics, Statistics, Polymer Physics, and Financial Markets, 3rd Ed.* World Scientific, 2004.

[59] S. R. Koti Ainavarapu, A. P. . Wiita, L. Dougan, E. Uggerud, and J. M. . Fernandez. Single-molecule force spectroscopy measurements of bond elongation during a bimolecular reaction. *Journal of the American Chemical Society*, 130(20):6479–6487, 2008. PMID: 18433129.

[60] H. A. Kramers. Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica*, 7(4):284–304, 1940.

[61] S. M. Kreuzer, T. J. Moon, and R. Elber. Catch bond-like kinetics of helix cracking: Network analysis by molecular dynamics and milestoning. *The Journal of chemical physics*, 139(12):121902, 2013.

[62] E. H. Lee, J. Hsin, M. Sotomayor, G. Comellas, and K. Schulten. Discovery through the computational microscope. *Structure*, 17(10):1295 – 1306, 2009.

[63] R. M. Levy, O. De la Luz Rojas, and R. A. Friesner. Quasi-harmonic method for calculating vibrational spectra from classical simulations on multi-dimensional anharmonic potential surfaces. *The Journal of Physical Chemistry*, 88(19):4233–4238, 1984.

[64] R. M. Levy, A. R. Srinivasan, W. K. Olson, and J. A. McCammon. Quasi-harmonic method for studying very low frequency modes in proteins. *Biopolymers*, 23(6):1099–1112, 1984.

[65] J. Liang and J. M. Fernandez. Mechanochemistry: One bond at a time. *ACS Nano*, 3(7):1628–1645, 2009.

[66] J. Liang and J. M. Fernndez. Kinetic measurements on single-molecule disulfide bond cleavage. *Journal of the American Chemical Society*, 133(10):3528–3534, 2011.

[67] G. Lipari and A. Szabo. Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 1. theory and range of validity. *Journal of the American Chemical Society*, 104(17):4546–4559, 1982.

[68] J. Lipfert, D. A. Koster, I. D. Vilfan, S. Hage, and N. H. Dekker. Single-molecule magnetic tweezers studies of type IB topoisomerases. *Methods in Molecular Biology*, 582, September 2009.

[69] F. Liu and Z.-c. Ou-Yang. Force modulating dynamic disorder: A physical model of catch-slip bond transitions in receptor-ligand forced dissociation experiments. *Phys. Rev. E*, 74:051904, Nov 2006.

[70] B. T. Marshall, M. Long, J. W. Piper, T. Yago, R. P. McEver, and C. Zhu. Direct observation of catch bonds involving cell-adhesion molecules. *Nature*, 423(6936):190–193, 2003.

[71] P. Metzner, C. Schtte, and E. Vanden-Eijnden. Transition path theory for markov jump processes. *Multiscale Modeling and Simulation*, 7(3):1192–1219, 2009.

[72] J. R. Moffitt, Y. R. Chemla, S. B. Smith, and C. Bustamante. Recent advances in optical tweezers. *Annual Review of Biochemistry*, 77(1):205–228, 2008. PMID: 18307407.

[73] P. Mjek and R. Elber. Milestoning without a reaction coordinate. *Journal of Chemical Theory and Computation*, 6(6):1805–1817, 2010. PMID: 20596240.

[74] K. Mller and L. Brown. Location of saddle points and minimum energy paths by a constrained simplex optimization procedure. *Theoretica chimica acta*, 53(1):75–93, 1979.

[75] A. Nitzan. *Chemical Dynamics in Condensed Phases: Relaxation, Transfer and Reactions in Condensed Molecular Systems: Relaxation, Transfer and Reactions in Condensed Molecular Systems*. OUP Oxford, 2006.

[76] H. Noji, D. Okuno, and T. Ikeda. Mechanochemistry of F1 motor protein. *Chem. Sci.*, 2:2086–2093, 2011.

[77] J. Nummela and I. Andricioaei. Exact low-force kinetics from high-force single-molecule unfolding events. *Biophysical Journal*, 93(10):3373 – 3381, 2007.

[78] J. Nummela, F. Yassin, and I. Andricioaei. Entropy-energy decomposition from nonequilibrium work trajectories. *The Journal of Chemical Physics*, 128(2):–, 2008.

[79] A. F. Oberhauser, P. K. Hansma, M. Carrion-Vazquez, and J. M. Fernandez. Stepwise unfolding of titin under force-clamp atomic force microscopy. *Proceedings of the National Academy of Sciences*, 98(2):468–472, 2001.

[80] R. W. Pastor, B. R. Brooks, and A. Szabo. An analysis of the accuracy of langevin and molecular dynamics algorithms. *Molecular Physics*, 65(6):1409–1419, 1988.

[81] P. Pechukas and J. Ankerhold. Agmon-hopfield kinetics in the slow diffusion regime. *Journal of Chemical Physics*, 107(7):2444, 1997.

[82] Y. Pereverzev and O. Prezhdo. Deformation model for thioredoxin catalysis of disulfide bond dissociation by force. *Cellular and Molecular Bioengineering*, 2(2):255–263, 2009.

[83] Y. V. Pereverzev and O. V. Prezhdo. Dissociation of Biological Catch-Bond by Periodic Perturbation. *Biophysical journal*, 91(2):L19–L21, July 2006.

[84] Y. V. Pereverzev and O. V. Prezhdo. Force-induced deformations and stability of biological bonds. *Phys. Rev. E*, 73:050902, May 2006.

[85] Y. V. Pereverzev, O. V. Prezhdo, M. Forero, E. V. Sokurenko, and W. E. Thomas. The two-pathway model for the catch-slip transition in biological adhesion. *Biophysical Journal*, 89(3):1446 – 1454, 2005.

[86] R. Perez-Jimenez and J. Alegre-Cebollada. Enzyme catalysis at the single-molecule level. In A. F. Oberhauser, editor, *Single-molecule Studies of Proteins*, volume 2 of *Biophysics for the Life Sciences*, pages 149–168. Springer New York, 2013.

[87] R. Perez-Jimenez, J. Li, P. Kosuri, I. Sanchez-Romero, A. P. Wiita, D. Rodriguez-Larrea, A. Chueca, A. Holmgren, A. Miranda-Vizuete, K. Becker, S.-H. Cho, J. Beckwith, E. Gelhaye, J. P. Jacquot, E. Gaucher, J. M. Sanchez-Ruiz, B. J. Berne, and J. M. Fernandez. Diversity of chemical mechanisms in thioredoxin catalysis revealed by single-molecule force spectroscopy. *Nat Struct Mol Biol*, 16(8):890 – 896, 2009.

[88] J. Ribas-Arino, M. Shiga, and D. Marx. Understanding covalent mechanochemistry. *Angewandte Chemie International Edition*, 48(23):4190–4193, 2009.

[89] M. Rief, M. Gautel, F. Oesterhelt, J. M. Fernandez, and H. E. Gaub. Reversible unfolding of individual titin immunoglobulin domains by AFM. *Science*, 276(5315):1109–1112, 1997.

[90] I. Rips and J. Jortner. The effect of solvent relaxation dynamics on outer-sphere electron transfer. *Chemical Physics Letters*, 133(5):411 – 414, 1987.

[91] F. Ritort, S. Mihardja, S. B. Smith, and C. Bustamante. Condensation transition in DNA-polyaminoamide dendrimer fibers studied using optical tweezers. *Phys. Rev. Lett.*, 96:118301, Mar 2006.

[92] M. Roy, G. Grazioli, and I. Andricioaei. Rate turnover in mechano-catalytic coupling: A model and its microscopic origin. *The Journal of chemical physics*, 143(4):045105, 2015.

[93] T. Sakamoto, M. R. Webb, E. Forgacs, H. D. White, and J. R. Sellers. Direct observation of the mechanochemical coupling in myosin Va during processive movement. *Nature*, 455(7209):128–132, 2008.

[94] K. K. Sarangapani, T. Yago, A. G. Klopocki, M. B. Lawrence, C. B. Fieger, S. D. Rosen, R. P. McEver, and C. Zhu. Low force decelerates l-selectin dissociation from p-selectin glycoprotein ligand-1 and endoglycan. *Journal of Biological Chemistry*, 279(3):2291–2298, 2004.

[95] M. Schaefer and M. Karplus. A comprehensive analytical treatment of continuum electrostatics. *The Journal of Physical Chemistry*, 100(5):1578–1599, 1996.

[96] M. Schlierf, H. Li, and J. M. Fernandez. The unfolding kinetics of ubiquitin captured with single-molecule force-clamp techniques. *Proceedings of the National Academy of Sciences of the United States of America*, 101(19):7299–7304, 2004.

[97] Y. Seol and K. Neuman. *Single-Molecule Measurements of Topoisomerase Activity with Magnetic Tweezers*, volume 778 of *Methods in Molecular Biology*. Humana Press, 2011.

[98] C. S. Sevier and C. A. Kaiser. Formation and transfer of disulphide bonds in living cells. *Nat Rev Mol Cell Bio*, 3(11):836–847, 2002.

[99] D. E. Smith. Single-molecule studies of viral DNA packaging. *Current Opinion in Virology*, 1(2):134 – 141, 2011.

[100] H. Sumi and R. A. Marcus. Dynamical effects in electron transfer reactions. *The Journal of Chemical Physics*, 84(9):4894–4914, 1986.

[101] Y. Suzuki and O. K. Dudko. Single-molecule rupture dynamics on multidimensional landscapes. *Physical review letters*, 104(4):048101, 2010.

[102] R. Szoszkiewicz, S. R. K. Ainavarapu, A. P. Wiita, R. Perez-Jimenez, J. M. Sanchez-Ruiz, and J. M. Fernandez. Dwell time analysis of a single-molecule mechanochemical reaction. *Langmuir*, 24(4):1356–1364, 2008. PMID: 17999545.

[103] M. Taranova, A. D. Hirsh, N. C. Perkins, and I. Andricioaei. Role of microscopic flexibility in tightly curved DNA. *J. Phys. Chem. B*, 118(38):11028–11036, 2014.

[104] W. Thomas. Catch bonds in adhesion. *Annual Review of Biomedical Engineering*, 10(1):39–57, 2008. PMID: 18647111.

[105] W. Thomas and V. Vogel. Biophysics of Catch Bonds - Annual Review of Biophysics, 37(1):399. *Annu Rev Biophys*, 2008.

[106] W. E. Thomas. Understanding the counterintuitive phenomenon of catch bonds. *Current Nanoscience*, 3(1), 2007.

[107] W. E. Thomas, E. Trintchina, M. Forero, V. Vogel, and E. V. Sokurenko. Bacterial adhesion to target cells enhanced by shear force. *Cell*, 109(7):913 – 923, 2002.

[108] P. Tompa. Intrinsically disordered proteins: a 10-year recap. *Trends in Biochemical Sciences*, 37(12):509 – 516, 2012.

[109] G. M. Torrie and J. P. Valleau. Nonphysical sampling distributions in monte carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics*, 23(2):187–199, 1977.

[110] D. G. Truhlar, B. C. Garrett, and S. J. Klippenstein. Current status of transition-state theory. *The Journal of Physical Chemistry*, 100(31):12771–12800, 1996.

[111] Y. Ueda, H. Taketomi, and N. Gō. Studies on protein folding, unfolding, and fluctuations by computer simulation. ii. a. three-dimensional lattice model of lysozyme. *Biopolymers*, 17(6):1531–1548, 1978.

[112] T. S. van Erp and P. G. Bolhuis. Elaborating transition interface sampling methods. *Journal of Computational Physics*, 205(1):157 – 181, 2005.

[113] T. S. van Erp, D. Moroni, and P. G. Bolhuis. A novel path sampling method for the calculation of rate constants. *The Journal of Chemical Physics*, 118(17):7762–7774, 2003.

[114] E. Vanden-Eijnden. Transition-path theory and path-finding algorithms for the study of rare events. *Annu. Rev. Phys. Chem.*, 61:391–420, 2010.

[115] E. Vanden-Eijnden and M. Venturoli. Markovian milestoning with voronoi tessellations. *The Journal of Chemical Physics*, 130(19):–, 2009.

[116] A. Warshel and M. Levitt. Theoretical studies of enzymic reactions: dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *Journal of molecular biology*, 103(2):227–249, 1976.

[117] A. Warshel and R. M. Weiss. An empirical valence bond approach for comparing reactions in solutions and in enzymes. *J. Am. Chem. Soc.*, 102(20):6218–6226, 1980.

[118] J.-D. Wen, M. Manosas, P. T. Li, S. B. Smith, C. Bustamante, F. Ritort, and I. T. Jr. Force unfolding kinetics of RNA using optical tweezers. I. effects of experimental variables on measured results. *Biophysical Journal*, 92(9):2996 – 3009, 2007.

[119] J. Wereszczynski and I. Andricioaei. On structural transitions, thermodynamic equilibrium, and the phase diagram of DNA and RNA duplexes under torque and tension. *Proceedings of the National Academy of Sciences*, 103(44):16200–16205, 2006.

[120] A. M. A. West, R. Elber, and D. Shalloway. Extending molecular dynamics time scales with milestoning: Example of complex kinetics in a solvated peptide. *The Journal of Chemical Physics*, 126(14):–, 2007.

[121] E. Wigner. The transition state method. *Trans. Faraday Soc.*, 34:29–41, 1938.

[122] A. Wiita, R. Perez-Jimenez, K. A. Walther, F. Grater, B. J. Berne, A. Holmgren, J. M. Sanchez-Ruiz, and J. M. Fernandez. Probing the chemistry of thioredoxin catalysis with force. *Nature*, 450(7166):124–127, 2008.

[123] A. P. Wiita, S. R. K. Ainavarapu, H. H. Huang, and J. M. Fernandez. Force-dependent chemical kinetics of disulfide bond reduction observed with single-molecule techniques. *Proceedings of the National Academy of Sciences*, 103(19):7222–7227, 2006.

[124] G. Wilemski. On the derivation of smoluchowski equations with corrections in the classical theory of brownian motion. *Journal of Statistical Physics*, 14(2):153–169, 1976.

[125] I. Wolfram Research. *Mathematica*. Wolfram Research, Inc., Champaign, Illinois, version 10.0 edition, 2014.

[126] C. Xing and I. Andricioaei. On the calculation of time correlation functions by potential scaling. *The Journal of Chemical Physics*, 124(3):–, 2006.

[127] H. Yu, L. Ma, Y. Yang, and Q. Cui. Mechanochemical coupling in the myosin motor domain. ii. analysis of critical residues. *PLoS Comput Biol*, 3(2):e23, 02 2007.

[128] C. Zhu, J. Lou, and R. P. McEver. Catch bonds: Physical models, structural bases, biological function and rheological relevance. *Biorheology*, 42(6):443–462, 2004.

[129] C. Zhu, J. Lou, and R. P. McEver. Catch bonds: Physical models, structural bases, biological function and rheological relevance. *Biorheology*, 42(6):443–462, 2005.

[130] C. Zhu and R. P. McEver. Catch bonds: physical models and biological functions. *Molecular & cellular biomechanics : MCB*, 2(3):91–104, Sept. 2005.

[131] R. Zwanzig. Rate processes with dynamical disorder. *Acc. Chem. Res.*, 23(5):148–152, May 1990.

[132] R. Zwanzig. *Nonequilibrium statistical mechanics*. Oxford University Press, USA, 2001.

Appendix A

Rate turnover in mechano-catalytical coupling: A model and its microscopic origin

A.1 Abstract

A novel aspect in the area of mechano-chemistry concerns the effect of external forces on enzyme activity, i.e., the existence of mechano-catalytical coupling. Recent experiments on enzyme-catalyzed disulphide bond reduction in proteins under the effect of a force applied on the termini of the protein substrate reveal an unexpected biphasic force dependence for the bond cleavage rate. Here, using atomistic molecular dynamics simulations combined with Smoluchowski theory we propose a model for this behavior. For a broad range of forces and systems, the model reproduces the experimentally observed rates by solving a reaction-diffusion equation for a "protein coordinate" diffusing in a force-dependent effective potential. The atomistic simulations are used to compute, from first principles, the parameters of the

model via a quasiharmonic analysis. Additionally, the simulations are also used to provide details about the microscopic degrees of freedom that are important for the underlying mechano-catalysis.

A.2 Introduction

Single-molecule manipulation techniques are increasingly often revealing important biomolecular conformational changes, one molecule at a time. Thereby, they enable one to identify intermediates and to characterize heterogeneity in conformational pathways, properties that otherwise would be masked by the averaging inherent in usual bulk experiments. Typical techniques include pulling by atomic force microscopy (AFM) and by optical or magnetic tweezers to probe individual folding in biomolecules or binding in biomolecular complexes [55, 23, 35]. Other examples include applying forces and torques to study supercoiled DNA [68, 97] and DNA-protein [24, 72] or DNA-nanoparticle complexes [91], to unzip [27, 49, 118] or generate novel forms of DNA[119, 21], to reveal the details of viral packing [99, 20], or to probe the interaction of proteins with lipid membranes [5]. An area of related work concerns understanding how chemical steps, such as ATP hydrolysis, can lead to the generation of force and to the movement of biomolecular machines. Single molecule techniques have here been crucial in detecting and estimating mechano-chemical coupling *i.e.*, the coupling between movement and the chemistry of ATP hydrolysis in molecular motors [3, 93, 127, 44, 52, 76, 31]. Application of external forces induces conformational motion and motion couples to chemistry. It is therefore relevant to seek ways in which applied forces modulate chemistry. This is precisely what has been explored recently by Fernandez and coworkers [66, 43, 59, 123, 87, 65], who, in a novel experiment, have studied how external forces affect the the quintessential chemical act of catalysis. The technique used was single molecule force clamp spectroscopy (SMFCS), a method of precise constancy in the force

application [22, 96, 79, 39], that has proved particularly useful previously in the characterization of the mechanical unfolding/refolding of proteins. They revealed a coupling between mechanically applied forces and the chemistry of bond cleavage in a *catalytic* reaction, i.e., the rate of force-catalyzed chemical reactions at the single molecule level. In particular, they studied the force dependence of the reduction of disulphide bonds in a protein substrate, titin when catalyzed by the enzyme thioredoxin [4] (reduction which occurs *in-vivo*), and when catalyzed by different small nucleophiles [123, 43, 66, 102]. Two opposing mechanical forces were applied via AFM to pull apart the C- and N-termini of the immunoglobulin-like domain number 27 (I27) of titin, which had an engineered disulphide bond between residues 32 and 75 (see Fig. A.1). The protein unfolded from the two termini up to the sequestered disulphide bond, which, from being buried inside the folded protein, now became both exposed to the nucleophilic moiety and stretched by the same mechanical force that caused the unfolding of the intervening protein backbone "handles" $1-32$ and $75-89$ (Figure A.1). The disulphide bond was subsequently reduced (cleaved) by thioredoxin or, in their subset of experiments, by the small nucleophiles, and the cleavage resulted in further extension of titin.

When the small nucleophiles were present, the disulphide reduction followed the kinetics of an ${S_N}^2$ reaction with a first order dependence of the reaction rate on the nucleophilic concentration and an exponential dependence on the force as in Bell's model [12],

$$\text{rate} = A \exp(-(E_a - F\Delta x_r)/k_B T)[\text{Nu}], \tag{A.1}$$

where [Nu] is the concentration of the nucleophile and E_a is the activation energy barrier, lowered by the external force, F by an amount $F\Delta x_r$, with Δx_r the distance to the transition state along the reaction coordinate, identified as the elongation of the disulphide bond [66, 89, 59]. By and large, the rate of reduction by small nucleophiles was exponentially accelerated by the force on titin. However, when thioredoxin was the catalyst, disulphide bond

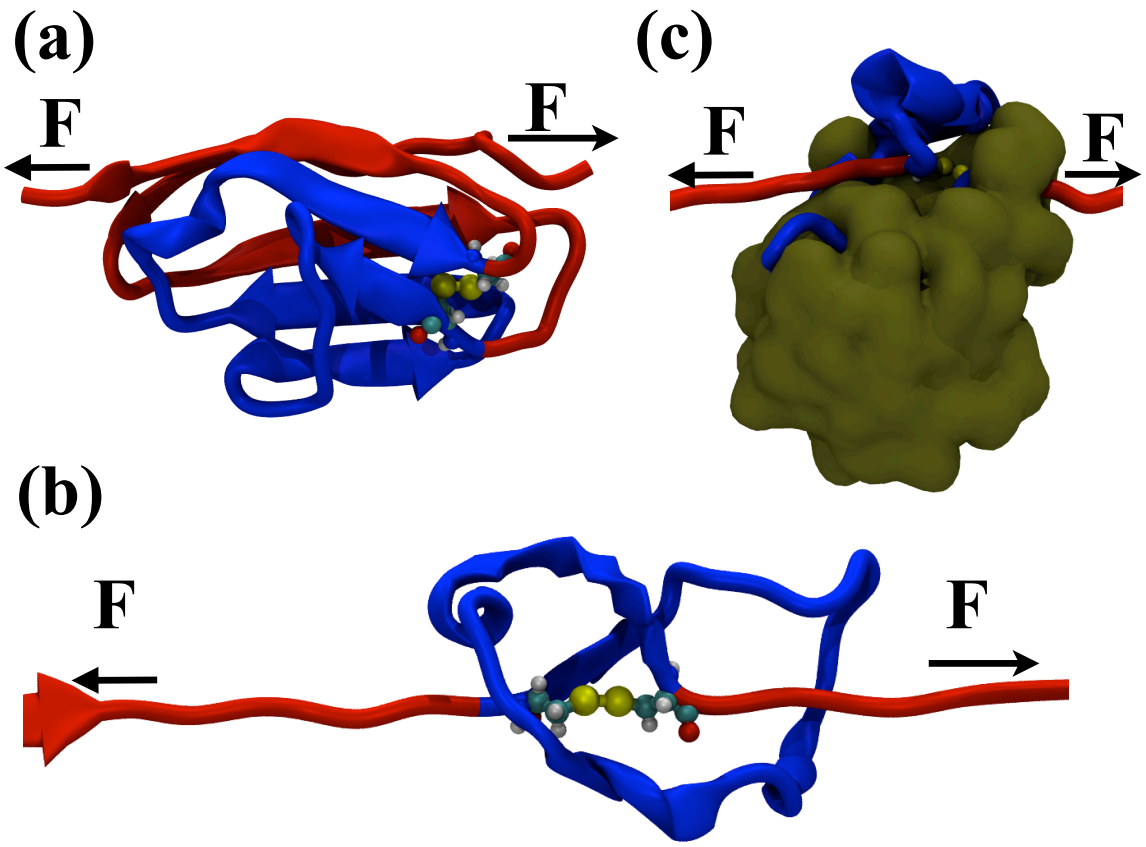

Figure A.1: Different stages of external force application (a) Force applied on the native titin I27 with an engineered disulphide bond at Cys32-Cys75. (b) Force applied to exposed disulphide bond through the extended C- and N-termini "handle" (viewed from the opposite direction relative to panel (a)). (c) Force applied on titin "handles" while the exposed disulphide bond is in close proximity to the active site of enzyme thioredoxin, shown in green (PDB : 1XOB).

reduction exhibited a biphasic force dependence [4, 122, 86] in the form of a turnover in the rate-force plots. Both eukaryotic and bacterial thioredoxins were studied. While all thioredoxins showed a negative force sensitivity at lower forces (the rate decreased with force), this was followed, at larger forces, by a force independent behavior for eukaryotic thioredoxins and an increase in the rate with the force for bacterial thioredoxin. The distinct chemical mechanisms underlying the catalytic activity of the two types of thioredoxin enzymes not seen in small nucleophiles was rationalized to be modulated by the highly conserved active site in the enzyme, defined by two "vicinal" cysteine residues at the 32 and 35 positions [48, 38], as well as the surface and depth of the substrate binding grove [87, 50, 9, 25].

Molecular dynamics simulation studies for thioredoxins of different origins attributed the biphasic kinetics in *E-coli* thioredoxin (bacterial) to the shallow binding grove controlling the chemistry of the reaction at lower forces [87] (substrate binding being the rate limiting in the absence of the force), according to a Michaelis-Menten mechanism. At higher forces, the reaction proceeds according to a simple ${S_N}^2$ mechanism, and the formation of the enzyme substrate complex is no longer the rate determining step. The force independent behavior in the case of eukaryotic thioredoxin can be explained on the basis of a single-electron transfer reaction [100] taking place irrespective of the orientation of the disulphide bond [90]. The experimental and molecular dynamics simulations of peptide bound enzymes also confirmed that the eukaryotic thioredoxins have a deeper binding grove which can lock the substrate disulphide bond, preventing further conformational variability. Even though the reduction rate was found to be force accelerated, following a Bell model (or generalizations thereof) in the case of the small nucleophile, this was not the case for the biphasic turnover in the force-dependence of the thioredoxin catalyzed chemical rate. This points to the idea that the force also modulates the behavior of the protein environment surrounding the cleaved disulphide.

The force dependence observed in titin's disulphide cleavage reaction is reminiscent of an otherwise unrelated class of complexes with similarly biphasic rate-vs-force profiles. The prime feature of this class is a so-called "catch-slip" transition, seen when pulling apart certain adhesive supramolecular complexes, such as the binding to P-selectin and L-selectin of the P-selectin glycoprotein ligand-1(PSGL-1) [10, 94, 104, 70] or the adhesion of the protein FimH to bacterial host cells [107]. The concept, introduced by Dembo *et al.* in 1988 [30], describes rates of dissociation of the ligand that, counterintuitively, first decrease with the pulling force, a range for which interactions are coined "catch bonds" (although no actual covalent bond exists). Subsequently, rates increase with force beyond a certain threshold, a force regime termed "slip bonds". Theories and phenomenological models explaining the dynamic transitions in catch-slip bonds have been developed, chiefly based on the existence

of an energy landscape with two bound states or two pathways [106, 129, 83, 11, 130, 105, 36, 85].

A natural framework –perhaps the simplest– to rationalize the qualitative change in the reaction rate for disulphide cleavage with the force applied to the protein is a two-dimensional reaction-diffusion model. An earlier incarnation of a related approach is the venerable Agmon-Hopfield model [2] (see also Ref. [8]), which describes a first-order kinetic process (more precisely, CO binding to a protein) whose rate depends on the "protein coordinate," i.e., on a variable that diffuses in time [81]. The protein coordinate can be thought of, in effect, as a displaced normal mode, or some linear combination of normal modes, of the protein [131]. An essentially similar description, with the chemistry being this time electron transfer, is the Sumi-Markus model [100]. In any case, in such models, the reaction coordinate, r, is the one along which chemistry occurs and it is coupled to an orthogonal coordinate, the conceptual protein coordinate, x, which evolves according to a Smoluchowski-based reaction-diffusion equation,

$$\frac{\partial \rho(x,t)}{\partial t} = D\frac{\partial^2 \rho(x,t)}{\partial x^2} + \frac{D}{k_B T}\frac{\partial}{\partial x}(\rho(x,t)\frac{\partial V}{\partial x}) - k(x)\rho(x,t), \tag{A.2}$$

where $\rho(x,t)$ is the probability density of x, D is the diffusion coefficient and the last term relates to the rate of the reaction.

Herein we pursue a description of the force-modulated kinetics of disulphide cleavage using a framework inspired by the Agmon-Hopfield or Sumi-Markus models embodied by Eq. (A.2). Our approach is, in spirit, similar to the treatment of catch-slip bond transitions proposed by Liu and Ou-Yang [69], who assume that the distribution of protein conformations involved in the adhesive complex is modulated by the external force, which also couples to the catch-slip detachment coordinate. In our case, the parameters of our model are derived from a set of all-atom molecular dynamics simulation of both the protein titin and titin-thioredoxin

complex under various applied forces. A quasiharmonic analysis of the substrate protein titin is used to evaluate candidate modes that can describe the collective motion of the protein coordinate. Our study highlights the importance of involving force-dependent protein modes in theoretical descriptions of mechano-chemical coupling (see Fig. A.2). The rest of the paper is organized as follows. We continue by introducing our novel treatment of the applied force in force-modulated enzyme-catalyzed disulphide bond reduction experiments where the force is represented as a vector with components in the reaction coordinate and protein coordinate, then we present the Smoluchowski-based formalism for reaction-diffusion in this two dimensional space, followed by a description of how we used atomistic simulations to simulate force unfolding and to compute the parameters for the force-modulated reaction-diffusion potential; we end with a concluding discussion.

A.3 Results

As described above, a diversity of mechanisms for the thioredoxin-catylized disulphide bond cleavage of titin, as it was subjected to pulling forces, were experimentally observed. There were two distinct rate-force dependencies observed, depending on the variety of thioredoxin present [87]. The two types of dependences corresponded to the two distinct families of thioredoxins studied, i.e., eukaryotic vs. bacterial. A prime role in this dichotomy was played by the binding groove, which was different in the two families, and which controlled the force application on the active site and hence on catalysis. However, a quantitative description of the rate remained to be obtained. Here, we investigate a model in which the diversity in the rates of disulphide cleavage is shown to result from a different force response of the protein coordinate, which in turn affects the reaction coordinate by modulating its energy barrier and transition state position. We simulated with molecular dynamics a family of four bacterial thioredoxin-titin complexes, used in the study by Perez-Jimenez et al. [87] to understand the

structural dynamics at the active site of thioredoxins in proximity of the substrate prior to bond cleavage that may affect rates of their catalytic reactions. A simple representation of the distribution of the protein coordinate is a Gaussian resulting from diffusion in an effective harmonic well. As such, we modeled it using quasi-harmonic analysis [53] of the protein-enzyme complexes sampled by the molecular dynamics trajectories. In this representation, the protein coordinate embodied collectively the global motions typically associated with large-amplitude, low-frequency modes of the complex. Such modes are often responsible for conformational change and function [15]. The difference in the modes and in their response to force are meant, in the model, to modulate the reaction rates for each complex. We note that while the protein coordinate moves in a harmonic well, anharmonicities in the protein motion are accounted for implicitly by averaging in the quasi-harmonic approximation [64, 63], which can also be used to estimate the effect of the force on entropy [7, 103].

The quasi-harmonic analysis of each of the four protein-enzyme complexes is featured in Fig. A.3, showing the normalized density of quasi-harmonic states as well as the cumulative density of states, $G(\omega)$ calculated for the $3N - 6$ modes for all complexes, plotted against the mode frequencies, ω.

To correlate the collective motions the enzyme-substrate complex to the biphasic kinetics of the enzymatic disulphide reduction reaction, we propose to use a force modulated diffusion model - a "bounded diffusion" orthogonal to the reaction coordinate, bound by a force-modulated harmonic potential. The conformational variations corresponding to low frequency vibrations give rise to a fluctuating intrinsic energy barrier, which being modulated by force is revealed as a biphasic behavior in the rate. The particular force-modulated reaction-diffusion equation utilized here is the generalization of the Agmon-Hopfield model [2] by Liu and Ou-Yang [69], which is in effect, an expanded form of the Smoluchowski

dynamics in an external field [124].

$$\frac{\partial \rho(x,t)}{\partial t} = D\frac{\partial^2 \rho(x,t)}{\partial x^2} + D\beta\frac{\partial}{\partial x}\rho(x,t)\frac{\partial V\,(x,F_\perp)}{\partial x} - k_{\text{off}}(x,F_\parallel)\rho(x,t) \tag{A.3}$$

where, $\rho(x,t)$ is the probability density of finding the value x at a time t, and the two components of the pulling force $\mathbf{F}$ along the protein and reaction coordinate are, respectively, given by

$$F_\perp = \mathbf{F}\cdot\hat{x} = F\sin\theta \tag{A.4}$$

$$F_\parallel = \mathbf{F}\cdot\hat{r} = F\cos\theta, \tag{A.5}$$

with θ the angle between the applied force and the reaction coordinate, r; D is the diffusion coefficient and β the inverse temperature. $V(x,F_\perp)$ is the force-modulated potential acting on the system, which is a function of both the position along the protein coordinate, x, and the perpendicular component of the applied force. The potential is expressed as:

$$V(x,F_\perp) = V_0 + \frac{1}{2}\kappa(x-x_0)^2 - F_\perp x, \tag{A.6}$$

with V_0 the minimum value of the potential of force constant κ, and x_0 the location of the minimum along the protein coordinate, x. The term $-F_\perp x$ models the amount by which the component of the force along the protein coordinate x tilts the energy along x. Herein we define the *protein coordinate* x as a conformational coordinate along which the motion of the system is orthogonal to the reaction coordinate r; x can be thought of as a linear combination of protein "breathing" modes. The final, sink term in the Smoluchowski equation, Eq. (A.3), consumes probability density directly proportionally to the reaction rate coefficient, k_{off}. The

rate coefficient is itself a function of the protein coordinate (via the $x-$dependent energy barrier height), as well as of the parallel component of the applied force:

$$k_{\text{off}}(x, F_{\parallel}) = k_0 \exp[-\beta(\Delta V^{\ddagger}(x) - F_{\parallel} r^{\ddagger})] \tag{A.7}$$

The component of the force parallel to the reaction coordinate r tilts the energy surface by the amount $-F_{\parallel} r^{\ddagger}$, with $r^{\ddagger}$ the distance to the barrier, namely the separation between the bound state and the energy barrier for disulphide cleavage. Following Liu and Ou-Yang [69], the shape of the reaction energy barrier height $\Delta V^{\ddagger}$ as a function of the protein coordinate x was modeled as a piecewise function, initially with a positive slope until an equilibrium distance, after which it assumed a zero slope. The Smoluchowski equation was then solved numerically for $\rho(x,t)$ using the partial differential equation solver in *Mathematica* [125], with Dirichlet boundary conditions $\rho(x_{max}, t) = 0$ and $\rho(x_{min}, t) = 0$, and the initial condition:

$$\rho(x, 0) = \frac{1}{\sqrt{\pi}} e^{-(x-1)^2}. \tag{A.8}$$

Although Neumann boundary conditions (which cancel the flux at the boundaries) are also possible, the simpler to implement Dirichlet boundary conditions were employed here without loss of precision by placing them at values of x_{min} and x_{max} which corresponded to states energetically inaccessible to the system at room temperature (we chose $x_{\text{min}} = -22$ and $x_{\text{max}} = 22$). Thus the potential itself keeps the system bounded in the x direction, and the boundary conditions merely serve to define zero-valued endpoints for the numerical method. The ultimate goal of generating the ρ(x,t) surfaces is to integrate them over all space and time, in order to generate τ, the disulphide bond lifetime:

$$\tau = \int_0^{\infty} \int_{-\infty}^{\infty} \rho(x, t) dx dt. \tag{A.9}$$

The sink term in the Smoluchowski equation ensures that the probability density decays to

zero at long times. In the numerical implementation, integrating the surface over infinite time was made tractable by setting the upper limit of the time integral to a value larger than the time t at which $\rho(x,t)$ has decayed more than a cutoff of 10^{-13} of the initial-time value integrated over all x. Since $\rho(x,t)$ decays to zero long before reaching the boundaries, integrating between the boundary conditions is equivalent to integrating over all space. The above process of numerically solving for $\rho(x,t)$ and then integrating it with Eq. (A.9) was repeated for increasing values of force from 0 to 600 pN, and the resulting values for lifetime τ were inverted to find the numerical values of the reaction rates. These reaction rates could then be plotted as a function of the applied pulling force and was fitted to the experimental data of Perez-Jimenez *et al.* [4] by collectively varying the distance to the reaction barrier $r^{\dagger}$, the rate constant for the reaction in the absence of an applied force k_0, and the force constant for the effective harmonic potential of the protein coordinate κ. Goodness of fit was monitored using a cost function which summed the squared differences between the experimentally measured rates and the calculated rates for each given force value. Varying the force constant κ of the force-modulated potential $V(x, F_{\perp})$ varied the rate, reinforcing the importance of the "softness" of the underlying protein coordinate. The matching protein coordinates were subsequently identified atomistically from the the low frequency modes of the quasi-harmonic analysis. *i.e.*, when the distribution of the protein coordinate was bound by a harmonic potential of different force constants κ for each enzyme-substrate complex. The variability in the low frequency quasi-harmonic modes in the inset to Fig. A.3 originates from the different protein environment engulfing the substrate disulfide in each binding grove pictured in Fig. A.2(b), and resulting in the variety of curves for the rates in Fig. A.4.

The two key parameters that strongly control the shape of the force-rate curves in Fig. A.4 are κ, the force constant defining the force modulated harmonic potential along the protein coordinate and $r^{\dagger}$, the distance along the reaction coordinate from the bottom of the reactant well to the transition state. The values for κ ranged from 8 pN nm^{-1} (2TRX) to 24 pN nm^{-1} (1XOB), and $r^{\dagger}$ ranged between 0.008 nm (1UVZ) and 0.02 nm (2TRX). Adding to

the information obtained through low frequency normal modes, we used molecular dynamics simulations to capture the conformational transitions within the substrate in the presence of force [42, 122] that occur in close proximity to the enzyme thioredoxin, proximity being a measure of r, the distance to the transition state. The force constant, κ, is the key parameter in establishing a connection between the Smoluchowski model for force-modulated chemistry and the atomistic simulations presented herein. Quasi-harmonic mode analysis of the protein/substrate system resulted in a matching distribution of low frequencies from less than 1 to around 30 cm^{-1}. Upon solving for κ from $\omega = \sqrt{\kappa/m}$, where ω is the frequency and m the reduced mass of the oscillator, a value of κ can be identified. This relates the force constant of the Smoluchowski model to a particular region of the spectrum of quaisharmonic modes. For example, for a total mass of the atomistic system in the simulation of roughly 15,000 amu, the larger reduced mass is $7500 \times 7500/15000 = 3570$ amu, which would correspond to a large scale oscillation of two portions of equal mass. This establishes the upper limit for reduced mass, with lesser values possible, (all the way down to 1 amu, corresponding to a single hydrogen atom oscillating against the rest of the system). As an example, if we assume that the normal mode which best corresponds to our protein coordinate is a low frequency mode approximated as motion involving about 5% of the full structure oscillating from the remainder of the structure, then we can use the previously stated relationships to establish a range for κ of around 10 pN/nm corresponding to frequency = 0.5 cm^{-1} to around 27 pN/nm for frequency = 0.8 cm^{-1}. In comparison, for the Smoluchowski equation used to generate the rate vs. force curves in Fig. A.4, the κ parameter was varied between 8 and 24 pN/nm to produce the fits. Although this matching strategy does not generate a unique mode identification, it does provide a reasonable picture of the protein coordinate, and it invites further exploration into ascribing physical meaning to motion along it using linear combinations of low-frequency normal modes.

At the atomistic level, we were interested in identifying the conformational changes at the site of catalysis induced by the application of the external force. To this end, we used all-

	κ (pN/nm)	$r^{\dagger}$ (nm)	θ (rad)	k_0 (s^{-1})
Blue	9.9798	0.008	0.1	6.0×10^6
Red	23.9798	0.01215	0.166	2.8×10^6
Black	11	0.016	0.1	2.0×10^6
Brown	8	0.02	0.1	1.1×10^6

Table A.1: Parameter values from the Smoluchowski equation used to fit the curves representing disulphide bond cleavage rate as a function of the applied force to the experimentally measured values for different forms of Thioredoxin, as shown in figure A.4.

atom simulations with force applied explicitly to the system (see Methods). At this higher resolution, we were able to identify two *atomic-level* parameters controlling the contribution of the protein coordinate at the substrate level. The first is the dihedral angle at the substrate disulphide bond and the second is the orientation of the disulphide bond with the applied force axis, both providing the microscopic response to applied force. During the first force pulse, the application of which initiated the unfolding of the substrate titin to expose the buried disulphide bond, the dihedral angle maintains an equilibrium value between 100° to 110°[17]. However, for larger applied forces, a flip to the other gauche conformation at 260°, an approximately 180° flip was observed (see Figure S1 (SI)). A 90° dihedral angle produces a staggered low energy configuration for the carbon atoms around the S-S bond, due to a favorable $\pi - \pi$ overlap, which in the protein environment has an equilibrium value of 100°-110°.

Interestingly enough, the dihedral angle distribution exhibits a different response to force when the disulphide is proximal to the oxidized state than it does when proximal to the reduced state of the thioredoxin enzyme. This points towards the influence of the enzyme in controlling the protein coordinate. The C-S-S-C dihedral bond distribution in titin is peaked away from the equilibrium conformation and assumes the transition state geometry at around 180°, a flip from its stable gauche- to the trans- state. This is observed mostly at low forces (100-200 pN), as the titin sulphur atom in CYS32 moves towards the sulphur atom in the active site of the enzyme (S32); the distance between these atoms is, roughly, the reaction coordinate for chemistry. This is seen more prominently when the cysteines

at the active site of the enzyme are in thiol form (as CYS32-SH and CYS35-SH) since a reduced state is capable of bringing about the thiol-disulphide exchange reaction [98] with the substrate disulphide bond. The force modulated behavior of the substrate coordinate is also reflected by the dependence of the root-mean-square deviation of the "trapped core" residues with force and by the close proximity of the enzyme. This underlies the role of these residues in the force-controlled reduction reaction (see also Fig.S3 [in SI]). Higher RMS fluctuations of residues adjacent to the disulphide bond (see Fig. S1(d)[in SI]) even in the absence of enzyme emphasize the fact that the force induced fluctuations at the single bond or dihedral level are, in fact, a consequence of large scale conformational changes or deformations [82, 84] of the substrate induced by force.

The second important atomic-level parameter observed through our force-dependent simulations is the orientation of the disulphide bond with respect to the force axis. The orientational flexibility of the disulphide bond at the active site is influenced by the protein coordinate, thereby modulating the reaction coordinate r. Through a study of the evolution of the binding grove of the enzyme thioredoxin and dwell time analysis techniques [102], it was proposed that at lower forces, the disulphide bond orients itself against the force axis to attain a linear configuration for the ${S_N}^2$ reaction, shortening its bond length and thereby the distance to the transition state. This results in a negative force dependency [4, 102], with the bond elongation projected onto the force axis being $\Delta x_F = \langle \mid \cos(\theta) \mid \rangle (b_{TS} - b)$, where b and b_{TS} are the reactant and transtion-state bond lengths and θ is the orientation angle of the disulphide bond with respect to the force axis, *i.e.*, the angle between the reaction coordinate and the applied force in our proposed model. Hence, the chemistry is likely controlled by the peptide binding groove in the enzyme requiring a rotation of the substrate disulphide bond with respect to the active site of the enzyme, in accord with the proposed Michaelis-Menten kinetics at low forces.

As the disulphide bond gets exposed upon initial unfolding, in our simulations of only the

substrate titin under a continued force application, the disulphide bond was observed to be aligned at either 40° or 140° (taking into account both direction of force vector) to the force vector (see Figure A.6). With increasing proximity of the substrate titin to the active site of the enzyme, while moving towards the enzyme active site (i.e., towards S32) along the reaction coordinate, the disulphide bond in our simulations is seen to divert from the equilibrium value and orient itself almost perpendicularly to the force axis at lower forces (100-200 pN), however, aligning back to its equilibrium state at 40° or 140° when subjected to higher forces. We hypothesize that at lower forces, and in proximity to the active state of the enzyme, the conformational dynamics of the substrate is influenced by the binding grove of the enzyme or, in other words, the protein coordinate of the enzyme. As the force increases, it wins over the control aligning the bond back along the force direction. This analysis prompts at the importance of conformational fluctuations in positioning the disulphide bond in certain orientations relative to the external pull. As such, it is important to generate a ensemble of conformation and analyze how the entirety of the conformations (whose distribution itself may be affected by force) may modulate the rate at different forces.

A.4 Concluding Discussion

Recent single molecule experiments on thioredoxin-catalyzed thiol-disulphide exchange unveiled a complex rate-force dependence. The rate initially decreased for forces below 200 pN and increased at higher forces. In principle, the initial decrease could formally be considered as a consequence of an effective barrier increase with force, modeled by Δx_r and F having opposite signs if the Bell model (Eq.(A.1)) is imposed [83]. This is would be then followed by a subsequent increase at larger forces, caused by a coupling of the force with the elongation in bond length (a regular, positive Δx_r). Such a force-reaction coupling modeled à la

Bell is necessarily one-dimensional. In the case of complex macromolecular systems, as are protein-enzyme complexes, the force applied to the substrate protein is more likely to act along directions other than the reaction coordinate. Hence the factors dictating the distance to the transition state and subsequent kinetics cannot be identified by a single bond elongation, Δx_r, but by a combination of several parameters. Here we offered evidence that such parameters arise naturally from internal protein coordinates, that vary or are modulated by force as the system progresses from the reactant to the transition state. We modeled an internal protein coordinate as a linear combination of low-frequency quasi-harmonic modes which differ in different active site environments, as represented by complexes of the same substrate with different enzymes. Accordingly, rates of different complexes exhibited a slight difference that was in agreement with experiments and the measured the biphasic behavior observed in all bacterial complexes. We successfully reproduced the biphasic force dependency of rates by simultaneously propagating the protein coordinate and reaction coordinate along the two components of a force modulated potential by solving a generalized version of a reaction-diffusion equation.

Via atomistic simulations we also studied the conformational dynamics of the partially unfolded core (i.e., when the handles are unfolded, but the core is held folded by the disulphide bond). These simulations further revealed the conformational ensemble of the dihedral angle geometries at distances close to the transition state. Not only was the disulphide bond affected by the direct application of the force, but also, importantly, the residues trapped behind the disulphide bond showed a higher degree of conformational variability with force in the presence of the enzyme, thus additionally impacting the average transition geometry of the disulphide bond. Orientation of the disulphide bond away from the force axis at lower forces and at closer distances to the enzyme emphasized the role of the peptide binding groove in controlling the dynamics of the reaction and the rate deceleration at low forces. The substrate was seen to be more impacted by higher forces, even in the presence of the enzyme, in accord with the fact that the reaction is accelerated at these forces and is

independent of the peptide binding grove [79].

While our reaction-diffusion equation used a simple model for the force-dependent kinetics, more detailed calculations can be set up to provide from first principles a description of a potential energy surface warping under the strain of an applied force. For example, Ribas-Arino et al. [88] used electronic structure calculations to describe the effects of applied forces on a simple bond cleavage reaction by directly computing force-transformed potentials. In principle, for enzymes the possibilities exist to estimate directly force-transformed potentials using QM/MM techniques [116, 40].

To conclude, we provided microscopic evidence of the protein conformational coordinate through our simulations and quahiharmonic mode analysis which successfully validated our model of force modulated diffusion of protein and reaction coordinate along two perpendicular dimensions. Similar descriptions are also relevant for the force and torque effects on the activity of enzymes on nucleic acid substrates during genetic transactions [6]. We expect our study to be important as more experimental examples of mechano-chemical coupling, new sono-chemical coupling [26, 47] or coupling to electrical fields [41] become available.

A.5 Methods

Protein Data Bank coordinates [PDB ID: 1TIT] were used for the substrate, the 89-residue β-sandwich protein titin-I27, responsible for the regulation of passive elasticity in muscles. The disulphide bond in titin-I27 was engineered by computational mutations at G32C and A75C, sequestering the amino acids behind the disulphide bond, residues 33 to 74, in a particular conformation that generated a "covalent trap" to impede complete unfolding, unless the bond was cleaved. The originally present cysteines 47 and 63 were mutated to alanine (in accord with the single molecule pulling experiment, where one needed to prevent complications

from unnecessary disulphide bond formation). All mutations were carried out with MMTSB package [*Michael Feig, John Karanicolas, Charles L. Brooks, III: MMTSB Tool Set (2001), MMTSB NIH Research Resource, The Scripps Research Institute*]. Constant force AFM simulations were carried out using the CHARMM27 force field in the ACE implicit solvent model [95]. Mimicking the experimental set up of a double force protocol, the simulations also used two-stage unfolding. The first stage of force application unfolded the native state of the protein up to the disulphide buried in the protein core. Since forces lower than the experimental maximum of 400 pN were unable to unfold the titin molecule from the native state in the allotted simulation time, higher forces in the range of $400 - 800$ pN had to be applied to initiate unfolding. Because in the empirical force field the disulphibe bond is modeled as a harmonic potential (with a force constant 173 kcal/mol/$\mathring{A}^2$), bond cleavage was neither possible nor expected. (Incidentally, it is experimentally established that forces less than nN cannot break a covalent bond [45].). While in principle an empirical valence bond model [117] may be used to effect bond breakage, the main aim here was to study the distribution of disulphide bond geometries assumed in the pulling experiment. We ran twenty 20-ns trajectories at each force. The residues trapped behind the disulphide bond $(33 : 74)$ hence remain in the folded state. For the second stage, the protein was pulled with a high force for a short duration until the disulphide bond is exposed while fixing the coordinates of the core residues from $32 - 75$ including the disulphide bond. It was then exposed to continued application of forces in the range $100 - 800$ pN (since no further unfolding was required). Twenty shorter trajectories of 10 ns each were run to sample the conformations of the single disulphide bond under a range of forces in the absence of the enzyme. To further model the conformational ensemble of the protein coordinate in the presence of the enzyme (thioredoxin) when the substrate is subjected to a range of external forces, thioredoxin [PDB ID: 1XOB] was placed at different distances from the substrate titin with the distance between titin-S32 and Trx-S32 sulphur atoms (active site) harmonically restrained at different distances representing the reaction coordinate. At each

force, 20ns trajectories were run for the force range 100-600 pN applied on the two C-α atoms of the termini (residues 1 and 89) of titin with the disulphide bond in the exposed state. Distributions were generated for time dependent dihedral angle variation at the substrate disulphide bond as well as the orientations of the disulphide bond with the force axis.

To identify the low frequency quasi-harmonic modes responsible for the protein conformational coordinate in the various titin-thioredoxin complexes, four thioredoxins of bacterial origin(human mitochondrial Trx2[PDB 1UVZ], *E. coli* Trx1[PDB 1XOB], *E.coli* Trx2 (PDB ID: 2TRX) and C.reinhardtII Trxm (PDB ID 1DBY) were complexed with the titin structure PDB ID: 1TIT and docked over a complex of human-TRX with transcription factor NFκB[PDB 1MDI] to get the correct transition state geometry. The distance between the sulphur atoms at the exposed disulphide bond of Cys-32 of titin and the active site of Cys-32/Cys-31 of the enzyme was restrained at 2.02 Åaverage distance using a harmonic potential. The structure was minimized in 200,000 steps of steepest descent followed by 200,000 steps of the adapted basis Newton- Raphson method, heated and equilibrated for 20 ns at 300 K using Langevin dynamics with implicit solvent. The two 'handles' in the partially unfolded form of titin (with the 'hidden' disulphide bond at 32-75 position being unsequestered), namely the residues $1-30$ and $78-89$ were deleted. Quasi-harmonic analysis was conducted on the protein core in complex with the enzyme. The trajectory was used to converge the $3N \times 3N$ covariance matrix of the fluctuations of the N atoms. The global translation and rotation was removed prior to the analysis by least square fitting to the reference structure using all heavy atoms. The average coordinates during the trajectory were used as reference. The transformation from the Cartesian to normal coordinates was performed by diagonalizing the mass-weighted covariance matrix [18], resulting in $3N$ eigenvectors identifying the mode motions and the corresponding 3N eigenvalues representing the frequency of each mode. The density of states, $g(\omega)$, as well as the cumulative density of states, $G(\omega)$ were computed and are shown in Fig. A.3 in semi-log or log-log plots against the mode frequencies.

A.6 Supplementary Information

Additional plots and table generated from atomistic simulations and corresponding theoretical calculation are included in the Supplementary Material.

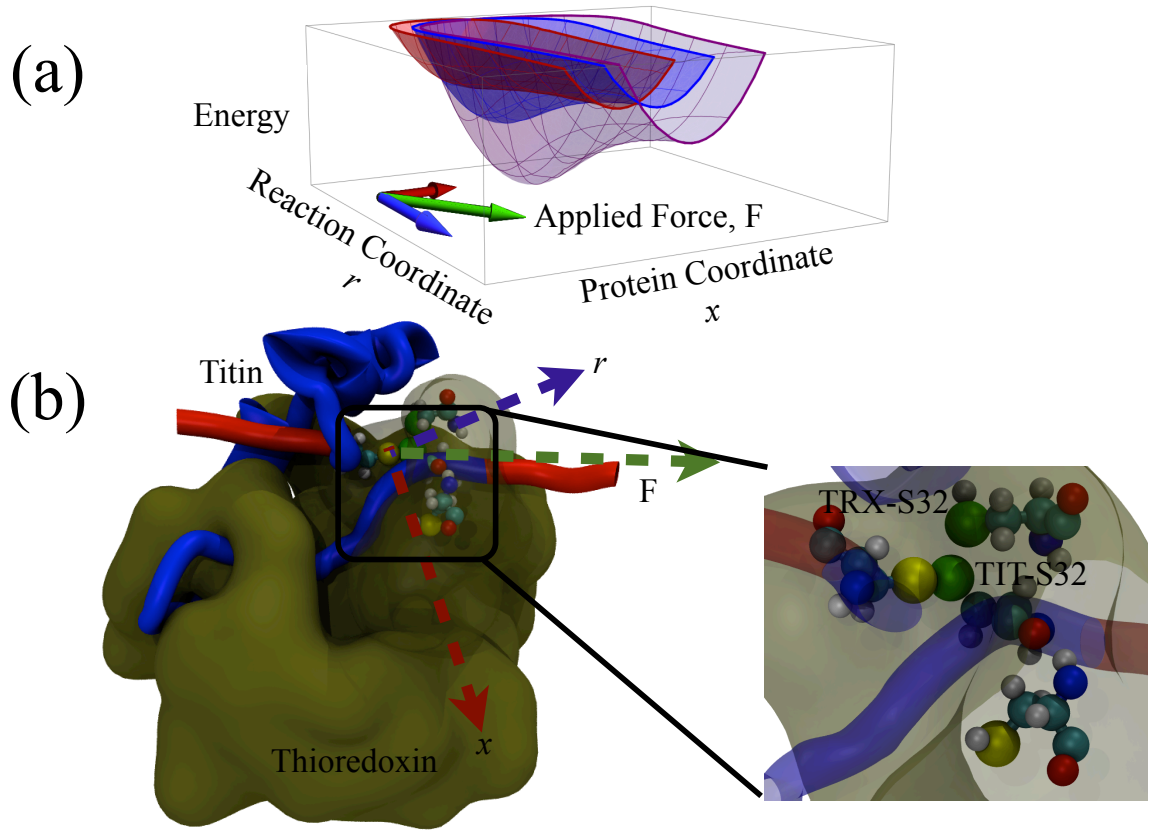

Figure A.2: (a) Sketch of a potential energy surface distorting under an applied force (green arrow) with vector components along the 'protein coordinate' (red arrow) and 'reaction coordinate' (blue arrow). The red surface corresponds to an applied force of smallest magnitude, with blue being larger, and purple being the largest. The small increase in applied force magnitude going from the red surface to the blue surface causes an increase in well depth for the reaction coordinate, but little change in energy for the non-bonded state along the reaction coordinate, while the highest magnitude applied force shows a pronounced decrease in energy for the non-bonded state (compare the positions of the parabolic cross-sections of the surfaces). This is analogous to the turnover behavior observed in force-modulated disulphide bond reduction, where smaller magnitude forces favor the bound state, while higher magnitude forces increase the rate of bond breakage, i.e., favor the unbound state. (b) The enzyme (thioredoxin) in complex with the substrate protein (titin, in cartoon representation) at the transition state as modeled during atomistic simulations(see Methods). The inset shows the detailed view at the active site with the reacting sulfur atoms CYS 32 of thioredoxin and CYS 32 of titin colored in green aligned along the reaction coordinate. The sulphur atoms of the proximal cysteines (CYS 75 in titin and CYS 35 in thioredoxin) are colored in yellow. The arrows represent the same orthogonal coordinates as in part (a), and employ the same color scheme.

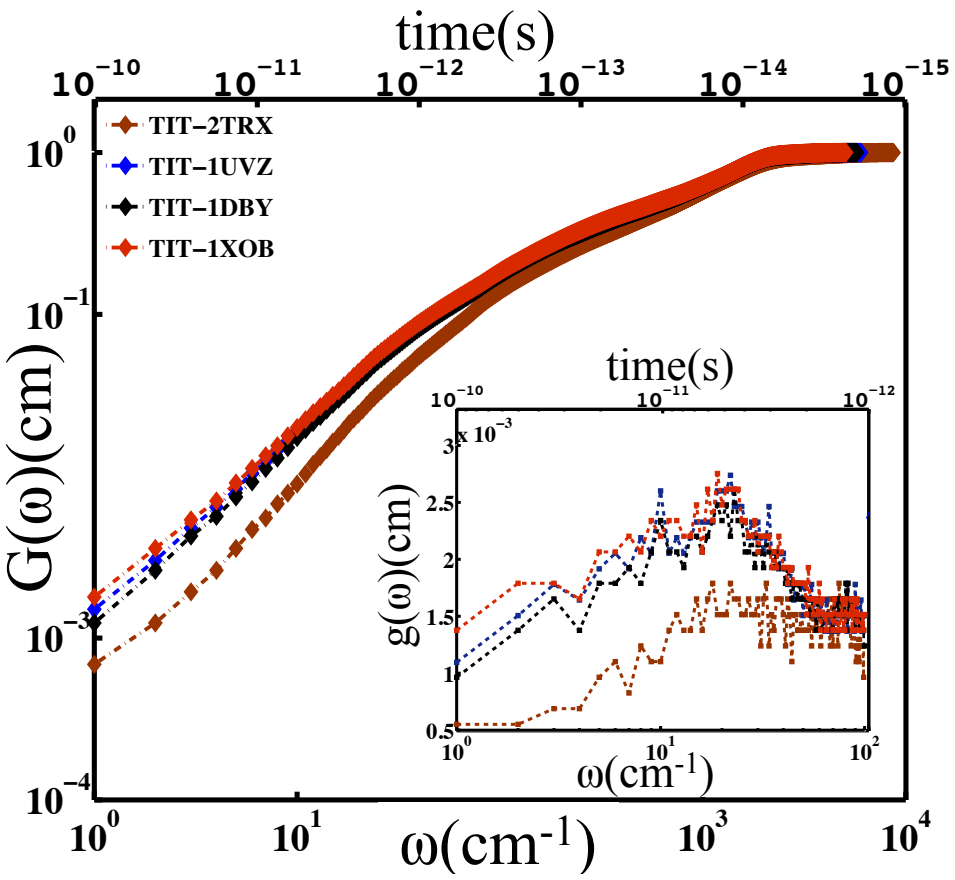

Figure A.3: Normalized G(ω), cumulative density of quasiharmonic modes for four enzyme-substrate complexes, i.e., substrate titin with four bacterial thioredoxins. Inset shows the semilog plot of the density of states g(ω) versus the mode frequencies, ω, (only low frequency modes) for all four thioredoxin-titin complexes. The color code remains the same as main figure.

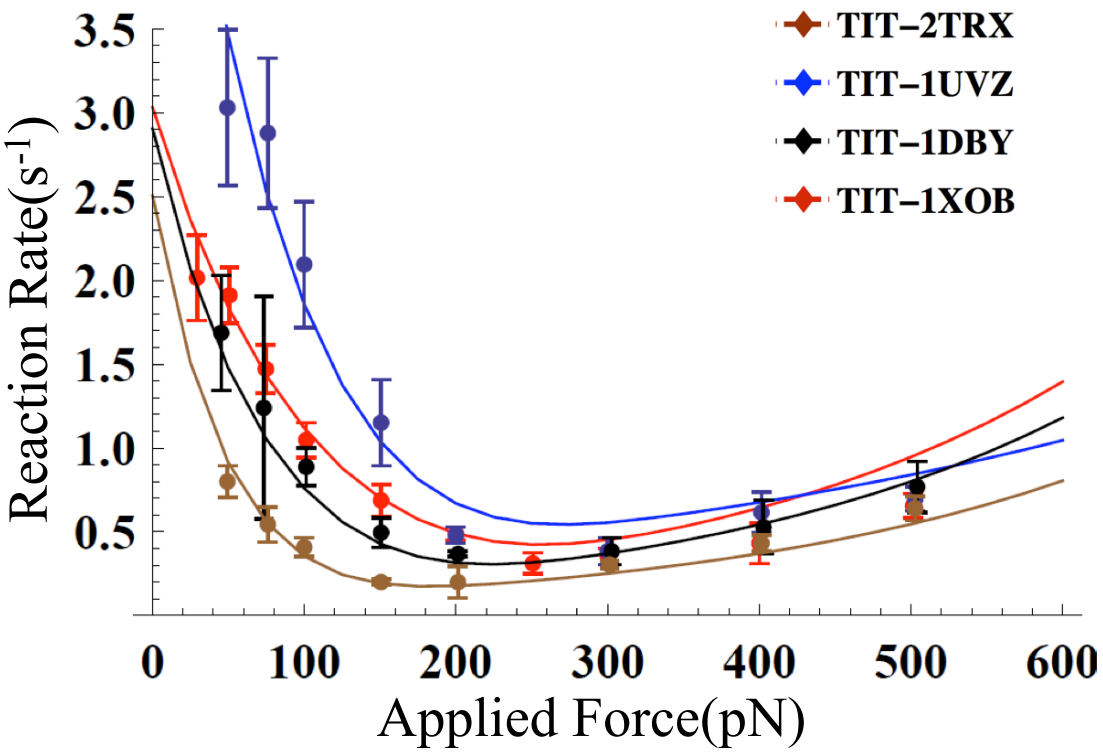

Figure A.4: Curves generated by numerically solving and integrating the Agmon-Hopfield-Smoluchowski equation for increasing values of applied force and fit to experimentally measured disulphide bond reduction rates from Perez-Jimenez et al. (fit parameters given in Table A.1). Color coding scheme same as in Fig. A.3.

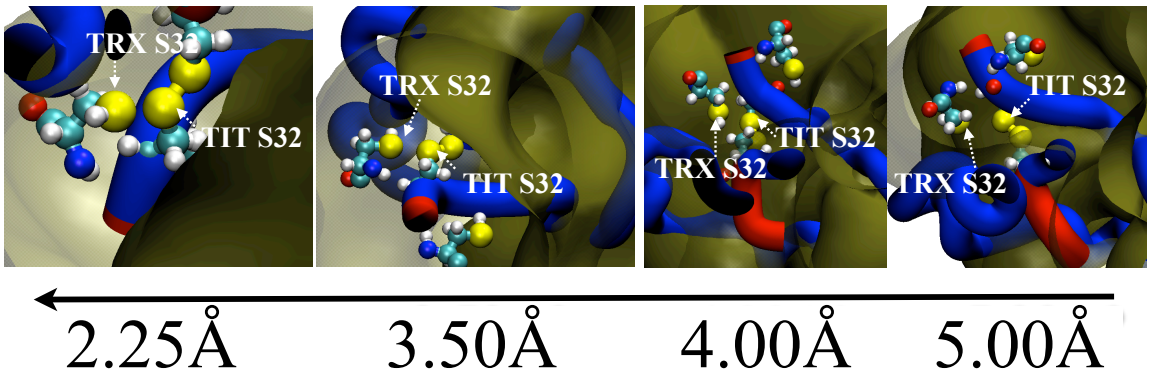

Figure A.5: Reaction coordinate shown as the distance between the sulphur atoms of two cysteine residues, one CYS (S32) in the titin substrate and another CYS (S32) at the active site of the thioredoxin enzyme (cf. labels in Fig. A.2).

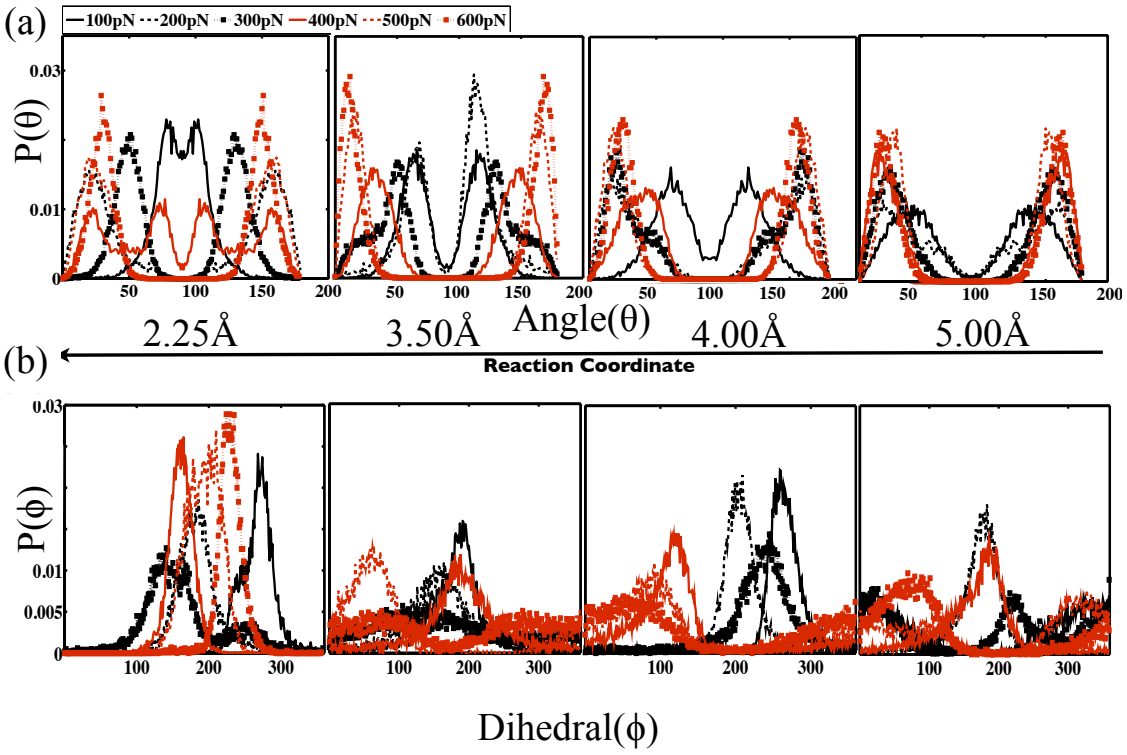

Figure A.6: In presence of the reduced form of thioredoxin (a) The substrate disulphide bond is seen to be aligned away and almost perpendicular to the force axis at lower forces as the substrate approaches the enzyme (b) Dihedral angle distribution along the reaction coordinate showing a flip to the trans- configuration with increasing proximity to the enzyme. The reaction coordinate is shown as the decreasing distance between the active site of the enzyme-S32 and the S32 atom of the titin substrate. At a distance of 2.25 $\mathring{A}$, it is the closest, corresponding to a transition state between substrate S32 and S32 at the active site of the enzyme.