Computational Approaches to Simulation and Analysis of Large Conformational Transitions in Proteins

by

Sean L. Seyler

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved November 2017 by the
Graduate Supervisory Committee:

Oliver Beckstein, Chair
Ralph Chamberlin
Dmitry Matyushov
Michael F. Thorpe
Sara Vaiana

ARIZONA STATE UNIVERSITY

December 2017

ABSTRACT

In a typical living cell, millions to billions of proteins—nanomachines that fluctuate and cycle among many conformational states—convert available free energy into mechanochemical work. A fundamental goal of biophysics is to ascertain how 3D protein structures encode specific functions, such as catalyzing chemical reactions or transporting nutrients into a cell. Protein dynamics span femtosecond timescales (i.e., covalent bond oscillations) to large conformational transition timescales in, and beyond, the millisecond regime (e.g., glucose transport across a phospholipid bilayer). Actual transition events are fast but rare, occurring orders of magnitude faster than typical metastable equilibrium waiting times. Equilibrium molecular dynamics (EqMD) can capture atomistic detail and solute-solvent interactions, but even microseconds of sampling attainable nowadays still falls orders of magnitude short of transition timescales, especially for large systems, rendering observations of such "rare events" difficult or effectively impossible.

Advanced path-sampling methods exploit reduced physical models or biasing to produce plausible transitions while balancing accuracy and efficiency, but quantifying their accuracy relative to other numerical and experimental data has been challenging. Indeed, new horizons in elucidating protein function necessitate that present methodologies be revised to more seamlessly and quantitatively integrate a spectrum of methods, both numerical and experimental. In this dissertation, experimental and computational methods are put into perspective using the enzyme adenylate kinase (AdK) as an illustrative example. We introduce Path Similarity Analysis (PSA)—an integrative computational framework developed to quantify transition path similarity. PSA not only reliably distinguished AdK transitions by the originating method, but also traced pathway differences between two methods back to charge-charge interactions (neglected by the stereochemical model, but not the all-atom force field) in several conserved salt bridges. Cryo-electron microscopy maps of the transporter Bor1p are directly incorporated into EqMD simulations using MD flexible fitting to produce viable structural models and infer a plausible transport mechanism. Conforming to the theme of integration, a short compendium of an exploratory

project—developing a hybrid atomistic-continuum method—is presented, including initial results and a novel fluctuating hydrodynamics model and corresponding numerical code.

To my Family and Future Self

ACKNOWLEDGMENTS

It's somewhat surreal to have found myself writing the acknowledgements of my dissertation, albeit not because I made it through the past five years on my own (I didn't). During my M.Eng. studies, I applied to a number of physics Ph.D. programs and, after receiving the final rejection letter, I began reconsidering whether I wanted to pursue a Ph.D. at all. On a whim, I expressed the despondency of my situation with my M.Eng. program director at Cornell University, Professor Manfred Lindau, who told me about a colleague at ASU who was (and is) an expert in MD simulations of proteins. Though at the time it was mid-April (and far past application deadlines), Professor Lindau insisted on connecting me with his colleague and the Physics Department Chair, Professor Peter Bennett. After an impromptu visit to ASU, I suddenly found myself in the physics Ph.D. program. The salience of my fortune has only continued to grow since the day that that mysterious MD expert became my new Ph.D. advisor. An extraordinarily kind, generous, thoughtful, and conscientious advisor, Professor Beckstein has given me so many opportunities and the support and guidance that have helped me come to a better understanding of myself, both as a young scientist and as a human being. Indeed, Professor Lindau and Professor Beckstein are the primary reasons I have the privilege of being able to write these acknowledgements—my gratitude can hardly be overstated.

Looking back on my doctoral studies, it is clear to me that this journey has been defined by the people who have helped me. I am grateful to Professor Dmitry Matyushov, Professor Sara Vaiana, Professor Michael Thorpe, and Professor Ralph Chamberlin for serving on my committee. I am also grateful to my office colleagues for the many wonderful conversations and discussions. I would like to specifically recognize several mentors, colleagues, and friends: Professor Chamberlin for his enthusiasm and willingness to discuss all kinds of physics I still don't quite understand; my mentee, Taylor Colburn, for every conversation that did not distinguish between physics, "the arts", consciousness, and philosophy; Jimmy Gallagher for demonstrating how to be a boss at physics; Paul Campitelli for his generosity and integrity of character that have served as beacons of spiritual reason; Avishek Kumar

for always having an open door and open mind; Ian Welland for many deep conversations about physics within and beyond biology; and my labmates who have helped make the Beckstein Lab a wonderful (virtual) place to work. I am also greatly indebted to Deanna Clark, Araceli Vizcarra, Ixchell Pape, Morgan Texiera, and Jaime Severson for going out of their way to make my life easier.

I would be remiss to forget my extended family from the ASU Club Tennis team who, over the past five years, has provided me a second home away from academia. I am especially thankful for Sherif Mansour, Amanda Moore, Naida Ortega, and Corey Rizzi-Wise who were not only amazing teammates, but my first real friends in Arizona when I moved to Tempe for graduate school. I would like to thank Eric Albertie for supporting me through so many of my existential hardships, and Kylie Southard for being by my side during these monumental life challenges over the past year. And I would like to thank my family: my brother, whose work ethic inspires me and who sacrificed many a serviceable evening to lend an ear over the phone; my mother, whose endless love and care has unequivocally helped me keep my life as a graduate student intact and in order; and my father, who has been not only a compassionate mentor during my studies, but also an inspiration as a scientist and physicist.

Lastly, I have received very generous support from numerous organizations. I am thankful for Wally Stoelzel for endowing the Wally Stoelzel Graduate Physics Fellowship; the Molecular Imaging Corporation Endowment; the Graduate College and GPSA for their support through ASU Summer Research Fellowships and an Outstanding Research Award. I am extraordinarily thankful for having had the privilege to attend the 66$^{\text{th}}$ Lindau Nobel Laureate Meeting for physics, and for my friend Anastasia Pervishko whom I met in Lindau, Germany. Finally, I am indebted to the University of Illinois at Urbana-Champaign, the National Center for Supercomputing Applications, the National Science Foundation, the Blue Waters project, and all who are part of those organizations for having given me—through the Blue Waters Graduate Fellowship—the magnificent opportunity to carry out an ambitious project of my own design.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

Chapter 1

INTRODUCTION: THE PROTEIN STRUCTURE-FUNCTION RELATIONSHIP

One of the fundamental ambitions of biophysics is to determine the function of proteins from knowledge about their structure: the so-called structure-function relationship [1–3]. Proteins comprise a very large class of biological macromolecules that in many ways behave like nanomachines, underlying the most fundamental processes of biology. By exploiting sources of thermodynamic free energy, proteins such as enzymes, molecular motors, and membrane transporters perform mechanicochemical work by cycling between multiple structural states called conformations. Determining the thermodynamic and mechanistic nature of such conformational transitions is therefore essential to understanding protein function. The tiny spatiotemporal scales over which these processes occur presents a formidable challenge, both experimentally and computationally. On the other hand, there is an ever-growing number of structures of increasingly high resolution from X-ray crystallography and NMR, and the advent of methods such as cryo-electron microscopy (cryo-EM) and serial femtosecond crystallography are enabling structural imaging under conditions that more closely represent the native environment and even the possibility of time-resolved structure determination. Interdisciplinary research by a growing computational community is driving innovation in numerical sampling and computational analysis methods that, coupled with the unrelenting advance of hardware capabilities and experimental techniques, is leading to exciting discoveries and new insights into protein function.

This unprecedented growth in experimental and computational power has led to the generation of enormous quantities of data, making the question of how to cultivate *meaning* from that data of the utmost importance. A rethinking of contemporary methodologies is necessary, implying the importance of not only incorporating experimental data into numerical simulation but deeper *quantitative* integration—using computational frameworks—of

many scales of description from disparate physical models. This thesis represents an exploration, both theoretical and applied, of various aspects of numerical sampling, and introduces one such analytical framework that can leverage data from a diverse range numerical simulation methods used to study conformational transitions in proteins.

This chapter first introduces the structure-function relationship to conformational transitions, then discusses experimental knowledge about the enzyme adenylate kinase (AdK) as a concrete illustration, and lastly sets the stage for numerical simulation and path sampling; it is based in part on the published review, **Sean L. Seyler** and Oliver Beckstein (2014). *Sampling large conformational transitions: adenylate kinase as a testing ground.* Molecular Simulation, 40: 855–877. [4] My contribution to this work was a majority of the research, synthesis, and writing. Chapter 2 then briefly surveys some of the advanced computational methods for sampling such transitions (including methods used in this thesis) followed by a summary of computational studies on AdK and the current state of knowledge. In Chapter 3, several general computational tools for analyzing simulations are described, then Path Similarity Analysis (PSA) [5] is introduced and illuminated with an application to a toy model. Chapter 4 demonstrates the viability of PSA in analyzing realistic transition paths generated by a variety of path-sampling methods, as well as relating back to discussions on AdK (Chapter 1 and Chapter 2) and path-sampling methods (Chapter 2). First, AdK is used as a testbed to apply a number of path-sampling methods (introduced in Chapter 2) to generate open to closed transitions. These transitions are compared using PSA along with two alternative techniques introduced in Chapter 3. Then, two sampling methods are analyzed in greater detail using ensembles of transitions generated from the AdK system and the diphtheria toxin (DT) protein. We demonstrate the concept of Hausdorff pairs analysis, which provides a direct way to connect PSA to the molecular detail. At the end of the chapter, the path-sampling methods comparison has been extended to include long-time, equilibrium MD (EqMD) transitions generated using the Anton supercomputer [6], followed by a brief discussion. Chapter 5 takes a look at a membrane transporter system, namely the borate transporter Bor1p; in particular, it is shown how molecular dynamics flexible fitting (MDFF)

can be combined with experimental data from cryo-EM to model new protein structures [7]. Chapter 6 is based on results from the author's project for the 2016 Blue Waters Graduate Fellowship, which motivates the role of hydrodynamic models and hybrid multiphysics methods in studying protein conformational transitions; preliminary findings—including a novel fluctuating hydrodynamics numerical method written in Fortran 90—are presented and a road map is given to help guide further development of the project in the future.

1.1 Proteins and conformational transitions

Conformational transitions are an inseparable aspect of protein function, as they are essential to driving reactions, membrane transport, and other processes at rates necessary to sustain life—far higher than would be possible in pure equilibrium. Concentrations of essential nutrients or neurotransmitters are maintained, for example, by the active transport of substrates into and out of cells by membrane transport proteins, which alternate between states that expose binding sites to either the cytoplasm or periplasm of a cell [8–10]. The catalytic cycling of enzymes, as another example, is responsible for maintaining appropriate biological concentrations of various chemical compounds, and it is generally accepted that conformational rearrangement acts to passively stabilize chemical transition states at the chemical step*, thereby indirectly lowering free energy barriers to reaction [15, 18]. However, there is also strong evidence that large-scale conformational changes and fluctuations enhance substrate fluxes into and out of binding sites by efficiently mediating reactant binding and product release [19–21].

To elucidate the physical mechanisms of conformational transitions—transitions between structural states—that underlie function, it is necessary to investigate protein structure. In principle, the 3D (static) structure of a protein should encode information about its biological function: the dynamics can be inferred from the structure, and the function can

*Although out of the scope of this dissertation, it bears mentioning that whether conformational motions also play a more active, direct role over the course of the chemical step is a topic of intensive debate and possible semantic confusion [11–17].

be inferred from the dynamics. Much like one can learn about the motion and function of a door by studying how its hinges allow it to swing between an open and closed position, one can, in principle, learn about the function of proteins by studying how their structures admit biologically relevant motions between two (or more) states.[†] X-ray crystallography, nuclear magnetic resonance, and, more recently, cryogenic-electron microscopy (cryo-EM) techniques are continuously contributing to a vast, growing database of protein structures (the Protein Data Bank [26]), many of which can be identified with functionally relevant states between which transitions take place. Such experimental methods have, however, only been able to capture static snapshots of proteins, limiting what can be deduced about biologically relevant motions [3].

Molecular dynamics (MD) simulation for many years has been the method of choice for generating atomistic-resolution movies from static experimental structures. The dream of watching a "movie" of a protein in motion is becoming a reality with the emergence of serial femtosecond crystallography—femtosecond-pulses from an X-ray free-electron laser (XFEL) pulses enable time-resolved structure determination [27, 28]. Although MD is playing an integral role in the development and validation of serial femtosecond crystallographic techniques, and the integration of both methods promises new horizons in the study of macromolecular dynamics and the structure-function connection, there are, at present, two underlying obstacles. The first is the equilibrium sampling problem (discussed in Chapter 2)—conformational transitions are *rare events* in the sense that typical transition events (on the order of nanoseconds) are disproportionately brief compared to the time spent between such events in metastable equilibrium states (milliseconds and beyond). Given that all-atom MD typically falls short by many orders of magnitude, approaches such as the well-known elastic network model (ENM) [29] or targeted MD (TMD) [30]

[†]To wit, there are experimental and computational studies where mechanical stresses or forcing are, literally, applied to a protein so as to measure its response and deduce possible connections between mechanical properties and function [22–24]. It is even possible to apply forces during MD simulations using haptic devices [25]; see also the Visualization and Simulation of Biomolecules Village on the Beckstein Lab website, an outreach project sponsored by BioXFEL.

have attempted to capture the essential physics using simple models or intelligent biasing, thereby trading accuracy for efficiency (cf. Chapter 2). The second problem is, however, intrinsic to both MD and time-resolved experimental techniques, both of which require a quantitative means to process, analyze, and interpret high-dimensional trajectory data—containing the position information of many thousands to millions of atoms—to be able to extract meaning from trajectories and, ultimately, construct an intelligible mechanistic description of the process of interest. Reducing the dimensionality of the data—so it is amenable to, perhaps, a two- or three-dimensional visual representation—without losing information in the process is *hard* (see Chapter 3, which briefly discusses dimensionality reduction).

To make full use of increasingly capable experimental and simulation methods and the increasingly massive quantities of (heterogeneous) data that will be generated, it will be essential to develop *quantitative* methodologies and analytical frameworks that can integrate these data, extract relevant structural-dynamical information with minimal bias, and facilitate interpretation and understanding. One such step in this direction is the Path Similarity Analysis (PSA) [5], which is explored in detail in Chapter 3 and Chapter 4.

1.2 Adenylate kinase as a model for large-scale conformational change

The intricacies of the current state of knowledge surrounding protein conformational transitions and the structure-function connection is perhaps more easily understood in the context of a suitable real-world example. To this end, we take a birdseye view of experimental and computational insights from the study of the enzyme adenylate kinase (AdK). The discussion concentrates on the apo enzyme (apo-AdK), as its dynamics lend insight into the structural changes involved in both ligand-binding and the full catalytic cycle. Indeed, the complexity of protein-ligand interactions makes accurate simulation difficult, so many computational studies have focused on the apo enzyme. We also primarily restrict our attention to the AdK enzyme of the mesophilic bacterium *Escherichia coli* (AdK_{eco}) as it has been also the focus of the majority of computational studies. The conceptual simplicity of

the apo-AdK transition is amenable to visualization and can serve as an intuitive guide to other protein systems with more complex conformational transitions. In some sense, apo-AdK has become a benchmark system for the development and application of an impressive array of experimental and computational methods, which affords a rather detailed picture of the uses and limitations of the various methods as well as how such methods may complement one another. Despite these efforts, a full mechanistic description of the relevant conformational changes is nevertheless still lacking, making AdK a worthwhile system to explore more deeply.

1.2.1 Background

Adenylate kinase (AdK, EC 2.7.4.3) is a relatively small enzyme that regulates the relative concentrations of ATP, ADP, and AMP and thereby assists energy homeostasis [31, 32]. AdK has three domains: a relatively stable *CORE* domain containing an α/β-sandwich fold typical of P-loop NTPases [33, 34]; one mobile domain called the *LID*, which binds the magnesium-bound ligand $Mg^{2+}{\cdot}ATP$ along with the P-loop in the CORE; and another mobile domain called *NMP*, which binds AMP. The hinge motions of the LID and NMP domains allow closing over the reactants, precisely positioning each substrate so as to catalyze phosporyl transfer while inhibiting the phosphate ion from diffusing away from the AMP molecule,

$$Mg^{2+} \cdot ATP + AMP \rightleftharpoons Mg^{2+} \cdot ADP + ADP. \tag{1.1}$$

The catalytic cycle is completed once the LID and NMP domains open up, allowing ADP to diffuse away and new reactants to bind. Fig. 1.1 depicts the three domains and the large conformational motions of the LID and NMP domains. NMR data suggest that the ADP-bound opening step of the holo enzyme is comparable to the overall catalytic rate, on the order of $300\,s^{-1}$, and is therefore rate-limiting [19].

That both the NMP and LID domains can move about respective labile regions relatively independently is supported by a preponderance of crystal structures exhibiting partially

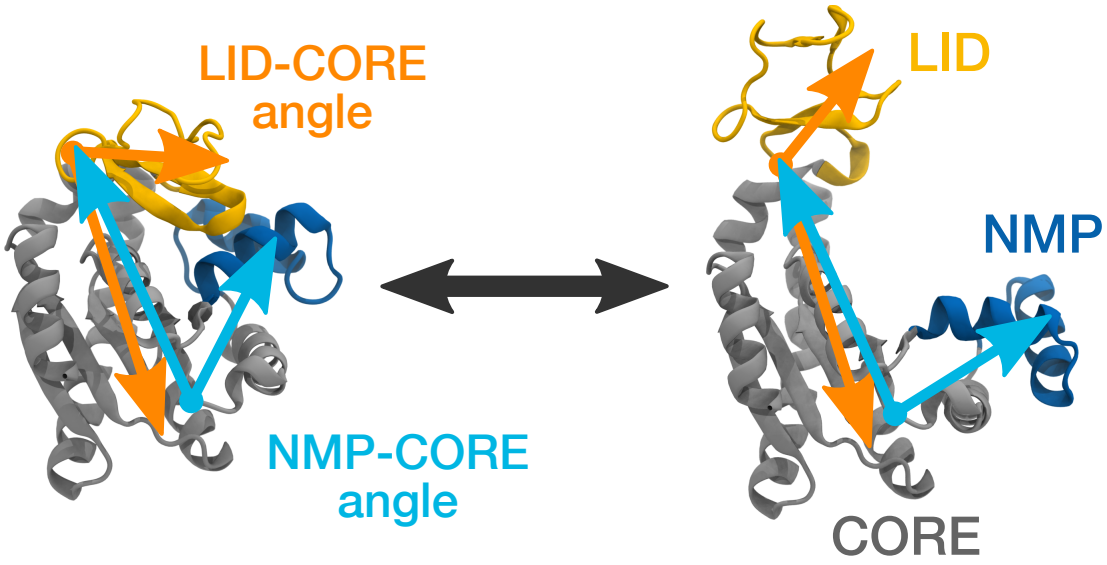

Figure 1.1: The closed ↔ open conformational transition and mobile domain angles of *E. coli* adenylate kinase (AdK) shown for the apo enzyme. AdK has three structural domains: a stable *CORE* domain containing an α/β-sandwich fold [33, 34]; a mobile *NMP* (Nucleotide MonoPhosphate-binding) domain that binds AMP; and a second mobile *LID* domain that binds Mg^{2+}·ATP. The LID–CORE angle, θ_{LID}, is formed by the geometric centers of backbone and C_β atoms in residues 179–185 (CORE), 115–125 (CORE–hinge–LID), and 125–153 (LID); the NMP–CORE angle, θ_{NMP}, is analogously formed by residues 115–125 (CORE–LID), 90–100 (CORE), and 35–55 (NMP).

open and closed conformations [35, 36]. Over 50 structures of AdK_{eco} and homologous proteins have been crystallized, predominantly in substrate-bound (holo), closed-like states and four in ligand-free (apo), open-like states (PDB IDs 4ake, 2rh5, 3umf, 3gmt [26]). An abundance of intermediate structures also suggestively spans a putative transition path between the open and closed states [37–40]. By comparison of both apo and holo X-ray structures, it has been conventionally thought that ligand binding induces a large-scale conformational change [37]. Efforts to go beyond the limitations of static structures, however, have led to a very different view, whereby the apo enzyme samples catalytically active closed-like conformations and substrate binding effects a dynamical reweighting of the relative distributions between open and closed states [41–43]. These limiting cases are referred to as, respectively, the *induced fit* model [44] and the *population shift* (or *conformational selection*) model [45].

Though a continuum exists between induced fit and conformational selection, there is an emerging consensus that intrinsic conformational flexibility is critical to enzymatic

function, necessitating a scrupulous examination of protein fluctuations and dynamics [3, 13, 14, 46–48]. We do not discuss the induced fit or conformational selection models in the remainder of this thesis but, rather, focus on the nature of large-scale conformational transitions and methods that can be used to investigate their role in protein function. Relatively recently, new computational and experimental techniques have enabled greater insight into the relationship between holo and apo dynamics, paving the way for a deeper understanding of the transition pathways, intermediate states, energetics, and kinetics that underlie the function of AdK. In the following section, we recapitulate key experimental results to summarize what is known about AdK and to build a basis for understanding the growing role of computational methodologies in studying large-scale conformational transitions.

1.2.2 Experimental insight into timescales and motions

Experimental techniques such as NMR spectroscopy, Förster resonance energy transfer (FRET), and small-angle X-ray scattering (SAXS) have revealed that the mobile domains of the apo enzyme can sample a range of conformations in equilibrium spanning open and closed states [41–43]. Numerous other experiments have further indicated that the conformational dynamics of ligand-binding strongly correlate with apo enzyme dynamics [49–52]. In the case of ligand-bound AdK_{eco}, NMR experiments carried out by Wolf-Watz et al. [19] estimated an opening rate of $k_{open} \approx 286\,s^{-1}$ and a much faster closing rate of $k_{close} \approx 1374\,s^{-1}$ in the presence of saturating concentrations of AMP, Mg^{2+}, and AMPPNP (a nonconvertible ATP analog). As the opening rate was comparable to the overall steady-state turnover rate ($k_{cat} \approx 263\,s^{-1}$), the ADP-bound opening step is effectively rate-limiting and conformational equilibrium is shifted toward a predominantly closed state [19].

Kinetic rate estimates from NMR and FRET experiments have observed that ensembles of conformations of apo AdK, with varying degrees of both NMP and LID closure, may undergo transitions between open and closed states on the millisecond timescale [42, 43]. The LID domain in particular appears to move relatively unimpeded between open and

closed conformations in the apo enzyme and has been a subject of investigation in several studies. Hanson et al. [42], for instance, used FRET labels on the LID and CORE domains of AdK_{eco} to specifically probe LID-CORE distances and estimate LID opening and closing rates, obtaining $k_{open} \approx 120\,s^{-1}$ and $k_{close} \approx 220\,s^{-1}$. SAXS has also been used to gauge the global flexibility of AdK_{eco} in solution and it was found that, under apo conditions, large-scale rigid-body LID movements contribute considerably to the scattering intensity [53]. K. A. Henzler-Wildman et al. [43] measured somewhat faster opening and closing rates of $k_{open} \approx 6500\,s^{-1}$ and $k_{close} \approx 2000\,s^{-1}$ for *Aquifex* AdK (from the hyperthermophile *Aquifex aeolicus*) at 293 K, though this was achieved using the separation of FRET labels located at I52 on the NMP domain and K145 on the LID as a 1D collective variable. Interestingly, Shapiro and Meirovitch [54] obtained significantly faster estimates for opening and closing rates of the apo AdK_{eco} enzyme using NMR spectroscopy, indicating that both the LID and NMP domains may move rapidly in the absence of ligands, possibly at rates of $1.9 \times 10^7\,s^{-1}$ at 293 K.

Although NMR techniques provide direct access to timescales with atomic resolution, a considerable advantage of FRET is that it can be used to measure time-resolved distance changes at the single molecule level, allowing a more intuitive visualization of the underlying opening and closing motions [3]. Distances between fluorophore labels also approximately correspond to C_α distances, which can be easily measured in NMR and directly extracted from crystal structures and numerical simulations. Numerous studies have used 1D FRET collective variables to directly measure overall opening and closing motions of apo AdK [3, 41–43, 54, 55] and such data comprise the primary evidence that the apo enzyme populates both open and closed states in conformational equilibrium. It was subsequently shown by Matsunaga et al. [56] that the 1D FRET pair I52–K145 used by K. A. Henzler-Wildman et al. [43] on *Aquifex* AdK reasonably tracks the overall conformational change, suggesting that the very fast rates obtained by Shapiro and Meirovitch [54] may underestimate apo enzyme transition times. Other the other hand, assigning thermodynamic states (such as "open" and "closed") to, say, multimodal distributions of

1D FRET distances, is a nontrivial problem, as these and other order parameters do not necessarily correspond to good reaction coordinates [40, 57, 58].

1.2.3 Experimental evidence of multiple pathways

Both NMR and FRET studies have indicated the presence of metastable intermediate states that are likely to be accessible through two different pathways [3, 41]. Fig. 1.2 schematically depicts two plausible closed $\leftrightarrow$ open pathways connecting intermediate states represented by half-open-half-closed AdK crystal structures. Specifically, starting from the closed state on the left, the top pathway depicts an NMP-opening trajectory that proceeds through an intermediate NMP-open-LID-closed conformation; the bottom sequence represents a LID-opening pathway whose intermediate state resembles a NMP-closed-LID-open conformation. Variations on the these two pathways are also possible, whereby several partial opening or closing steps may take place over the course of a closed $\rightarrow$ open or open $\rightarrow$ closed transition. It is notable that Adén and Wolf-Watz [41], by analyzing chemical shift perturbations from solution NMR, found evidence for unidirectional allosteric communication—from the ATP binding site in the LID to the distant NMP domain—observing that such energetic coupling could be rationalized if the ligand-free enzyme samples open and closed conformations with a bias toward the open state. On the other hand, the LID-CORE distances examined with FRET by Hanson et al. [42] led to the unanticipated result that the LID in the apo enzyme favors a closed state in equilibrium. Despite the apparent conceptual simplicity of rigid-body LID and NMP domain motions in the closed $\leftrightarrow$ open transition of apo AdK, discrepancies in kinetic rate estimates remain and there is as of yet no clear consensus regarding the relevant transition states and pathways.

Given the possibility of hinge-like behavior for both domains, it is likely that at least two observables will be necessary to fully resolve independent NMP and LID motions. Two FRET pairs between residues A55–V169 (NMP–CORE, [55]) and A127–A194 (LID–CORE, [42]) have been shown to measure distances that track the major conformational change

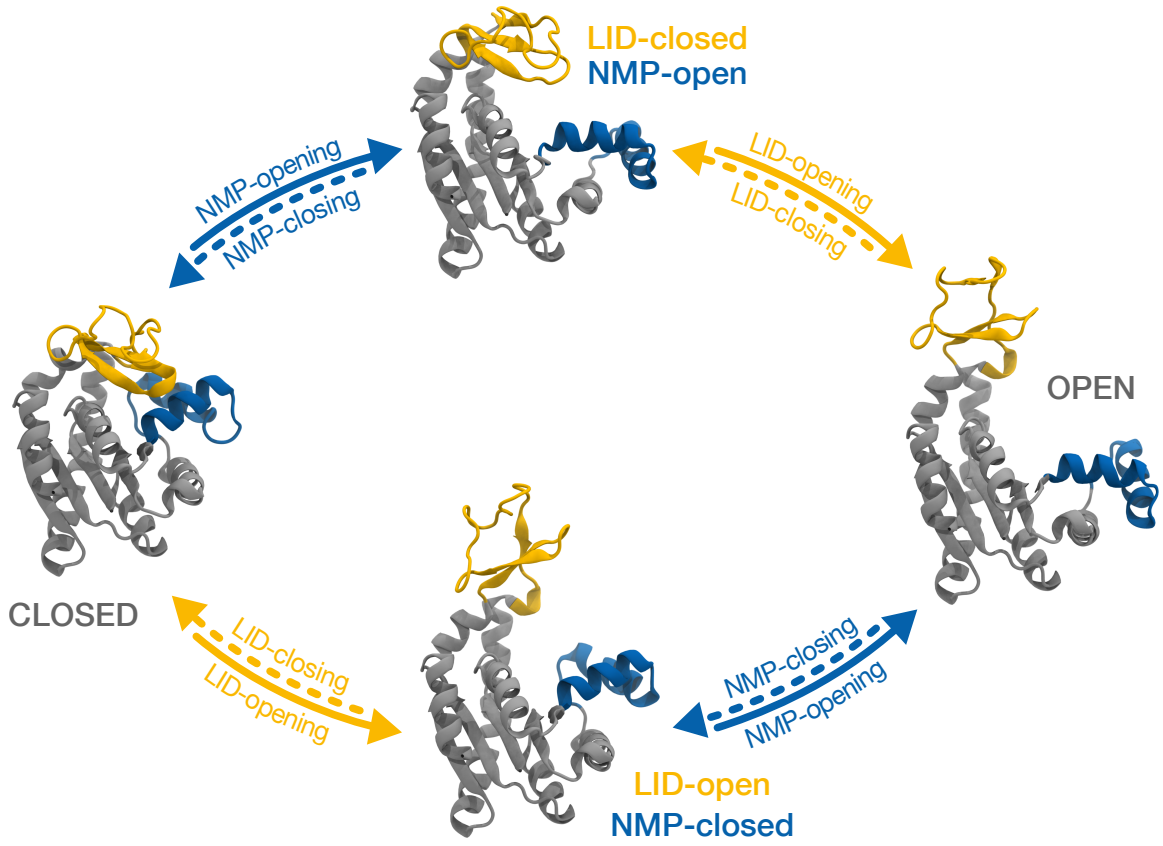

Figure 1.2: Two plausible pathways for the conformational change between the closed (left, represented by PDB ID 1ake) and open (right, represented by PDB ID 4ake) conformations, during which the LID domain (orange) and NMP domain (blue) move sequentially. The top path in the closed → open direction (solid lines) is the NMP-opening pathway, placing AdK in one possible intermediate state (top, represented by PDB ID 1ak2), followed by opening of the LID domain to the 4ake open state. The bottom pathway in the closed → open transition starts with a LID-opening motion, occupies a LID-open/NMP-closed conformation (bottom, represented by PDB ID 2ak3), and completes with the opening of the NMP domain. The dashed lines and arrows indicated the closing order of the domains for the open → closed (reverse) transition.

in (*E. coli*) apo-AdK in an almost orthogonal manner.[‡] Indeed, a 2D space spanned by such observables can shed light on the conformational dynamics of AdK at the level of individual domain motions while at the same time providing a direct connection between experimental and computational techniques. Questions still remain, however, about the order of opening or closing motions of the mobile domains and the degree to which those motions are correlated, particularly in the apo enzyme. To understand the mechanistic

[‡]Lou and Cukier [59], based on principal component analysis (PCA) of MD simulations, suggested that the residue pair L40–L141 may be a better probe of LID motions than A55–V169, at least on the nanosecond timescales that were sampled.

role of the conformational motions involved and to reconcile apparent discrepancies in experimentally measured timescales, further investigation will undoubtedly be necessary. Naturally, the ready availability of AdK_{eco} starting structures and experimental data have paved the way for powerful computational methodologies to contribute new insights.

1.3 Connecting experiment and simulation

In talking about enzymes in general, the presence of multiple metastable intermediates connected by multiple pathways appears to be a ubiquitous mechanistic aspect [14, 60]. Depending on the enzyme and reaction, billion- to trillion-fold rate enhancements (possibly up to 10^{19}!) over uncatalyzed reactions can be achieved [61], predominantly by a lowering of the (quasi)thermodynamic free energy barrier [62, 63]. As the rate enhnacement of an enzyme increases, so does its binding affinity for antagonists in the *transition* state (TS), but not the unactivated state [64, 65]. That the binding affinity of an enzyme for the inhibiting substrate correlates with the substrate's resemblance to the TS is indicative of the relevance—to the catalytic process—of those structural rearrangements involved in passing through the TS [17, 63, 66]. From this transition-state theory (TST) [67] of enzymes, the motivation for studying protein structure as it pertains to such enormous enhancements of catalytic rates, and thus enzymatic function, is transparent. Going one step further, it is clear that a physical model that can accurately estimate kinetic rates would reveal much about the underlying physico-structural features of enzymatic catalytic competence and, perhaps, other protein conformational transitions. Being a direct test of our mechanistic intuitions, kinetic rate prediction is thus arguably a principal connection between experiment and simulation.

The equilibrium sampling problem (cf. Chapter 2) makes brute-force rate calculations—a task necessitating an explicit solvent environment and atomistic detail to account for all relevant physical interactions—effectively impossible. The thermodynamic free energy difference between, say, the open and closed states of apo-AdK moreover does not provide kinetic (rate) information and brute-force calculations are impracticable for the same reasons.

The *ratio* of the forward and reverse reaction rates between two states is, however, related to the population ratio between those states through the equilibrium constant, K. In the case of AdK, the primary evidence for there being an accessible open and closed states in equilibrium comes from 1D FRET data. The opening and closing rates of apo-AdK in equilibrium

$$\text{closed} \underset{k_{\text{close}}}{\overset{k_{\text{open}}}{\rightleftharpoons}} \text{open} \tag{1.2}$$

are related to the open and closed populations ($[\text{AdK}_{\text{open}}]$ and $[\text{AdK}_{\text{close}}]$, respectively) by

$$K = \frac{k_{\text{open}}}{k_{\text{close}}} = \frac{[\text{AdK}_{\text{close}}]}{[\text{AdK}_{\text{open}}]}, \tag{1.3}$$

which, in turn, relates to the free energy difference between open and closed conformations,

$$\Delta G = G_{\text{open}} - G_{\text{close}} = -k_{\text{B}}T \ln K = -k_{\text{B}}T \ln \frac{k_{\text{open}}}{k_{\text{close}}}. \tag{1.4}$$

In situations where, for instance, the opening rate constant is known—perhaps the opening rate is easier to measure than the closing rate—the closing rate can be calculated from Eq. 1.4 if the free energy difference between open and closed states is also known.

In principle, both the kinetics (rates) and thermodynamic properties (free energies), as well as pathways, can be calculated in a single, sufficiently long all-atom, explicit-solvent EqMD simulation. As mentioned previously, the equilibrium sampling problem renders such an approach effectively impracticable (see expanded discussion in Chapter 3). In many situations, free energy differences between two states can be computed without knowledge of the pathways that connect them, but kinetic rates are comparatively expensive to compute since they require that the full space of reactive pathways and intermediate states be explored [62, 68].

Advanced computational methods have been devised to directly calculate rates without resorting to free energy calculations (some are mentioned in Chapter 3), although a more common strategy is to divide the problem into separate calculations for the free energies and kinetic rates using complementary methods. Although free energy and rate calculations are not presented in this dissertation, it is worth noting that, for instance, one can compute the

free energy, or *potential of mean force* (PMF[§]), along a putative pathway to estimate the barrier height at the transition state (TS) and then apply a kinetic rate theory (e.g., using transition state theory [TST] or Kramers' theory [75]). Accurate rate estimates can be obtained from such an approach when it is feasible to: (1) identify relevant kinetic pathways and (2) perform sufficient sampling along the pathways to compute a converged PMF. Kinetic rate theories effectively transform the problem of rate estimation into one of locating the reaction pathways with good collective variables (CVs). The difficult task of selecting a satisfactory set of CVs generally falls under the purview of dimensionality reduction (cf. Chapter 3), though it is possible, at least in principle, to choose arbitrary combinations of CVs when using computational methods. An ideal, comprehensive computational approach might be imagined to employ dimensionality reduction and enhanced sampling to identify a kinetic pathway, construct a PMF by sampling free energies along the pathway, then generate a rate estimate by applying a kinetic rate theory to the PMF.

The appeal of a PMF lies in the simplicity of the description, which depicts a free energy profile and barrier between two states as a function of a single coordinate. Although it may be tempting to infer a mechanism from a PMF, it is imperative that one keep in mind two failure points. Specifically, PMFs are: (1) subject to the usual concerns about obtaining sufficient sampling for convergence; (2) constructed from a selected set of collective variables that may not yield an adequate description of the underlying transition process. In a situation where (1) and (2) can be addressed with both thorough sampling and an unbiased selection of CVs describing a functionally relevant pathway, a PMF may be able to yield a meaningful estimate of the transition state barrier and help connect to the underlying kinetics. However, such a task is complicated by the fact that large-scale conformational transitions can take place over multiple pathways proceeding through functionally relevant intermediate states. That reactive pathways may not be described by

[§]PMFs can be computed, for example, using umbrella sampling [69] along with an unbiasing technique like the weighted histogram analysis method (WHAM) [70, 71] or the multistate Bennett acceptance ratio (MBAR) [72] estimator, or using specialized enhanced sampling algorithms like metadynamics [73] or the *free energies from adaptive reaction coordinate forces* (FEARCF) method [74].

obvious coordinates choices and that EqMD is generally unsuitable for this kind of search is the primary motivation of numerical path sampling techniques [57].

Chapter 2

NUMERICAL SIMULATIONS AND ADENYLATE KINASE

This chapter is based in part on the published review, **Sean L. Seyler** and Oliver Beckstein (2014). *Sampling large conformational transitions: adenylate kinase as a testing ground.* Molecular Simulation, 40: 855–877. [4] My contribution to this work was a majority of the research, synthesis, and writing. This chapter reflects a contemporary perspective that incorporates the literature published since the review, as well as other articles not previously mentioned, including many recent developments in path-sampling methodology [76–88] and a number of new studies involving AdK [68, 84–86, 89–97].

A concise overview of numerical sampling approaches is given to set the stage for discussing computational studies on AdK. Many of these studies employ fast path-sampling methods that overcome the timescale limitations inherent to equilibrium molecular dynamics (MD) simulations. To this end, the first part of this chapter begins with an overview of MD simulation since this method has been used to produce results in Chapter 4 and Chapter 5 and is also discussed in Chapter 6. Next, there is a discussion of challenges to sampling conformational transitions with equilibrium MD due to the so-called equilibrium sampling problem, motivating the reason that fast path-sampling algorithms have been used to produce transition paths. To provide a concrete example of one such fast path-sampling algorithm, the Dynamic IMportance Sampling MD (DIMS-MD) algorithm is briefly described; DIMS-MD was also extensively used to produce the results presented in Chapter 4, so this example also serves as a backdrop for later discussion. Then, a general overview of various types of path-sampling approaches then puts into context many of the computational studies on AdK, which are discussed in final part of this chapter.

2.1 Molecular dynamics

Equilibrium molecular dynamics (EqMD) simulation [98–101], much like an atomistic resolution computational microscope [102, 103], is a robust and popular method that has been valuable for investigating the link between protein structure and function [104–106]. In the *molecular mechanics* (MM) approach, mutual forces between particles are defined by a classical force field; to generate the dynamics, experimental structures (often from X-ray crystallography) are used as putative initial conditions and the equations of motion of classical mechanics (i.e., Newton's second law) are numerically discretized and integrated forward in time [100, 101]. Symplectic numerical integrators such as Verlet methods* are widely used for MD of Hamiltonian systems [101, 108] because they possess a number of desirable properties, including time-reversibility and global energy conservation†

Classical force fields used in biomolecular and protein simulations—CHARMM [112], AMBER [113], GROMOS [114], OPLS [115], and others—define a pairwise-additive potential energy in terms of bonded and non-bonded energy terms with a general (or similar) form [116–118]

$$U(\mathbf{X}) = \sum_{\text{bonds}} k_b(l - l_0)^2 + \sum_{\text{angles}} k_\theta(\theta - \theta_0)^2 \tag{2.1a}$$

$$+ \sum_{\text{dihedrals}} k_\chi \left[1 + \cos\left(n\chi - \delta\right)\right] + \sum_{\text{impropers}} k_{imp}(\phi - \phi_0)^2 \tag{2.1b}$$

$$+ \sum_{\text{n.b. pairs}} \left[\left(\left(\frac{A_{ij}}{r_{ij}} \right)^{12} - \left(\frac{B_{ij}}{r_{ij}} \right)^{6} \right) + \frac{q_i q_j}{\epsilon r_{ij}} \right]. \tag{2.1c}$$

In Eq. 2.1a, the first term is a summation is over atom pairs with bond length l and the second term is taken over three consecutive atoms (1–3 interactions) that form an angle θ; in Eq. 2.1b, the third (sinusoidal) and fourth (harmonic) terms involve four consecutive

*Variational, or Higher-Order Harmonic (HOH), integrators introduced by Predescu et al. [107] have many of the desirable properties as Verlet methods and may be alternatives worth considering.

†More precisely, symplectic integrators conserve a "shadow Hamiltonian"—an asymptotic expansion in powers of the time step—that closely approximates the real Hamiltonian, with energy errors that depend on the timestep size and integration order [109–111].

atoms (1–4 interactions), which are, respectively, dihedral rotations with "torsion" angle χ and improper dihedrals with the out-of-plane angle ψ. The final sum (Eq. 2.1c) is taken over non-bonded pair interactions, for atoms i and j with separation r_{ij}, that fall within a predefined cutoff distance[‡] The first non-bonded term is the 6-12 Lennard-Jones (LJ) potential, which approximates van der Waals (vdW) dispersion interactions and short-range repulsion between atoms i and j with separation r_{ij}. The second non-bonded term is recognizable as the Coulomb potential; due to its long-range nature, truncation errors can be avoided by computing electrostatic forces when atoms i and j (with respective charges q_i and q_j) fall within the cutoff, while long-range interactions are computed using special methods[§]

The parameters k_b, k_θ, k_χ, and k_ψ (bonded parameters) and A_{ij}, B_{ij}, and ϵ (non-bonded parameters) in Eq. 2.1 are often calibrated to reproduce a set of experimental observables (e.g., thermodynamic equations of state or transport coefficients), but molecular mechanics parameters may also be derived from quantum-mechanical electronic structure calculations, especially when reliable empirical data are scarce [128]. By comparison, *ab initio* MD (AIMD) [129, 130] determines forces from first principles by calculating the electronic structure on the fly, avoiding excessive parameter fitting that can lead to model bias. AIMD can also capture the chemistry and polarization of materials. Despite such appeal, the typically large number of electronic degrees of freedom places severe limits on the number of atoms that can be simulated ($\lesssim 10^4$) and the physical timescales that can be accessed ($\lesssim 100\,\text{ps}$) [131, 132]. For all-atom equilibrium simulations of biological macromolecules, some of which which may contain upwards of 10^6 atoms including solvent atoms, molecular mechanics (i.e., EqMD) is still the most viable approach.

[‡]It is typical to choose a single cutoff—on the order of 8 Å to 12 Å for biomolecular simulations—for both van der Waals and electrostatic terms [119].

[§]Mesh-based Ewald methods run in quasilinear time and are common for systems with periodic boundary conditions [120–123], but techniques such as the multilevel summation method (MSM) [124] and the fast multipole method (FMM) [125] that have comparable accuracy and handle general boundary conditions are becoming potentially competitive alternatives to [126, 127].

Force fields remain a significant, though manageable, source of systematic error in EqMD simulations [93, 133–138]. Indeed, the task of quantifying *all* sources of error, both statistical and systematic, is imperative if systematic errors are to be mitigated and the results of computer experiments are to be taken seriously [138], especially with unprecedented access to long timescales pushing the limits of current force fields [103]. Yet empirical biomolecular force fields have often been within the accuracy of the experimental method [139] and there has been considerable progress in recent years [103, 140–143]. Numerous millisecond-timescale EqMD simulations indicate that biomolecular force fields are sufficiently accurate to reproduce experimental quantities relevant to, for example, protein folding [134, 144, 145] and ongoing efforts to improve water models [137, 146] and polarizable force fields [147–149] suggest that the molecular mechanics approach is far from obsolete. Development efforts that extend the reach of conventional EqMD are underway—e.g., constant-pH MD [150–152]. It is also becoming increasingly realistic to implement and optimize more capable biomolecular force fields, such as anharmonic force fields [153–155] that go well beyond the form of Eq. 2.1, using novel optimization techniques [155, 156] and machine learning approaches [157, 158].

2.2 Why path sampling?

A primary advantage of applying EqMD to sampling large-scale conformational change is that the dynamics from the (unperturbed) Hamiltonian is expected to remain faithful to the microscopic dynamics while reproducing the expected thermodynamic observables and their fluctuations. As such, it should be possible in principle to produce a true ensemble of atomistically detailed equilibrium trajectories that sample a putative transition pathway. On the other hand, most conformers extracted from an EqMD simulation necessarily correspond to metastable states rather than transition states of interest. To be able to observe large-scale transition events, which tend to take place among the low-frequency modes of a protein [159], EqMD simulations must be able to sample timescales on the order of at least tens to hundreds of microseconds if not well beyond milliseconds. Currently, all-atom

EqMD simulations with explicit solvent can access the microsecond regime using modern commodity GPUs and GPU-enabled MD codes [160–164] and continual improvements in computational power will eventually allow access to even longer timescales. However, all-atom EqMD requires timesteps on the order of femtoseconds for stable integration, necessitating at least billions of timesteps to access the microsecond regime. EqMD ultimately falls short of relevant timescales—often by many orders of magnitude—where one would observe most large-scale transition events, let alone achieve statistically valid sampling [165, 166]. Even when EqMD can generate transitions in practice, such disproportionate sampling of metastable states is a severe hindrance computational efficiency, limiting its competitiveness as a method for generating conformation transitions [167].

It nevertheless seems that a detailed, mechanistic understanding of conformational transitions should be essential to forming a complete picture of such protein functions as enzyme catalysis, ion channel gating, membrane transport, and other vital molecular-cellular processes. Computational scientists have devised specialized algorithms, known generally as enhanced sampling (or rare-event sampling) methods, to overcome the equilibrium sampling problem (cf. [4, 29, 166, 168–171]). Some enhanced sampling techniques accelerate the overall rate at which phase space is explored, while others incorporate additional knowledge (about a dynamical system) so as to intelligently restrict sampling to relevant regions of phase space. A wide variety of enhanced sampling algorithms have been developed for protein systems to generate conformational transitions—henceforth referred to as *path-sampling* methods—using a fraction of the computational effort of equilibrium MD [4, 166, 171]. On the one hand, enhanced sampling generally requires the introduction of nonequilibrium forces/effects so as to overcome (equilibrium) free energy barriers; on the other hand, ideal enhanced sampling trajectories should at least represent, if not fully reproduce, equilibrium transition ensembles. By generating plausible transition paths very efficiently, many path-sampling methods enable the study of fast processes that elude equilibrium approaches, though whether these methods can reproduce physically realistic transitions remains an open question [4, 5].

2.3 Dynamic Importance Sampling MD

To provide an illustrative example of an enhanced (path-)sampling algorithm, we discuss the Dynamic IMportance Sampling MD (DIMS-MD) [172–174] method, a biased-MD approach that can be used to generate ensembles of plausible transition paths by biasing trajectories from an initial state toward a known final state (e.g., X-ray crystal structures). DIMS is used in Chapter 4 to generate ensembles of closed $\rightarrow$ open for apo-AdK and diphtheria toxin (DT) transitions that are analyzed using the path similarity analysis (PSA) method introduced in Chapter 3. As such, several relevant aspects of the DIMS-MD method are reproduced here (cf. [174] for complete details) to reinforce intuition before introducing other path-sampling approaches in the subsequent section.

DIMS-MD (or DIMS) combines MD with a Maxwell-demon-like biasing approach that allows a trajectory to be stochastically guided from an initial state toward a known final state (e.g., X-ray crystal structures). DIMS is currently implemented in the CHARMM MD program and employs Langevin dynamics. Biasing is achieved through a combination of a suitable progress variable and a choice of biasing algorithm. A simple progress variable implemented in this dissertation is the root mean square distance (RMSD) to the target,

$$d_{\mathrm{RMS}}(\mathbf{X}|\mathbf{X}^F) = \sqrt{\frac{1}{M}\sum_{i\in\mathbf{X}} m_i \left(\mathbf{x}_i - \mathbf{x}_i^F\right)\cdot\left(\mathbf{x}_i - \mathbf{x}_i^F\right)}, \tag{2.2}$$

where $\mathbf{X}$ is set of 3D coordinates for a selection of atoms (e.g., backbone atoms), $\mathbf{x}_i$ are the 3D coordinates of the i^{th} atom in $\{\mathrm{atoms}\}$ and m_i its mass, M is the total mass of the atoms in $\{\mathrm{atoms}\}$, and $\mathbf{x}_i^F$ are the corresponding atomic coordinates in the final state $\mathbf{X}^F$. If the coordinates at the n^{th} MD step are $\mathbf{X}^n$, then the change in RMSD relative to F at step $n+1$ is given by

$$\Delta\phi(\mathbf{X}^n \rightarrow \mathbf{X}^{n+1}) = d_{\mathrm{RMS}}(\mathbf{X}^{n+1}|\mathbf{X}^F) - d_{\mathrm{RMS}}(\mathbf{X}^n|\mathbf{X}^F). \tag{2.3}$$

Thus, when $\Delta\phi$ is negative (positive), the step is considered to be toward (away from) the final state F. In the case of the soft-ratcheting algorithm[¶] (SRA), MD steps toward the

[¶]A normal-mode biasing scheme is also available, though this scheme is not used in this thesis; references to DIMS also refer to the SRA implementation unless explicitly stated. See [174] for more details.

target state are always accepted whereas steps away from the target are rejected with finite probability (Fig. 2.1), leading to a Metropolis-like acceptance criterion

$$p_{\mathrm{acc}}(\Delta\phi) = \begin{cases} 1, & \Delta\phi \leq 0 \\ \exp\{-|\Delta\phi/\phi_0|^2\}, & \Delta\phi > 0 \end{cases} \tag{2.4}$$

where the soft-ratcheting parameter ϕ_0 determines the softness of the ratcheting algorithm; when $\phi_0 \to \infty$ only forward steps are accepted, while $\phi_0 = 0$ recovers equilibrium (unbiased) dynamics.

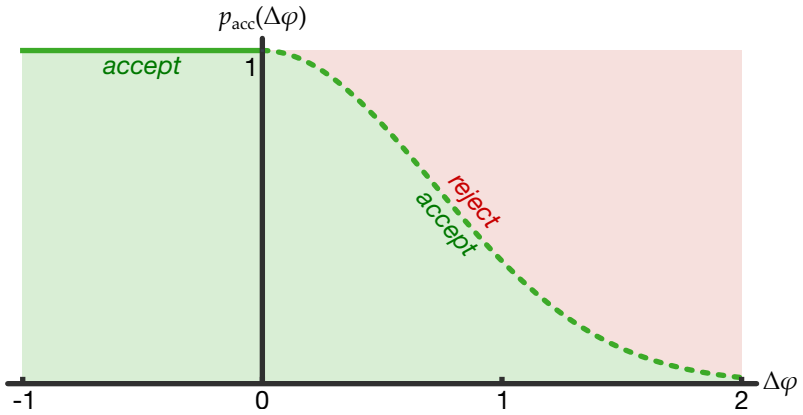

Figure 2.1: Schematic depiction of the acceptance and rejectance regions for DIMS-MD steps using the soft-ratcheting algorithm. Steps toward the target state, $\mathbf{X}^F$, lead to a decrease in the progress variable, ϕ, such that $\Delta\phi \leq 0$ (by definition) and are always accepted (i.e., $p_{\mathrm{acc}} = 1$). Steps away from the target increase the progress variable (i.e., $\Delta\phi > 0$) such that $p_{\mathrm{acc}} = \exp\{-|\Delta\phi/\phi_0|^2\}$, meaning that large backward steps are less likely than small steps.

During a DIMS simulation, a trial step is taken by computing the interatomic forces (from the force field) and stochastic forces (from the noise term in Langevin dynamics) and $\Delta\phi$ is calculated. When a step is rejected (finite ϕ_0), (only) the stochastic forces are resampled and $\Delta\phi$ calculated until a step is rejected. A notable advantage of SRA biasing over TMD, for instance, is that SRA supplies what is essentially an "entropic force"—thermal fluctuations that tend to progress a trajectory toward the target state are selected preferentially—thus obviating the need to perturb the system's native Hamiltonian with an external bias potential. Furthermore, the softness of the acceptance criterion gives energetically "stuck" trajectories a chance to circumnavigate large barriers via backtracking.

2.4 Path-sampling methods

An impressive array of physical models and algorithmic techniques have been developed and exploited with the purpose of sampling large-scale conformational transitions otherwise inaccessible to EqMD. Generally speaking, path-sampling techniques aim to accelerate the overall rate at which configuration space is explored (e.g., reducing the number of degrees of freedom via coarse-grained or implicit solvation models) [29, 169], restrict the space of sampling by incorporating prior information into a biasing algorithm (e.g., targeting an end state represented by a known experimental structure or driving sampling away from previously explored regions), or employ some combination thereof [4, 166, 168, 170, 171]. Perhaps the simplest class of approaches is structural morphing based on linear interpolation [175], including the linear adiabatic mapping algorithm [176] and nonlinear morphing methods like Climber [177]. Many path-sampling methods are not specific to protein systems and may be applied to an arbitrary dynamical system, while others have in mind the specific goal of, for instance, generating protein conformational transitions between two states (e.g., two-state elastic network models). One key function of path-sampling is to locate a reaction coordinate that passes through the transition state (TS) to obtain an accurate estimate of the free energy barrier (e.g., from a PMF calculation) [178]. The barrier height (and width) will, however, depend on the choice of coordinate along which the PMF is sampled, and poor reaction coordinates can lead to misleading pictures of the underlying physical process [101].

As path-sampling methods may differ in their level of spatial detail, whether temporal information is included (e.g., dynamical trajectories versus minimum energy paths), their suitability for extracting thermodynamic information, their overall computational efficiency, an so on. However, a path-sampling method may integrate a number of techniques to enhance sampling; for instance, two MD-based methods may differ primarily in their biasing algorithms, or normal mode information from an elastic network model (ENM) [29] might be used to bias many distinct physical models (e.g., a model based on an all-atom

molecular mechanics force field, another on a Gō model [179], and a third using only stereochemical considerations [180]). It is clear that such combinatorics frustrates attempts to find a consistent classification of all known path-sampling methods, especially since many approaches hybridize several methods and physical models. In the synopsis that follows, we focus on the classes of algorithms that are representative of the path-sampling methods comparison performed in Chapter 4.

2.4.1 Enhanced dynamical sampling

Enhanced path-sampling methods based on dynamical algorithms (e.g., MD) retain temporal information about the system and generate time-dependent trajectories [30, 172, 181–186]. Sampling can be enhanced over equilibrium MD either by reducing the computational cost per time step (e.g., eliminating degrees of freedom via implicit solvent or coarse-grained models), by minimizing the total number of time steps needed for sampling (e.g., adding a bias that guides sampling toward a known target), or both [166]. Many exploit the use of a low-dimensional collective variable (CVs) space in which sampling is performed and, for some methods, in which the free energy surface can be estimated.

Examples of methods that combine a bias algorithm with a known target state include essential dynamics sampling MD [187], dynamic importance sampling MD (DIMS-MD) [172, 173, 188], adiabatic bias MD (ABMD) [189, 190], and targeted molecular dynamics (TMD) [30, 191] and restricted perturbation TMD (RP-TMD) variants [192, 193]. Other methods use compensatory biases to minimize time spent in metastable equilibria. Adaptive biasing force (ABF) methods [194, 195] act to dynamically adjust an external biasing force so as to exactly cancel an on-the-fly free energy (PMF) gradient estimate, thereby flattening the free energy surface (FES). Metadynamics [73, 196], on the other hand, deploys history-dependent biasing potentials in a (predefined) CV space to avoid regions where sampling has already occurred, and whose effects on the equilibrium (Boltzmann) distribution can be "undone" to compute the FES in the CV space; a recent adaptation known as *infrequent metadynamics* enables direct kinetic rate calculations as well [78, 83]. Inspired by

metadynamics, temperature-accelerated MD (TAMD) [197, 198] (which is fundamentally different from TAD) is an extended-system approach, where a chosen set of CVs are promoted to dynamical variables and artificially high temperatures are used to hasten the exploration of the free energy surface over those CVs.

In the case of coarse-grained molecular dynamics (CG-MD) simulations that employ structure-based models of the potential energy function, speed gains up to a few orders of magnitude over atomistic MD are possible due to reduction in computational costs at each step (i.e., fewer degrees of freedom are needed to compute forces) in addition to smoothing the potential energy surface (PES) and increasing the rate of configuration space exploration (i.e., removing roughness in the PES by removing full atomic detail) [169, 199]. Structure-based models incorporate interactions between atom or residue pairs and, using microscopic or macroscopic mixing models, merge the energetic terms from two end states into a unified PES [200]. Structure-based models are coarse-grained but retain important local and global features of the landscape like metastable intermediate states and even the atomic-scale roughness of the full PES. Several studies using structure-based models to perform CG-MD have been applied to the AdK closed $\leftrightarrow$ open transition [201–204].

2.4.2 Methods based on elastic network models

The elastic network model (ENM) assumes a harmonic energy approximation about a known structural configuration (e.g., an atomic crystal structure) and has found many applications due to its conceptual and computational simplicity [29, 205–212]. The ENM is often implement at a coarse-grained level (e.g., a C_α representation) making the calculation of slow, low-frequency modes with normal mode analysis (NMA) a computationally inexpensive way to predict large-scale, collective protein motions. The first example was the atomic resolution elastic network introduced by Tirion (the Tirion model) [205] and the idea of a spring "network" applied to C_α atoms in the coarse-grained Gaussian network model (GNM) by [206] less than a year later [29]. Soon after, Hinsen's ENM [213, 214] and then the similar, well-known anharmonic network model (ANM) [207] extended the GNM

description of 1D fluctuations to 3D. More recently, a dissipative electro-elastic network model (DENM) was introduced that includes charge-charge interactions and overdamped Langevin dynamics on the normal modes [76], including a variation called the solvated DENM (sDENM) that re-normalizes the Hessian to account for solvent interactions with ionized surface residues [77].

In the context of conformational transition paths, the ENM can be used to generate transitions using a number of different approaches. One approach that can roughly capture intermediate states away from local minima is to iteratively advance a structure along the slowest eigenmodes while gradually decreasing the distance to a target structure [215–217]. Another approach—used by many of the ENM-based methods examined in Chapter 4—has been to construct two-state potential energy functions; from ENM representations of two known end states, a "mixing function" unifies the harmonic wells and a plausible transition path can be extracted by finding, for example, a minimum energy path (MEP) [38, 208, 209, 218, 219]. It is common for the ENM (and NMA) to be hybridized with other methods as a means of guiding various kinds of geometry-based or dynamical sampling methods along a few slow normal modes [216, 220–223].

2.4.3 Information-based approaches

There are a growing number of methods that exploit "prior information" to perform an intelligent search of configuration space, usually without regards to the underlying dynamics. One such method is the geometrical targeting algorithm FRODA [180], which is used in Chapter 4 in testing Path Similarity Analysis (PSA) and comparing various path-sampling methods. Other methods include motion planning algorithms, such as the rapidly exploring random trees (RRT) algorithm [224, 225], the probabilistic road map (PRM) [226, 227], and the "mining-minima" algorithm [228]. Some methods eschew potential energy functions by incorporating basic geometric considerations (e.g., preserving protein stereochemisty by avoiding atomic clashes, respecting loop-closure constraints,

etc.), including geometry-based essential sampling [229], geometric-based RRT [230], and geometric targeting [180].

2.4.4 Other methods

There are a substantial number of methods that bear mentioning, although the details are beyond the scope of this dissertation. Ensemble-type approaches are based on the notion of an equilibrium trajectory ensemble between two states. Such methods are often directly suitable for or facilitate kinetic rate calculations, though some require an order parameter or reaction coordinate as initial input. Many well known examples take the approach of dividing up the region between the end states into a series of interfaces, including transition path sampling (TPS) [57], transition interface sampling (TIS) [231, 232], forward flux sampling [233–235], weighted ensemble dynamics [82, 236–238], and milestoning algorithms [81, 239].

String-type methods place an ensemble of simulations in sequence, linking the members by a virtual "string" to form a path; an initial path is defined as input and then relaxed by performing energy minimization or MD on each member concurrently so as to locate a MEP or minimum free energy path (MFEP). Those that generate an MFEP include the finite-temperature (FTS) string method [240, 241] and FTS based on CVs [242, 243], "strings with swarms" [244]. Methods that generate an MEP include the zero-temperature string (ZTS) method [245], nudged elastic band methods [246–248], and conjugate peak refinement [249]. A disadvantage of these methods is that changing the initial path can alter the predicted low energy pathway.

A few other methods include replica exchange MD (REMD) [182, 250], hyperdynamics [80, 181, 251] and accelerated MD (aMD) [183], and temperature-accelerate dynamics (TAD) [252, 253].

2.5 Computational methods and AdK

The overview of AdK computational studies that follows serves two purposes. The first objective is to revisit the concrete example of the AdK closed $\leftrightarrow$ open transition in order to provide a lens through which the role of computational methods—in addressing the protein structure-function problem—can be put into proper context, as well as connecting back to experimental AdK studies (Chapter 1). In particular, it is hoped that this synopsis reveals extant gaps in our understanding so as to motivate the need for some kind of integrative framework that can facilitate quantitative comparisons between path-sampling methods, "gold-standard" MD, and, in principle, experimental data, thus motivating the Path Similarity Analysis (PSA) approach introduced subsequently in Chapter 3. The second objective is to provide the necessary background for Chapter 4 in which the AdK closed $\rightarrow$ open transition is used as a model problem to compare a number of fast path-sampling methods with the PSA framework. To be clear, it is not the intention of the author to provide a thorough review of either path-sampling algorithms nor the entire body of literature on AdK; however, every effort has been made to provide a fair, accurate, complete, and up-to-date account of these broad, interesting areas of discourse while also concentrating on general findings that are most relevant to this dissertation.

Convincing experimental evidence from several NMR and FRET-based studies [3, 19, 41] indicates a connection between the motions of substrate-bound AdK during the full catalytic cycle and the intrinsic dynamics of the apo form. Since apo-AdK is simpler to simulate but is still likely to follow the dynamics of its substrate-bound forms, it has been appealing system to study using computational approaches and, in particular, has motivated meticulous investigations into the closed $\leftrightarrow$ open apo transition. An impressive number of computational studies have since attempted to elucidate the mechanism underlying the apo transition, has helped to turn apo-AdK into a de facto testbed. Though the picture of AdK's full catalytic cycle remains incomplete, the ever-increasing number of computational

studies, along with the advent of novel computational algorithms, are helping to shed light on AdK and enzyme function in general.

2.5.1 Metastable states and energy landscapes

Computational studies are divided on whether the closed or open state has a lower relatively free energy. Though a majority indicate that the closed state is energetically favorable [39, 40, 203, 204, 220, 254, 255], several studies explicitly predict a lower open state free energy from 1D umbrella sampling [56, 58, 256, 257] and a few have observed a preference for open conformations [184, 187, 258].

The 2D free energy landscapes computed by Beckstein et al. [40], Whitford, Onuchic, and Wolynes [202], Q. Lu and J. Wang [203], and Bhatt and Zuckerman [255] all indicate that at least two intermediates, resembling LID-open-NMP-closed or LID-closed-NMP-open conformations, are plausible locally stable states. One study hypothesized that a so-called half-open/half-closed (HOHC) state—a LID-semi-open and NMP-semi-closed conformation similar to the 1ak2 crystal structure (Fig. 1.2, top)—should be directly involved in the catalytic cycle [259]. Using a coarse-grained structure-based representation of AdK, Y. Wang et al. [260] identified a population of LID-closed-NMP-open intermediates similar to 1dvr:A as well as the HOHC intermediate, though the apparent instability of the latter was suggested to have been due to the absence of explicit protein-solvent interactions in their model. A number of other computational studies also imply the existence of intermediate states [184, 219, 220, 222, 257, 258, 261].

The existence of one or more metastable states—as distinct from closed and open populations—in the apo ensemble entails a more intricate picture of the dynamics than what is likely to be captured by a single order parameter or collective variable. Substantial experimental and computational evidence has led to the contemporary view that the native fluctuations of apo AdK encode the substrate-bound dynamics. There is a lively discourse concerning the order of apo state domain motions and functional pathways due in large part to the relative ease of ligand-free computational modeling. A variety of simulation

methods have been used to produce several distinct pathways that recapitulate, to varying degrees, the schematic representation of LID-opening and NMP-opening pathways and intermediate states in Fig. 1.2.

Numerous computational studies have generally found that the LID can sample a range of conformations in a relatively flat free energy landscape spanning open and closed states [40, 56, 220, 254, 256, 259]. These results are in agreement with experimental SAXS data that indicate large-scale, rigid-body movements of the LID comprise the dominant scattering feature of the apo enzyme [53]. Such innate flexibility in the ATP-binding LID is consistent with a conformational selection mechanism. Chemical shift perturbation analyses from solution-state NMR support the idea that, much like the apo enzyme, the ATP-bound complex comprises a transient structural ensemble that populates closed and open states in similar proportions [41]. This dynamic equilibrium picture contrasts with an induced fit model that implies a well-defined ATP-AdK complex is induced by ATP binding. A particularly interesting result was that Whitford, Onuchic, and Wolynes [202] demonstrated the viability of a mixed mechanism—LID motions are described by conformational selection and NMP motions are ligand-induced—leading to a picture where LID opening (closing) always precedes NMP opening (closing).

Most computational studies find clear evidence of a LID-opening pathway where the LID opens in advance and independently of the NMP domain [38, 40, 203, 222, 224, 255, 258, 259, 261, 262]. On the other hand, an appreciable number of studies predict an NMP-opening (or LID-closing) pathway, with most finding evidence for an initial step defined by a closed LID and a *partial* opening of the NMP domain, resulting in a LID-closed-NMP-half-open (or similar) intermediate state. It is not known whether this state is truly metastable, though several studies suggest that a state similar to LID-closed-NMP-open conformation may be catalytically competent [259, 263] including simulations using a structure-based model [260] and a recent study involving extensive QM/MM (quantum mechanics molecular mechanics) simulations [97]. The opening motions following partial NMP-opening were

in some instances characterized by a complete opening of the LID [261] and in other cases defined more or less by a concurrent opening of both mobile domains [184].

2.5.2 What does EqMD have to say about AdK?

Large-scale conformational changes are statistically rare events, but there is evidence that the closed $\rightarrow$ open transition of apo AdK may be accessible to equilibrium MD (EqMD) simulations. The relatively early study by Pontiggia, Zen, and Micheletti [264] used the OPLS-AA force field, running simulations from both the closed and open states (50 ns each), observed appreciable LID closing from the open state. Brokaw and Chu [258] performed EqMD simulations of AdK from the closed and open states both with and without ligands using all-atom CHARMM27 (CHARMM22 + CMAP) force field, observing a near-complete opening initiated by the LID and followed by NMP over the course of 40 ns in a 100 ns simulation. H. D. Song and Zhu [257], using the all-atom CHARMM36 force field and TIP3P water, reported multiple spontaneous closed $\rightarrow$ open transitions within $\sim$ 20 ns and several within 100 ns to 200 ns in repeat simulations; most runs took a LID-opening pathway, although two runs exhibited NMP-opening after the LID initially opened slightly. More recently, D. Li, M. S. Liu, and Ji [90] EqMD with Amber03 force field observed distinct LID-opening and NMP-opening pathways well within 100 ns, the former characterized by relatively rapid opening by the LID and the latter by an initial partial opening of NMP into a semi-open conformation. To compare with previous studies, D. Li, M. S. Liu, and Ji [90] also produced 500 ns CHARMM27 and OPLS-AA simulations from the open state, finding that the open conformation was stabilized with CHARMM27 but LID-closed states were reachable on the order of 100 ns with OPLS-AA.

The question of the impact of force field choice on transition mechanisms was the subject of interest in a recent study by Unan, Yildirim, and Tekpinar [93] where AdK was simulated from the closed conformation for a total of 8 µs using Amber99, CHARMM27, Gromos53a6, and OPLS-AA (2 µs per force field). Despite relatively extensive sampling, one should be cautious in drawing firm conclusions about the energy landscape or functional pathways

(predicted by a given force field); however, their results were consistent with previous studies, showing that CHARMM27 simulations were largely restricted to the vicinity of the 4ake structure, while Amber and OPLS-AA simulations could adopt closed conformations and access LID-closed-NMP-open states relatively easily. Up to this point, EqMD simulations (regardless of force field choice) have not been observed to sample conformations in the immediate vicinity of the 1ake closed crystal structure ($\gtrsim 2$ Å), which is likely accessible on the millisecond timescale [43]. However, simulations using OPLS-AA and, to a lesser extent, Amber force fields have consistently been able to reach virtually closed conformations starting from either closed or open states, while those using CHARMM force fields clearly stabilized the open state. It is worthwhile to consider that chemical-shift NMR strongly suggests that fully closed conformations require *both* substrates to be bound [41]; in an MD study by Delalande, Sacquin-Mora, and Baaden [265] on guanylate kinase (GK), a related enzyme with three domains and similar lid-like and substrate-binding moieties, physics-based mechanical resistance analyses support the idea that magnesium ion (Mg^{2+}) coordination of a network of select residues (and possibly water molecules[||]) is necessary to stabilize and "lock" the closed state. Nevertheless, the precise mechanostructural correlates of cofactor-induced active-site coordination or even different MD force fields involved in the closure (or opening) of AdK remain to be discovered.

2.5.3 Summarizing remarks

Although apo-AdK appears to represent a conceptually simple system whose closed $\leftrightarrow$ open transition appears to be straightforward to investigate, there is as of yet no clear consensus regarding many aspects of the free energy landscape, intermediate and transition states, functional pathways, or kinetic rates—it is possible that the underlying dynamics

[||] Based on their observation of several localized water molecules near the Mg^{2+}·ATP chelate, Delalande, Sacquin-Mora, and Baaden [265] hypothesized that something like a water "channel" or "bridge" may also take part in coordinating enzyme closure. Given the large negative charges of ATP, ADP, and AMP under physiological conditions, it is plausible that multivalent metal cofactors like Mg^{2+}—along with water molecules and charged residues—may be necessary to facilitate the functional closure and/or nucleotide binding and release of AdK, GK, or other kinases.

are richer than originally suspected. In many cases it is still unclear whether either of the mobile domains are occupying fully open or fully closed states. This question could be resolved by computing a free energy landscape using a suitable set of collective variables that can fully characterize the accessible substates that span the open and closed conformations for both mobile domains. However, 2D projections such as angle-angle or domain center-of-mass coordinates may only be effective descriptions under the assumption that mobile domains behave as rigid-bodies; such projections may differ in their sensitivity to twisting or bending motions, or they may not be able to capture such motions at all, which presents a challenge when precise distinctions are needed to resolve intermediates in *close* proximity to closed and open state populations. An important consideration that has been somewhat overlooked is the precise definition of the closed state; more carefully analyses will be needed to elucidate any important differences between closed-like apo and holo conformations and the conformation adopted by the inhibitor-bound 1ake crystal structure.[**]

It should be reiterated that the depiction in Fig. 1.2 ultimately represents a convenient but artificial dichotomy. Indeed, a number of studies describe transitions that do not categorically delineate a LID-opening or NMP-opening pathway. For instance, Korkut and Hendrickson [269] used the virtual atom molecular mechanics (VAMM) algorithm to compute reversible closed $\leftrightarrow$ open paths mimicking apo, ATP-bound, and AMP-bound structures; the apo transition, viewed in the closed $\rightarrow$ open direction, shows that a slight NMP opening preceding a dominant LID-opening step is followed by the more or less concurrent opening of both domains. It is likely that the disagreement among computational studies is likely due to, in part, the fact that PMFs strongly depend on the chosen path (or CV space) along which sampling is performed, as well as on obtaining sufficient sampling for convergence.

[**]MD-based mutation studies [266] and investigating the molecular mechanism of thermophile stability [267, 268] may shed light on the role of salt bridges, hydrogen bonding, and other electrostatic interactions (among ions, charged residues and substrates, and water) in protein stability and kinase catalysis.

That generated pathways can sensitively depend on user choice is demonstrated by Chu and Voth [219], where a double-well ENM (DWNM) potential—effectively a network of 1D double-wells that capture the "roughness" of the potential landscape—in which dissimilar MFEPs between the closed and open states were computed by initiating the refinement process in the DWNM potential using different initial paths as input; refining a path based on simple linear interpolation produced an NMP-closing (LID-opening) pathway in the DWNM, with a barrier located halfway between the end states. By contrast, starting from a MFEP that was initially refined in a PNM-based (smooth) energy landscape led to a LID-closing (NMP-opening) pathway in the DWNM with a barrier positioned near the closed state. Making this point even more explicit, Pan et al. [270] found clear evidence that certain collective variable choice prevented two tested enhanced sampling methods, namely the string method and steered MD (similar as used to TMD), from producing accurate paths. Given such difficulties in arriving at a consensus for a (relatively) simple enzyme system such as apo-AdK, there is a need: (1) for a reconsideration of whether apo-AdK is in reality an ideal testbed; (2) for reliable, automated methods for locating and validating good collective variables [271]; (3) for computational methods that can quantify differences between transition paths and their originating methods, as well as integrating data from multiple methods. In the following chapter, we introduce the Path Similarity Analysis (PSA) method [5] as a possible solution to point (3).

Chapter 3

A NEW APPROACH TO QUANTIFYING CONFORMATIONAL CHANGE

This chapter is based in part on the published study, **Sean L. Seyler**, Avishek Kumar, Michael F. Thorpe, and Oliver Beckstein (2015). *Path Similarity Analysis: A Method for Quantifying Macromolecular Pathways.* PLoS Comput Biol 11(10): e1004568. [5] My contribution to this work involved the conception, design, and performance of the experiments, analysis of the data, writing of the paper, and design of the PSA software.*

With Chapter 1 and Chapter 2 establishing a need for new computational methodologies—capable of integrating heterogeneous data generated by disparate methods—the quandary of high-dimensional systems and dimensionality reduction is summarily presented, followed by a detailed introduction to Path Similarity Analysis (PSA). To strengthen intuition, the chapter culminates in a comprehensive mini-study of transition paths (generated by a toy model based on Brownian dynamics) using the PSA approach; the mini-study also provides a tangible means of verification for the PSA methodology and corresponding Python module implemented in the MDAnalysis Python library [272, 273] (see `mdanalysis.analysis.psa`; the data are available via the PSA tutorial on GitHub under the GNU General Public License, v3 DOI 10.5281/zenodo.31457). These analyses lay the necessary groundwork for real-world applications of PSA using representative protein systems in Chapter 4.

The toy model, which includes a simple but consistent coarse-graining scheme, reproduces several relevant aspects of biomacromolecular systems such as large-scale transitions, thermal fluctuations, dimensionality changes under coarse-graining, and barrier-crossing

*The detailed author contributions are as follows. Conceived and designed the experiments: SLS OB. Performed the experiments: SLS AK. Analyzed the data: SLS AK MFT OB. Contributed reagents/materials/analysis tools: SLS AK. Wrote the paper: SLS AK MFT OB. Designed the software: SLS.

events; full details, including the parameter settings and a mathematical derivation of the coarse-graining scheme are provided in Appendix B.

3.1 Dimensionality reduction and collective variables

The concept of dimensionality reduction is essential to interpreting data, in both experiment and numerical simulation, generated by a high dimensional system. The degree of difficulty in quantifying and visualizing a dynamical system grows very quickly as the dimensionality increases beyond three. The purpose of dimensionality reduction is to minimize the number of independent variables necessary to distinguish between relevant states of the system. In the case of a protein consisting of N-atoms, the raw spatial coordinates collectively represent a $3N$-dimensional (configuration) space. Conformational transitions, represented by a sequence of conformers, are therefore $3N$-dimensional configuration space paths. To quantify large-scale collective motions, it is useful to define a *reaction coordinate* (RC), order parameter, or collective variable (CV); these are similar concepts that take on slightly different meanings in different contexts (cf. [57]). For brevity, the following distinctions will suffice: CV is the most general term, RCs and order parameters are both examples of CVs, and an RC is a special case of a *discriminating* order parameter. An order parameter, q, simply assigns states to (one of two) basins of attraction in an energy landscape—it is a discriminating order parameter if, given a transition state (TS) at $q = q^*$, it correctly assigns the states where $q < q^*$ ($q > q^*$) to the initial (final) basin I (F). By contrast, a (good) reaction coordinate is fundamentally determined by the underlying reactive pathway; order parameters are not necessarily good reaction coordinates, though discriminating order parameters can sometimes serve as reasonable approximations. However, a discriminating order paramter can still obscure the actual sequence of events along the true reactive pathway (cf. Figs. 8 and 9 in Bolhuis et al. [57]).

Most generally, a collective variable (CV) is a single coordinate that describes the collective motion in a system, potentially having many degrees of freedom, as a possibly

nonlinear combination of the spatial coordinates. Formally, the collective variable ξ is

$$\xi(\mathbf{X}) = \xi(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N), \tag{3.1}$$

where $\mathbf{x}_i$ is the 3D coordinate vector for particle i and N is the total number of particles in the system. The ensemble-averaged probability density function (pdf), $\langle\rho\rangle$, describing the system along ξ is a Boltzmann-weighted average over the spatial coordinates $\mathbf{X}$, so that

$$\langle\rho(\xi)\rangle = \frac{\int \mathrm{d}\mathbf{X}\, \delta\big(\xi'(\mathbf{X}) - \xi\big) e^{-U(\mathbf{X})/k_\mathrm{B}T}}{\int \mathrm{d}\mathbf{X}\, e^{-U(\mathbf{X})/k_\mathrm{B}T}} \tag{3.2}$$

To find the potential of mean force (PMF) along a collective variable ξ (relative to some reference state, ξ_I), we take the natural logarithm of the ratio of the ensemble-averaged pdfs to obtain

$$W(\xi_F) = W(\xi_I) - k_\mathrm{B}T \ln \frac{\langle\rho(\xi_F)\rangle}{\langle\rho(\xi_I)\rangle} \tag{3.3a}$$

$$\Delta W(I \to F) = -k_\mathrm{B}T \ln \frac{\langle\rho(\xi_F)\rangle}{\langle\rho(\xi_I)\rangle}, \tag{3.3b}$$

where $\Delta W(I \to F) = W(\xi_F) - W(\xi_I)$ is the effective free energy difference between states I and F *as parameterized by* ξ [101]. Rearranging Eq. 3.3b, it can furthermore be seen that the ratio of the ensemble-averaged pdfs is the Boltzmann-weighted free energy difference $\Delta W(I \to F) = W(\xi_F) - W(\xi_I)$ such that

$$\frac{\langle\rho(\xi_F)\rangle}{\langle\rho(\xi_I)\rangle} = e^{-\Delta W(I \to F)/k_\mathrm{B}T}. \tag{3.4}$$

In many situations it is not obvious how a CV such as ξ can be chosen so as to faithfully characterize a transition (between states I and F); in the worst case, it is possible to chose a CV along which the computed PMF provides a misleading characterization of the underlying kinetic process.[†] Perhaps up to three or four CVs can be used for visualization of a given dynamical process before plotting the CVs becomes laborious or unintuitive—a

[†]See the papers by Bolhuis et al. [57] and Dellago et al. [178] for some of the earlier discussions about order parameters vs. reaction coordinates. Zuckerman [274] also has a nice write-up on his blog that draws attention to some of the dangers of PMF calculations.

good CV in general should be descriptive and intuitive. An ideal dimensionality reduction algorithm would be able to identify the minimum number of CVs that correspond to the intrinsic low-dimensional dynamical manifold [275] for an arbitrary system as input. In reality, identifying CVs that adequately span the relevant dynamical space, e.g., capture a conformational transition, is not only highly nontrivial, but often system-dependent [276, 277].

A common technique is to perform principal component analysis (PCA) on the Cartesian coordinates of a system evolving in time. Since the eigenvalue corresponding to a given principal component is a measure of the variation that it describes, one can project the full dynamics onto the subspace spanned by the first few principal components so as to capture most of the variation [278, 279]. Another general approach is native contacts analysis (NCA), which has found frequent use in characterizing protein folding pathways [218, 280]. A native contact is one that exists in a known reference structure; a relatively general analysis approach involves measuring, for each conformer in a putative transition path, the fraction of contacts that are present in (shared with) a native state—the native contacts. Given initial and target conformations (e.g., from known crystal structures), native contact fractions can then be plotted on a 2D space to produce a projection into native contact (NC) space. 2D NC projections are particularly useful when good CVs are not known a priori.

Heuristic collective variables typically trade generality for an intuitive description of a given system. Choosing ad hoc CVs usually requires strong intuition about the system in question, limiting the utility of this approach when studying unfamiliar systems. However, for relatively simple systems, one or more sets of CVs may be available as effective low-dimensional descriptions; in the case of the AdK closed $\leftrightarrow$ open transition, several descriptive collective variables are known (cf. Section 1.2.1 and Table 1 in S. L. Seyler and Beckstein [4]). Previous studies, for example, have used angle-angle coordinates to quantify the degree and order of opening of the LID and NMP domains relative to the CORE domain during the closed $\leftrightarrow$ open transition [40]. The LID-CORE angle, θ_{LID}, is defined as the angle between the backbone and C_β atoms in residues 179-185 (CORE), 115-125 (CORE-hinge-

LID), and 125-153 (LID), while the NMP-CORE angle, θ_{NMP}, is formed by the geometric centers of the backbone and C_β atoms in residues 115-125 (CORE-LID), 90-100 (CORE), and 35-55 (NMP). Projecting full trajectories into the 2D space defined by (θ_{NMP}, θ_{LID}) provides an intuitive means to visualize the conformational changes involved. Such descriptive projections are, however, not always available for many protein systems and may also fail to describe such subtle movements as twisting, bending, or breathing modes.

One cannot guarantee that dimensionality reduction or heuristic approaches will identify a descriptive set of CVs for an arbitrary system—a set that is both descriptive (i.e., the CVs have intuitive or physical interpretations) and complete (i.e., the CVs capture the essential dynamical motions). It should be reiterated, however, that the primary reason for projecting dynamics in CV spaces is to visually distinguish and quantify differences between multiple transition paths or pathways. Thus, in comparing paths generated by path-sampling methods, a good set of CVs should be able to distinguish paths in situations where those paths are meaningfully separated in configuration space; importantly, such a set of good CVs can serve to identify the originating method of a given path when the physical model used to produce it leads to a distinct dynamical pathway.

3.2 A new approach: Path Similarity Analysis

3.2.1 The general idea

To address the general problem of quantitatively comparing $3N$-dimensional macromolecular transition paths, we describe the *Path Similarity Analysis* (PSA) approach introduced in S. L. Seyler et al. [5]. PSA is a quantitative framework for the analysis of conformational transition paths. The central idea is to employ a suitable distance function, a *path metric*, in order to compute a distance between (i.e., geometric similarity of) two transition paths. Once distances between all pairs of paths are calculated, cluster analyses can be used to group paths by similarity. Furthermore, we can identify the structural determinants that give rise to differences between any two transition paths at the atomic

level by exploiting properties of the underlying metric. In other words, PSA allows us to measure the degree to which any two transition paths are different and identify the structural origins of those differences. A key advantage of the PSA approach is that it is system-agnostic and is particularly effective for assessing transitions with well defined initial and final states. A typical minimal procedure for applying PSA is to (1) represent the transition paths as mathematical objects, (2) quantify the pairwise similarities between the paths using a suitable distance function, and (3) use cluster analysis to extract groups of similar paths.

Configuration space trajectories. To start, we treat a transition path (e.g., a simulation trajectory) for a protein system of N atoms mathematically as an ordered sequence of points in $3N$-dimensional configuration space; if it is imagined that successive points are connected by straight lines, a transition path is thus a high-dimensional piecewise-linear, or polygonal, curve. A set of transition paths can be meaningfully compared only if they exist in the same $3N$-dimensional configuration space, though it is possible to choose a reduced representation for all paths (e.g., backbone atoms, only C_α atoms, etc.).

Path metrics. Next, a path metric is selected to quantify differences between paths; the chosen path metric, δ, must accept two polygonal curves as input and return a non-negative real number representing a distance between the curves. In particular, given three polygonal curves A, B, and C, δ must satisfy

$$\delta(A, B) \geq 0 \tag{3.5a}$$

$$\delta(A, B) = 0 \iff A = B \tag{3.5b}$$

$$\delta(A, B) = \delta(B, A) \tag{3.5c}$$

$$\delta(A, C) \leq \delta(A, B) + \delta(B, C). \tag{3.5d}$$

Together, the identity property (Eq. 3.5b) and non-negativity (Eq. 3.5a) imply, for two curves A and B, where B undergoes a continuous deformation (Fig. 3.1) such that $\delta(A, B)$ is monotonically decreasing, that B becomes increasingly similar to A in such a way that

B is *identical* to A when $\delta(A,B) = 0$. The commutative property (Eq. 3.5c) is clear from the definition. The triangle inequality (Eq. 3.5d) generalizes the transitive property and preserves our usual intuition about the concept of closeness—if two curves A and B are "close" to C, then A and B are also close in the sense that their mutual distance is bounded from above by $d(A,C) + d(B,C)$. Many clustering algorithms furthermore require the triangle inequality to hold, so the use of proper metric functions is advisable.

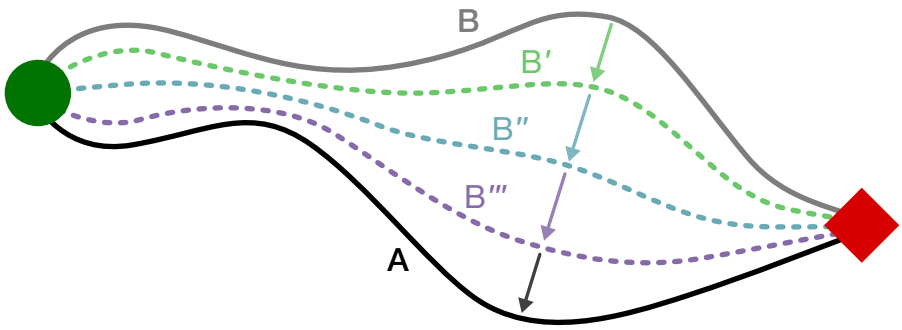

Figure 3.1: Schematic representation of two curves A and B that begin and end at the same states (green circle and red diamond, respectively). B undergoes a continuous deformation through several intermediate curves $B \to B' \to B'' \to B''' \to A$ (resp. green, blue, and purple dashed lines).

In this dissertation, we apply PSA to a toy model and realistic conformational transition paths using two path metrics—the Hausdorff distance [281–283] and the discrete Fréchet distance [284, 285] (defined below)—and assess any relevant differences. One important property of these metrics is that they do not account for dynamics, nor any physical time scales in a transition path; they are only sensitive to their geometry. Definitions are given in the sections that follow.[‡]

Distance matrix. Finally, given a set of N_{path} transition paths, we perform an all-pairs comparison, whereby the distances between all unique path pairs are stored in a symmetric (Eq. 3.5c), traceless (Eq. 3.5b) *distance matrix* that has $N_{\text{path}}(N_{\text{path}} - 1)/2$ unique elements (in the upper triangle). Equivalently, the distances may be stored in a *distance vector* of length $N_{\text{path}}(N_{\text{path}} - 1)/2$. Many clustering methods utilize a distance matrix or vector—generated

[‡]Measures of "distance" between polygonal curves that do not strictly satisfy the triangle inequality are not true distance functions, though some may be true metrics under restricted conditions; average-type Fréchet and Hausdorff distance functions are briefly explored in Appendix F.

from a distance *metric*—as input and return the distance matrix with its rows and columns re-ordered to reflect the clustering. Though technically classified as unsupervised learning, partitional clustering methods, such as the centroid-based k-means algorithm, are not used since the number of clusters, k, must be specified as an input parameter [286]. In order to avoid imposing a-priori constraints on the number of clusters—a quantity that we seek to *discover* in the data—hierarchical (agglomerative) clustering was used previously [5], as well as in this dissertation.[§] Since the clustering is performed using pairwise comparisons, it is important that the triangle inequality (Eq. 3.5d) hold for the metric used in generating the distance matrix.[¶] Other robust approaches that have not been explored here include Density-Based Spatial Clustering of Applications with Noise (DBSCAN), which can be found in the scikit-learn Python library (see `sklearn.cluster.DBSCAN` in v 0.19.0). Appendix E details some considerations that went into choosing the clustering methods applied throughout this dissertation.

3.2.2 Hausdorff distance

The Hausdorff distance is a measure of similarity between two distinct sets of points that, in general, need not have any internal structure (e.g., a directed polygonal curve). However, it is useful to define the Hausdorff distance in the context of transition paths given its current application in this dissertation.

Given an N-atom protein, a path P in its $3N$-dimensional configuration space may be represented by a sequence of points $\{(p_k)_{k=1}^{n} \mid p_k \in \mathbb{R}^{3N}, k = 1, \ldots, n\}$. Similarly, defined

[§]The k-means algorithm, which is designed for Euclidean distances, also has the disadvantage that the cluster centroids are not elements of the dataset. Thus, a centroid of a hypothetical cluster of transition paths will not a member of the set of paths forming that cluster, meaning that conformers in a "centroid path" will in general not have realistic structural features (stereochemically or otherwise). Medoid-based clustering (e.g., k-medoids) overcomes this limitation by seeking, for each cluster, a *medoid*—a "middle" or representative object (e.g., path) of a set (e.g., of paths). There are many other advanced algorithms for clustering and data discovery, though such discussion is well beyond the scope of this dissertation.

[¶]See, for instance, Baraty, Simovici, and Zara [287].

another path Q, $\{(q_k)_{k=1}^m \mid q_k \in \mathbb{R}^{3N}, k = 1, \ldots, m\}$. We first define

$$\delta_h(P \to Q) = \max_{p \in P} \min_{q \in Q} d(p, q) \tag{3.6}$$

as the *directed* Hausdorff distance from P to Q, where d is a distance metric on $\mathbb{R}^{3N}$ for single points [281]—$d(p, q)$ is measure of similarity between conformers p and q. For a given point p in P, define the nearest neighbor of p as the nearest point, $q*$, in Q. It can then be stated informally that $\delta_h(P \to Q)$ measures the nearest neighbor distances for all points in P and selects the maximum distance among them. Note that $\delta_h(P \to Q) \neq \delta_h(Q \to P)$. The Hausdorff distance can then be defined as

$$\delta_H(P, Q) = \max\left\{\delta_h(P \to Q), \delta_h(Q \to P)\right\}, \tag{3.7}$$

which is now commutative as required by metric property Eq. 3.5c. Calculating the exact Hausdorff distance can be naively done in $\mathrm{O}(pq)$ time, though an "early break" algorithm [288] that has a nearly-linear runtime was recently implemented in SciPy 0.19.0 [289] (see the `directed_hausdorff` function in `scipy.spatial.distance` and SciPy 0.19.0 RC1 for more information).

Loosely speaking, the Hausdorff distance between transition paths P and Q is the upper bound on their mutual nearest neighbor distances. We define a *Hausdorff pair*, or δ_H-pair, as the unique pair[||] of conformers, (p^*, q^*), corresponding to a Hausdorff distance. Fig. 3.2 schematically illustrates the Hausdorff pair concept, showing that the structural similarity of the Hausdorff pair defines, in some sense, where P and Q are maximally dissimilar, since all other conformers and their nearest neighbors are by definition structurally more similar: $d(p, q) \leq d(p^*, q^*) \,\forall\, p, q \in P, Q$.

In comparison to pairs of conformers chosen at random, a Hausdorff pair should, in principle, reveal the salient structural features that underlie differences in path geometry. As such, a Hausdorff pair analysis enables the extraction of structural indicators designating

[||] It is possible to obtain a *set* of Hausdorff pairs under certain situations; in practice, a Hausdorff pair will typically be unique.

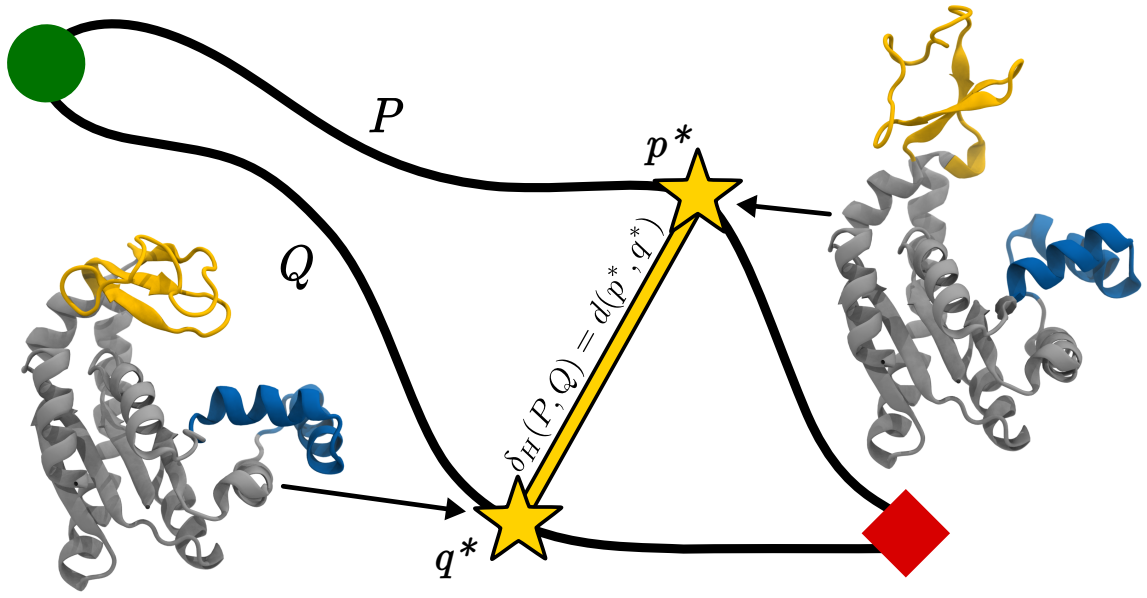

Figure 3.2: Two paths P and Q begin and end at the same states (green circle and red diamond, respectively). The Hausdorff distance, $\delta_H(P,Q)$, is indicated by the length of the yellow line and corresponds to the *point* distance between the Hausdorff pair (conformers p^* and q^*). The point-wise distance may be any structural similarity measure (i.e., distance metric), such as the RMSD**, that reports a distance between two conformers in configuration space.

"where to look" when searching for structural differences between two paths. While a Hausdorff pair establishes an upper bound on the structural dissimilarity of two paths, it may be useful to examine other "nearest neighbors". Using Eq. 3.6, we can define the *nearest neighbor distance* for a point p_k on path P to a second path Q as $\delta_{\mathrm{nn}}(k; P \to Q) := \delta_{\mathrm{nn}}(k; P \to Q) := \min_{q \in Q} d(p_k, q)$; similarly, the corresponding nearest neighbor distance for a point q_k on path Q to path P is $\delta_{\mathrm{nn}}(k; Q \to P)$. These distances are in general not symmetric, i.e., $\delta_{\mathrm{nn}}(k; P \to Q) \neq \delta_{\mathrm{nn}}(j; Q \to P)$ for any conformations j, k. By jointly plotting $\delta_{\mathrm{nn}}\big(k(\xi); P \to Q\big)$ and $\delta_{\mathrm{nn}}\big(j(\xi); Q \to P\big)$ as a function of a suitable order parameter ξ, it is possible to quantify the change in nearest neighbor distances over various segments or sections of the trajectories. Such an approach can be valuable when a more precise geometric picture of the manner in which two paths differ is needed. Hausdorff pairs and nearest neighbors can be easily extracted from an all-pairs calculation of Hausdorff distances in a transition path ensemble using the PSA module in the MDAnalysis Python library [272, 273], which provides a simple and efficient means to scrutinize structural patterns—down to the

atomistic level—across many sampled paths and, thus, identify the molecular-structural determinants that encode geometric differences between full transition pathways.

3.2.3 Discrete Fréchet distance

Whereas the Hausdorff distance is insensitive to the order of points in a path, Fréchet metrics are sensitive to orientation (i.e., direction of traversal). Indeed, when viewed as trajectories in configuration space, conformational transition paths are directional in time and, in principle, Fréchet metrics could be used to capture meaningful aspects of time dependence. We first discuss the *continuous* Fréchet distance [284] to provide an intuitive context for the discrete Fréchet variant.

The canonical way of visualizing Fréchet distance goes as follows [285]. First imagine a human and a dog, connected by a taught, extendable leash, walking along two separate paths P and Q, respectively, according several rules: (1) they both start at their respective initial points on P and Q, and they must eventually end at their respective final points; (2) backward steps are not allowed, so they can only stand still or move forward; (3) they are allowed to move independently of one another. Then, under these rules, we visualize all legal ways the human and dog can go from start to finish. Finally, remembering that they are connected by an extendable, taught leash, we identify an optimal protocol that minimizes the extension of the leash—the length of the shortest leash is defined as the Fréchet distance between P and Q, $\delta_F(P, Q)$. For a formal description, see Alt and Godau [285] and the discussion in S. L. Seyler et al. [5]. We note that for two polygonal paths, each with p and q vertices, exact algorithms for the continuous Fréchet distance take $\mathrm{O}\left(pq \log pq\right)$ time [285]; faster, but approximate, algorithms have also been explored [290, 291].

To describe the *discrete* Fréchet distance, δ_{dF}, we return to the picture of paths in $3N$-dimensional configuration space (i.e., polygonal, or piecewise-linear, curves) as in the discussion of the Hausdorff distance. Given two such paths P and Q, the leash analogy can be modified for discrete Fréchet in the following manner. The human and dog: (1) are restricted to discrete forward jumps to consecutive vertices (i.e., conformers) along a path;

(2) can move in succession so that one jumps while the other remains in place, or both can jump simultaneously to their respective consecutive vertices. When the typical distances between consecutive conformers are small (i.e., when the segments of a polygonal curve are short) relative to the characteristic distance between paths, the discrete Fréchet closely approximates the continuous metric. Eiter and Mannila [292] explored a dynamic algorithm that computes δ_{dF} in $\mathrm{O}(pq)$ time, providing a faster, simpler alternative to the continuous metric.

From the leash analogy, it is intuitively clear that a wandering path that backtracks should tend to make both Fréchet distances larger than the Hausdorff distance (Fig. 3.3). In particular, given the conceptual connection between a random walk and thermal fluctuations, dynamical path-sampling methods with explicit or stochastic solvent degrees-of-freedom may produce transition paths that backtrack many times. An intuition for some of the ways in which the Fréchet and Hausdorff metrics might differ can be understood in terms of upper and lower bounds.

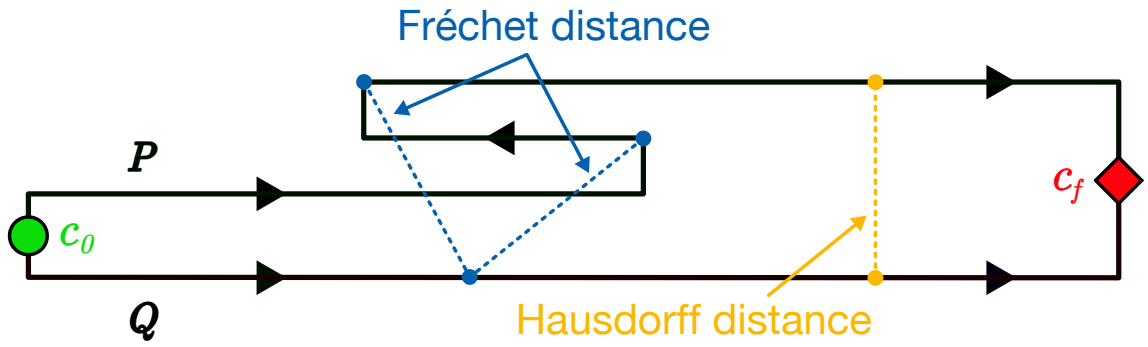

Figure 3.3: Two paths P and Q begin at state c_0 (green circle) and end at state c_f (red diamond) with direction indicated by the arrows. The Fréchet distance, δ_F, and Hausdorff distance, δ_H, are given by the lengths of the blue and orange lines, respectively. The two blue lines are the same length and correspond to the least extended Fréchet "leash"; the orange line spans a pair of points separated by the Hausdorff distance (only one is shown as there are infinitely many pairs of points with the same δ_H, since P and Q remain parallel after the backtracking). The backtracking of path P toward state c_0, combined with the monotonicity constraint of Fréchet (i.e., no backward motion along a path), leads to $\delta_F > \delta_H$. In this depiction, there are two Fréchet pairs (where the conformer Q is present in each pair) and an infinite number of Hausdorff pairs. [Reprinted from S. L. Seyler et al. [5], Copyright © (2015), under the terms of the Creative Commons Attribution License / Colors and fonts adapted from original.]

First, the continuous Fréchet distance is a lower bound on the discrete Fréchet distance, $\delta_F \leq \delta_{dF}$, since it is always possible to emulate a discrete jump using continuous

movements. In the leash analogy, smooth movements allow the possibility of advancing to the next pair of vertices using a shorter leash than would be required by discrete jumps—which can always be emulated as a last resort. It is furthermore intuitively clear that the difference between the discrete and continuous versions should be exacerbated when successive vertices are far apart. Second, the discrete Fréchet distance is also bounded from above, in that it will differ from continuous Fréchet by no more than the length of the longest line segment in the pair of polygonal paths in consideration. Thus, $\delta_{dF}(P,Q) \leq \delta_F(P,Q) + \max\{d_{\max}(P), d_{\max}(Q)\}$ for polygonal curves P and Q, where $d_{\max}(P) \equiv \max_{i=1,\ldots,p-1} d\left(p_i, p_{i+1}\right)$ is the length of the longest segment in P [292]. Third, the Hausdorff distance is a lower bound on the *continuous* Fréchet distance, $\delta_F \geq \delta_H$, for any pair of polygonal curves [293]. In fact, the Hausdorff and continuous Fréchet distances are equal for two *convex* polygonal curves [294], though it is possible to make the Fréchet distance arbitrarily larger than Hausdorff for certain curve geometries [290]. Thus, $\delta_H \leq \delta_F \leq \delta_{dF}(P,Q) \leq \delta_F(P,Q) + \max\{d_{\max}(P), d_{\max}(Q)\}$. In this dissertation, the discrete (rather than continuous) Fréchet distance is used exclusively; for brevity, "Fréchet distance" and the symbol δ_F henceforth refer to the discrete version, unless explicitly stated. We also mention for completeness that, in analogy to the definition of Hausdorff pairs given above, Fréchet pairs can also be defined, although we choose to omit further discussion since only Hausdorff pairs are used in this dissertation.

3.2.4 RMSD as a structural similarity measure

The Hausdorff metric (Eq. 3.7) and Fréchet metric (cf. Eiter and Mannila [292]) are both defined in terms of a point metric $d(p,q)$ on $3N$-dimensional configuration space that measures the distance (i.e., similarity) between conformations p and q. A common choice is to employ the root mean square distance (RMSD), which we define in the usual way as

$$d_{\mathrm{RMS}}(p,q) = \sqrt{\frac{1}{N}\sum_{i=1}^{3N}\left(p_i - q_i\right)^2}, \tag{3.8}$$

where N is the number of atoms, and $\{p_i\}_{i=1}^{3N}$ and $\{q_i\}_{i=1}^{3N}$ define the configuration space coordinates of conformations p and q, respectively. It is common to compare the mutual distance among several structures using the best-fit RMSD, where the RMSD is minimized on a pairwise basis by finding the optimal translation and rotation. Though such pairwise comparisons are intuitive, it is important to note that the best-fit RMSD is generally *not* a metric on the configuration space of structures—it can fail to obey the triangle inequality because the optimal superposition of conformer A onto conformer C and conformer B onto C does not generally result in the optimal superposition of A and B [295]. As an alternative, it is possible to use the RMSD as a metric by choosing a *common* (i.e., shared) reference structure, fitting all conformers to the reference, then calculating the RMSD between each pair (see Section 3.2.4). This global alignment approach is employed in Chapter 4 to guarantee that the path metric properties hold.

It should also be noted that Hausdorff and Fréchet metrics can be defined in terms of point metrics other than what is used in this dissertation. It is thus possible to deliberately emphasize differences in specific structural features of macromolecules by selecting a suitable point metric. As a simple example, one could choose to quantify the percentage of shared contacts between conformers as a putative measure of similarity. Another approach would be to use information-based metrics [296] to measure the similarity of protein ensembles, which are currently available in the ENCORE analysis module [297] through the MDAnalysis Python library [272, 273]. For reasons of simplicity and familiarity, however, the analyses presented in Chapter 4 exclusively use the RMSD using the global alignment approach based on a predefined reference structure.

3.3 A Path Similarity Analysis mini-study

In this section, the Path Similarity Analysis (PSA) method, using the Hausdorff and Fréchet metrics, is introduced by applying PSA to the simple case of assessing transition paths generated by a low-dimensional Brownian dynamics model. The trajectories are generated by driving toy molecules in one direction using a ramp-like global potential

function, while stochastically-driven fluctuations mimic temperature effects. Transition progress is measured using the center of mass position along the ramp; a transition is complete once a molecule's center of mass crosses a predefined threshold.

3.3.1 Overview and setup

Each toy "molecule" is an N-particle cluster, where each particle has the same mass and is connected to all others by springs as illustrated in Fig. 3.4. All inter-particle forces are harmonic with an equilibrium at zero separation, while a global "double-slide potential" drives particles with a constant force along the z-direction; the y-direction has the double-well shape depicted in Fig. 3.4 with an energy barrier equal to $2\ k_{\mathrm{B}}T \approx 5\,\mathrm{kJ\,mol^{-1}}$ at $T = 300\,\mathrm{K}$, while a harmonic potential restrains motion along the x-direction. Individual particles move according to Brownian dynamics subject to the global and interparticle spring forces.

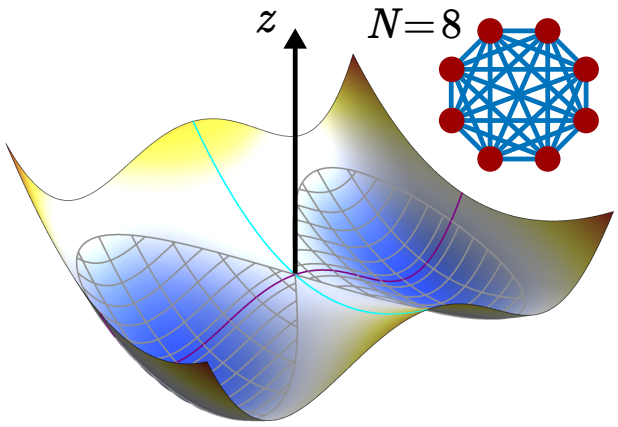

Figure 3.4: The toy model consists of an N-particle molecules and a double-well potential in the xy-plane. An eight-particle molecule (not drawn to scale) depicts the spring connections (blue) between particles (red). For the double-well potential, red (blue) regions correspond to high (low) energies—the cyan and purple lines show, respectively, the parabolic shape in the x-direction and double-well shape the y-direction, their intersection marking a saddle point. Combined with a linear "ramp" potential along the z-direction, two "slides" or "tubes" (gray crosshatching) are separated by central barrier. Motion in this landscape is biased toward the center of either slide (where the energy is low), but transitions between slides are possible at finite temperature. [Reprinted from S. L. Seyler et al. [5], Copyright © (2015), used under the terms of the Creative Commons Attribution License / Fonts adapted from original.]

Trajectories were initialized by placing molecules at the top of the ramp and giving each a slight position bias toward one of the two slides. The starting state for a molecule

was defined as having a center-of-mass located at $z \leq 0\,\text{nm}$ and the final state as having a center of mass position of $z \geq 4\,\text{nm}$. We simulated single particles and eight-particle molecules to get, respectively, $3D$ and $24D$ transition paths to provide a simple way to test how the path metrics behave when the configuration space dimensionality is varied by way of coarse-graining.

A relatively consistent coarse-grained model was designed so as to handle a general N-particle molecule, the requirements of which we describe here briefly.[††]. First, our prescription was simplified under the requirement that all particles within a molecule be initialized at the same point, thus eliminating interparticle forces at 0 K. We then demanded that the zero-temperature dynamics be identical for any N-particle molecule subject to an arbitrary global potential energy function and, at finite temperature, that all molecules have the same effective diffusion coefficient (of their center of mass), assuming a flat potential landscape. Simultaneous satisfaction of both constraints leads to a rescaling of the friction coefficient and the coupling strength to the global potential proportional to $1/N$, which ensured that the mean transition time at any given temperature was invariant under coarse-graining. Interparticle spring coefficients were determined somewhat arbitrarily, the only requirement being that the springs be strong enough relative to the shape and height of the center central barrier to prevent "straddling".

3.3.2 Results and Discussion

Simulations were performed at temperatures from 0 K to 600 K in 50 K increments, resulting in two eight-trajectory ensembles (for both single particles and eight-particle molecules) at each temperature; for each ensemble, four simulations were initialized towards one slide at $(x_0, y_0) = (0\,\text{nm}, 0.4\,\text{nm})$ and four towards the other slide at $(0\,\text{nm}, -0.4\,\text{nm})$. Results are shown for one- and eight-particle molecules at $T = 0\,\text{K}$ and 250 K (Fig. 3.5). The zero-temperature simulations served as a reference point since, in the absence of any interparticle

[††]A complete discussion of the double-slide toy model and coarse-graining scheme are provided in Appendix B.

separation or thermal fluctuations, particles were only subjected to forces from the double-slide potential. The initial conditions guaranteed two distinct groups of zero-temperature transitions such that all four paths in a group were identical.

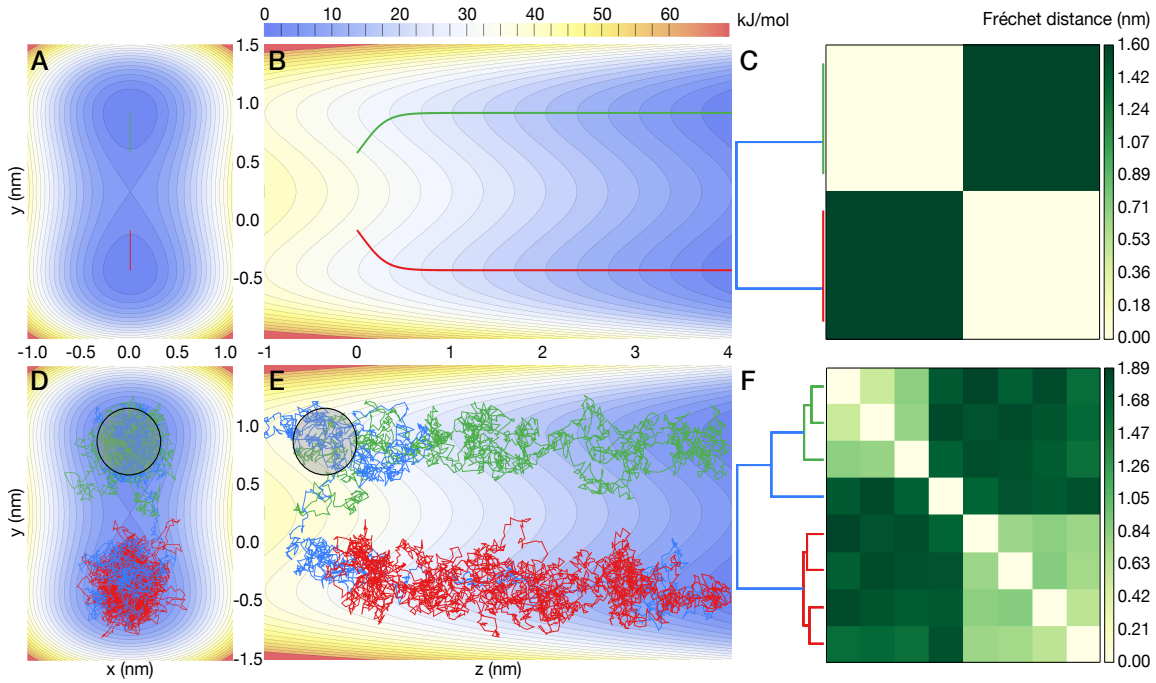

Figure 3.5: Double-barrel potential energy landscape projected onto the xy-plane and yz-plane. Groups of point masses (clusters) mutually connected by harmonic springs move under the influence of a transition-inducing ramp potential in the positive z direction and the two low-energy minima of the "barrels" at $y = \pm 0.8$ nm. Colored lines depict the center of mass trajectories for each cluster. (A–C) trajectories at 0 K. (D–F) trajectories at 250 K. (A, D) Projection of paths onto the xy-plane together with the double-barrel potential. (B, E) Projection of paths onto the yz-plane. (C, E) Clustered heat maps summarize the Fréchet distances for all pairs of trajectories; dendrograms record cluster distances according to the Ward criterion. Trajectory colors in each row match the corresponding path(s) in the dendrogram. The trajectory-averaged radius of gyration for clusters at finite temperature is 0.35 nm (black circles). [Reprinted from S. L. Seyler et al. [5], Copyright © (2015), used under the terms of the Creative Commons Attribution License.]

Fig. 3.5A,B shows the xy and yz projections of eight-particle molecule simulations at 0 K with one group of trajectories (green line) following the gradient of the top "tube" at $y \geq 0$ and the other group (red line) following the bottom tube $y \leq 0$, consistent with expectation. A hierarchical clustering of the trajectories clustered by their Fréchet distances revealed two distinct clusters containing four trajectories each, as shown in Fig. 3.5C. In particular, both the structure of the dendrogram and contrasting colors within the heat map reflected the similarity of the paths within a group as well as the overall difference

between the groups. At 250 K, thermal fluctuations substantially influenced the character of the trajectories (Fig. 3.5D,E). The clusters in Fig. 3.5F were much less clearly defined than in the zero-temperature case. While there still appeared to be roughly four trajectories per cluster (red trajectories in one, green/blue in the other), blue underwent a transition across the central barrier near $z = -0.5\,\text{nm}$ and was as an outlier of the cluster with three green trajectories. Results did not substantially differ between the single-particle and eight-particle cases, and meaningful differences in the center of mass motions were not discernible. As a check, we found that using the center-of-mass trajectories for the 250 K eight-particle trajectories (Fig. 3.5D–F) did not result in a different clustering (data not shown). Overall, thermal fluctuations appeared to be a more dominant effect (Fig. 3.6).

We also repeated PSA using the Hausdorff distance and identified two distinct pathways for temperatures below 300 K, similar to PSA with Fréchet. However, Hausdorff and Fréchet distances became substantially uncorrelated between 350 K and 500 K (Fig. 3.7). Decorrelation was expected to some degree since appreciable backtracking should occur when the typical energy of thermal fluctuations becomes sufficient climb up the slope of the slide potential (in the backward, negative z direction). Furthermore, thermal fluctuations also became comparable to the central barrier height (2 $k_\text{B}T$ at 300 K), and barrier-crossing events become probably. The Fréchet metric was expected to be sensitive to increased backtracking, although backtracking among trajectories within the same slide was not expected to be substantial since the trajectories were confined a narrow region. However, it was somewhat unforeseen that minor backtracking could exacerbate the effect on Fréchet distances in cases where two trajectories are in, or transition to, opposite slides, which was the case for the blue trajectory in Fig. 3.5D–F. Such an effect is at play across intermediate temperatures (i.e., the upper temperature range used here) where transition events become important, and thermal fluctuations and backtracking are not so dominant as to render the downward slope of the slide insignificant. Indeed, we observed that correlations between Fréchet and Hausdorff (Fig. 3.7) returned to near-unity going from 500 K to 600 K, thus it can be stated that thermal fluctuations were significant enough to wash out the central

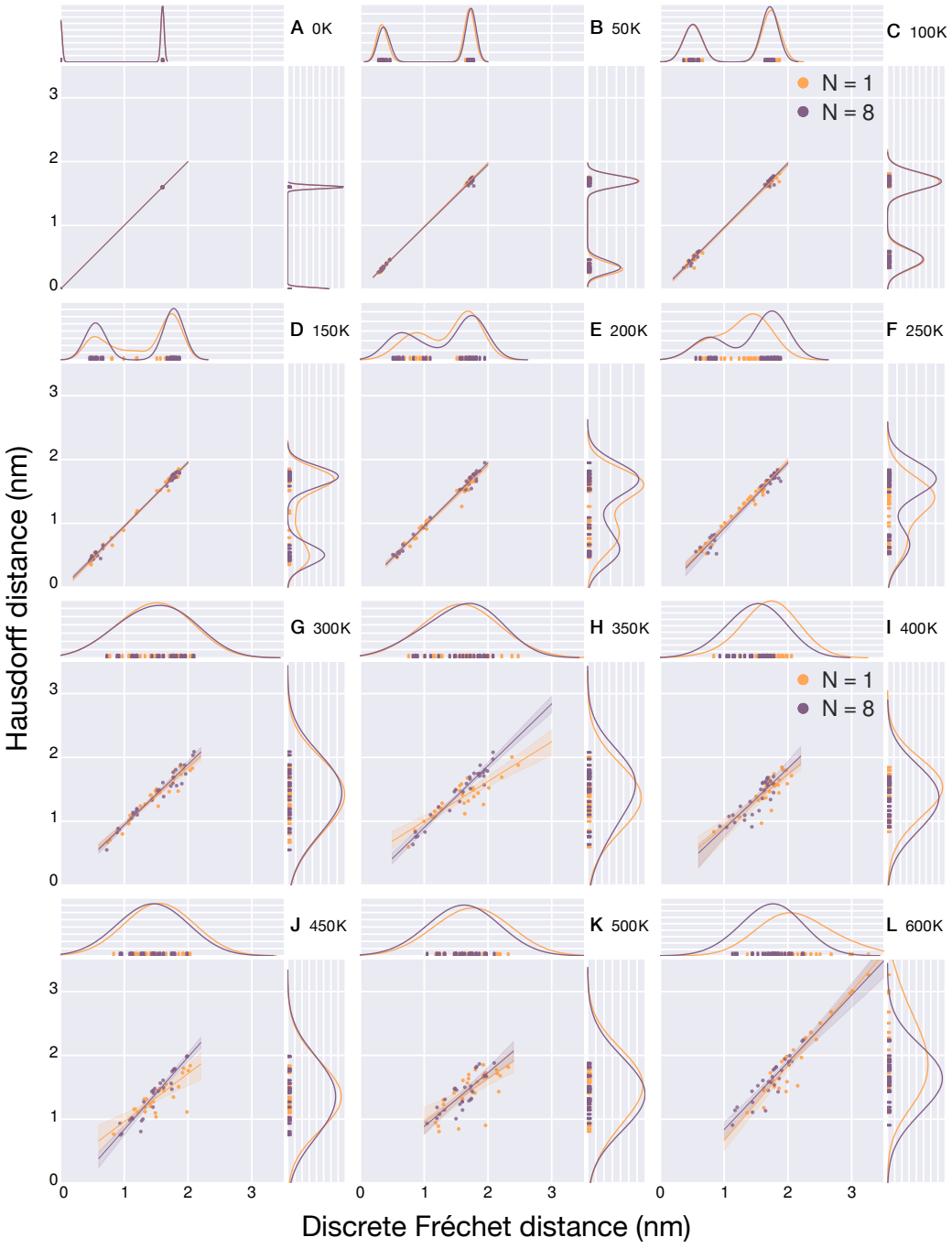

Figure 3.6: Correlation analysis of Fréchet (abscissa) and Hausdorff (ordinate) distances for double-slide model using one- and eight-particle (resp. orange and purple) runs for temperatures between 0 K to 500 K in 50 K increments (panels A–K) and at 600 K (panel L). Scatter points correspond to distance measurements in nm RMSD. Shading about the regression lines corresponds to a 95% confidence interval. Marginal distributions from kernel density estimates (KDE) are shown for each (N, T) pair; bandwidths for each pair are assigned so as to unveil bimodal behavior at low T and unimodality at high T. [Reprinted from S. L. Seyler et al. [5], Copyright © (2015), used under the terms of the Creative Commons Attribution License / Fonts adapted from original.]

barrier but not so large that backtracking on its own could further decorrelate Fréchet and Hausdorff. As temperatures are continuously increased beyond 600 K, Fréchet and Hausdorff should progressively decorrelate as sizable backward steps become more and more probable.

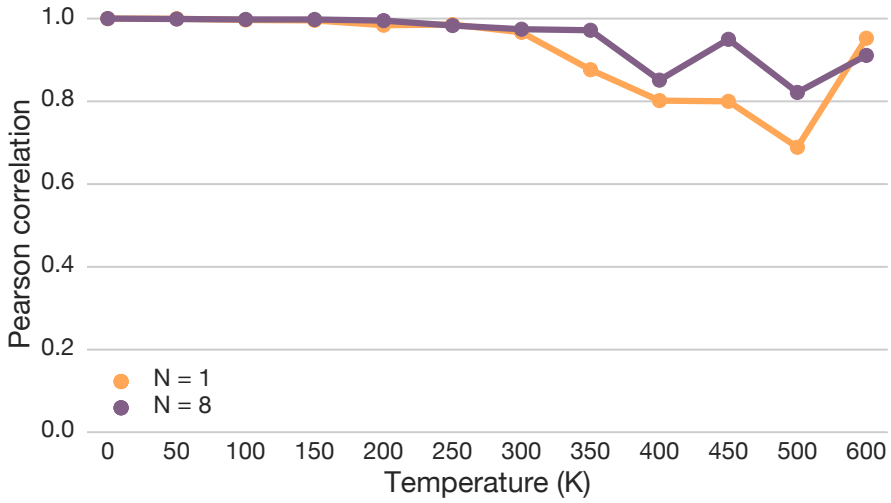

Figure 3.7: Coefficients of the Pearson correlation between Hausdorff and Fréchet distances for one- and eight-particle simulations plotted as a function of temperature. Path distances remain well correlated up to 300 K and are least correlated at 500 K, with the one-particle simulations exhibiting a substantially larger drop in correlation. At the highest temperature the central barrier becomes negligible and the simulations start to equally sample a single tube dominated by the steep repulsive walls. Therefore, paths increase in similarity between $N = 1$ and $N = 8$ particles and the correlation coefficient increases. [Reprinted from S. L. Seyler et al. [5], Copyright © (2015), used under the terms of the Creative Commons Attribution License / Fonts adapted from original.]

Unexpectedly, the aforementioned coordination between backtracking and barrier-crossing events manifested as a dynamical transition as the temperature was increased, where the bimodal distributions of Hausdorff and Fréchet distances (Fig. 3.6) for temperatures below roughly 250 K became increasingly unimodal beyond 300 K. With a central barrier height of $2k_{\mathrm{B}}T$ at $T = 300$ K, it is reasonable to expect thermally-driven transitions between slides to effect such bimodal behavior. Fig. 3.7 shows that Hausdorff and Fréchet distances also began to decorrelate significantly at temperatures between 350 K to 500 K, reaching a minimum at 500 K (Pearson correlation coefficients were ≈ 0.75 for $N = 1$ and ≈ 0.81 for $N = 8$); however, the correlation appeared to rebound strongly (> 0.9) near the 600 K mark. Indeed, as the temperature approached 600 K, simulations increasingly

sampled both slides as though they were a single pathway and the central barrier was effectively washed out by thermal fluctuations. Fig. 3.8 depicts the means and standard deviations of Hausdorff and Fréchet distances as a function of temperature, where a rough cutoff was used to divide the measurements for $T \leq 250\,\mathrm{K}$ to more clearly delineate the bimodal behavior.

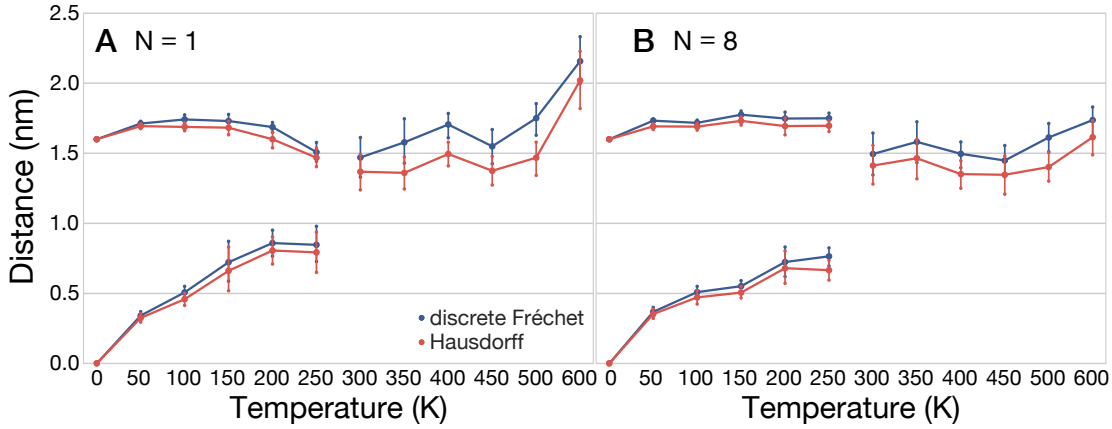

Figure 3.8: Means and standard deviations of the discrete Fréchet (blue) and Hausdorff (red) distances for double-barrel simulations of one particle (A) and eight particles (B) are shown as a function of temperature. Measurements for simulations at 250 K and below are divided into two distributions by dividing distance measurements above and below a 1.25 nm cutoff. Above the cutoff, all measurements are grouped into one distribution. Both the Fréchet and Hausdorff metric lose the ability to distinguish between the two slides as paths begin to wander out of well-defined pathways when the temperature is comparable to the central barrier energy ($2k_\mathrm{B}T$ at 300 K). At higher temperatures, thermal perturbations begin to dominate, allowing molecules to explore the full width of the double-slide and generate a unified pathway. [Reprinted from S. L. Seyler et al. [5], Copyright © (2015), used under the terms of the Creative Commons Attribution License / Fonts and layout adapted from original.]

3.3.3 Conclusions

Our analysis of a simple toy model of transitions demonstrates that PSA can clearly distinguish between two pathways (the upper and lower slides) when the thermal energy is small relative to the energy scale of distinguishing features (e.g., the central barrier) in the potential energy landscape. Thermal fluctuations were a more dominant factor than the dimensionality of configuration space. Results for the Hausdorff and Fréchet metrics were in close agreement overall, with the Fréchet metric appearing to be somewhat more sensitive to reversals of trajectory progress at higher temperatures. The Pearson correlation

between Hausdorff and Fréchet distances as a function of temperature was an indicator of a dynamical shift from a two well-defined pathways to a single pathway spanning the width of both slides. While the Pearson correlation coefficient between Hausdorff and Fréchet distances was nearly unity at the low and high temperature extremes, there was significant decorrelation between 350 K to 500 K with the correlation coefficient reaching a minimum at 500 K for both the single- and eight-particle simulation. The distinguishing feature of intermediate-temperature simulations was the presence of a small number of barrier-crossing events (on the order of 1 over the course of a transition). However, barrier-crossing events within the duration of a single transition are also possible at temperatures below 250 K, given a sufficiently long slide; in the long-slide limit where many crossings take place, it is unlikely that any two paths will take the same meandering trajectory in such a way that their Hausdorff and Fréchet distances will be small. The current prescription of path metrics and hierarchical clustering may be challenging to apply to cases where there is an abundance of complicated, perhaps nonlinear, transition pathways, especially when the intrinsic manifold of the dynamics is relatively high dimensional (e.g., in the extreme case of pure, unbounded diffusion).

It would be worthwhile in the future to explore the effects of more complex potential landscapes and dynamics, which may reduce the ability of PSA to differentiate between paths and functional pathways. Such complications notwithstanding, it should be reiterated that PSA (and the path metrics) can only identify paths as being "similar" insofar as the researcher understands that the path metrics are functions of a point metric—in this case, RMSD—which serves as a putative structural similarity measure. For example, if RMSDs below 0.5 nm for two structures imply such a degree of structural similarity that the structures, for hypothetical purposes, can be considered effectively indistinguishable, then Fréchet or Hausdorff distances below 0.5 nm can be interpreted to mean that the paths in question are also indistinguishable. The identity property (Eq. 3.5b) and triangle inequality (Eq. 3.5d) provide a rigorous sense in which two paths can be said to be indistinguishable when the path metrics return small values, as long as the point metric has a sensible

interpretation. When the distances measured by path metrics, on the other hand, are large, then there is no guarantee that a cluster of paths are truly similar in any meaningful sense—strong statements about similarity can only be made when distances are *small*. One should be especially prudent when the dimensionality of the system in question is large, since large distances become increasingly difficult to interpret (i.e., the curse of dimensionality). However, in situations where Fréchet, Hausdorff, and RMSD prove inadequate for some set of complicated transitions, distance measurements will almost certainly *not* be small.[‡‡]

[‡‡]This statement has clearly not been formally proven and there are known contrived situations where the Hausdorff distance may become arbitrarily small while the Fréchet distance becomes arbitrarily large (cf. Fig. 1 in article by Alt, Knauer, and Wenk [293]); the Fréchet distance can serve as a check for the specific scenario explored by Alt, Knauer, and Wenk [293], and it is probably wise to use both Hausdorff and Fréchet together when possible as a simple control. The author acknowledges that this dissertation has not considered all corner cases (albeit, such systems are unlikely to be amenable to conventional analyses, let alone PSA, in the first place) and, as with any analysis, advises common sense in applying PSA.

Chapter 4

APPLICATIONS OF PATH SIMILARITY ANALYSIS

This chapter is based in part on the published study, **Sean L. Seyler**, Avishek Kumar, Michael F. Thorpe, and Oliver Beckstein (2015). *Path Similarity Analysis: A Method for Quantifying Macromolecular Pathways.* PLoS Comput Biol 11(10): e1004568. [5] My contribution to this work involved the conception, design, and performance of the experiments, analysis of the data, writing of the paper, and design of the PSA software.* The Python module is implemented in the MDAnalysis Python library [272, 273] (see `mdanalysis.analysis.psa`; the data are available via the PSA tutorial on GitHub under the GNU General Public License, v3 DOI 10.5281/zenodo.31457). The studies presented in this chapter build on Chapter 3 and can be broken down into the three following case studies.

In the first case study—Section 4.1—a comparison of closed $\rightarrow$ open transitions of AdK is presented, integrating PSA and two collective variable (CV) approaches to assess a number of paths generated by a variety of fast path-sampling methods. The heatmap-dendrogram approach of PSA is presented first, followed by two analyses based on 2D CV projections—native contacts analysis (NCA) and angle-angle coordinate projection—where transition paths are visually assessed. A final discussion shifts the focus toward the apo transition itself (rather than the methods) and serves to consolidate previous discussion in light of all the data from PSA, NCA, and angle-angle coordinates.

In the second case study—Section 4.2—the analyses focus on two path-sampling methods, namely DIMS-MD and FRODA, which are applied to transition ensembles of closed $\rightarrow$ open transitions for two proteins: the diphtheria toxin (DT) protein, as a more difficult example, as well as apo-AdK. In the ensemble-based study of apo-AdK, the usual heatmap-

*The detailed author contributions are as follows. Conceived and designed the experiments: SLS OB. Performed the experiments: SLS AK. Analyzed the data: SLS AK MFT OB. Contributed reagents/materials/analysis tools: SLS AK. Wrote the paper: SLS AK MFT OB. Designed the software: SLS.

dendrogram approach has been supplemented with a detailed Hausdorff pair analysis that has been successfully used to distinguish DIMS and FRODA pathways by identifying the underlying structural differences between conformers produced by each method.

The final case study revists the path-sampling methods comparison with the inclusion of four long-time equilibrium MD (EqMD) trajectories generated with the Anton supercomputer [6]. Three of these "Anton transitions" have been integrated into a full heatmap-dendrogram analysis with the original methods from the first case study (the fourth way left out so as to keep all methods, three paths apiece, on an equal footing). All four EqMD Anton transitions have been projected into 2D NC and angle-angle space so as to enable a direct visual comparison with the original fast path-sampling approaches. The chapter culminates with a forward-looking assessment of how direct comparisons with EqMD trajectories can be carried out, taking into consideration the various issues that may arise. For instance, the problem of finding a consistent definition of a "transition" is discussed in the context of an equilibrium trajectory; consideration is also given to the overall utility of the path metrics when comparing fast path-sampling methods (characterized by more direct paths between states) and EqMD trajectories (characterized by a broad range of sampling).

4.1 Case study 1: assessing many path-sampling methods

To generate closed $\rightarrow$ open transitions, the initial closed state (chain A of PDB ID 1AKE) and final open state (chain A of PDB ID 4AKE) are used as inputs to the methods. Eight tested methods were available on publicly accessible servers [176, 208, 209, 218, 223, 298, 299]. Local computational resources were used to generate NAMD-based targeted MD (rTMD) trajectories [191], dynamic importance sampling MD (DIMS) trajectories with CHARMM [174], and Framework Rigidity Optimized Dynamics Algorithm (FRODA) trajectories [180]. Table 4.2 classifies each method by its means for generating a path (i.e., dynamics + biasing-to-target, minimum energy path, etc.), while Table 4.1 summarizes the energetic interactions modeled by each method. The path-sampling methods comparison

has also been extended to include a new, direct PSA comparison between the path-sampling methods and three transitions extracted from equilibrium MD simulations generated with the Anton supercomputer [6]. As it would be useful to be able to assess the quality of transitions against a transition reference standard (e.g., leveraging equilibrium sampling from MD) using the tools provided by PSA, the results presented in this thesis are intended to establish the general viability of such an approach and help identify prerequisites for more robust comparisons in the future. The tested methods are described in greater detail in Appendix C, including the input parameters used to generate the closed $\rightarrow$ open apo-AdK transitions in this study.

4.1.1 Overview of tested path-sampling methods

We demonstrate a realistic application of path similarity analysis (PSA) by generating apo-AdK closed $\rightarrow$ open transitions using a variety of path-sampling methods. To retain focus on PSA rather than directly evaluate the performance of the sampling methods, several steps were taken: (1) each method was employed with the highest allowable resolution—C_α representations were used for those that did not have atomistic resolution; (2) three unique paths were generated for each method—stochastic algorithms were simply re-run, while a single parameter (typically ENM cutoff distance) was varied between runs for deterministic methods; (3) default input parameters were used wherever reasonable, unless stated explicitly.

Three repeats were generated for DIMS, FRODA, and MDdMD due to their stochastic nature. rTMD is also stochastic since we employed Langevin thermostatting, however we performed three repeats using a slow pulling speed and three more with fast pulling (six total). GOdMD was run using three different relaxation windows (20 ps, 50 ps, and 100 ps). To obtain three distinct paths for each ENM-based method, we varied the spring cutoff distance, using the default cutoff along with a larger and smaller cutoff. Morph trajectories were produced by modifying the settings for structural pre-alignment and

energy minimization. A single "LinInt" path (i.e., simple linear interpolation between 1AKE and 4AKE) was included for reference.

Since the C_α representation corresponds the lowest model resolution among the tested methods, all analyses were performed using C_α atoms to provide a fair basis for comparison. All generated transition paths were furthermore aligned to the same reference structure, which was produced by aligning and averaging the CORE C_α coordinates of the 1AKE:A and 4AKE:A structures to mitigate biases toward a particular state and ensure the RMSD satisfies the metric requirements in eqs. (3.5a) to (3.5d) (see Appendix D in the Supporting Information for the detailed alignment procedure).

4.1.2 Methods

In all, 12 fast path-sampling methods were tested: 37 transition paths were generated in total, including three transitions each for 11 of the methods, six transitions for rTMD, and a single path representing simple linear interpolation. Paths were first analyzed with the PSA approach introduced in Chapter 3 using the Hausdorff distance (Eq. 3.7) and Fréchet distance (cf. Eiter and Mannila [292]) and several hierarchical clustering linkage algorithms. To help establish the consistency of PSA results, each path was subsequently analyzed using 2D native contacts analysis (NCA) followed by projecting each method onto 2D angle-angle coordinates. Finally, the results from all three methods were integrated into a unified discussion of the apo-AdK closed $\rightarrow$ open transition itself. For further discussion of the parameters used for the path-sampling methods, see Appendix C.

Path Similarity Analysis

Using the Hausdorff and Fréchet metrics, distance matrices were generated from an all-pairs distance comparison and subsequently clustered using Ward's minimum variance criterion to produce a heat map-dendrogram representation (Chapter 3); supplementary results were also obtained for average-type Hausdorff and Fréchet path distance functions, which are not metrics (see Appendix F). All transition paths were initially reduced (if not

already) to a C_α representation after which C_α RMSD alignment to a reference structure (cf. Appendix D) was performed on all conformers in each path. As noted above, the C_α RMSD was used in the Hausdorff and Fréchet metrics to measure the structural similarity between (aligned) conformers. A Pearson correlation analysis was used to compare the Hausdorff and Fréchet results.

Native contacts analysis

NCA was chosen because the approach is both simple and general in that can be applied to any system when one or more native structures are known. A contact was defined as a C_α pair falling with an 8 Å cutoff; using this definition, a *native contact* (NC) was a contact present in a native structure (i.e., either 1AKE, 4AKE, or both). Given an arbitrary conformer, the fraction of native contacts, Q_i, is the fraction of contacts the conformer shares with native structure i [300]. Given our two native states, 1AKE and 4AKE (respectively $i = 1, 2$), we project each conformer (and, therefore, each path) onto 2D Q_1-Q_2 (NC) space to obtain the fraction of native contacts relative to the closed starting state ($Q_{1\text{ake}}$) and to the open target conformation ($Q_{4\text{ake}}$) as collective variables. The dynamic relationship of contact formation and breaking for each method, i.e., the shape of paths in NC space, can be used to assess putative transition states. Specifically, a positive slope (e.g., $Q_{1\text{ake}}$ and $Q_{4\text{ake}}$ are both increasing or both decreasing) implies that contacts are either simultaneously breaking or forming, which can be taken to be indicative of passage through a transition state that is distinct from either end state conformation.

2D angle-angle coordinates

The 2D angle-angle space projection—in (θ_{NMP}, θ_{LID}) coordinates—for the closed $\leftrightarrow$ open AdK transition provides a natural visualization of LID and NMP domain motions and connections to previous studies [40]. In particular, the (θ_{NMP}, θ_{LID}) projection manifests the order in which the LID/NMP domains open and close, potentially revealing important differences in how various path-sampling methods explore different pathways and

intermediate states. The AdK LID-CORE angle, θ_{LID}, is defined as the angle between the C_α atoms in residues 179-185 (CORE), 115-125 (CORE-hinge-LID), and 125-153 (LID), while the NMP-CORE angle, θ_{NMP}, is defined as the angle between the geometric centers of the C_α atoms in residues 115-125 (CORE-LID), 90-100 (CORE), and 35-55 (NMP).[†]

Table 4.1: Energetic models among tested path-sampling methods.

Res[a]	Name	Force field, potential[b]	Mixing potential, other energetics[c]	T_{sim}[d]	Solvent interactions[e]
AA	DIMS[174]	CHARMM27	–	Y	ACS/ACE2 IS
	rTMD[191]	CHARMM27	–	Y	Generalized Born IS
	MDdMD[298]	dMD UA	bond/angle* + vdW/U_E[†]	Y	Lazaridis-Karplus IS
	FRODA[180]	stereochemical	bond/angle[‡] + overlap/H-bond[‡]	–	hydrophobicity[‡]
	Morph[176]	CHARMM	adiabatic mapping	–	–
	LinInt	–	–	–	–
C_α	GOdMD[223]	bond*+Gō-like	–	Y	–
	ANMP[299]	two-well ANM	$U_{\mathrm{mix}} = \min\{U_i, U_f\}$	–	–
	iENM[209]	two-well ANM	$U_{\mathrm{mix}} = F(U_i, U_f)$ (any) + U_{coll}	–	–
	MAP[218]	two-well ANM	min OM action $\rightarrow$ ODEs + BCs	N	OD Langevin
	MENM-SD[208]	two-well ANM	$U_{\mathrm{mix}} = \beta^{-1} \ln\left[e^{-\beta(U_i+\epsilon_i)} + e^{-\beta(U_f+\epsilon_f)}\right]$	N	–
	MENM-SP[208]	two-well ANM	$U_{\mathrm{mix}} = \beta^{-1} \ln\left[e^{-\beta(U_i+\epsilon_i)} + e^{-\beta(U_f+\epsilon_f)}\right]$	N	–

MD-based methods, Morph, and LinInt use atomistic resolution. Except for MAP, ENM-based methods generate two-well (double-well) potentials via mixing function U_{mix} from initial/final state anisotropic network models (ANMs). MAP solves two ODEs (from minimizing the Onsager-Machlup action using ANMs for end states) using position/velocity continuity at the ANM interface; thermal effects from Langevin dynamics, but T_{sim} is not adjustable. MENM-SD/SP assumes weak mixing, $T_m \sim 300\,\mathrm{K}$ ($\beta = 1/k_{\mathrm{B}}T_m$ is adjustable parameter, but not via server); in the no-mixing limit $T_m \rightarrow 0^+$, $U_{\mathrm{mix}} = \min\{U_i, U_f\}$ reduces to the ANMP double-well.

[a] AA, all-atom; C_α, alpha-carbon representation.

[b] dMD UA, discrete MD united atom force field (see [301]); pseudo CHARMM/X-PLOR energy; ANM, anisotropic network model.

[c] vdW, van der Waals potential; U_E, electrostatic potential energy; overlap, avoids steric clashes; H-bond, hydrogen bond; U_{mix}, mixing potential; U_i (U_f), initial (final) state potential; U_{coll}, collision penalty; OM, Onsager-Machlup.

[d] Adjustable simulation temperature? Y, yes; N, no, but thermal effects are present.

[e] IS, implicit solvent; OM, Onsager-Machlup; OD, overdamped Langevin (via OM).

[*] Bonded interactions use an infinite square well potential.

[†] Two-step square wells for attractive interactions; soft barrier for repulsive interactions.

[‡] Constraints on bond distances/angles, H-bonds, and hydrophobic contacts; no solvent model.

[†]Note that [40] used the geometric centers of the backbone and C_β atoms; since some path-sampling methods only use a C_α representation, C_α atoms are used in this thesis to keep an even footing.

Table 4.2: Approaches to transition path generation among tested path-sampling methods.

Type	Name	Dynamics	Path propagation & biasing[a]	Rev[b]	TS/Stoch[c]	Progress variable
MD with biasing	DIMS[174]	Langevin NVT	SR	N	Y/Y	RMSD-to-target
	rTMD[191]	Langevin NVT	moving harmonic restr.	N	Y/Y	RMSD-to-target
	MDdMD[298]	discrete MD	SR + essential dynamics	N	Y/Y	SSD-to-target[†]
	GOdMD[223]	discrete CG-MD	SR + metadynamics	N	Y/Y	SSD-to-target[†]
geometric targeting	FRODA[180]	–	stepwise-enforced RMSD constraint*	N	Y/(Y/N)	RMSD-to-target
CG-ENM	ANMP[299]	–	SD via SP cusp to minima	Y	N/N	–
	iENM[209]	–	parametric SP/fixed-pt eq.	Y	N/N	–
	MAP[218]	–	OM minimum action path	Y	N/N	–
	MENM-SD[208]	–	SD from SP to minima	Y	N/N	–
	MENM-SP[208]	–	parametric SP/fixed-pt eq.	Y	N/N	–
linear interp.	Morph[176]	–	adiabatic mapping	Y	N/N	–
	LinInt	–	linearly interpolated snapshots	Y	N/N	–

DIMS, rTMD, MDdMD, and GOdMD are all non-deterministic MD-based methods. DIMS and rTMD employ a conventional force field and Langevin dynamics in the canonical ensemble; the discrete MD algorithms used by MDdMD and GOdMD assume ballistic particle motion until a collision occurs—along with the depth of the interatomic square wells, momentum and energy conservation are used to determine outgoing momenta without explicitly computing forces. FRODA uses a non-physical dynamical algorithm to path-search stereochemically correct regions of configuration space. CG-ENM methods generate transitions by constructing low-energy paths in the potential energy landscape. Morph and LinInt linearly interpolate the position of each atom between the initial and final states.

[a] SR, soft ratcheting; SD, steepest descent; SP, saddle point; OM, Onsager-Machlup.

[b] Is the method exactly reversible?

[c] Is the algorithm based on a (physical or non-physical) time step? Is it stochastic?

[*] At each step, RMSD reduced by fixed amount while simultaneously enforcing other constraints.

[†] SSD, sum of squared distances to target (includes weighting that varies between MDdMD and GOdMD.

4.1.3 Results and Discussion

Direct comparison using PSA and clustering

Fig. 4.1 shows the heat map-dendrogram generated by a hierarchical clustering of Fréchet distances using the Ward linkage algorithm. Nearly all methods formed their own clusters, implying that two transitions produced by a single method tend to have greater similarity than two transitions produced by two different methods. In fact, the only exceptions are Morph (3)—which shares a cluster with LinInt and the other Morph paths, all of which are generated by linear interpolation—and the outlier cluster formed by the GOdMD paths—which are substantially different from all other methods ($\delta_F > 3\,\text{Å}$). Given that the Morph and LinInt paths are nearly identical ($\delta_F \leq 0.5\,\text{Å}$), the additional features

implemented in Morph, such as checking for steric overlaps, may not be relevant for the closed → open AdK transition.

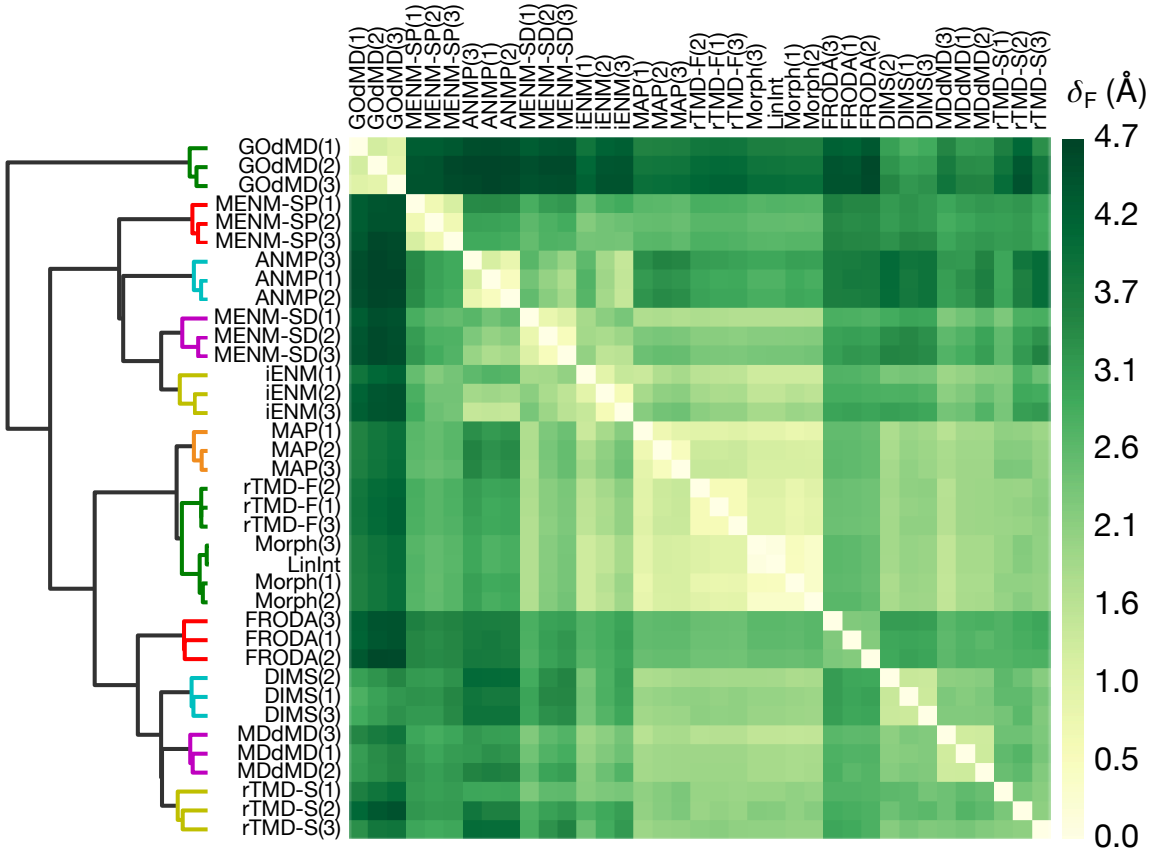

Figure 4.1: Path similarity analysis of apo-AdK closed → open transitions from various path-sampling methods. Each method was used to generate three paths, except for rTMD where six were generated (rTMD-F/S) and a single LinInt path. Fréchet distances, δ_F, are in Å and correspond to a structural C_α RMSD in accordance with the RMSD point metric. Smaller distances (lighter colors) correspond to greater similarity. The dendrogram depicts a hierarchy of clusters where smaller node heights of parent clusters indicate greater similarity between child clusters. See text for a description of the methods. S5FIG contains the same data annotated with numerical values of δ_F. [Reprinted from S. L. Seyler et al. [5], Copyright © (2015), used under the terms of the Creative Commons Attribution License.]

The dendrogram also reveals that methods with commonalities between their physical models tended to cluster with one another. The two MD-based importance sampling methods, DIMS and MDdMD, formed a cluster with the Langevin MD-based rTMD at slow pulling velocity ("rTMD-S"; Fréchet distance 2.1 Å $\leq \delta_F \leq$ 2.7 Å). Correspondingly, the ENM-based methods (iENM, MENM-SD/SP, and ANMP) another cluster, where MENM-SD and iENM were the most similar ($\delta_F \leq$ 2.3 Å) while MENM-SP and ANMP were the

most different ($\delta_F \leq 3.4$ Å). In the case of FRODA, there is no potential energy function—only stereochemical constraints are enforced—so its trajectories are pseudo-dynamical. The absence of full energetics notwithstanding, FRODA forms a cluster of dynamical methods with DIMS, MDdMD, and rTMD-S (2.6 Å $\leq \delta_F \leq$ 3.1 Å).

Remarkably, fast-pulling rTMD ("rTMD-F") and MAP were very similar to both Morph and LinInt paths ($\delta_F \approx 1$ Å) despite having fundamentally different physical models: whereas rTMD-F uses MD and a physics-based atomistic force field, MAP uses a double-well ENM for the energetics and generates paths by minimizing the Onsager-Machlup action (without any linear interpolation). The similarity between rTMD-F and linear interpolation is sensible, since, given that trajectories were driven by rapidly pulling a harmonic RMSD restraint toward the target (i.e., heavy-atom RMSD to the target was the biasing order parameter to match DIMS), rTMD-F simulations have little time to sample orthogonal directions. It is not obvious that MAP should be in a different cluster than the other ENM-based methods. However, given that the action-minimized paths generated by MAP are minimum *free* energy paths (the other ENM-based methods produce, in effect, minimum energy paths), it is not unreasonable that MAP may generate paths bearing greater resemblance to those generated the dynamical methods. Indeed, the MAP/rTMD-F/Morph sub-cluster and the sub-cluster of dynamical algorithms (DIMS, MDdMD, rTMD-S, and FRODA) form a separate cluster from the ENM-based methods. At first glance, the MAP, rTMD-F, and Morph paths somewhat resemble iENM and MENM-SD paths ($\delta_F \leq 2.5$ Å). However, closer examination of the heat map—in particular, the patterns of Fréchet distances evident in the heat map's striping—reveals that MAP/rTMD-F/Morph paths and DIMS/MDdMD/rTMD-S/FRODA paths bear qualitative resemblances in how they all compare to iENM/MENM-SD. In other words, the qualitatively similar way in which these methods all differ from iENM/MENM-SD paths is such that the "Morph-like" and dynamical methods form a distinct cluster apart from the MEP-like ENM methods.

To check the effect of the linkage algorithm on the hierarchical clustering results,

we tested four other linkage algorithms (single[‡], complete, average, and weighted—see Fig. 4.2A–D, respectively). The cluster composed of MAP/Morph/LinInt/rTMD-F appears particularly robust since it is reproduced by Ward, complete, average, and weighted linkages. The DIMS/MDdMD/rTMD-S cluster generated by Ward is also produced by complete and average linkage, and all three methods are contained in a larger cluster along with the "Morph-like cluster" for weighted linkage. Interestingly, complete linkage assigns FRODA to the dynamical methods cluster (DIMS/MDdMD/rTMD-S) like Ward, whereas average linkage places FRODA in a unique cluster with MENM-SP and weighted places both FRODA and MENM-SP outside of the cluster formed by nearly all other methods (except ANMP and GOdMD). That FRODA should be grouped with DIMS/MDdMD/rTMD-S thus seems less robust than, say, the cluster formed by DIMS with MDdMD, which is produced by the Ward, complete, and average linkages. All linkage algorithms agree that GOdMD paths are outliers, though it is curious that average and complete linkage clusterings suggest ANMP is also somewhat of an outlier while ANMP is placed with in the "ENM cluster" when Ward or complete linkage is used. Results were consistent among the linkage algorithms, with the Ward and complete linkages appearing to be relatively robust. (See Appendix E and Husic and Pande [302] for more in-depth analyses of clustering algorithms).

The overall picture did not substantially change when using Ward linkage with the Hausdorff distance (Fig. 4.3A), which reproduced the same clusters as with Fréchet; the Pearson correlation coefficient between δ_H and δ_F was essentially unity (Fig. 4.3B). For completeness, several alternative path distance functions were examined, namely the average-type Fréchet and Hausdorff distances (which are, however, not proper metrics, as proved in Appendix F). Both average-type path metrics reduced the amount of detail in the clustering without substantially changing the overall picture, ultimately amalgamating the clusters into one large "dynamical methods cluster" (TMD-S, DIMS, MDdMD, GOdMD,

[‡]The single linkage is known to be susceptible to measurement uncertainty and "chaining", though it is included for completeness.

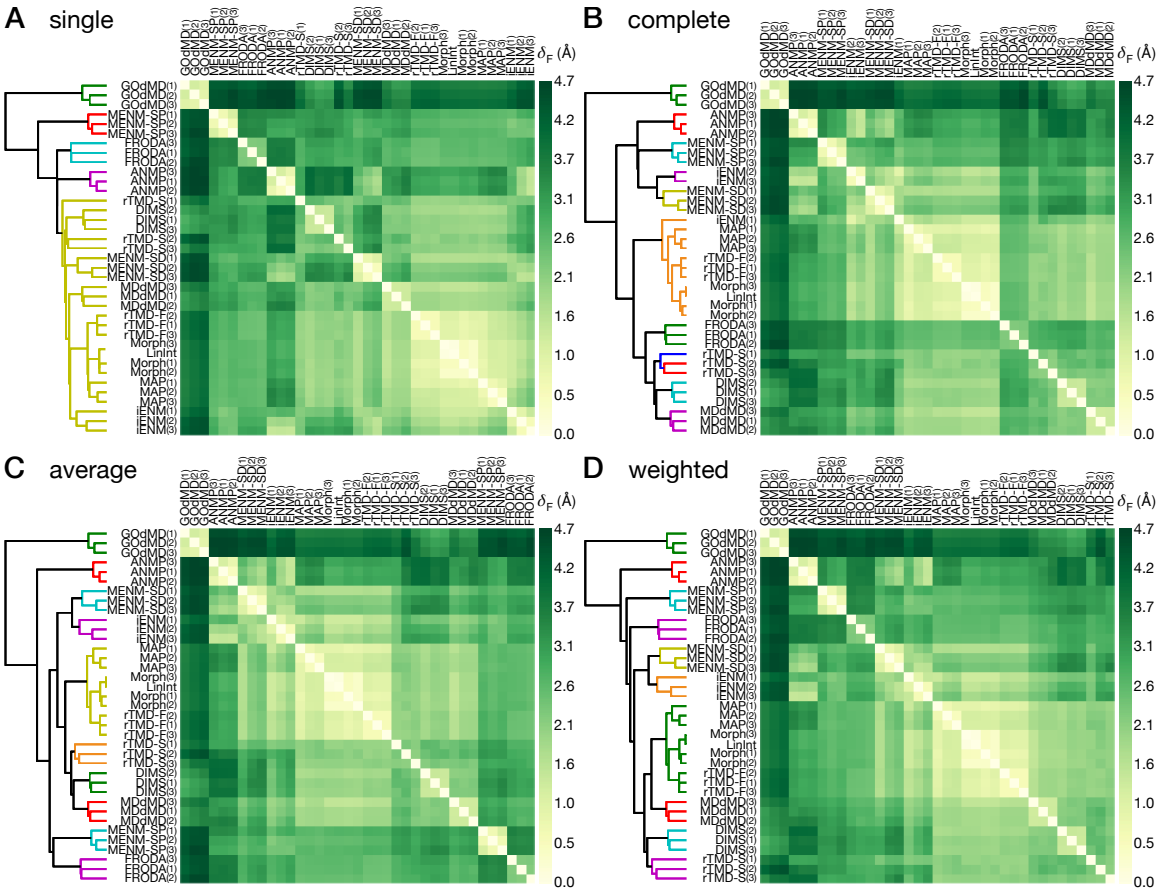

Figure 4.2: Path similarity analysis of apo-AdK closed → open transitions from various path-sampling methods using alternative linkage algorithms for hierarchical clustering. Dendrograms for each heat map correspond to the hierarchical clustering produced by the single (A), complete (B), average (C), and weighted (D) linkage algorithms, depicting a hierarchy of clusters where smaller node heights of parent clusters indicate greater similarity between child clusters. [Reprinted from S. L. Seyler et al. [5], Copyright © (2015), used under the terms of the Creative Commons Attribution License.]

FRODA), a "Morph-like cluster" (Morph, LinInt, TMD-S, MAP), and an "ENM cluster" (ANMP, iENM, MENM-SP/SD). In the sections that follow, the robustness of the PSA results is assessed in the context of NCA and angle-angle projections.

Native contacts analysis

Native contact fractions relative to the closed starting state (Q_{1ake}) and to the open target conformation (Q_{4ake}) were computed for each closed → open transition path. Fig. 4.4A shows that most trajectories began on or near the right vertical axis (green circle), corresponding to the first conformers of the paths having essentially 100% of their contacts in

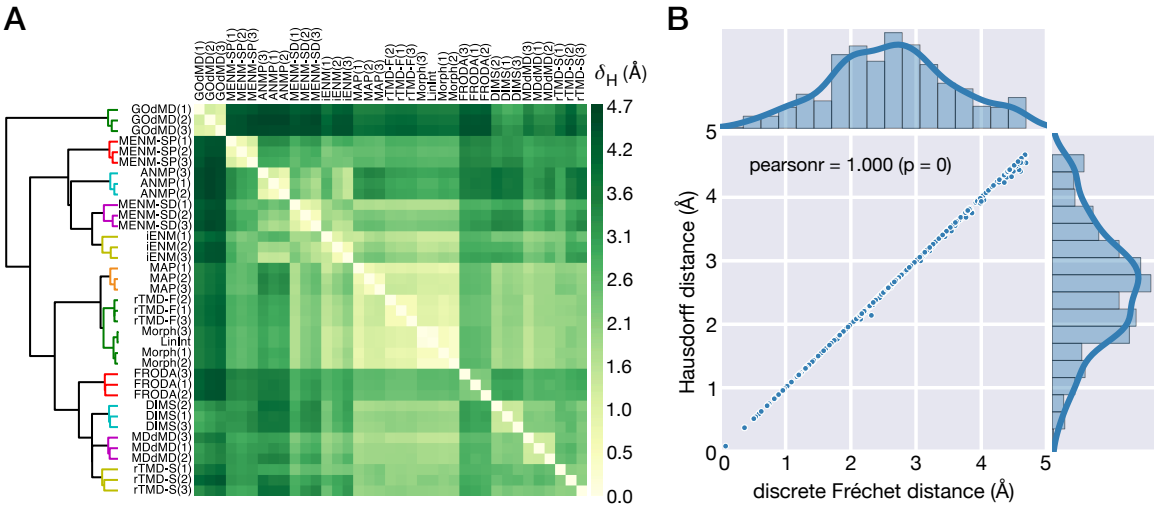

Figure 4.3: (A) Heat map for path-sampling methods for the apo-AdK closed → open transition of Hausdorff distances produced using the Ward algorithm. Clusters are identical to the Ward clustering for Fréchet distances in Fig. 4.1. (B) Correlation and joint distributions between (discrete) Fréchet versus Hausdorff distance measurements (in Å rmsd) for the AdK closed → open methods comparison. Strong linear correlation indicated by the scatter plot, with a Pearson correlation coefficient very close to unity, indicates that either metric could have been used to perform the path-sampling methods analysis with essentially identical results. A slight deviation of the scatter points below the line of unity slope is consistent with the fact that Fréchet distances are bounded from below by corresponding Hausdorff distances. [Adapted from S. L. Seyler et al. [5], Copyright © (2015), used under the terms of the Creative Commons Attribution License.]

common with the native contacts of the initial 1ake closed state ($Q_{1ake} = 1.00$) and around 95% of the final 4ake open state contacts ($Q_{4ake} = 0.95$).[§] Correspondingly, most trajectories terminated near the top horizontal axis (red diamond), the final conformers containing nearly 100% of the 4ake native contacts ($Q_{4ake} = 1.00$) and around 93% of 1ake contacts ($Q_{1ake} = 0.93$). Both DIMS and MDdMD paths did not finish precisely at the final state since both use a soft-ratcheting biasing algorithm; since precise convergence to the target is difficult, simulations are terminated once a conformer falls within a cutoff radius of the final 4ake final state (0.5 Å heavy atom [non-hydrogen] RMSD for DIMS; 1.5 Å C_α RMSD for MDdMD). DIMS and MDdMD trajectories broke a comparable number of contacts relative to both native states (around 8-9% and 9-10%, respectively), though MDdMD failed to reform any 4ake contacts ($Q_{4ake} < 0.94$) while DIMS reformed most ($Q_{4ake} \leq 0.98$).

[§]Close examination of DIMS trajectories would show that they start with 96% of 1ake contacts, which is to be expected given that the 1ake structure must be energy-minimized and equilibrated prior to performing DIMS-MD.

Paths generated by the dynamical methods (DIMS, rTMD, FRODA, MDdMD, and GOdMD) were relatively noisy compared to the non-dynamical methods; fluctuations were primarily along lines of positive slope and orthogonal to progress in the direction of the final state. The NC space paths from FRODA exhibited the greatest degree of contact breaking ($Q_{1\text{ake}} = 0.82$, $Q_{4\text{ake}} = 0.80$) of all methods tested, though this behavior is not unexpected given the nature of FRODA's stochastic algorithm.[¶]

Overall, NCA qualitatively recapitulates the general dichotomy between the dynamical and non-dynamical methods seen in PSA. NC paths that took more direct paths toward the final state (negative slopes) were typically from non-dynamical methods; paths that moved away from *both* native states before heading to the final state (positive slopes) were mostly from dynamical methods. With the exception of GOdMD (which more closely resembles Morph), NC paths from dynamical methods also exhibited a "V"-shaped structure indicative of a transition state region. In contrast, non-dynamical methods tended to show the opposite pattern, as ANMP, iENM, Morph, and LinInt paths wandered relatively little on the way to the final state.

The strong qualitative similarities among DIMS, MDdMD, and rTMD-S in NC space is consistent with robust cluster they form in PSA; moreover, that FRODA formed a looser cluster with DIMS/MDdMD/rTMD-S (among the various linkage algorithms) is unsurprising given the degree to which it broke contacts. The V-shape of rTMD-F paths was less pronounced than rTMD-S paths, indicating that faster rTMD pulling velocity directly limited the extent of contact breaking and formation. PSA accordingly clustered rTMD-S among the dynamical methods and rTMD-F in the robust rTMD-F/MAP/Morph/LinInt cluster; in the latter case, though neither MAP nor rTMD-F resembled Morph/LinInt in NC space, both were not entirely different from one another since both break excess 1ake native

[¶]Stochasticity in FRODA is achieved at each step by a random displacement and rotation of one rigid unit of the molecule (rigid units for a protein are constructed at sub-amino acid level), after which stereochemical constraints are enforced. Limited only by geometric constraints, rigid units containing C_α atoms can therefore move in a manner that may otherwise be prohibited by a realistic force field, leading to potentially larger fluctuations in the C_α distances, which in turn is likely to magnify contact breaking and (re-)formation behavior.

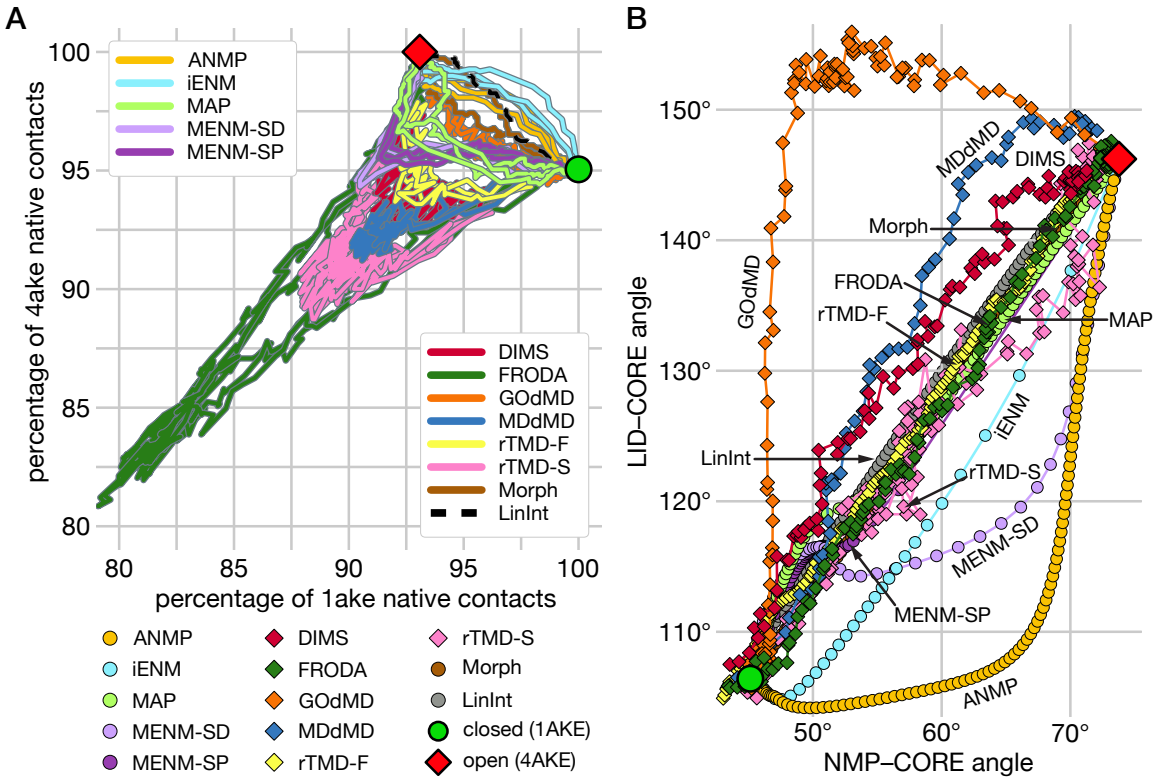

Figure 4.4: Projections of the AdK closed → open transitions from each path-sampling method onto low-dimensional collective variables. The initial 1AKE:A (final 4AKE:A) structure is shown in each plot by the green circle (red diamond). (A) Projection of all paths from the path-sampling methods onto NC space. The abscissa (ordinate) corresponds to the percentage of contacts shared with the initial (final) state. The top-left legend identifies EN-based methods and the remaining methods are listed on the bottom-right. LinInt is shown a broken black curve. (B) Projection of run #2 of each method onto NMP angle (θ_{NMP}) *vs* LID angle (θ_{LID}). In B and C, trajectories generated by the dynamical methods (DIMS, rTMD, FRODA, MDdMD, GOdMD) are plotted with diamonds and non-dynamical method trajectories with circles. [Adapted from S. L. Seyler et al. [5], Copyright © (2015), used under the terms of the Creative Commons Attribution License / Adapted from original.]

contacts ($Q_{\text{1ake}} \approx 0.92$) that reform toward the final state ($Q_{\text{1ake}} \approx 0.93$). Indeed, MAP and also MENM-SD/SP were clear exceptions to the non-dynamical methods by way of contact breaking/formation. What is particularly interesting is that MENM-SD/SP paths not only broke the most excess 1ake contacts among the non-dynamical paths ($Q_{\text{1ake}} \approx 0.90$–$0.91$), but also exhibited the characteristic "V" shape produced by dynamical methods.

Angle-angle space projection

Projecting representative trajectories from the path-sampling methods onto angle-angle (AA) space (Fig. 4.4B) immediately reveals considerable variation in the apparent order in

which the mobile domains open (or close). A majority of the transitions lie in the vicinity of the linear-interpolation path (LinInt), which is a reflection of the fact that apo AdK has little in the way of structural obstructions that would otherwise prevent a direct path in going from its closed state (green circle) to its open state (red diamond). There are also a number of transitions that clearly deviate from the direct path. As with PSA and NCA, the projection onto AA space (Fig. 4.4B) reproduces the overall dichotomy between dynamical and non-dynamical methods. ENM-based methods appeared to prefer a NMP-opening pathway (i.e., NMP-opening takes place prior to LID-opening) on the whole, though Morph, MAP, and MENM-SP tracked LinInt closely. In contrast, dynamical methods generally sampled a LID-opening pathway, including DIMS, MDdMD, and GOdMD, though the FRODA path sampled the vicinity of LinInt.

DIMS and MDdMD appeared qualitatively very similar in AA space, in full agreement with both PSA and NCA. The exaggerated LID-opening of the GOdMD path was unlike any of the other methods and its classification as an outlier by PSA is sensible (Fig. 4.1). In the case of rTMD-S, its similarity to DIMS and MDdMD in PSA seems reasonable since it remained in the vicinity of the LinInt path despite being fairly erratic. When rTMD is used with a very high pulling speed, the system moves almost exclusively in the direction of the restraint potential which in turn follows the order parameter (used for targeting); with an RMSD-to-target restraint, the restraint potential exactly follows the LinInt path. It is thus clear that rTMD-F should behave more like Morph/LinInt and less like equilibrium MD with an perturbing potential, which is in turn reinforced by the agreement between PSA and projection in AA space. FRODA also predominantly followed LinInt, though its overall behavior was somewhat erratic due to the stochastic nature of the algorithm.

The ANMP path displayed unambiguous NMP-opening behavior with the NMP domain opening most of the way before much LID-opening took place; the iENM path also showed initial NMP-opening, but was shortly followed by a simultaneous opening of both the LID and NMP domains. Given that the ANMP path exhibited dramatic NMP-opening relative to any other method, it is easy to see why ANMP was somewhat of an outlier in PSA

(average and weighted linkages). MENM-SP was the most distant member in the cluster of ENM methods in PSA (Fig. 4.1), in addition to being grouped with FRODA (average linkage) and being an outlier (weighted linkage). Careful inspection of the MENM-SP path revealed a sizable gap between the penultimate conformer (located in the first half of the transition) and the 4ake final state. Though two MENM-SP paths with well-aligned gaps may not produce a discernible effect on their Hausdorff or Fréchet distance, it is clear that an MENM-SP path will tend to have larger distances to another paths from a different method—there will likely be at least one conformer in the other path (in the part of the transition where the MENM-SP gap occurs) whose nearest neighbor is either the final or penultimate conformer in the MENM-SP path but is relatively far from both.

Further observations

It is compelling that the MAP and MENM-SP/SD paths initially exhibit slight LID-opening until around 116°, but MENM-SD, in contrast with MENM-SP and MAP, exhibited a partial reversal in the LID-opening pathway whereby the NMP domain was nearly completely opened. The presence of this "wiggle" in MAP and MENM-SD/SP but not others in AA space is compatible with the contact breaking behavior seen in MAP/MENM-SD/MENM-SP but not in the other non-dynamical methods in NCA. Such qualitative resemblances notwithstanding, MENM-SD consistently clustered with iENM while MENM-SP forms somewhat inconsistent clusters across all linkage algorithms. In the case of ANMP, its paths appeared quite similar to iENM and Morph in NC space; in PSA, ANMP and iENM were somewhat similar ($1.4\,\text{Å} \leq \delta_F \leq 2.7\,\text{Å}$ in Fig. 4.1)—though ANMP was sometimes an outlier—but ANMP and Morph were very different ($2.8\,\text{Å} \leq \delta_F \leq 3.1\,\text{Å}$). Furthermore, NCA did not appear to recapitulate the close correspondence between MAP, rTMD-F, and Morph paths. On the other hand, the ANMP paths, which were reasonably similar to iENM in PSA ($1.4\,\text{Å} \leq \delta_F \leq 2.7\,\text{Å}$ in Fig. 4.1) but fairly different from Morph ($2.8\,\text{Å} \leq \delta_F \leq 3.1\,\text{Å}$), appeared fairly similar to both iENM and Morph in NC space.

It is difficult at this stage to pinpoint the exact reasons for apparent discrepancies

between PSA and NCA results. The division between dynamical and non-dynamical methods is consistent with PSA, including some subdivisions such as the inconsistent grouping of FRODA in comparison with the otherwise close-knit DIMS/MDdMD/rTMD-S cluster. rTMD-F and MAP generated NC paths that occupied an intermediate region between most paths and also displayed qualitative similarities to both dynamical and non-dynamical paths, suggesting a partial explanation as to why MAP and rTMD-F consistently clustered with Morph/LinInt. However, the clear "V" shape of MENM-SP/SD paths cannot be explained directly from the clusterings in PSA. NCA did not offer clear hints as to why PSA subdivided ANMP, iENM, and MENM-SD/SP is it did, nor did it reveal why GOdMD always appeared as an outlier.

Whether the "wiggles" in AA space coincide with each path's cusp-like region in NCA space, and whether these features predict a physical transition state, cannot be inferred from the data. If the gap in the MENM-SP paths were filled, PSA might cluster with the larger group formed by MAP and DIMS/MDdMD/TMD-S; the MENM-SD path is distinct and somewhat resembles both iENM and ANMP, so it is reasonable that MENM-SP and iENM form a robust cluster. The heatmap-dendrogram analysis alone did not furnish this somewhat nuanced picture derived from considering NCA and AA projections. We anticipate that a Hausdorff pair analysis (cf. Section 3.2.2 and Section 4.2) could provide the necessary structural insight to help answer these questions.

4.1.4 Summary of path-sampling methods study

PSA was shown to be an effective means of quantitatively comparing conformational transition paths generated by algorithmically distinct path-sampling methods. For this investigation, we used AdK—reviewed in Chapter 1 (experimental studies) and Chapter 2 (computational studies)—as a testbed protein to generate closed $\rightarrow$ open transitions from the 1ake closed state to the 4ake open state. Using PSA, we were able to distinguish broad features among the transition paths, linking our observations to their originating physical models and algorithms. Paths generated by a given method were more similar to one

another than two paths from different methods, and there was a consistent distinction between dynamical and non-dynamical algorithms and the apparent domain-opening pathways (generally LID-opening and NMP-opening, respectively). This analysis required only C_α trajectories as input along with the familiar C_α RMSD as a structural similarity measure. It should be made clear that the comparison presented in this section was not intended to represent a comprehensive assessment of all path-sampling methods that have been applied to the apo-AdK closed $\leftrightarrow$ open transition.

One important avenue of exploration would be to examine more sophisticated approaches that do not require a predefined order parameter or progress variable as input, like transition path sampling (TPS) [57], those that find a MFEP by refining an initial path, such as the finite-temperature string method [303], or an interface-based method like weighted ensemble sampling [82, 255]. Alternatively would be to use metadynamics-based [91] approaches (made possible through the PLUMED plugin [304, 305]) in conjunction with MD-based biasing methods (e.g., DIMS, TMD, etc.) to assess the degree to which pathways and energetics are affected by differing choices of CVs (e.g., 1D RMSD-to-target, 2D angle-angle variables, or 2D RMSD-to-end-states; cf. Table 1 in S. L. Seyler and Beckstein [4]). It would also be useful to design a comparison whereby different aspects of physical models—electrostatic forces, solvent interactions, etc.—can be toggled so as to systematically examine how each type of interaction contributes to the pathway geometry.

4.2 Case study 2: transition ensemble analyses

We first apply PSA to the DT transition ensemble using the same general approach as in the path-sampling methods comparison in Section 4.1. For clarity, only the PSA results for the Fréchet distance are presented here as the Hausdorff distance results lead to identical conclusions. In the second part of our ensemble study, we revisit AdK and repeat the usual PSA approach for the closed $\rightarrow$ open transition ensemble. Then, we extend this analysis by generating Hausdorff pairs for the entire transition ensemble, extracting representative Hausdorff pairs, and examining their projections in angle-angle space. To help make the

analysis more transparent, we conceptually divided the full set of Hausdorff distance measurements into: (1) distances between pairs of DIMS paths, (2) distances between pairs of FRODA paths, and (3) inter-method distances measured between a DIMS and a FRODA path. Since the full ensemble consists of $N = 400$ paths, we generated $N(N-1)/2 = 79800$ δ_H-pairs in aggregate. A total of six representative δ_H-pairs were then selected by extracting the δ_H-pairs associated with the median and maximum Hausdorff distances for each of the segregated comparisons (1), (2), and (3) above.

4.2.1 Methods

The overall flow of the analyses goes as follows: first, the standard heatmap-dendrogram protocol is straightforwardly applied to the diphtheria toxin protein to set the stage. Then, we illustrate a thorough investigation of DIMS and FRODA closed $\rightarrow$ open transitions of apo-AdK. Finally, a realistic Hausdorff pair (δ_H-pair) analysis is presented for the apo-AdK closed $\rightarrow$ open transitions; in particular, due the availability of relatively good CVs (i.e., LID-CORE, NMP-CORE angle-angle coordinates) for the AdK transition, we show the δ_H-pairs projected in AA space to provide an intuitive visual reference.

Hausdorff pair analysis

Though PSA is effective for analyzing arbitrary systems using (up to) the full $3N$-dimensional configuration space, the global heatmap-dendrogram analysis presented thus far does not provide any direct connection to the molecular-structural details. To extract physically relevant differences at the molecular level, a Hausdorff pairs (or Fréchet pairs) analysis can be used to single out a pair of conformers that are relatively likely to harbor structural features giving rise to physically relevant differences between the paths in question. We only performed a Hausdorff, rather than Fréchet, pairs analysis since the Hausdorff and Fréchet results were found to be strongly correlated in the AdK ensemble comparison.

The concept of a Hausdorff pair (δ_H-pair) follows from the description of the Hausdorff

distance in Section 3.2.2 and is recapitulated briefly here. Given two paths, the Hausdorff distance represents the maximum nearest neighbor distance between a conformer in one path and a conformer in the other; we call this pair of conformers a Hausdorff pair. A Hausdorff pair should, in principle, reveal salient structural differences giving rise to their mutual dissimilarity more frequently than an pair of conformers chosen at random. Hausdorff pairs can be easily extracted from an all-pairs calculation of Hausdorff distances in a transition path ensemble using the MDAnalysis Python library [272, 273], providing an efficient means within the PSA framework to scrutinize structural patterns—down to the atomistic level—across many sampled paths and, thus, identify the molecular-structural determinants of differences between full transition pathways.

AdK and DT closed → open ensembles from DIMS and FRODA

To demonstrate a different mode of analysis within the PSA framework, we concentrated on two distinct, stochastic methods—DIMS-MD and FRODA—to generate ensembles of transitions containing hundreds of trajectories. For this exercise, the closed → open transition of apo-AdK was supplemented by a second example, namely a closed → open transition of the diphtheria toxin (DT) protein, which exists in inactive (closed) and active (open) conformations, from the bacterium *Corynebacterium diphtheriae*. It is thought that DT—a potent exotoxin that disrupts protein synthesis via inactivation of the eukaryotic elongation factor-2 (EEF-2) protein—leverages such a conformational change to gain access to susceptible mammalian cells through receptor-mediated endocytosis [306, 307]. There are three domains, depicted in Fig. 4.5, in a DT monomer: the N-terminal, catalytic (C) domain (resids 1–190), which disables EEF-2 by catalyzing ADP-ribosylation (of EEF-2); and the transmembrane (translocation or T) domain (resids 191–378), which assists the C domain in traversing the cytoplasmic membrane of a host cell; and the C-terminal, receptor-binding (R) domain (resids 379–535), which initiates endocytosis by binding to specific cell surface receptors [307]. After binding, a conformational change maneuvers the C- and T-domains (toward the bilayer, away from the R-domain) into position for membrane insertion and

translocation; visual inspection of the R-domain, which differs between closed and open states by both an extension and $\sim 180^\circ$ rotation relative to the C- and T-domains (Fig. 4.5), hints at the motions involved in the transition.

A closed DT conformation was solved in a monomeric form by Bennett and Eisenberg [308], though the open conformation was found in a domain-swapped dimeric structure [309] where the R-domain of each monomer has an interdomain interface with the other monomer. It was proposed by Carroll, Barbieri, and Collier [310] that monomeric DT converts to an open conformation upon a decrease in pH, which in turn leads to dimerization via monomer-monomer hydrophobic interactions enabled by domain swapping.

Previously, the open $\rightarrow$ closed transition was generated computationally by Krebs and Gerstein [176] using morphing (adiabatic mapping) and declared "a hard or impossible case" due to the unphysical rearrangement of atoms along the linearly interpolated path; it was speculated that unfolding and refolding (of the R-domain) could yield a physical pathway, though Farrell, Speranskiy, and Thorpe [180] demonstrated that FRODA could generate a plausible open $\rightarrow$ closed path without unfolding/refolding by enforcing DT's stereochemistry. To generate DT transitions compatible with the aforementioned studies, we extracted a closed initial state from the monomeric structure (chain A of PDB ID 1MDT [308]) and an open final state from the domain-swapped dimeric structure (chain A of PDB ID 1DDT [309]).

DIMS and FRODA were used to generate the transitions as both methods have the necessary stochasticity to generate meaningful ensembles and use similar progress variables to induce transitions. Specifically, FRODA attempts to gradually decrease the C_α RMSD to the target, while DIMS uses a heavy-atom RMSD-to-target progress variable for soft-ratcheting. The purely geometric landscape employed by the FRODA algorithm then serves to contrast with the atomistic MD force field used in DIMS without unnecessarily confounding the analyses with dissimilar progress variables. We generated four separate ensembles, with one ensemble per method per protein; in total, 800 transitions were generated, with 200 transitions per method per protein. All transition paths were subjected

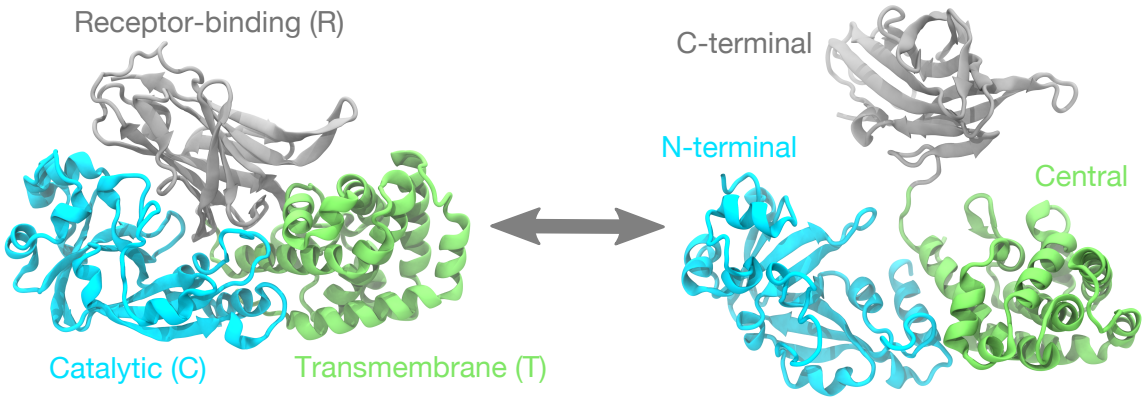

Figure 4.5: Diphtheria toxin (DT) accesses closed and open conformations, similar to the two depicted crystallographic structures, that are linked by a closed ↔ open transition. Comparing the two crystallographic structures suggests that the transition should involve a combination of rolling, twisting, and extending motions of the C-terminal, receptor-binding (R) domain (shown in gray) about the N-terminal, catalytic (C) domain (cyan) and the translocation (T) domain (green).

to a careful alignment protocol so as to minimize source of bias (details for both AdK and DT are provided in Appendix D of the Supporting Information).

4.2.2 Results and discussion

DT closed → open transition

The first pass of PSA on the DT transitions unveiled nine FRODA trajectories having Fréchet distances up to 12 Å from most other paths in the ensemble, indicated by the dark green bands in Fig. 4.6 showing Fréchet distances upwards of 10 Å. Visual inspection of the transition paths in question found that several trajectories terminated very short of the open DT target structure and a few exhibited large, unrealistic fluctuations about the final state. These erroneous trajectories were trimmed from the distance matrix and clustering was repeated, producing the heat map and dendrogram in Fig. 4.7. The DIMS and FRODA transitions formed two distinct clusters with no intermixing between the methods. The upper left section of the matrix showed that DIMS generated relatively similar paths, with darker stripes corresponding to a number of outliers. By comparison, however, the FRODA

paths (seen in the lower right part of the matrix) were much more similar to one another and, with the exception of a few outliers as well, appeared less diverse overall.

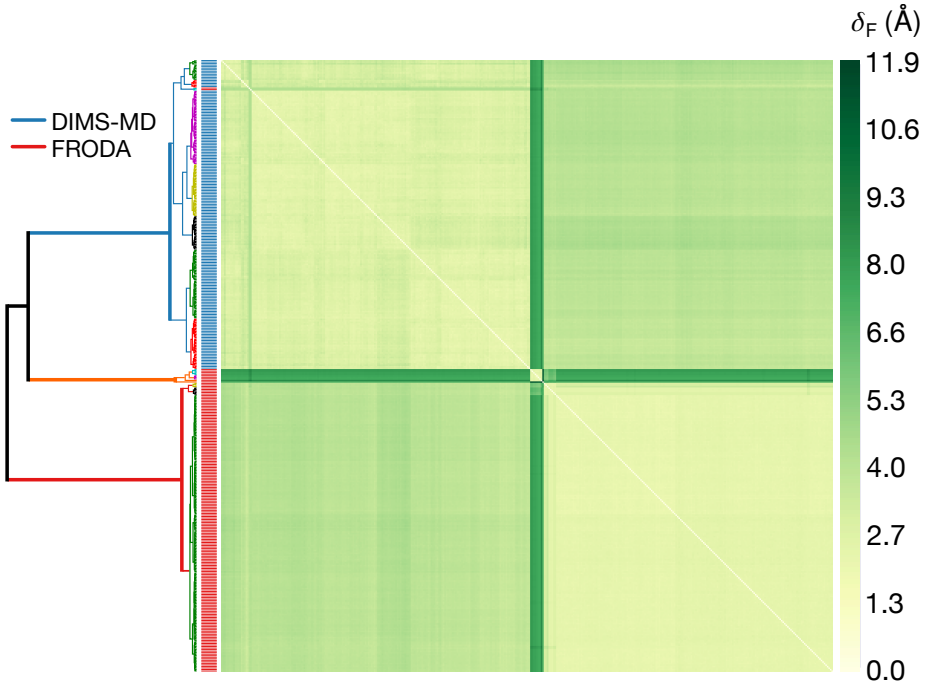

Figure 4.6: The raw, clustered heatmap-dendrogram generated by PSA from all-pairs calculations of the (discrete) Fréchet distances, δ_F, for all 200 DIMS (blue bars) and 200 FRODA (red bars) transitions of the diphtheria toxin protein (1MDT:A to 1DDT:A). Nine paths generated by FRODA (orange cluster) were very different from every other path in the comparison. Visual assessment confirmed that the erroneous paths were generated by unsuccessful FRODA runs and were subsequently filtered from the ensemble. [Adapted from S. L. Seyler et al. [5], Copyright © (2015), used under the terms of the Creative Commons Attribution License.]

We quantified these observations by histogramming DIMS–DIMS, FRODA–FRODA, and DIMS–FRODA distances for both the Fréchet and Hausdorff metrics and performing a Pearson correlation analysis. Fig. 4.8A shows near perfect correlation between Fréchet and Hausdorff distances (for corresponding pairs of paths), indicating that backtracking was unlikely a contributing factor. The distance distribution for FRODA paths (green) was both smaller on average and narrower than the DIMS distribution (red), consistent with a visual inspection of the heat map. However, whereas the bulk of the FRODA distribution was qualitatively more symmetric than the DIMS distribution, which was somewhat positively skewed, there were several outlier FRODA paths that were relatively dissimilar to most other paths (reflected in the dark stripes in Fig. 4.7). Moreover, it was

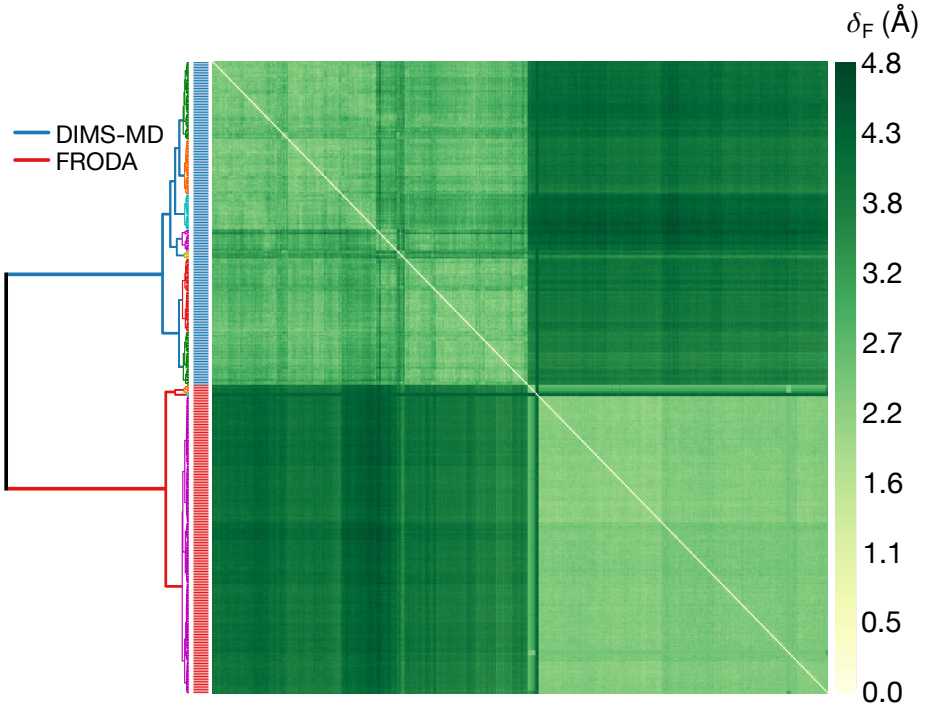

Figure 4.7: Clustered heat map comparing ensembles of diphtheria toxin (1MDT:A to 1DDT:A) transition pathways produced by DIMS (blue bars) and FRODA (red bars) using the (discrete) Fréchet distance, δ_F. Hierarchical cluster analysis was produced using the Ward linkage criterion using ascending distance order; incomplete trajectories were filtered and not displayed (see text).

observed that the distribution of DIMS–FRODA distances overlapped with the upper tail of the DIMS distribution as well as FRODA outliers. Thus, although DIMS and FRODA paths are grouped into distinct clusters, it can be concluded that the width of the transition tubes sampled by each method are comparable (via the triangle inequality of the metrics) to their separation in configuration space.

AdK closed → open transition

As with the DT ensemble, PSA generates two distinct clusters separating DIMS and FRODA paths with no intermixing (Fig. 4.9). In contrast with the DT results, however, FRODA generated paths that were more dissimilar to one another ($\langle\delta_F\rangle = 2.2(1)$ Å) than DIMS paths were from each other ($\langle\delta_F\rangle = 1.4(2)$ Å). Plotting the distributions of DIMS–DIMS, FRODA–FRODA, and DIMS–FRODA Fréchet and Hausdorff distances (Fig. 4.8B) shows that the DIMS–DIMS and FRODA–FRODA distributions have indeed swapped

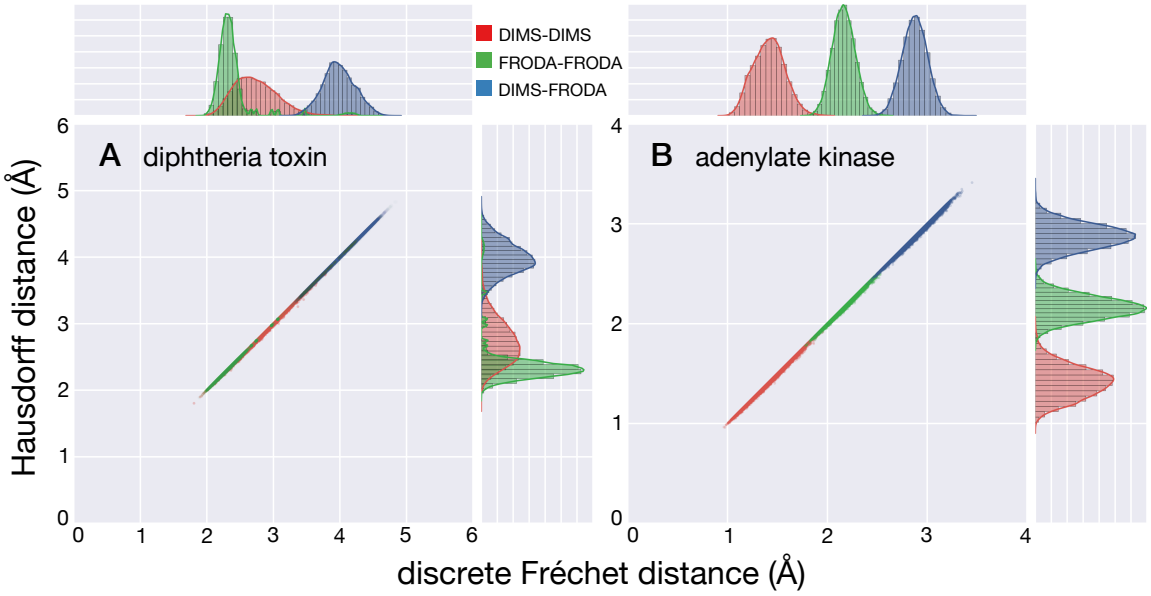

Figure 4.8: Correlations and joint distributions of (discrete) Fréchet versus Hausdorff distances (in Å RMSD) of the AdK (B) and DT (A) ensemble analyses are shown. Measurements are divided into three separate distributions: (1) mutual distances for DIMS paths (red), (2) mutual distances for FRODA paths (green), and (3) inter-method distances (blue) measured between a DIMS and a FRODA path. Both scatter plots indicate near perfect correlation between the path metrics for all distributions, with Pearson correlation coefficients equal to unity and p-values equal to zero to two decimal places, indicating that one path metric could be substituted for the other with affecting results. A slight deviation of the scatter points below the regression line (of unity slope) is consistent with the fact that Fréchet distances are bounded from below by their corresponding Hausdorff distances. DIMS simulations exhibited less variation than FRODA in the AdK transition, but had a larger average variation in the DT transition. For both DT and AdK, inter-method DIMS-FRODA distances are substantially larger than distances among paths from a given method. [Adapted from S. L. Seyler et al. [5], Copyright © (2015), used under the terms of the Creative Commons Attribution License.]

places, while all three distributions also showed little overlap. Furthermore, the mean Fréchet distance between DIMS and FRODA paths was $2.9(2)$ Å, significantly larger than the mean Fréchet distances within the FRODA and DIMS ensembles, indicating that transition tube sampled by each method is narrower than their separation in configuration space.

To connect to the atomistic structural differences between DIMS and FRODA conformers that gave rise to apparent differences in the sampled pathways, we explicitly extracted several δ_H-pairs and examined their projections in angle-angle space. Specifically, for each of the comparisons in DIMS–DIMS, FRODA–FRODA, and DIMS–FRODA, we extracted the δ_H-pairs corresponding to the median and max Hausdorff distances, giving the six total pairs plotted in Fig. 4.10A (red, blue, and purple circles and lines, respectively). Taking the median δ_H-pair among DIMS–FRODA comparisons and projecting the per-atom

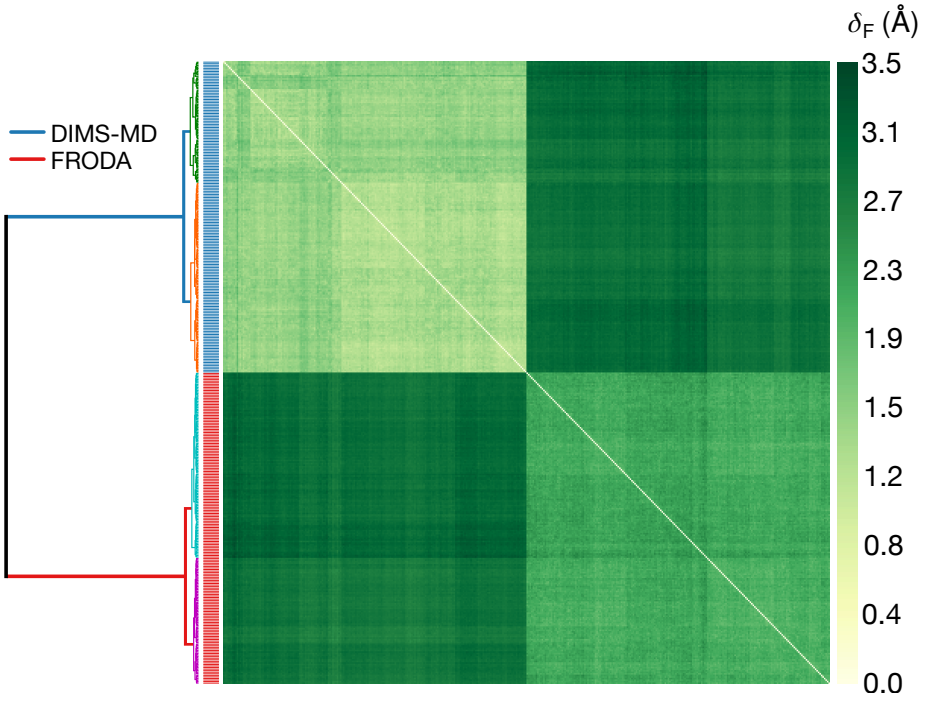

Figure 4.9: Clustered heat map comparing ensembles of adenylate kinase (1AKE:A to 4AKE:A) transition pathways produced by DIMS (blue bars) and FRODA (red bars) using the (discrete) Fréchet distance, δ_F, measured in Å. Lighter values indicate similar paths, while dark green corresponds to greater dissimilarity. Cluster analysis was produced using the Ward linkage criterion using ascending distance order. The DIMS and FRODA clusters are distinct, with two clusters were formed within both the DIMS ensemble (green and orange branches) and FRODA ensemble (magenta and cyan branches). [Adapted from S. L. Seyler et al. [5], Copyright © (2015), used under the terms of the Creative Commons Attribution License.]

displacements between the two structures (Fig. 4.10A), it became clear that the NMP domain in DIMS transitions had a number of evolutionarily conserved salt bridges (D33–R156, R36–D158, D54–K157) that tended to remain intact well into the transition. A direct examination of the DIMS conformer in the δ_H-pair indicated that the strong electrostatic interactions between the acidic and basic moieties [40] (Fig. 4.10B) were preferentially impeding the opening motions of the NMP domain. Full opening of the NMP domain required all salt bridges to be broken, while the LID domain could move relatively unhindered since the salt bridges are primarily located on its side. In the case of FRODA, which operates using purely geometrical influences, has no notion of the charge-charge interactions necessary for salt-bridge formation or breaking. Indeed, the FRODA conformer from the median δ_H-pair (with DIMS) did not have intact salt bridges (Fig. 4.10C) and did not reproduce the

"salt-bridge zipper" [40] evident in DIMS transitions. The manifest difference between DIMS and FRODA trajectories in angle-angle space for the apo-AdK closed $\rightarrow$ open transition was therefore a predominantly LID-opening pathway for DIMS simulations (Fig. 4.10A, blue circles), whereas FRODA paths (Fig. 4.10A, green circles) exhibited roughly simultaneous LID/NMP-opening in the vicinity of LinInt (Fig. 4.10A, black dashed line).

An alternative way to visualize differences between two paths is to consider their corresponding sets of nearest neighbor distances where, by construction, δ_H-pairs coincide with the maxima of the sequences of nearest neighbor distances. Fig. 4.11 depicts sequences of nearest neighbor distances for the three median δ_H-pairs shown in Fig. 4.10A. The DIMS and FRODA paths in question differed considerably from one another along the initial $\sim 60\%$ of the transition, corresponding to the slightly favored LID-opening pathway explored by the DIMS trajectory. The two FRODA trajectories differed by $\sim 2\,\text{Å}$ near the middle portion of the transition but were essentially coincident at the end points, demonstrating that FRODA, even with its stochastic component enabled, can systematically connect two given end-point structures. The DIMS trajectories differed from each other in a nearly uniform manner, $\lessapprox 1.3\,\text{Å}$, for the duration of the transition, suggesting that both effectively sampled the same pathway up to perturbations induced by thermal fluctuations.

4.2.3 Transition ensemble analyses in summary

To test whether PSA was amenable to ensembles containing hundreds of trajectories, we used the closed $\rightarrow$ open transitions of both AdK and DT as test systems to which the DIMS-MD and FRODA methods were applied. DT, representing a "hard" transition case Krebs and Gerstein [176], was previously shown to be amenable to a geometric targeting approach (FRODA) [180]. Initially, the heatmap-dendrogram approach in PSA allowed us to straightforward disregard erroneous trajectories without having to manually validate individual trajectories in the ensemble by direct inspection. A re-clustering of the Fréchet distances showed that DIMS and FRODA effectively sampled distinct pathways as indicated

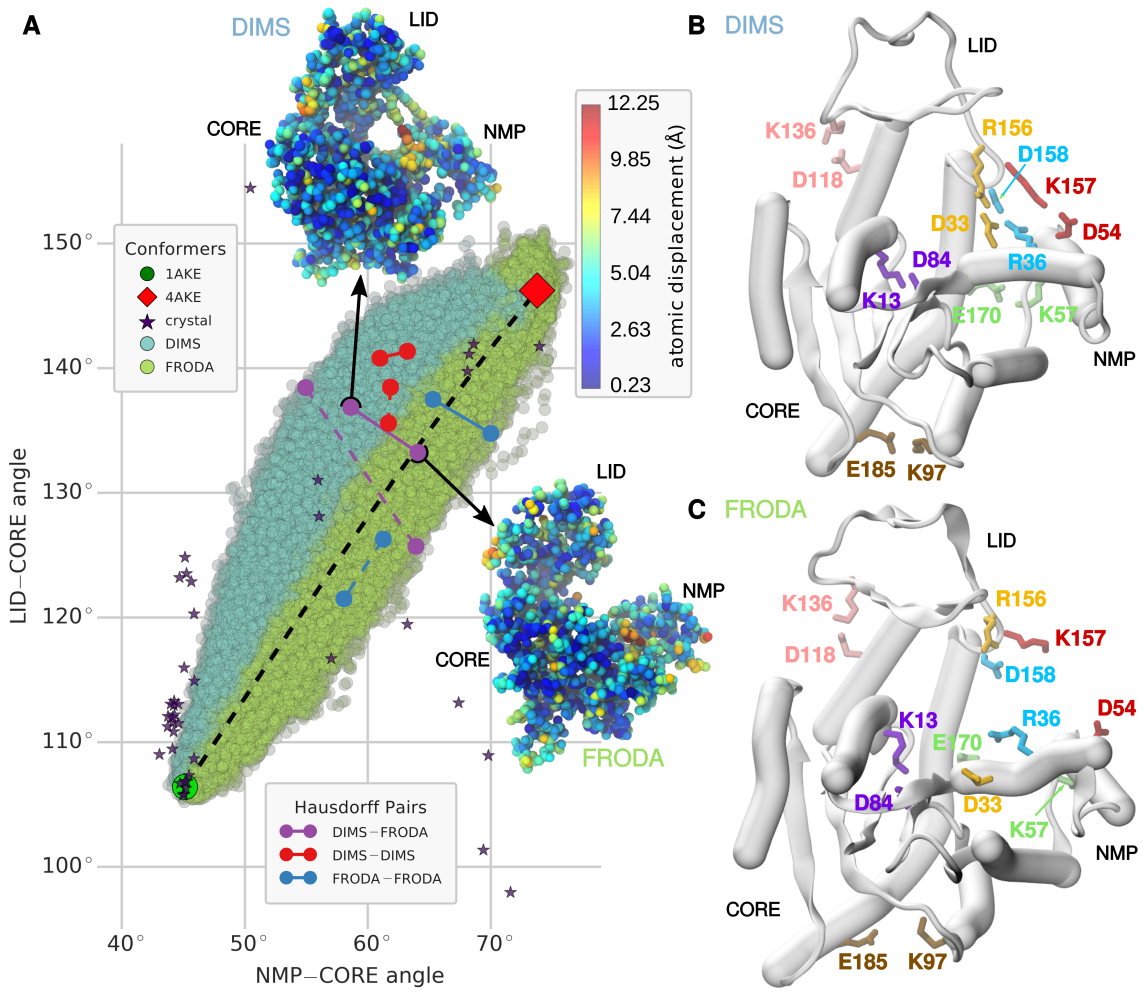

Figure 4.10: "Hausdorff pairs" (δ_H-pairs) analysis using 200 DIMS (cyan) and 200 FRODA (light green) trajectories projected into AA space. Hausdorff distances were computed for all unique path pairs. (A) Conformer pairs—corresponding to the δ_H-pairs with the median and maximum Hausdorff distances (solid and dashed lines, respectively)—are projected onto domain angle space for the following comparisons: DIMS–FRODA (purple), DIMS–DIMS (red), and FRODA–FRODA (blue). Experimental crystal structures, including some intermediates, are shown as stars [40], with further details available in S1Tab. Insets: Two heavy-atom representations for the median δ_H-pair structures between a DIMS and FRODA path, corresponding to conformers from their respective trajectories. The magnitudes of the per-atom displacement vectors between the two conformations are projected onto each conformer, with per-atom displacements given in Å according to the color bar. The initial and final conformations (green circle and red diamond, respectively) are provided for reference along with the linear interpolation path (LinInt – black dashed line). (B,C) Salt bridges in the DIMS and FRODA conformers from the DIMS–FRODA median Hausdorff pair. Three LID–NMP salt bridges (R156–D33, D158–R36, and K157–D54) and a CORE–NMP salt bridge (E170–K57) are intact in the DIMS structure (B) that are broken in the FRODA structure (C). The locations of the residues responsible for these salt bridges are towards the base of the LID and so the moment arm with respect to the NMP domain is substantially larger. [Reprinted from S. L. Seyler et al. [5], Copyright © (2015), used under the terms of the Creative Commons Attribution License.]

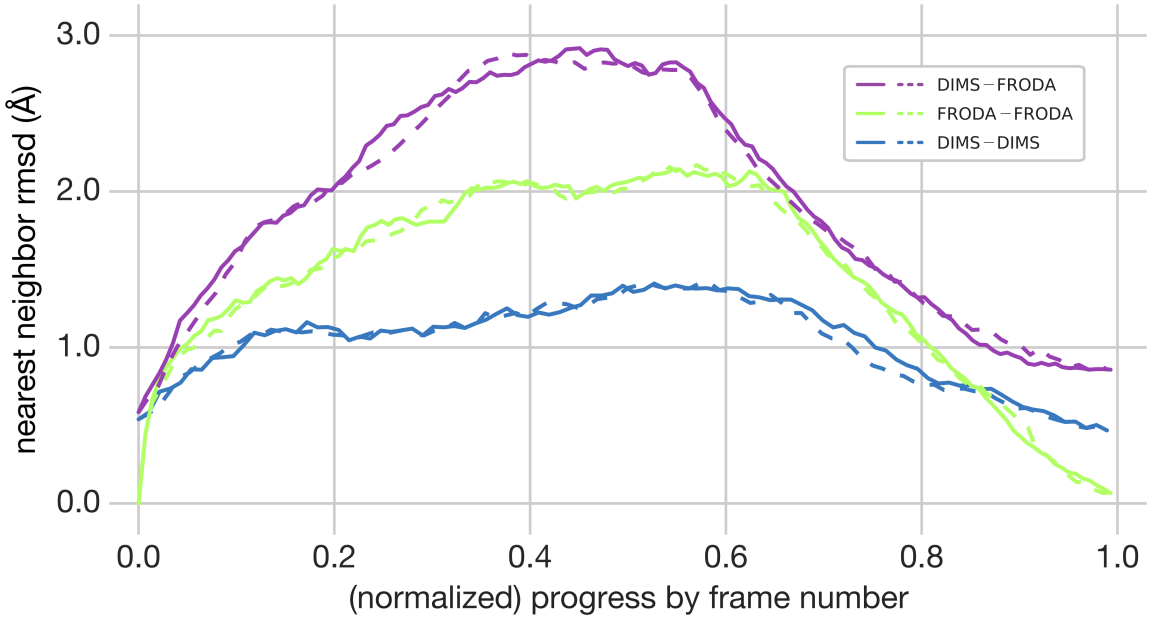

Figure 4.11: The nearest neighbor distances $\delta_{\mathrm{nn}}(k; Q \to P)$ (solid line) and $\delta_{\mathrm{nn}}(k; P \to Q)$ (dashed line) between pairs of paths P/Q belonging to the three median Hausdorff pairs in the AdK ensemble comparison (Fig. 4.10A) are shown for DIMS–FRODA (purple), DIMS–DIMS (blue), and FRODA–FRODA (green). The largest value $\max_{k,j}\{\delta_{\mathrm{nn}}(k; Q \to P), \delta_{\mathrm{nn}}(k; P \to Q)\}$ is the actual Hausdorff distance. For illustrative purposes, nearest neighbor distances are parameterized by plotting them as a function of frame number, k, normalized to the interval $[0, 1]$ (i.e., $k/|P|$), where 0 (1) corresponds to the first (last) frame. In general, it is more appropriate to use a one-dimensional order parameter or, ideally, a good reaction coordinate. [Reprinted from S. L. Seyler et al. [5], Copyright © (2015), used under the terms of the Creative Commons Attribution License.]

by the absence of intermixing. Similar results were obtained for AdK, with both methods again forming separate clusters. Interestingly, there was larger variation among paths in the DIMS AdK ensemble than the variation in the FRODA AdK ensemble, which was opposite to the results obtained for DT.

A Hausdorff pairs analysis of the AdK transition ensemble demonstrated that δ_H-pairs were effective indicators of pathway differences arising at the molecular level. Specifically, extracting median δ_H-pairs enabled us to implicate charge-charge interactions—fully modeled by the atomistic CHARMM27 force field in DIMS but completely omitted in FRODA—within a set of conserved salt bridges as the molecular-structural origin of differences between the predominantly LID-opening pathway sampled by DIMS and the relatively linear pathway sampled by FRODA. Notwithstanding the use of the C_α RMSD effectively limited the resolution of the analyses to the backbone or residue level, the Haus-

dorff pairs and nearest neighbor analyses could still resolve the salient molecular-structural features.

It might have been anticipated that the sampling space accessible to a DIMS simulation should be more limited in scope, relative to coarse-grained potentials, by the detailed all-atom energetics of its force field and should, in principle, be a subset of the space delimited by the purely *stereochemical* constraints in FRODA. However, no overlap between the transition tubes sampled by DIMS and FRODA was observed in either the AdK or DT ensembles, indicating that the overall extent of sampling can be substantially modulated by the character of a given biasing scheme and, if applicable, any progress variables. In the case of AdK transition ensembles, the Hausdorff pairs analysis revealed that the inclusion of electrostatics, resulting in salt-bridge breaking in DIMS-MD, was sufficient to explain the inclination toward a LID-opening pathway.

4.3 Case study 3: path-sampling methods and equilibrium MD

In studying conformational transitions using path-sampling methods rather than equilibrium MD (EqMD), we are routinely interested in the extent to which a given transition path is indistinguishable from an ensemble of unbiased transitions generated in equilibrium. Long-time simulations using unbiased, equilibrium MD with modern all-atom force fields can, in principle, be used to generate unbiased transition ensembles and calculate select observables to a sufficient degree of precision [166], though such efforts are frustrated by the sampling problem. However, with continual advancements in software and hardware, relatively simple conformational transitions of small proteins, like apo-AdK, may soon be amenable to such "gold-standard" equilibrium simulations. Using these systems as de facto test beds, one could then employ a suitable metric to directly compare fast path-sampling methods to gold-standard MD, thus allowing one to anticipate the performance of a fast path-sampling approach when applied to more complicated transitions that are otherwise inaccessible via equilibrium approaches. Special-purpose machines for fast MD simulations, such as the Anton [6] and Anton 2 [311] supercomputers, have enabled

millisecond-timescale simulations and may eventually be able to provide gold-standard transitions.

To demonstrate the general idea, we used apo-AdK system as the basis for long-time equilibrium simulations using the Anton supercomputer. In total, we consider 4 µs of sampling starting from the 1ake closed state and 7.5 µs of sampling from the 4ake open state (data not shown[||]). Transitions were only observed in the closed $\rightarrow$ open direction in spite of there being nearly 7.5 µs of simulation time starting from the open state. All four closed $\rightarrow$ open runs, on the other hand, made appreciable progress toward the open state within the first couple hundred nanoseconds. To get an idea of the time scales involved, the C_α RMSD relative to the 4ake state was measured over the course of each simulation and a closed $\rightarrow$ open transition was considered to have taken place once the RMSD from the open state dropped below 1.5 Å. Runs 1 through 4 reached the 1.5 Å cutoff within 17 ns, 278 ns, 138 ns, and 63 ns, respectively. These four trajectories were then truncated at their respectively defined transition cutoff times to create "Anton transitions" which we examined using PSA, angle-angle space projections, and 2D NCA alongside the original path-sampling methods from the comparison in Section 4.1. We note again that, for consistency with the original path-sampling methods comparison, only three runs (1, 3, and 4) were used; Run 3 was specifically chosen for exclusion since it exhibited the largest fluctuations (occurring around the 4ake open state) among the four simulations initiated at the closed state. This is an interesting observation on its own, though we omit further discussion to keep the analyses brief.

The heatmap-dendrogram comparison combining Anton transitions with the original paths (Fig. 4.12) from the path-sampling methods comparison did not, apart from the Anton transitions themselves, actually change the hierarchical clustering (Fig. 4.1) of the original path-sampling methods when using the Ward linkage criterion. Several changes in the clustering were observed when using the complete, average, and median linkage algorithms

[||] Sampling from the 4ake open state was primarily to check whether a spontaneous open $\rightarrow$ closed transition could be observed despite no such events having taken place.

were used (data not shown), but the general patterns also did not change substantially. The reason for similar clustering is sensible, because the Anton transitions were substantially *less* similar to all other paths than the paths from any other methods were to non-Anton paths—since the hierarchical clustering is performed *agglomeratively*, building up from the most similar (smallest Fréchet distances) transitions to the least similar (greatest Fréchet distances), the Anton transitions are clustered nearly last. Despite being dissimilar to GOdMD based on the heatmap coloring (paricularly Run 3), the Anton cluster and GOdMD cluster formed a cluster among themselves, though this only implies that the Anton cluster (of three) is more similar to the GOdMD cluster than it is to the very large cluster composed of all other methods, at least according to the Ward criterion.

In the angle-angle space projection (Fig. 4.13), Anton transitions unambiguously explored a LID-opening pathway. Given the evidently large region of angle-angle space that they sampled, it is remarkable that the Anton transitions were considerably more similar to one another ($\lesssim 3.5\,\text{Å}$) than they were to nearly any other method (typically $\approx 4\,\text{Å} to 6\,\text{Å}$), which is an indication that there are relevant structural relationships among the conformers from Anton transitions that are not present in the other methods. Fig. 4.13 suggests that relatively extensive sampling in the large basin surrounding the 4ake open region is partly responsible for such differences, though it seems that those differences would also arise *among* the Anton transitions as well. It is thus probable that the intermediate conformers comprising the LID-open-NMP-closed metastable basin sampled by the Anton transitions are primarily responsible for the observed differences between Anton transitions and fast path-sampling methods. That Run 1, which appears qualitatively similar to GOdMD transitions in angle-angle space, is also the most similar overall to other methods, and Run 4 to a lesser degree. Run 3 extensively sampled a LID-open-NMP-closed metastable basin, clearly delineating putative boundaries of such a state, and was the least similar to all methods in PSA. All Anton transitions suggest the a LID-open-NMP-closed intermediate state to

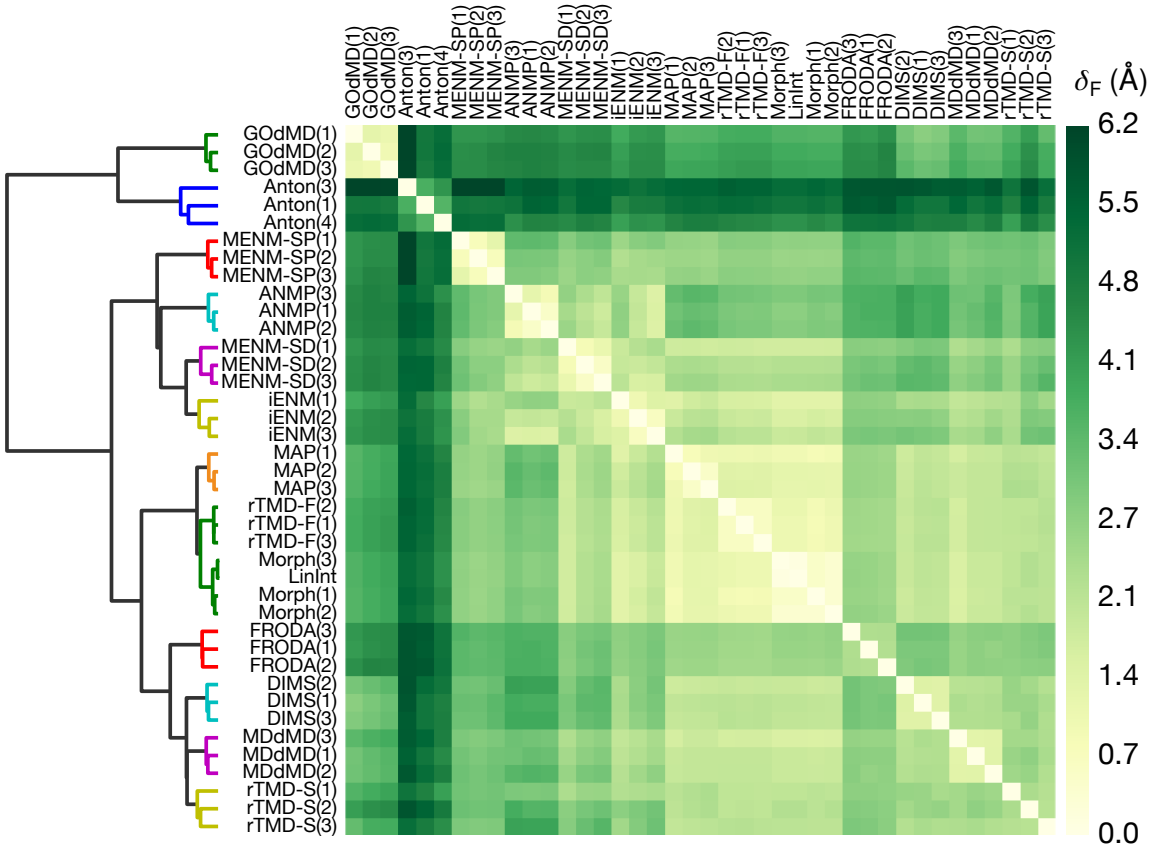

Figure 4.12: Path similarity analysis is used to compare three Anton closed → open "transitions" for apo-AdK with the path-sampling methods from Fig. 4.1. Method labels are unchanged from the original comparison. Fréchet distances, δ_F, are in Å and correspond to a structural C_α RMSD in accordance with the RMSD point metric. Smaller distances (lighter colors) correspond to greater similarity. The second Anton trajectory was excluded to avoid obfuscating the comparison by otherwise relatively large Fréchet distances it generated (refer to Fig. 4.13, the red conformers are scattered over a very broad region in the general vicinity of the open 4AKE:A state).

different degrees, with Run 4 being the least convincing, Run 3 being unequivocal, and Runs 1 and 2 definitely consistent with such a metastable basin.

It should first be noted that the Anton trajectories rapidly moved away from the 1ake state almost as soon as backbone restraints were released during equilibration (dashed lines and stars in Fig. 4.13). It is clear that the relatively compact, originally substrate-bound structure represented by 1ake played a role in the way in which EqMD trajectories tended to move away from the closed state, especially since closed state simulations are initialized by deleting the AP_5A inhibitor coordinates from 1ake prior to minimization/relaxation/production. Crystal contacts and other 1ake native contacts affected by the inhibitor are

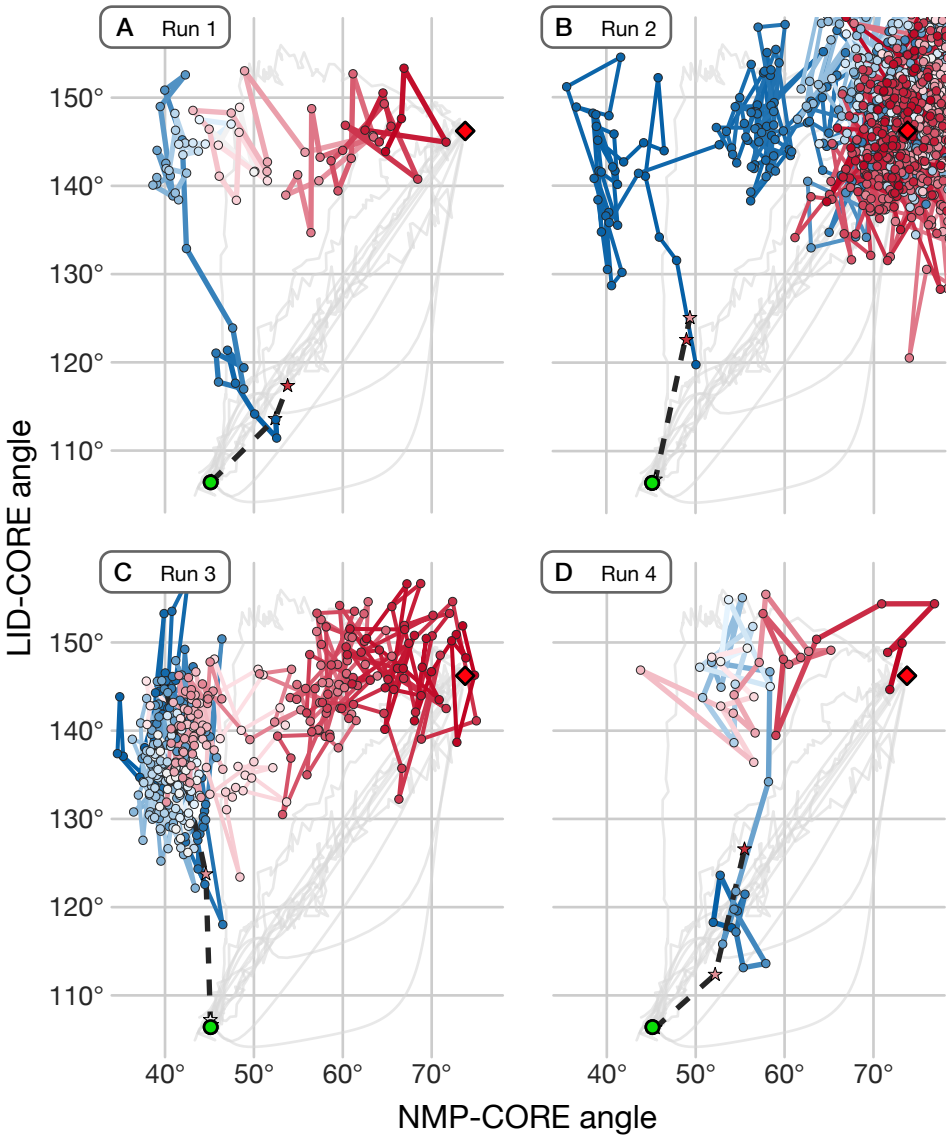

Figure 4.13: Projections of closed → open transitions from four 1 µs equilibrium MD trajectories generated with the Anton supercomputer—onto 2D NMP-CORE and LID-CORE angles (θ_{NMP}, θ_{LID}). The initial 1AKE:A (final 4AKE:A) structures are shown as green circles (red diamonds). Stars indicate the location of conformers at various stages of a multi-stage equilibration procedure (to achieve stable Anton runs); red (pink) stars indicate the final (penultimate) stage prior to production (stars corresponding to earlier stages are mostly obscured by green circles at the initial state). The forward temporal progress of each Anton transition is encoded by a gradual change in color from dark blue (early times) to dark red (late times), with light colors indicating intermediate times; successive conformers (circles) are separated by 240 ps. Transitions from the fast path-sampling methods are included for comparison and de-emphasized for clarity (translucent gray). (A–D) First through fourth Anton runs, respectively; the second run was excluded from Fig. 4.12.

possible causes of such behavior, which has been observed in a number of other EqMD studies of AdK that used the CHARMM27/36 force field [257, 258]. Such behavior was also

evident in NC space (Fig. 4.14), which showed Anton trajectories reaching the "V-shaped" cusp (discussed in Section 4.1.3 in the original path-sampling methods comparison) before the proper production runs were initiated. Once reaching the region of NC space roughly delineated by 87.5%–92.5% of 1ake contacts and 91%–95% of 4ake contacts, the Anton trajectories did not progress closer to the 4ake state. Notably, they are occupied the putative transition region sampled by MENM-SD/SP (and MAP, to a lesser extent) paths, perhaps indicating that there are relevant structural commonalities that could be explored. It is difficult to determine not only the relevance of the native state contacts from 1ake given the aforementioned concerns about contacts arising from crystallization and/or the AP_5A inhibitor, but possibly also the relevance of 4ake native contacts, which remain $\approx 5\%$ different from all of the Anton trajectories, even when including the full 1 µs trajectories of each transition (data not shown). In the future, a straightforward comparison among EqMD simulations using different all-atom force field may help resolve the question of whether such structural contacts are significantly force field-dependent and should be included in a future study.

4.4 Conclusions and Recommendations

We have shown that path similarity analysis (PSA) and the components comprising it—path metrics, hierarchical clustering-based heatmap-dendrogram approach, and Hausdorff pairs analysis—can be used to quantitatively investigate conformational transition paths of proteins. In the first study, path-sampling methods were compared and we verified that PSA (via heatmap-dendrogram visualization) was, at minimum, consistent with the overall picture provided by sensible collective variable projections, specifically AdK domain-angle space and NC space. The visual presentation of the data facilitated the identification of general patterns among the paths, their parent methods, and the underlying physical model that produced them. The heatmap-dendrogram analysis was also robust against various combinations of the path metrics and (clustering) linkage algorithms. The "trajectory ensemble" is a useful statistical mechanical picture for conceptualizing the dynamics of

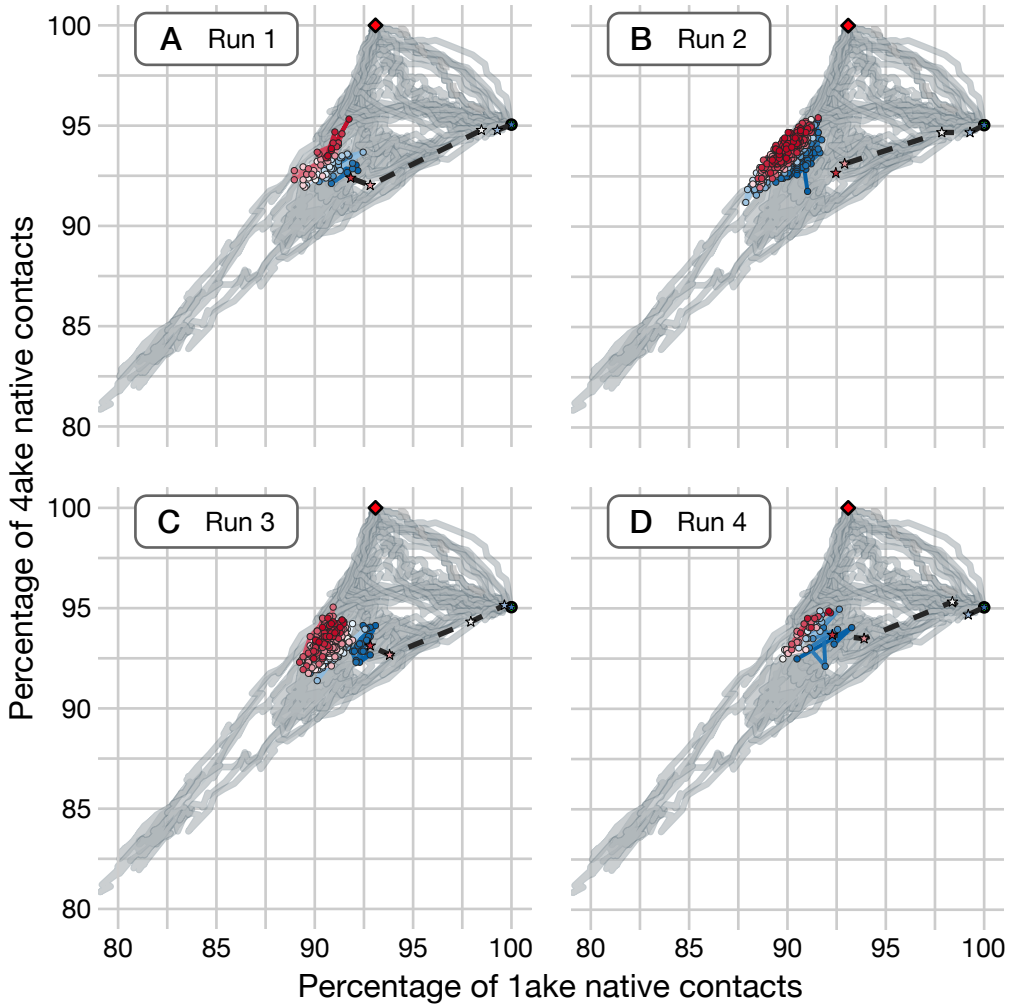

Figure 4.14: Native contacts analysis of closed → open transitions—extracted from four 1 µs equilibrium MD trajectories generated with the Anton supercomputer. Conformers are projected onto the 2D native contacts space where abscissa (ordinate) corresponds to the percentage of contacts shared with the 1AKE:A initial (final 4AKE:A) state shown as a green circle (red diamond). Stars indicate the location of conformers at different stages of an extensive equilibration procedure (to achieve stable runs on Anton); red (pink) indicates the final (penultimate) stage prior to the production run (stars from backbone restrained equilibration during earlier stages are mostly obscured by the green circles at the initial state). For production trajectories, each conformer is represented by a colored circle, with 240 ps separating successive conformers. The forward progress of each trajectory in time is encoded by a gradual change in color from dark blue (early times) to dark red (late times) with light colors indicating intermediate times. For clarity, Anton transitions are plotted separately (red/blue circles and lines) and transitions from the original comparison of fast path-sampling methods are de-emphasized (translucent gray) for clarity. (A) First Anton run. (B) Second Anton run. (C) Third Anton run (excluded from Fig. 4.12). (D) Fourth Anton run.

high-dimensional dynamical systems, including conformational transitions in proteins. As such, we demonstrated that PSA is a viable means of analyzing hundreds of stochastic trajectories having potentially hundreds or thousands of frames. In particular, this statistical

approach combined with Hausdorff pair analysis revealed that the electrostatics modeled in DIMS (but not FRODA) created the conditions for a force imbalance between the mobile LID and NMP domains—caused by charge-charge interactions in several salt bridges connecting the mobile domains—that induced a LID-opening pathway bias in DIMS simulations.

The methods and ensemble studies strongly support the idea that PSA would be a good tool to assess the influence of a biasing protocol—as well as biasing strength and progress variable selection—on the overall extent of (orthogonal) sampling. Though it was not immediately obvious how one could go about tuning one path-sampling method so as to emulate the sampling behavior of another method, we observed, for example, that pulling the rTMD restraint potential with fast and slow speeds led to quantitatively (Fig. 4.1) and qualitatively (Fig. 4.4) different behavior; quickly moving the restraint potential along RMSD-to-target progress variable created paths resembling LinInt, while a slow pulling speed produced paths comparable to dynamical methods like DIMS and MDdMD. In general, stronger biasing potentials or fast-moving protocols should be expected to restrict sampling to a narrower tube, while "soft" biasing approaches (e.g., soft-ratcheting) should be less likely to discourage sampling orthogonal to the overall course of progress. Moreover, the ensemble study suggests that it would be instructive to carefully examine how transition pathways are transformed when specific aspects of a physical model are changed (e.g., electrostatic forces). In principle, one should be able to test whether adding realistic charge-charge interactions in FRODA would be sufficient to generate DIMS-like LID-opening pathways. To take an even deeper look at charge-charge and solvent-protein interactions, one might adapt a version of the solvated dissipative ENM (sDENM) [76, 77] for the task of generating transitions between two input states. By direct comparison with other ENM-based path generating approaches, a two-state sDENM could allow that electrostatic effects be isolated and quantified at the molecular-structural level using heatmap-dendrogram analysis, Hausdorff pair analysis, and other integrative analyses.

A key objective in studying protein structure-function is to identify fast path-sampling methods—and the physical models and sampling schemes therein—that consistently repro-

duce (important aspects of) realistic conformational transitions. Eventually, quantitative assessments of path-sampling accuracy will have to be devised so as to build confidence in the various physical models and algorithms that have been discussed. As a first step toward this goal, we extended the original methods comparison to include a preliminary comparison with long-time equilibrium MD trajectories produced with the Anton supercomputer. "Anton transitions" generally favored a LID-opening pathway (Fig. 4.13), thus recapitulating the distinguishing qualitative feature among the dynamical path-sampling methods, and three transitions were consistent with an extant LID-open-NMP-closed intermediate state, with one simulation (Fig. 4.13C) unmistakably demarcating a region of metastability. All of the original methods were, however, quantitatively more similar to one another than to any of the equilibrium runs (Fig. 4.12), consistent with the very broad extent of sampling evinced by the angle-angle projection (Fig. 4.13). Although precise conclusions about the performance of each method cannot yet be drawn from this comparison, the mutual similarities among the Anton transitions indicate that there are quantifiable structural patterns that distinguish them from the fast path-sampling methods.

Chapter 5

COMPUTATIONAL INSIGHT INTO THE BOR1P TRANSPORT MECHANISM

This chapter is based on the published study, Nicolas Coudray, **Sean L. Seyler**, Ralph Lasala, Zhening Zhang, Kathy M. Clark, Mark E. Dumont, Alexis Rohou, Oliver Beckstein, and David L. Stokes (2017). *Structure of the SLC4 transporter Bor1p in an inward-facing conformation.* Protein Science, 26: 130–145. [7] My contribution to this work was the design of the protocol for generating an outward-facing model, execution of MD flexible fitting (MDFF), performance of equilibrium MD simulations, and the assessment of structural viability and the transport mechanism via the analysis of MDFF and MD simulation data.

5.1 Introduction

The family of SLC4 transporters is part of the Amino acid-Polyamine-organoCation (APC) superfamily, which, next to the Major Facilitator Superfamily (MFS), is the second largest superfamily of secondary transporters [312]. Members of the SLC4 family are solute-carrying secondary active transporters that play important roles in pH homeostasis and, in animals, acid-base transport in the kidney, stomach, pancreas, and other organs [313]. SLC4 transporters regulate the flux of bicarbonate (or a similar species) across a cell membrane by coupling its transport to ions. In humans, the well-known anion exchanger 1 (AE1) antiporter is essential to, among other things, the uptake of CO_2 by red blood cells via the electroneutral exchange of chloride and bicarbonate, which cannot otherwise permeate the plasma membrane [313, 314]. Other SLC4 members are electrogenic (i.e., transport net charge across a membrane), such as the sodium bicarbonate cotransporters (NBC) and the boron transporter from yeast (Bor1p), the latter being the focus of this chapter. Bor1p has been implicated in the transport of borate (BO_3^-) and protons [315] and it has been suggested that the closely homologous BOR1 boron exporter facilitates the tolerance of yeast to boron [316], which is toxic in high concentrations. Boron has been known to be an

essential micronutrient in plants (for instance, in maintaining the integrity of cell walls); in animals, boron plays a role in the regulation of steroid hormone levels and the activity of certain enzymes, and the recent discovery of the mammalian homolog NaBC1 by Park et al. [317] has suggested that animals may also require detailed control of boron concentrations [318].

Studies of the SLC4 transport mechanism have been limited by a lack of experimental structural data. However, the fold revealed by the X-ray crystal structure of AE1's membrane domain [319] bears a close resemblance to the uracil transporter UraA and uric acid-xanthine permease UapA from the NCS2 family, as well as the fumarate transporter SLC26Dg from the SLC26 family, all of which are members of the APC superfamily. In particular, the arrangement of transmembrane (TM) helices in AE1, UraA, UapA, and SLC26Dg is characterized by 7 + 7 (referring to the TM helices) inverted structural repeats that mesh to form a "gate" domain (often involved in dimerization and helping to anchor the protein in the membrane) and a "core" domain (implicated in substrate binding and transport). Fig. 5.1 depicts the secondary structure topology of the Bor1p homology model, which delineates two inverted repeats (defined by the purple dashed boxes), gate domain helices (dotted cylinders), and core domain helices (solid cylinders). Such structural congruence is suggestive of a common transport mechanism, which has been hypothesized to involve the relative movements of the gate and core domains [320–322]. There are two different mechanistic pictures consistent with the alternating-access transport model [10], whereby the stoichiometry of the transport cycle is preserved during the conformational transition by alternatively exposing the binding site to the cytoplasm and periplasm.

The first picture—the so-called elevator mechanism—proposes that a core or transport domain translocates vertically relative to both the membrane plane and the gate or scaffold domain anchoring the protein in the bilayer. NapA [323], apical sodium-dependent bile acid transporters (ASBT) [324], CitS [325], and GltPh transporters [326] have been shown to involve large vertical movements of the core domain; though these transporters are not themselves in the APC superfamily, they bear an analogous core/gate architecture that led

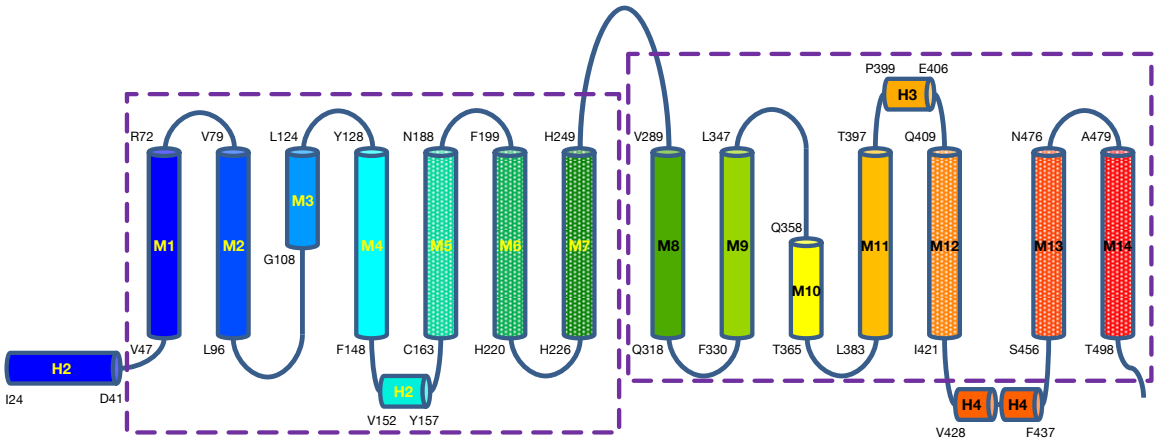

Figure 5.1: Secondary structure of the homology model for Bor1p. Dashed boxes are drawn around the two pseudo-symmetric, repeats, which have an inverted topology with respect to the membrane. Dotted helices correspond to the gate domain and solid helices correspond to the core domain. [Reprinted from Coudray et al. [7], Copyright © (2016), with permission from Wiley / Adapted from Fig. S4A.]

Alguel et al. [321] to propose that the elevator mechanism may extend to the NCS2 and AE families. In the second picture—the rocking-bundle mechanism—the transport cycle does not involve a significant vertical translocation of the substrate binding site relative to the membrane, but rather a rigid-body pivoting motion of the core domain about the binding site. The rocking-bundle picture describes the conformational motions of several transporters from the APC superfamily, including Mhp1, BetP, AdiC, vSGLT, and LeuT [8, 9], which all share a 5 + 5 inverted repeat topology. Despite the similarity in the folds among APC superfamily members, it is not known whether the rocking-bundle description can be extended to the 7 + 7 topology of the NCS2, SLC26, and AE1 families, particularly because their evolutionary relationship is not fully understood [312]. Additionally, given that AE1 has only been resolved in an outward-facing conformation, determining the nature of the transport mechanism has been challenging.

To provide structural insight into the function of SLC4-like transporters, our collaborators solved the structure of Bor1p from the yeast *Saccharomyces mikatae* in an inward-facing conformation. Cryo-electron microscopy (cryo-EM) was used to generate an electron density map of helical crystals of membrane-bound protein at $\sim$ 6 Å resolution. We used molecular dynamics flexible fitting (MDFF) to flexibly fit an atomistic homology model

based on the AE1 structure into the density map, after which several MD simulations in an explicit solvent/membrane environment were performed on the model to assess its stability and study its dynamics. Water accessibility calculations suggest the structure represents an inward-facing conformation. An outward-facing model of Bor1p was generated by using MDFF to guide the structure into a pseudo-density map produced from the outward-facing AE1 structure. Further MD simulations of the outward-facing Bor1p model did not exhibit the large vertical shift expected of an elevator mechanism, but were instead more consistent with the rocker-switch model that describes other members of the APC superfamily.

5.1.1 Structural determination and cryo-EM

The determination of 3D protein structures has been essential to our understanding of how proteins work. X-ray crystallography has been one of the most successful techniques, enabling the study of many macromolecules at near-atomic resolution ($\lesssim 4$ Å) when crystallization is possible. Growing high quality crystals of a protein is a difficult task, however, and one must account for structural artifacts or crystal contacts that may not reflect physiological conformational states. In the challenging case of membrane proteins whose native environment is a lipid bilayer, large portions of their hydrophobic surface must be stabilized (often using detergents) to achieve crystallization, a process hindered by the presence of detergents [327]. Solution nuclear magnetic resonance (NMR) can be applied to aqueous solutions of proteins, though the approach is less suitable for membrane proteins since good resolution requires the molecules to be reoriented rapidly in solution. Solid-state NMR has been applied to study time-averaged conformations of membrane proteins in a bilayer, but atomistic resolution remains a challenge.

Cryo-electron microscopy (cryo-EM) refers to a set of several subdisciplines—cryo-electron tomography, single-particle cryo-electron microscopy, and electron crystallography—that combine cryogenic freezing techniques with electron microscopy [328]. Cryo-EM has the advantage over other microscopy techniques in that specimens can be observed in native-like environments and has been used in many contexts for

biological structure determination. By rapidly reducing a protein sample to cryogenic temperatures using liquid ethane, the aqueous environment is vitrified without inducing crystallization that not only can damage the sample but also create diffraction peaks that obscure the signal [329]. In particular, cryo-EM is better suited than other methods to characterize membrane proteins in a lipid bilayer and can reveal important lipid-protein interactions in the transmembrane region of a protein [330, 331]. Single-particle cryo-EM of detergent-solubilized membrane proteins leads to 3D electron density maps that, in principle, represent equilibrium ensembles of one native-state conformations; a variety of techniques can be used to process and cluster the structural ensembles to reconstruct a 3D image. Lipid nanodiscs offer one means of mimicking a bilayer and avoiding the problems with detergent-based techniques [332]. Electron crystallography, the approach our collaborators used to generate EM density maps, can also be used to grow 2D arrays of helical or tubular crystals of transmembrane proteins [330].

Using single-particle cryo-EM, *de novo* structure determination is possible at resolutions below $\sim 3.5\,\text{Å}$, where side-chains can be resolved clearly. Though the best achievable resolutions were typically limited to $4\,\text{Å} \lesssim 7\,\text{Å}$, especially for larger macromolecules that admit higher signal-to-noise ratios [333], recent advances in both electron detection and image processing were sufficient to build the first *de novo* atomic structure using single-particle cryo-EM [334], namely an icosahedral virus solved by X. Zhang et al. [335]. In the case of electron crystallography, it is generally difficult to obtain ordered crystals, and the achievable resolutions are typically insufficient to build atomic structural models *de novo*. However, computational "fitting" methods can integrate moderate-resolution EM density data, for instance, into MD flexible fitting (MDFF) simulations and be used to generate plausible atomic structural models.

5.1.2 Molecular dynamics flexible fitting

Atomistic models of a protein can be taken directly from existing structures (perhaps one or several conformations from X-ray crystallography) or they can be built through

homology modeling techniques when such structures are unavailable. Flexible fitting methods, including MD, coarse-grained, and structure-based fitting approaches, are designed to aid the refinement of such models—particularly when only medium-resolution density data is available [336–342]. In the study described in this chapter, the molecular dynamics flexible fitting (MDFF) approach is used, which accepts an EM map as input and defines a global "fitting" force field that guides an MD simulation into higher density regions of the EM map [336].

Where high-resolution data are available, MDFF variants such as cascade MDFF (cMDFF) and resolution exchange MDFF (ReMDFF) can help to overcome difficulties in fitting high resolution maps (sub-5 Å) that often contain sharp peaks and valleys where structures can be trapped [343]. By combining several methods with different physical models, consensus flexible fitting can be used to improve the confidence level of the global and local features of a fit, offering an automated means to enhance the overall interpretation of cryo-EM data [344, 345].

5.2 Methods

5.2.1 IF and OF Bor1p models from EM maps and MDFF

Starting with a homology model of Bor1p based on the AE1 atomic structure (PDB 4yzf)—built by our collaborators using an online version of MODELLER (https://toolkit.tuebingen.mpg.de/modeller) [346, 347]—MDFF was used to fit the model into the experimental EM density maps (cf. Table 5.1). To generate the 3D density maps, our collaborators first produced batches of helical Bor1p crystals and eventually identified a promising batch having two populations of tubes with radii 164 Å (Type 1, Fig. 5.2A) and 157 Å (Type 2) [7]. 3D reconstructions were then generated from these tubes using either Fourier-Bessel reconstruction or iterative real-space refinement. From the estimated 9–10 Å resolution of the Fourier-Bessel reconstruction (Fig. 5.2B), the architecture of the TM helices was difficult to resolve, though the membrane boundaries were

well-defined. In the real-space reconstruction implemented in the Frealix program [348], the TM helices were clearly defined by the tubes of density spanning the full height of the membrane (Fig. 5.2C–D). The maps from Fourier-Bessel and real-space reconstruction were used to generate the inward-facing (resp. IF/1 and IF/2 for Fourier-Bessel and Frealix real-space) models, respectively (see Table 5.1).

Table 5.1: Summary of models used in this study.

Name	Start model	Map	Method
AE1	PDB 4yzf	N/A	Published X-ray structure
AE1-based homology	AE1	N/A	MODELLER using sequence alignment for threading
IF/1	AE1-based homology (manual manipulation to match map)	Fourier-Bessel	MDFF
IF/2	AE1-based homology	Frealix, real-space	MDFF
OF/1	IF/1	AE1 simulated density	MDFF
OF/2	IF/2	AE1 simulated density	MDFF

Using the colores tool in SITUS [350], each inward-facing model was initially placed into the experimental density maps via rigid-body docking. The docked models and their corresponding density maps were then used as inputs to molecular dynamics flexible fitting (MDFF) calculations as implemented in VMD [351] and NAMD [352]. MDFF simulations used the CHARMM27 force field [112, 139] to model atomic interactions in addition to the global fitting forces derived from the density map. The protocol for each model closely followed the recommended settings described in the tutorial (http://www.ks.uiuc.edu/Training/Tutorials/science/mdff/tutorial_mdff-html/), which included constraints on secondary structure, cis peptides, and chirality during the fitting process. Two MDFF steps were performed for each model, the first of which being a Langevin dynamics simulation performed in vacuo for 106 steps using a 1 fs timestep with the map-derived forces using a scaling factor of 0.3; the second step involved a final energy minimization using a stronger scaling factor of 2. Other scaling factors were tested, though the values recommended by the tutorial produced satisfactory results.

To supplement our analyses, we generated outward-facing Bor1p models (Table 5.1, OF/1 and OF/2) based on the outward-facing conformation of AE1 and the structural

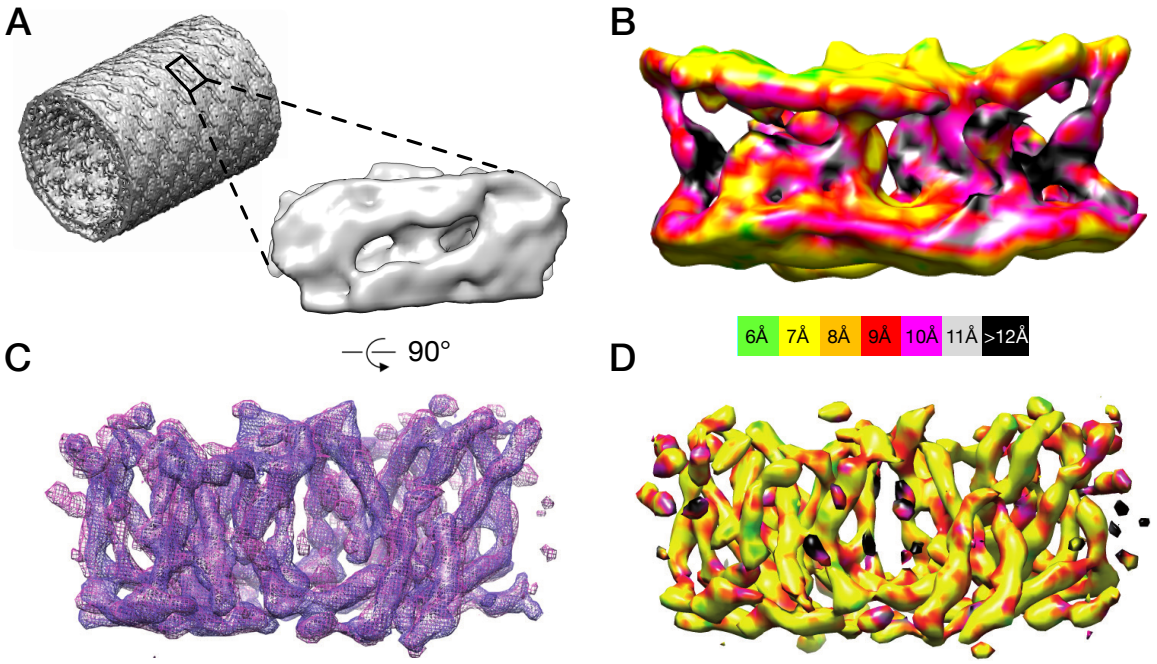

Figure 5.2: Fourier-Bessel (FB) and Frealix-based real-space reconstructions from Bor1p tubular crystals. (A) 3D reconstruction from Type 1 tubes showing the tubular morphology. The bilayer is visible as an almost continuous density around the perimeter of the tube. The lack of extra-membranous domains in Bor1p means that the surfaces of the tubes are relatively smooth, though a regular lattice is detect-able. One unit cell has been outlined. The zoomed view of the outer surface depicts a top-down view of the Bor1p dimer extracted from Type 1 tubes and is rotated 90° with respect to the structures depicted in B–D. (B) Fourier-Bessel reconstruction of the Bor1p dimer as viewed along the membrane plane. The continuous densities at the top and bottom correspond to the boundary of the membrane, with poorly resolved TM densities running in between. The surface of the map has been colored according to estimates of local resolution as determined by ResMap [349]. (C) Overlay of independently determined Frealix maps from Type 1 and Type 2 tubes (blue and purple mesh), showing the close correspondence between these structures. (D) Real-space map of the Bor1p dimer from Type 1 tubes viewed parallel to the membrane plane shows well resolved TM densities. The surface has been colored according to the local resolution as determined by ResMap [349]. [Reprinted from Coudray et al. [7], Copyright © (2016), with permission from Wiley / Adapted from Figs. 2 and 3.]

information of the EM maps from the MDFF-derived inward-facing models. The final frames of the MD-equilibrated (after MDFF) inward-facing models served as starting structures, where IF/1 (IF/2) was used as the starting point for OF/1 (OF/2). The docking step was again performed with colores to place the IF/1 and IF/2 models into the density maps prior to MDFF simulation. To build the OF/1 model, two simulated EM density maps were first generated from the AE1 dimer at 7 Å and 15 Å resolution using the MDFF plugin in VMD. Then, the IF/1 starting structure was run through a first round of MDFF with Langevin dynamics to drive the structure into the lower resolution simulated map at 15 Å using 250 ps of simulation time and a scaling factor of 0.3; in the second step, MDFF

with Langevin dynamics was performed again for 750 ps using the same scaling factor to further fit the model to the higher resolution map (7 Å to match the Fourier-Bessel IF map resolution); using an scaling factor of 2, a final energy minimization was performed as was done to create the IF/1 model. The purpose of the additional MDFF step was to more gently guide the initial IF/1 structure into the simulated map to avoid producing artifacts (that could be caused by fitting a model of one protein, Bor1p, to a simulated map from a homolog, AE1). Such artifacts did not arise as reversing the order of the target maps in the first two MDFF Langevin dynamics steps produced comparable results. Thus, to create the IF/2 model, we only performed one MDFF Langevin dynamics step by fitting to an additional AE1-based simulated map at 6 Å (to match the Frealix IF map resolution) for 1 ns (again, using a scaling factor of 0.3) after which a final minimization (using a scaling factor of 2) was performed.

5.2.2 Equilibration protocol for MD

CHARMM-GUI [353, 354] was used to embed the MDFF-derived models in a membrane composed of a mixed 1-Palmitoyl-2-oleoyl-sn-glycero-3-phosphoethanolamine (POPE): 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphoglycerol (POPG) bilayer with a 4:1 ratio, which approximates the E. coli lipids present in the tubular crystals, and generate input parameters for Gromacs. Atomistic MD simulations were performed in explicit solvent and explicit membrane using Gromacs 5.1 [355] using the all-atom CHARMM36 force field [356] and CHARMM TIP3P water model. The systems contained between 127,995 and 169,563 atoms in a hexagonal unit cell as described in Table 5.2. Leap-frog integration was used with a time step of 2 fs. Production runs were performed at $T = 310\,\mathrm{K}$ and $P = 1\,\mathrm{bar}$ in the isobaric-isothermal (NPT) ensemble using stochastic velocity rescaling thermostat [357] (with time coupling constant (1 ps) and semi-isotropic Parrinello-Rahman barostat [358] with coupling constant (5 ps). Verlet neighbor lists were updated every 20 time steps using a 1.2 nm cutoff. Force-switching was used for Lennard-Jones potentials from 1.0 nm to the cutoff at 1.2 nm. Electrostatic interactions had a real-space cutoff of 1.2 nm; the reciprocal space

contribution was computed with the smoothed particle-mesh Ewald (SPME) method [359], utilizing an FFT grid with 0.12 nm spacing and fourth-order spline interpolation. SETTLE was used to constrain the TIP3P water molecules and covalently-bonded hydrogen atoms were constrained using P-LINCS with fourth-order expansion and two LINCS iterations. Given the sensitivity of lipids to non-bonded parameters, the above settings were matched those used in [323] which were shown to reproduce the original CHARMM36 values from [356].

Table 5.2: Summary of molecular dynamics simulations performed for this study.

Name[a]	Conformation	Map[b]	Atoms[c]	Lipids	Water	Na/Cl	a (Å)	c (Å)	Time[d] (ns)
IF/1.1	Inward-facing	FB[e]	127,995	445	19,117	125/54	140.96	78.4	1100
IF/1.2									1100
IF/1.3									1100
IF/2.1	Inward-facing	Frealix	153,914	480	26,058	154/74	145.25	89.25	433.5
IF/2.2									451.1
IF/2.3			162,339	480	28,861	162/82	145.25	94.25	433.3
IF/2.4									430.1
OF/1.1	Outward-facing	FB	138,378	450	22,231	134/63	140.11	84.61	729.7
OF/1.2		MDFF							731.1
OF/1.3		AE1	146,010	450	24,763	140/69	140.11	89.61	702.1
OF/1.4									687.2
OF/2.1	Outward-facing	FB	169,563	482	31,267	167/87	144.6	97.9	522.1
OF/2.2		MDFF							524.8
OF/2.3		AE1							512.1

[a] Numbers after the decimal indicate the repeat number. Models are summarized in Table 5.1.
[b] Summary of map and approach used to build the model serving as initial structures for MD simulation.
[c] System composition (number of atoms, lipids, water molecules, and ions as well as the a and c unit cell parameters for the hexagonal simulation cells; values varied between simulations but were kept identical for repeat simulations.
[d] The total equilibrium simulation time was 9.5 µs.
[e] Fourier-Bessel.

To achieve stable equilibrium MD simulations, a careful minimization and equilibration process was used to relax various components of the protein-membrane systems. The default protocol for generated by the bilayer builder in CHARMM-GUI was used as a rough template. Restraint definitions were initially identical to those suggested by CHARMM-GUI: lipid phosphorus positions (1000 kJ mol^{-1} nm^{-2}), lipid chain double-bond dihedral angles (1000 kJ mol^{-1} nm^{-2}), protein backbone positions (4000 kJ mol^{-1} nm^{-2}), and side chain carbon positions (2000 kJ mol^{-1} nm^{-2}). Energy relaxation was carried out with restraints using the steepest-descent method followed by conjugate-gradient mini-

mization until the maximum force in the system fell below 500 kJ mol^{-1} nm^{-1}. The first five stages of equilibration were carried out at constant temperature, using the weak coupling Berendsen thermostat with small time steps (1 fs), followed by further NPT equilibration, using the Berendsen barostat with 2 fs time steps. By the sixth stage, all restraint forces had been gradually reduced to zero over the previous equilibration steps, with the exception of position restraints on the protein backbone (500 kJ mol^{-1} nm^{-2}); a full 5 ns of protein backbone-restrained equilibration using the stochastic velocity-rescaling thermostat and Parrinello-Rahman barostat was performed to allow the bilayer to relax around the protein. Two final equilibration steps were performed to gradually reduce the backbone restraints to zero. Each production simulation was initialized from the final frame of restrained equilibration, varying in length from 433 ns to 1100 ns for a total of 9.5 µs of sampling (cf. Table 5.2).

5.2.3 Analysis of MD simulations

RMSDs, RMSFs, and water densities were computed for the MD simulation systems using code based on MDAnalysis [272]. Most structural alignments and RMSD calculations were restricted to the C_α atoms in TM helices and other secondary structure regions (i.e., loop regions were excluded). Secondary structure assignments from DSSP [360] were collected with the Gromacs do_dssp tool [355] and subsequently processed with code that utilized the GromacsWrapper package (https://doi.org/10.5281/zenodo.437705). The degree of water molecule penetration in the models was evaluated from equilibrium MD simulations by mapping the locations of water molecules to a 1 Å grid and histogramming the water-oxygens using simulation frames sampled every 10 ps; to improve the water density sampling, intermediate conformers from the MD trajectories were superimposed onto the TM helices of chain A (used as a reference structure for RMSD fitting). In some instances, only the final half to a third of the trajectory was used for water density calculations to ensure that the protein-membrane systems were sufficiently equilibrated. As a consistency

check, we alternatively employed the HOLE program [361] to locate solvent-accessible cavities within the static structures taken from the final steps of MDFF simulations.

To ascertain domain motions characteristic of a possible elevator-like transport mechanism, the distributions of the differences in core and gate domain positions relative to the membrane horizontal were computed as described in previous work [323] and summarized briefly here. Specifically, we calculated the centers-of-mass of the core and gate domains—using only TM helices and excluding loops (to reduce noise due to their fluctuating behavior)—relative to the z-position center-of-mass of the membrane for both the IF/2 and OF/2 simulations (cf. Table 5.2) at each time step. Time series of the gate and core z-displacements were obtained by first subsampling at 1 ns intervals* then applying a Gaussian kernel density estimator to generate smooth distributions, $f(z)$, of the z-coordinate for each domain. Data from each chain (A and B) from all IF/2 and OF/2 simulations were extracted independently and later combined to improve statistics. An effective total of 3.50 µs of IF data ($N_{IF} = 2\,^{\text{chains}}/_{\text{frame}} \times 1748$ frames at 1 ns subsampling) and 3.12 µs of OF data ($N_{OF} = 2\,^{\text{chains}}/_{\text{frame}} \times 1559$ frames at 1 ns subsampling) were used. The difference in domain displacements between inward-facing and outward-facing simulations, $f(Z_{OF} - Z_{IF})$, were calculated by convolving the individual distributions to generate the joint distributions. A rough estimate of the standard error of the mean for the joint difference distribution was calculated from $\sigma\sqrt{N_{IF}^{-1} + N_{OF}^{-1}}$, where σ is the standard deviation of the distribution, and N_{IF} (N_{OF}) is the number of independent samples for the IF (OF) simulations.

*The longest correlation time of the domain position data was estimated to be < 199 ps using a single exponential fit to the autocorrelation function, so the intervals could be treated as independent samples.

5.3 Results and Discussion

5.3.1 MDFF

Fig. 5.3A,B shows the final snapshots of MDFF simulations for both the Fourier-Bessel and Frealix real-space [348] maps from which the IF/1 and IF/2 models (cf. Table 5.1) were generated, respectively. A strong correspondence between the 14 TM helices and their respective densities is evident, and the resulting cross-correlations over the course of MDFF showed relatively fast convergence. Given the uncertainty of the membrane orientation based on Fourier-Bessel reconstruction, we ran MDFF simulations with the starting model in both orientations (after manually adjusting several helices to better match the densities) and found that the cross-correlations did not indicate an unambiguous distinction. While the cross-correlations from MDFF (Fig. 5.3C) using the real-space map also did not reflect the sharper structural details that were clear from direct visual inspection, the real-space reconstruction nevertheless allowed us to unequivocally determine the orientation of the model in the bilayer, implying that the cross-correlation is not necessarily a reliable measure of the validity of MDFF results. In particular, the density revealed the characteristic shape of the M13-M14 hairpin on the side of each monomer, as well as the N-terminal helix H1 on the cytoplasmic side of the membrane. From the density corresponding to the H4 surface helix (between M12 and M13), which is not reflected in the inverted-repeat (Fig. 5.1) symmetry, we were also able to distinguish between the surface helices in M4-H2-M5 and M11-H3-M12.

Though the CHARMM27 force field used in MDFF provides a reasonable model of intramolecular forces, MDFF introduces an artificial map-derived fitting potential and performs dynamics *in vacuo*; to more realistically evaluate protein-solvent interactions and the dynamics, a proper solvent and membrane environment—matching the constitution of the *E. coli* lipids used in the tubular cryo-EM crystals—is necessary. As such, using the final structures of the MDFF simulations, the IF/1 (from MDFF with Fourier-Bessel map)

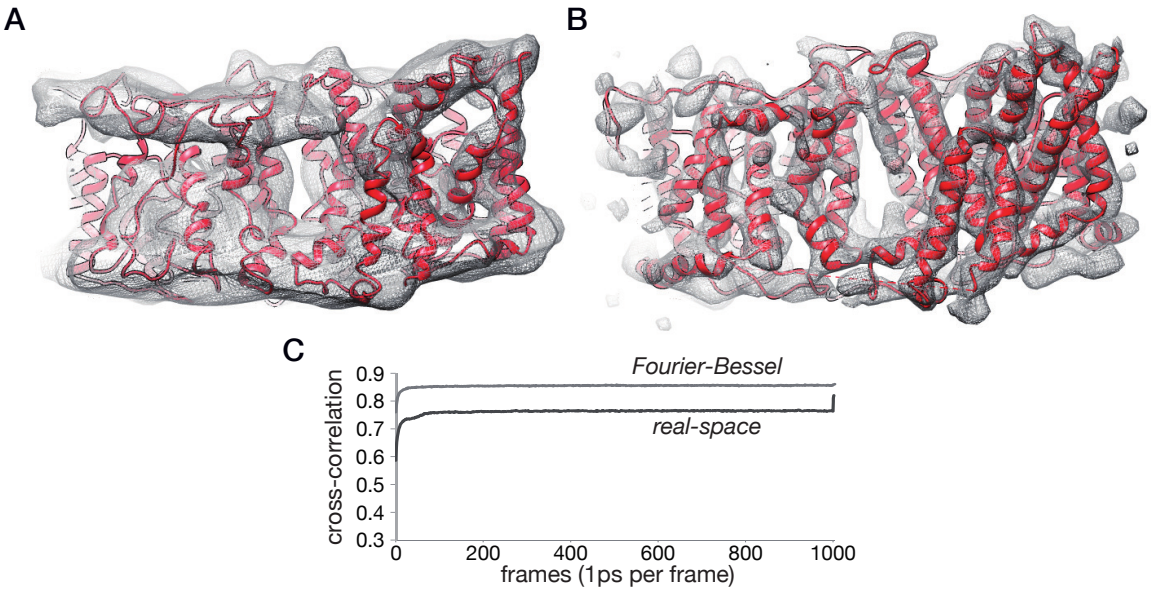

Figure 5.3: MDFF fitting to Fourier-Bessel and real-space reconstructions. (A) MDFF model (IF/1) superimposed on the Fourier-Bessel reconstruction from Type 1 tubular crystals (see Supporting Information Movie 1 and 3). (B) MDFF model (IF/2) superimposed on real-space reconstruction from Type 1 crystals (see Supporting Information Movies 2 and 4). (C) Cross-correlation coefficients showing the course of the MDFF fitting to both Fourier-Bessel (light grey) and real-space maps (dark grey). The spike at the end of the latter reflects the final energy minimization step with higher scaling factor. [Reprinted from Coudray et al. [7], Copyright © (2016), with permission from Wiley.]

and IF/2 (from MDFF with real-space map) were evaluated in an all-atom explicit solvent, explicit membrane environment using equilibrium MD and the CHARMM36 force field.

5.3.2 Structural stability and flexibility in equilibrium MD

Three MD simulations of the IF/1 model, each 1100 ns long (IF/1.1–IF/1.3 in Table 5.2), exhibited significant structural relaxation, the C_α RMSD (relative to the final MDFF structure) exceeding 5 Å (data not shown) over the course of each run. The source of the large overall C_α RMSD was reflected in the per-residue RMSF calculations (Fig. 5.5A), which revealed large fluctuations in the loops regions between M5-M6, M6-M7, M7-M8, M8-M9, M10-M11, and H4-M13. The C_α RMSD was thus recalculated among only secondary structure elements (i.e., non-loop regions), which included primarily the TM helices (M1-M13), as well as the surface helices (H2-H4) and non-helical S1/S2 β-strand regions near the putative substrate binding sites. The resulting RMSD timeseries ranged between 3 Å to

5 Å, though it did not appear to converge after even 1 µs of equilibrium MD. The lack of convergence notwithstanding, the IF/1 TM helices retained their α-helical character to a reasonable degree (Fig. 5.6A).

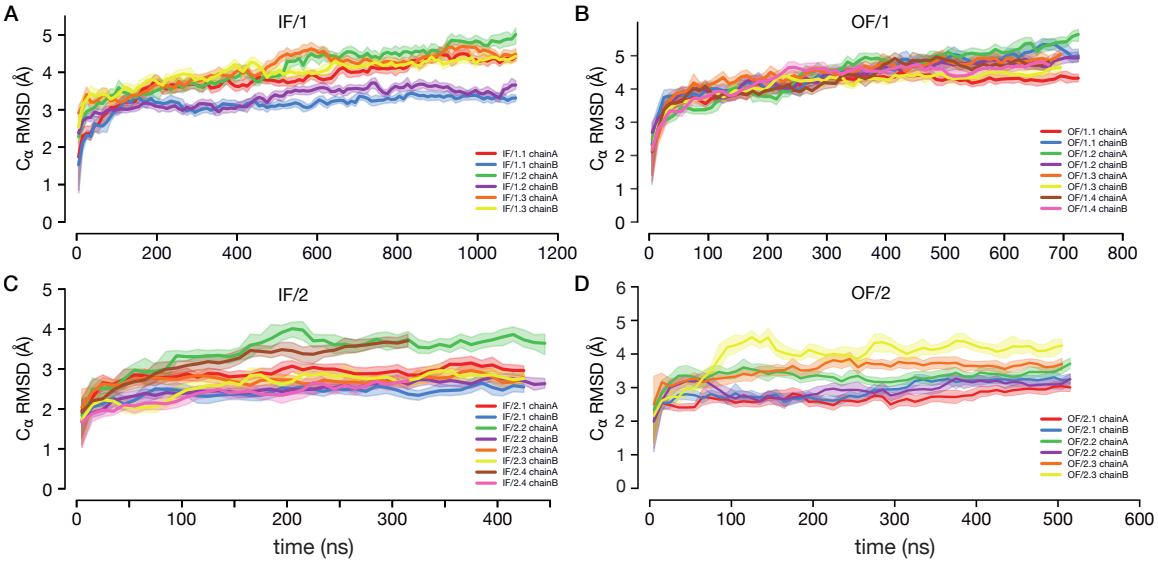

Figure 5.4: Structural drift during MD simulations as measured by RMSD relative to the starting frame. All frames in the MD trajectory where superimposed on the initial frame by minimizing the RMSD of the C_α atoms in all the TM helices. The RMSD of these atoms to the initial frame, which is close to the starting model, was plotted as a function of simulated time. Bands show the 95% confidence interval of the data around the mean (solid line) over short blocks of 10 ns duration. (A) Simulations the IF/1 model, based on the inward-facing Fourier-Bessel-generated model. (B) Simulations the IF/2 model, based on the inward-facing Frealix-generated model. (C) Simulations the OF/1 model, based on the outward-facing model generated from IF/1. (D) Simulations of the OF/2 model, based on the outward-facing model generated from IF/2. Separate lines for independent simulations and the two chains forming the dimer are shown. More information about these simulations is contained in Table 2. These data suggest that IF/2 and OF/2 simulations are well equilibrated, whereas IF/1 and OF/1 require longer simulations. [Adapted from Coudray et al. [7], Copyright © (2016), with permission from Wiley.]

For the IF/2 model derived from MDFF on the Frealix real-space map, we performed four MD simulations for at least 430 ns each (IF/2.1–IF/2.4 in Table 5.2). Despite shorter simulations, the non-loop (secondary structure) C_α RMSD (Fig. 5.6C) showed reduced structural drift $\sim$ 3 Å and improved convergence, while the TM helices were better preserved overall (Fig. 5.6B), even gaining additional helical turns in the case of H3, M12, and M13. Interhelical loops still exhibited large fluctuations in the per-residue RMSF (Fig. 5.5C), though the TM helices displayed remained comparatively stable in the explicit solvent-membrane environment. Overall, the IF/2 model based on the real-space construction

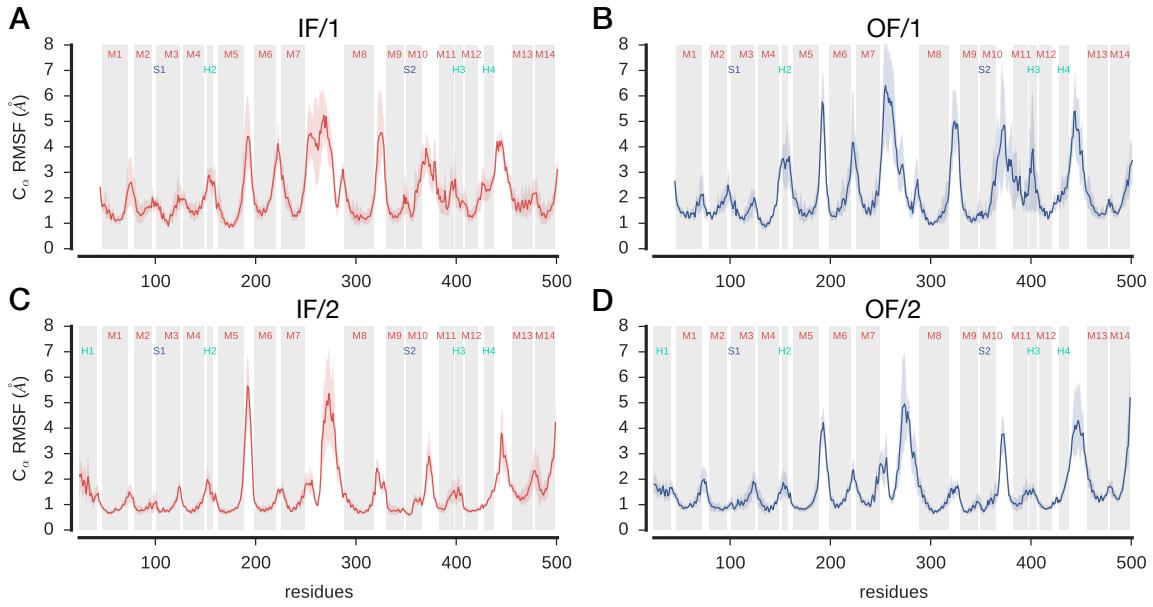

Figure 5.5: Fluctuation of residues during MD simulations of the model derived from the Fourier-Bessel and real-space EM maps. The root mean square fluctuation (RMSF) for C_α atoms was computed after superposition of all TM helices with RMSD fitting. Transmembrane helices and other secondary structure elements noted in Fig. 5.1 are indicated as shaded regions. Data represent averages over equivalent simulations and over both chains of the dimer, and are representative of at least two simulations from a given model. Heavy lines correspond to the mean, whereas bands indicate an interval containing 95% of the data, as determined by bootstrapping. (A) Data from inward-facing model based on Fourier-Bessel reconstruction (simulations IF/1.1–IF/1.3, see Table 2). (B) Data from outward-facing model based on IF/1 and generated by MDFF (simulations OF/1.3 and OF/1.4). (C) Data from inward-facing model based on Frealix reconstruction (simulations IF/2.1 and IF/2.2, see Table II). (D) Data from outward-facing model based on IF/2 and generated by MDFF (simulations OF/2.1–OF/2.3). [Adapted from Coudray et al. [7], Copyright © (2016), with permission from Wiley.]

retained secondary structures better than IF/1 throughout the simulations, especially in the β-strand geometry of the S1 and S2 regions near the binding sites (Fig. 5.6A,C).

In the case of the OF/1 model, the secondary structure C_α RMSD did not appear to converge over the course of the four $\gtrsim 700$ ns MD simulations, approaching 5 Å much like the IF/1 simulations. Similarly, the per-residue RMSF (Fig. 5.5B) and secondary structure profile from DSSP (Fig. 5.6B) of the OF/1 simulations were similar in magnitude and structure to the IF/1 simulations, especially in the lack of helical structure in the M11 and H3 helices and β-strand geometry in the S1 and S2 regions. On the other hand, OF/2 simulations appeared well-converged, with RMSDs remaining $\lesssim 3.5$ Å (Fig. 5.4D) over the course of three ~ 500 ns runs. OF/2 simulations also had a similar RMSF profile to IF/2 (Fig. 5.5C,D), both displaying reduced overall fluctuations across the interhelical loops,

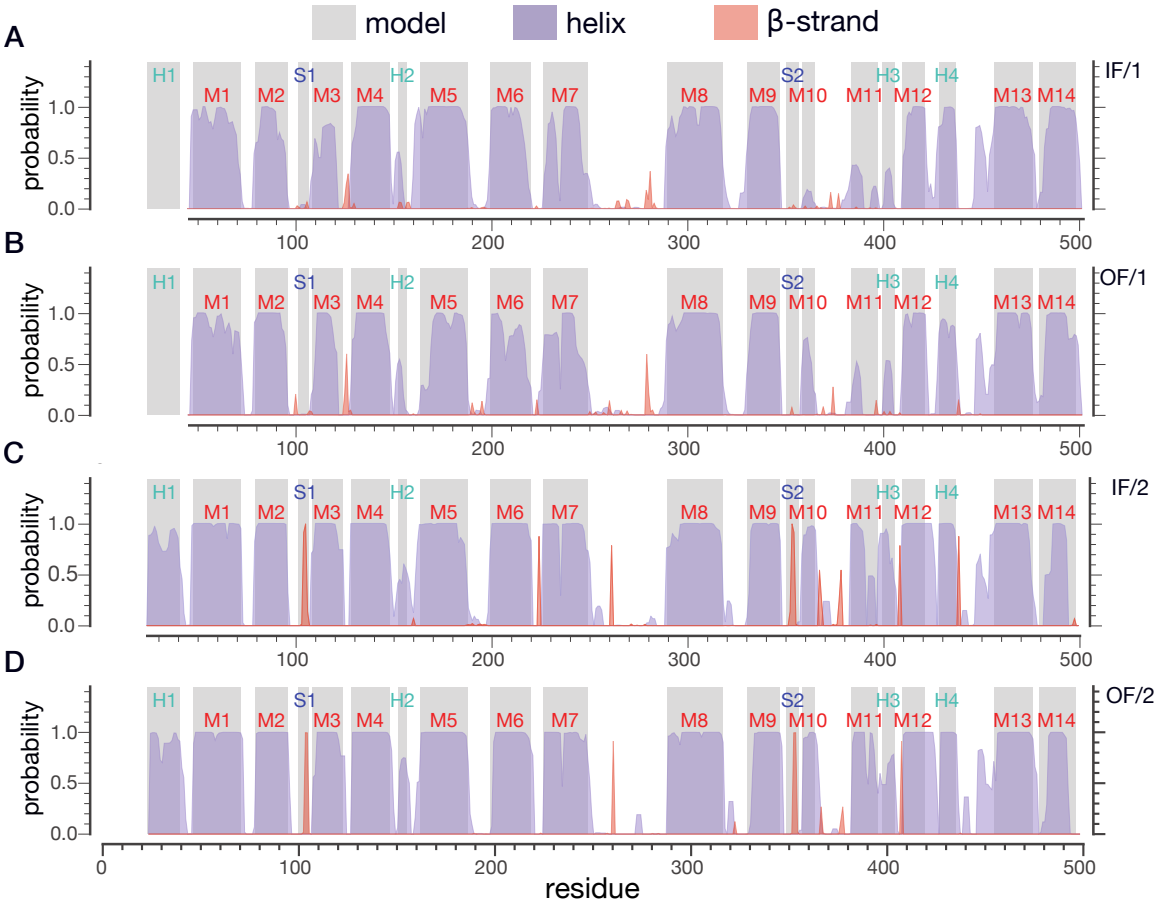

Figure 5.6: Probability of observing secondary structure during MD simulations. The secondary structure was assigned to each residue with DSSP for trajectory frames sampled every 0.1 ns and the probability calculated as the number of frames with the observed secondary structure divided by the total number of frames. Data were averaged over all chains and repeat simulations. (A) IF/1 model (Fourier-Bessel model), average of simulations IF/1.1, IF/1.2, and IF/1.3. (C) OF/1 model, average over OF/1.1 and OF/1.2. (C) IF/2 model (Frealix), average of simulations IF/2.1 and IF/2.2. (D) OF/2 model, average over OF/2.1, OF/2.2, and OF/2.3. [Adapted from Coudray et al. [7], Copyright © (2016), with permission from Wiley.]

while the DSSP profile indicated much greater stability in the S1 and S2 β-strand regions as well as the H3 and M11-H4 surface helix regions (Fig. 5.6C,D). The consistency of these analyses indicates not only that the (IF/2-based) OF/2 model retained the structural features of the Frealix real-space reconstruction and was of comparable quality to IF/2, but also that the real-space models were of superior quality as compared to the Fourier-Bessel-based IF/1 and OF/1 models.

5.3.3 Water accessibility analysis

In order to assess solvent accessibility, we computed water densities by mapping the locations of water molecules to a grid and also searched for cavities in the PDB models using the HOLE program; both approaches produced consistent results across the four models. As mentioned in Section 5.2, the water density calculations for chains A and B were consolidated by fitting chain B onto chain A using the non-loop C_α RMSD. Fig. 5.7A,C indicates the presence of intracellular funnels with higher-than-bulk density between the gate and core domains of the IF/1 and IF/2 models. Although there also appeared to be a funnel in the water densities on the periplasmic side, HOLE detected a deeper water-accessible cavity on the cytoplasmic side (Fig. 5.8A,C); however, a continuous water pathway was not visible from either method, indicating that IF/1 and IF/2 both represent inward-facing conformations.

Water densities from the OF/1 and OF/2 simulations were somewhat more ambiguous, exhibiting what appeared to be water-accessible funnels on both the intracellular and extracellular sides (Fig. 5.7B,D). Although OF/1 had somewhat higher densities than OF/2 in the extracellular region, a continuous water pathway through the protein was visible, symptomatic of the lower quality of the OF/1 model. In the case of OF/2, the cavity found by HOLE (Fig. 5.8D) was in agreement with the clear intracellular funnel in the AE1 structure (Fig. 5.8B). HOLE also also revealed (Fig. 5.8D) that water accessibility on the intracellular side of OF/2 was not as well-defined as in the IF/2 model. Taken together, the presence of intracellular and extracellular solvent-accessible cavities in the IF/2 and OF/2 models, respectively, along with the absence of continuous water pathways in both models, is fully compatible with the alternating access model for secondary transporters. Our analyses further support that the Bor1p EM density represents an inward-facing conformation rather than an outward-facing state as in the AE1 X-ray structure [319]. It should be noted, however, that both structures were solved in ligand-bound states (DIDS in AE1 and borate in Bor1p), which generally induces a partial closure of the binding site

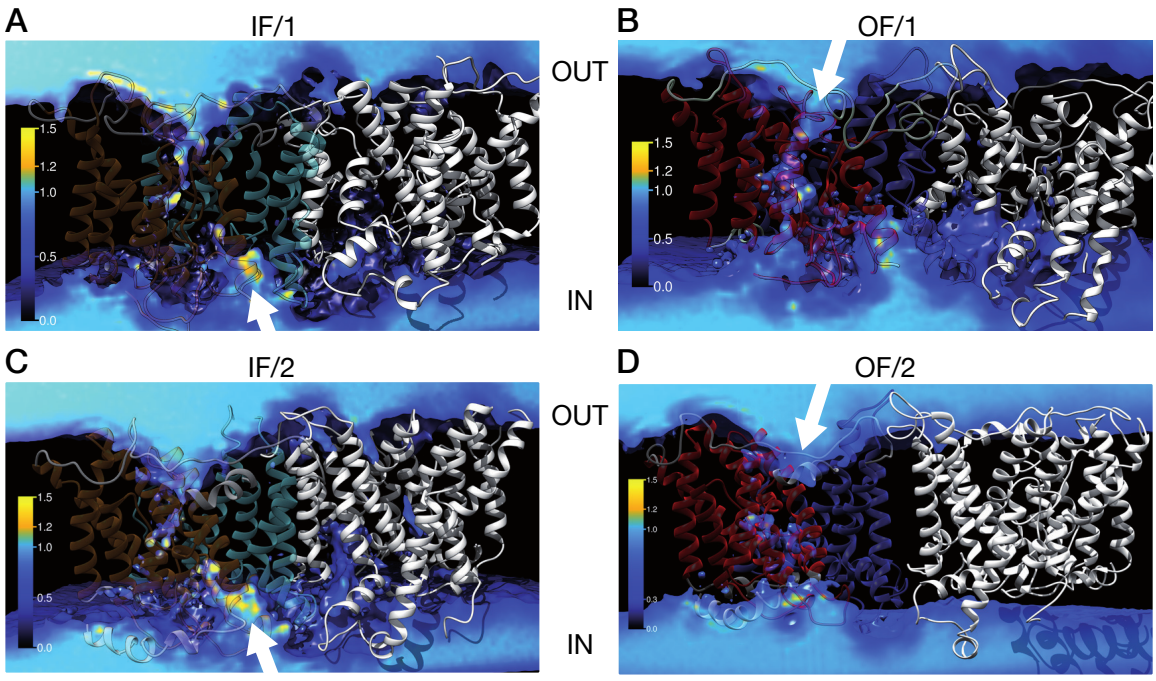

Figure 5.7: Maps of the density of water molecules during equilibrium MD simulations of Bor1p models measured relative to the bulk density of water. (A) Water density from MD simulations of IF/1, showing accessibility to the interior of the model from the cytoplasmic side ("IN") of the membrane. This map represents simulations IF/1.1 and IF/1.2 for which data from chains A and B were averaged over the second half of the trajectories. (B) Water density from MD simulations of the OF/1 model from the periplasmic side ("OUT") of the membrane determined from the last $\sim$ 270 ns of simulations OF/1.1 and OF/1.2 from both chains A and B. (C) Water density from MD simulations of the inward facing model of Bor1p (last $\sim$ 100 ns from simulations IF/2.1 and IF/2.2). White arrows indicate the entrance to the funnel. (D) Water density from MD simulations of the OF/2 as determined from the full simulations OF/2.1, OF/2.2, and OF/2.3 from both chains A and B. The density is measured relative to bulk water density at ambient conditions with the color code shown to the left. Transmembrane helices for gate and core domains are colored similar to panel A. [Adapted from Coudray et al. [7], Copyright © (2016), with permission from Wiley.]

more typical of an occluded state. Indeed, it is possible that the core domain may undergo larger rocking motions than the $\sim 10°$ rotation seen by direct comparison of the AE1 and Bor1p structures.

5.3.4 Domain movement and transport

By tracking the core and gate domain centers-of-mass during MD simulations of the IF/2 and OF/2 models, we were able to assess the putative domain movements that are likely to occur during an inward-facing to outward-facing transition. Adopting the membrane plane as a frame of reference, the z-positions of the domains fluctuated between 4 Å to 6 Å

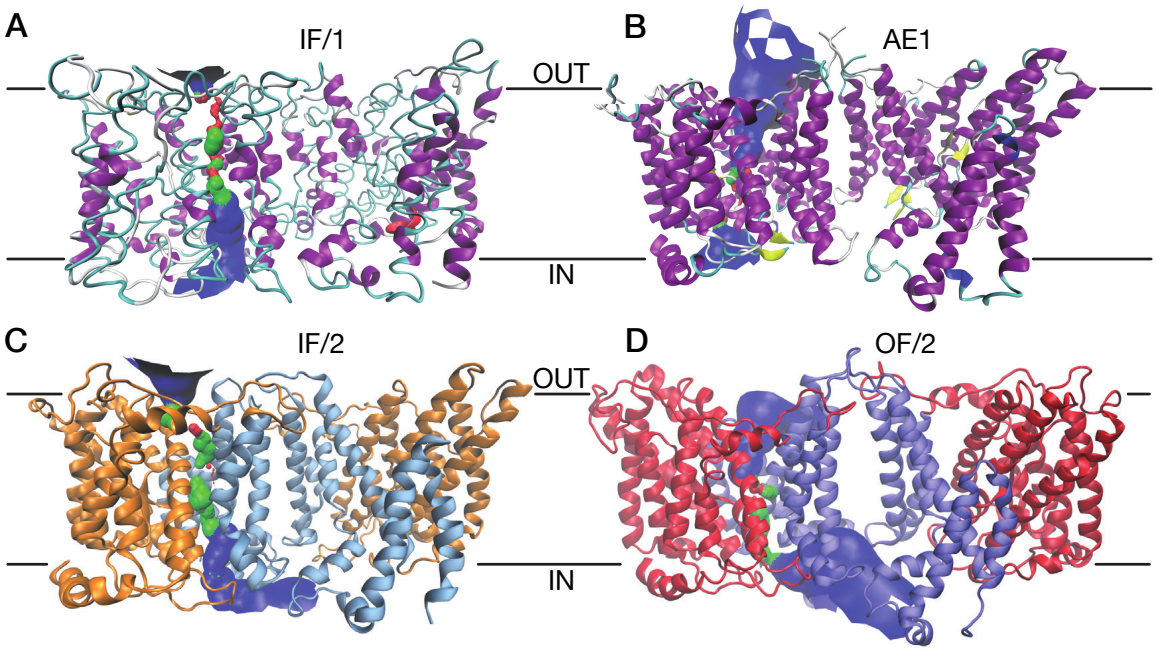

Figure 5.8: Solvent accessibility of Bor1p models as determined by HOLE [361]. (A) Water accessibility for the IF/1 model as determined by HOLE. The surface of a water accessible cavity is shown with colors indicating its diameter at each point along a putative pore: blue when radius is twice the minimum for a single water molecule, green when there is room for a single water molecule, red when the cavity is too narrow for water molecules. (B) Water accessibility for the AE1 X-ray structure (PDB 4yzf) determined by HOLE after removing the DIDS ligand from the structure. The large funnel from the extracellular side of the membrane is consistent with the outward-facing state. (C) Solvent accessible surface for inward-facing model of Bor1p (IF/2) determined by HOLE. In this case, the gate domain is light blue and the core domain is orange. (D) Solvent accessible surface for outward-facing model of Bor1p (OF/2) as determined by HOLE, which shows a funnel leading to the extracellular side of the membrane. The surface is colored blue (pore radius $R > 2.3$ Å), green (1.15 Å $\leq R \leq$ 2.3 Å) or red ($R \leq 1.15$ Å) depending on the width of the channel at each point. The gate domain is colored blue and the core domain red. [Adapted from Coudray et al. [7], Copyright © (2016), with permission from Wiley.]

about well-defined equilibrium positions (Fig. 5.9A,B). The equilibrium positions varied by conformation and in general can depend on protein-membrane interactions, as seen previously for the elevator-type NapA transporter [323]. In Fig. 5.9C, relative domain movements—calculated from the difference of position distributions between the inward- and outward-facing states after aggregating all simulation data over chains A and B—were very small, with mean shifts (and standard errors) of (-0.6 ± 0.0) Å and (-0.2 ± 0.0) Å for the gate and core domain, respectively.

As an elevator-like transport mechanism would require a large translocation of a domain across the membrane (≥ 6 Å), such small shifts are more consistent with a rocking-bundle

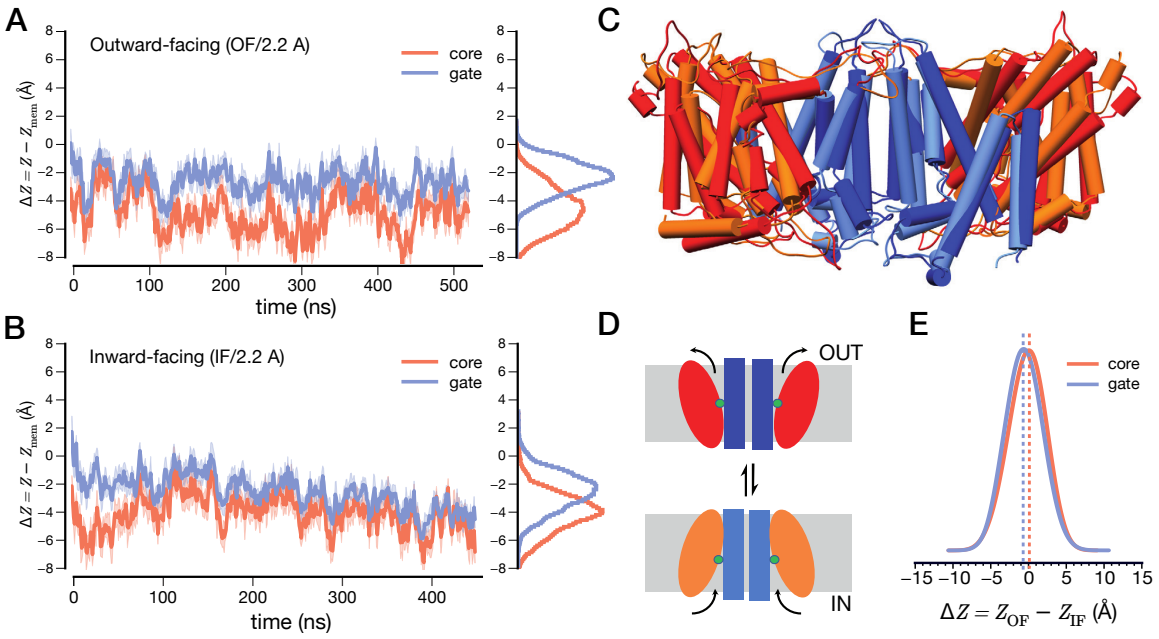

Figure 5.9: Domain shifts between inward-facing (IF) and outward-facing (OF) conformations of Bor1p calculated from core/gate domain z-displacements. (A) Representative time series (left) and histogram (right) of domain displacements of the OF/2.2 MD simulation (chain A) showing the the center of mass z-displacements of the core/gate (orange/blue) domain relative to the instantaneous center of mass of the lipid membrane. (B) Analogous data from chain A of the IF/2.2 simulation. (C) Overlay of IF and OF conformations; the gate domains (blue) are closely aligned, whereas the core domains of the IF/OF conformations (red/orange) show different inclinations without any noticeable vertical displacement. (D) Schematic illustration of the proposed (core) tilting during interconversion between IF and OF conformations. (E) Distribution of differences of gate/core displacements between IF and OF simulations (aggregated over both chains and all IF/2 and OF/2 MD runs). Both the core and gate shift by less than 1 Å relative to the membrane between IF and OF conformations and even less relative to each other. [Adapted from Coudray et al. [7], Copyright © (2016), with permission from Wiley / Adapted from Figs. 7 and S9.]

mechanism. The correspondence of the folds among SLC4 transporters and other members of the APC superfamily—several of which having solved structures for multiple states—provides insight into the putatively related conformational changes underlying the transport mechanism [8, 9]. In particular, several members of the APC superfamily, including the well-studied Mhp1, vSGLT and AdiC transporters (characterized by a 5 + 5 TM fold), exhibit shifts involving rocking-bundle motions of the substrate-binding (core) domain about to the opposing scaffold (gate) domain, similar to the rigid-body domain-rocking of both Bor1p and AE1. Indeed, APC superfamily transporters are generally presumed to operate by a rocking-bundle transport mechanism [9], which is consistent with our analysis of Bor1p domain motions based on equilibrium MD simulations (Fig. 5.9). In the case of

UraA, UapA, and SLC16Dg, superposing their gate domains on the outward-facing AE1 structure suggests that a displacement of the TMs 1, 3, 8, and 10 in the core domain of up to 10 Å is possible [321]. Identifying the transport mechanism is nevertheless difficult because, despite sharing the 7 + 7 inverted repeat topology of AE1 and Bor1p, their folds exhibit a different helix organization in the gate domain and their structures have only been solved in inward-facing conformations, making their alignment to AE1 is rather imprecise (as compared to Bor1p) [7]. Currently, the core domain translations that would be necessary for elevator-like transport in UraA, UapA, and SLC16Dg have yet to be directly observed, and our results support the notion that APC superfamily generally operate by a rocking-bundle transport mechanism.

5.4 Conclusions

An inward-facing structure of Bor1p from yeast was solved at 6 Å resolution within a reconstituted lipid membrane using cryo-EM. The arrangement of well-resolved TM helices in the cryo-EM map allowed the construction of a high fidelity homology model based on the atomistic outward-facing structure of AE1, a homologous SLC4 transporter with which Bor1p shares a 23% sequence identity. At this resolution, surface loops tend to be poorly defined, though the orientations of TM and several surface helices was sufficient to unambiguously determine the inward-facing orientation of Bor1p in the membrane. Two different types of tubular crystals were used to produce two completely independent EM density maps, after which MDFF was used to refine and prepare the structures for all-atom equilibrium MD simulations with an explicit membrane and solvent. Both MDFF-refined models had essentially the same α-helix orientations. Over the course of multiple microseconds of equilibrium simulation, substantial flexibility was observed in the loop regions; the TM helices and structural core of the model, however, remained sufficiently stable and well-defined throughout, thus retaining the substantial conformational difference between the Bor1p EM density and the outward-facing AE1 structure.

Recently, the structure of Bor1 was solved in an inward-facing conformation in the

absence of a bound ligand by X-ray crystallography and has been suggested to operate by an elevator-like mechanism [362]. As a bound substrate can trigger a shift to an occluded-like state, it is possible that the larger domain shifts necessary for an elevator mechanism would be evident in structures without bound ligands. Thurtle-Schmidt and Stroud [362] did not, however, simulate the equilibrium dynamics of the Bor1 structure and, without such dynamical information, it is difficult to draw conclusions about the conformational changes involved. In the absence of a bound ligand, AE1 and Bor1p are likely to adopt conformations that differ from the somewhat occluded states observed thus far. However, our results based on multi-microsecond equilibrium MD simulations are consistent with the rocking-bundle model, and so it remains to be determined whether an elevator-like mechanism describes SLC4 transporters and other members of the APC superfamily.

Given the wide range of dynamical time scales involved in secondary transport combined with the complexity of the solvent-bilayer environment, ascertaining the conformational motions involved remains a serious challenge to both experimental and computational approaches. Although the rapid, continuous improvement of cryo-EM-based imaging techniques has already enabled new insights into the equilibrium ensembles of transporters in a native-like environment, we anticipate that computational methods like flexible fitting and atomistic equilibrium MD simulations will become increasingly important rather than secondary to experiment. For instance, generating and examining plausible conformational transition paths will become an increasingly viable approach to studying transporters as the number of transporters solved in multiple conformations inevitably grows. Not only will such methods continue to provide intuition that can help interpret and guide experiments, atomistic simulations remain the only viable approach to accessing the dynamics involved in transport with atomistic spatiotemporal detail. Thus, the confluence of rapidly improving experimental and computational approaches will likely provide the greatest opportunity for insight.

Chapter 6

DEVELOPMENT OF A HYBRID ATOMISTIC-CONTINUUM METHOD

Recently, there has been growing interest in multiphysics simulation in which two or more disparate physical models are integrated into a unified numerical method that balances accuracy and efficiency [363]. One example is the *hybrid atomistic-continuum* (HAC) method, which applies an atomistic numerical method (e.g., MD or Monte Carlo [MC] [101]) to a restricted subdomain requiring atomistic resolution (e.g., an atomistic model of a protein or other solute) while using a relatively frugal fluctuating hydrodynamics (FHD) model to replaces some, or all, of the solvent [363–365]. By avoiding a fully explicit solvent representation whose water-water interactions would otherwise comprise most of the computational expense in explicit solvation MD, HAC approaches have the potential to capture crucial features of solute-solvent interactions and extend the reach of atomistic simulations to longer timescales. The research project described in this chapter—the development of a hybrid atomistic-continuum numerical simulation method with a view toward biomacromolecular simulation—was carried out for the 2016 Blue Waters Graduate Fellowship (BWGF) and was composed of two phases: (1) development of a custom FHD code for modeling dense fluids (i.e., aqueous solutions) based on the Landau-Lifschitz Navier-Stokes (LLNS) equations [366], a system of stochastic hydrodynamic equations—describing mass, momentum, and energy transport—subject to thermal fluctuations in the viscous stress tensor and heat flux vector; (2) design and implementation of a viable hybrid simulation code combining the FHD model developed in phase one with a flexible, open-source MD engine suitable for biomolecular simulation (e.g., LAMMPS [367], NAMD [352], OpenMM [164], or GROMACS [355]).

Though the implementation of a full HAC method was not expected to be achieved within a year-long timeframe, there were a number of tangible outcomes at the culmination of the BWGF. A novel FHD model and viable numerical solver was derived from a Fortran

90 code called PERSEUS, a numerical model of the eXtended MagnetoHydroDynamics (XMHD) equations for high-energy-density plasma simulation [368]. The PERSEUS-derived hydrodynamics code has been named HERMESHD (Hyperbolic Equations and Relaxation Model for Extended Systems of HydroDynamics) to distinguish it from original PERSEUS code. The HERMESHD code organically developed from a basic implementation of the LLNS system into fully fledged extension of LLNS designed to better capture physics expected to arise at the nanoscale; in addition, a Python-based interface was developed for the Fortran-based HERMESHD code that allowed FHD simulations to not only be executed from a Python environment, but also integrated with the Python interface to LAMMPS to implement a proof-of-concept HAC method. The HERMESHD model and numerical code are detailed in Section 6.2. Perhaps the most valuable outcome of the project was a deeper intuition of hydrodynamic effects at the nanoscale in dense fluids and, more generally, the relationship between atomistic and continuum perspectives (i.e., the continuum limit). This chapter begins by briefly summarizing the history of the PERSEUS numerical code and motivates the purpose of HERMESHD in going beyond the LLNS formulation. recapitulates important project milestones and an in-depth discussion of essential theoretical and numerical aspects of the HERMESHD model, including several detailed derivations. A rough road map is given that serves to motivate future investigation.

6.1 Motivation

Proteins (and other biological macromolecules) largely exist in ionic aqueous conditions and, in the case of membrane proteins, a phospolipid bilayer. Restricting attention to the case of globular proteins like enzymes, for simplicity, explicit solvent all-atom EqMD (using the molecular mechanics approach) arguably remains the most robust method equilibrium simulation method that can capture solute-solvent interactions with atomistic detail [369]. At the cost of reduced detail and, presumably, some degree of accuracy, an implicit solvent model can be used in lieu of explicit representations of the water molecules, which would

otherwise be specified by a *water model* [370–372]. Implicit solvent models* are typically static continuum representations that attempt to capture the mean influence of the solvent on the solute and can accelerate simulations by up to one or two orders of magnitude, depending on the physical process being studied [378].

To generate constant-temperature dynamics for equilibrium implicit solvent simulations, particles are typically driven by Langevin dynamics—a dynamical model balancing viscous dissipation and stochastic forcing—to account for the eliminated solvent degrees of freedom and thus simulate an idealized heat bath. When correctly implemented, Langevin dynamics can offer an effective means of temperature control and equilibrium sampling in the absence of a solvent [379, 380], though care must be taken in making judgements about nonequilibrium processes since larger-scale motions tend to be overdamped compared to EqMD [381, 382]. Given the limitations of implicit solvents (e.g., static representation, secondary structure biases, poor reproduction of synthetic FRET diagnostics [383–386]) and Langevin dynamics (e.g., linear model restricted to a local equilibrium assumption, damping of larger-scale motions [381, 382]), a sensible approach may be to augment the static continuum model with one that captures nonlinear flows and the resultant correlations arising from hydrodynamic effects.

The extent to which hydrodynamics and hydrodynamic fluxes are important in describing such processes are subtle. As the small length and velocity scales encountered in biomacromolecular systems place the dynamics well into the Stokes regime, where the Reynolds number, Re, falls well below unity, diffusive dynamics would seem to be dominant. However, it has been known for some time that inertial forces can be dominated by solute-solvent coupling at small scales, as evidenced by observations of "long-time tails" (i.e., power-law decay) in the velocity autocorrelation function [387–389], as well as solvent influences the slow internal dynamics of proteins suggested by the comparison of MD

*More specifically, implicit solvent models estimate the solvation free energy from either geometric properties of the solute, like the solvent-exposed surface area (SASA), or using a continuum electrostatics approach and solving the Poisson-Boltzmann equation [373] (or the linearized Generalized Born [GB] theory) to calculate the electrostatic solvation free energy [369, 374–377].

simulations of solvated and unsolvated lysozyme [390].; for short-time velocity fluctuations, such coupling can also overshadow steady-state Stokes damping and necessitate the inclusion of hydrodynamic interactions in Brownian dynamics [389]. Indeed, the assumption of Stokes damping is only valid for flows that have relaxed to steady, constant-velocity motion and does not account for the energy of the fluid set in motion by the acceleration of an immersed particle—the memory of a particle's acceleration retained by the vorticity field of a viscous fluid [391, 392]. For an accelerating spherical particle under the assumption of creeping flow, additional terms are required to account for differences in its velocity and that of the surrounding fluid, leading to a transient hydrodynamic force known as the Boussinesq-Basset force [393, 394]. In particular, one study found that thermal motions of DNA are significantly affected by inertial effects, particularly with respect to configurational transitions in equilibrium [395].

6.1.1 The hybrid atomistic-continuum approach

Multiphysics simulation approaches can provide a compromise between explicit and implicit solvent representations in terms of accuracy and efficiency, helping to extend the accessible spatiotemporal scales [364]. The hybrid atomistic-continuum (HAC) method can be used to replace most of the solvent in MD simulations with a relatively frugal fluctuating hydrodynamics (FHD) model that retains much of the solute-solvent physics; in the subdomain containing, say, a protein, an all-atom force field can be used to capture the heterogeneities in its structure. The separate atomistic and hydrodynamic codes must also have some sort of overlap region where the simulations can exchange boundary conditions; the question of how this should be handled is nontrivial and is a topic of intensive research [363, 365, 396].

HAC methods have already become valuable tools to study microelectromechanical systems (MEMS) such as microfluidic devices [397, 398], polymer melts and shear flows [399, 400], and macromolecular dynamics (e.g., enzyme dynamics or molecular motors) [401–403]. The simplest HAC methods employ Brownian dynamics, though such models

are incapable of reproducing many effects[†] seen in experiment and all-atom simulations, prompting a need to incorporate hydrodynamic effects [388, 402, 404]. The basic idea is conveyed in Fig. 6.1, which schematically shows one possible scheme for hybridizing particle-based (orange lines, boxes) and continuum-based (purple lines, boxes) numerical codes using a top-level driver or *umbrella* program that facilitates communication and manages data between two independent codes. In Section 6.3, we revisit Fig. 6.1 in greater detail as it pertains to the HAC method that was developed for the BWGF project.

6.1.2 LLNS and beyond

To construct a proper HAC method for protein simulation, where length scales are on the order of nanometers, it is necessary to augment the Navier-Stokes (NS) equations—a *macroscopic* hydrodynamic model—with the effects of thermal fluctuations [402, 405, 406]. This procedure was first explored by Landau and Lifschitz [366], where stochastic hydrodynamic equations describing mass, momentum, and energy transport were obtained by incorporating fluctuations in the viscous stress tensor and heat flux vector with correlations obeying linear fluctuation-dissipation [366, 407, 408]. This fluctuating hydrodynamics model, commonly known as the Landau-Lifschitz Navier-Stokes (LLNS), contains the familiar linear constitutive relations of the macroscopic Navier-Stokes system: Newton's law of viscosity (stress is proportional to velocity gradients) and Fourier's law (heat flux is proportional to the temperature gradient). Hydrodynamic timescales in the Landau-Lifschitz hydrodynamic theory are assumed to be "long" compared to the microscopic mean collision time, which leads to Gaussian white noise fluctuating terms that are spatiotemporally uncorrelated.

Intuitively, it can be seen that LLNS will have a limited domain of applicability as the nanoscale is approached: if one substitutes the aforementioned constitutive relations

[†]Brownian dynamics implicitly assumes a large timescale separation between momentum and configurational degrees of freedom—i.e., a separation between the molecular (kinetic) collision timescale and the hydrodynamic timescale—that predicts exponential relaxation in the velocity autocorrelation function [**Muser2013**].

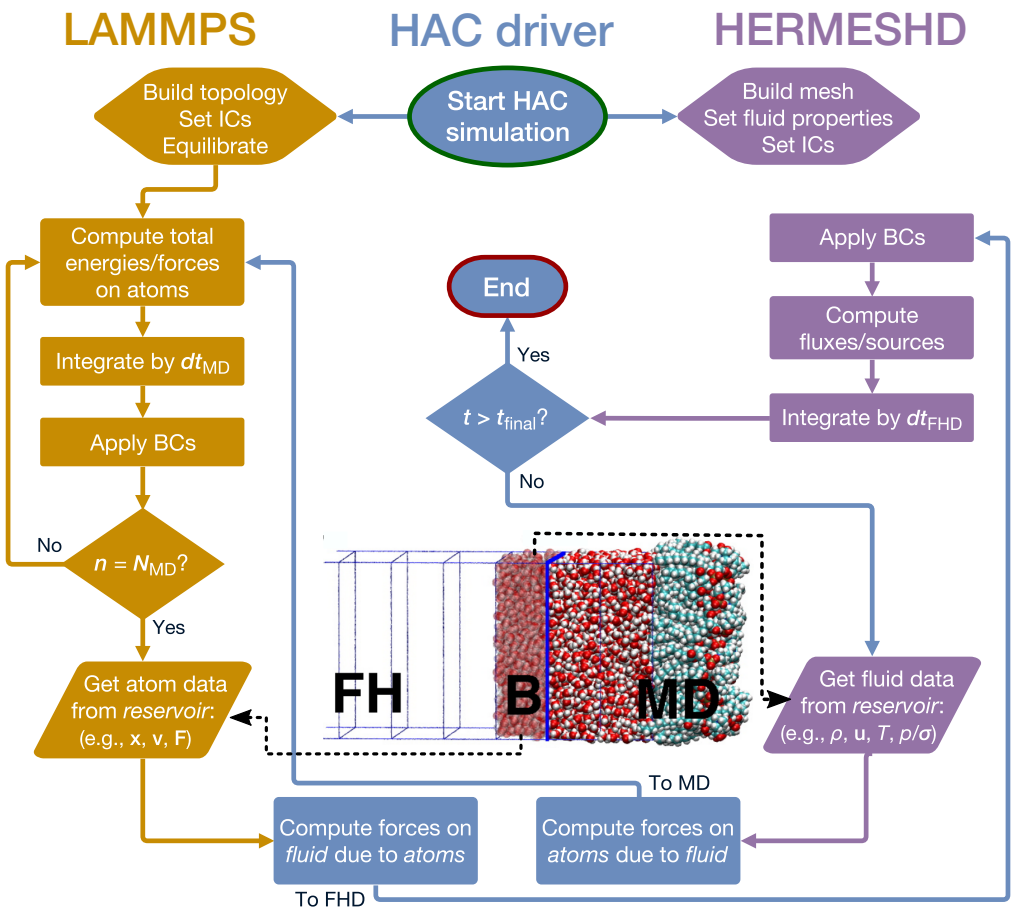

Figure 6.1: A schematic depiction of a hybrid time step involving the coupling of HERMESHD to an MD code, such as LAMMPS, via a HAC driver/umbrella code. The Pythonic HAC driver runs a hybrid simulation by communicating with and managing the exchange of data between MD (LAMMPS) & FHD (HERMESHD); atomistic and field (fluid) data in the reservoir region (buffer, B) is extracted and sent by respective codes to HAC driver for exchange. [Image of FH–B–MD schematic (with fluid cells and overlapping MD particles) from Rafael Delgado Buscalioni, *Hybrid MD setup: Molecular Dynamics coupled to Fluctuating Hydrodynamics*; downloaded from http://www.uam.es/personal_pdi/ciencias/rdelgado/setupnew.jpg on November 2017.]

into the momentum and energy equations, it can be seen that the resulting LLNS system contains second-order spatial derivatives, making the LLNS system semi-parabolic. In particular, the NS and LLNS contain parabolic laws of heat conduction and shear diffusion, leading to a paradox in which localized perturbations in temperature and shear velocity spread instantaneously throughout the entire spatial domain [409, 410]. Such a situation is not usually a problem for many physical systems encountered by physicists and engineers; however, when modeling mesoscale or nanoscale phenomena such as those involved in

protein dynamics, transport speeds should be expected to appear finite and, possibly, comparable to other characteristic physical scales.

Though LLNS has proven a powerful approach to modeling many nanoflows, aforementioned assumptions are expected to break down for dense fluids, especially as simulation grid cells approach nanometer dimensions (i.e., a water molecule and hydrodynamic timescales of interest become comparable to collision times. The main scaling parameter quantifying this breakdown is the Knudsen number,

$$\mathrm{Kn} = \lambda / L, \tag{6.1}$$

which is the ratio of the mean free path, λ, to a characteristic length scale, L (i.e., the dimension of a simulation grid cell); Navier-Stokes is valid in the limit $\mathrm{Kn} \ll 1$, though it will begin to fail as a description of certain processes as Kn approaches unity [411–413]. Indeed, the mean free path of liquid water is comparable to the length scales of interest, implying that LLNS may be an inadequate hydrodynamic description for HAC applications.

In extending the NS and LLNS equations in the subsequent section, it will be useful to keep ideas from kinetic theory in view, namely that the hydrodynamic equations can be interpreted as *statistical moments* in the velocity distribution of the Boltzmann transport equation [414–416]. In particular, the five hydrodynamic equations for (scalar) mass, (vector) momentum, and (scalar) energy transport in NS/LLNS can be derived from kinetic theory by taking, respectively, the zeroth, first, and second statistical moments over the velocities of the Boltzmann transport equation (BTE). Thus, higher-order (in a sense) systems of hydrodynamic systems can be constructed by taking higher-order statistical velocity moments of the BTE. A relatively well known example is the 13-moment approach pioneered by Grad [417], which includes (beyond the five Navier-Stokes equations) five additional equations describing viscous stress transport (i.e., off-diagonal tensorial components from the second velocity moment, apart from the scalar energy) and three further equations for the transport of heat flux (taking the third moment) [413, 417, 418]. Grad's classical moment equations have the property of *hyperbolicity* (at least, in the near-equilibrium regime), which

guarantees finite speeds of propagation [409, 419–421]. Grad's approach is not unique, and the 13-moment equations (G13, for brevity) have been obtained through other techniques, such as the Chapman-Enskog expansion [413, 422–424].

6.2 From the Euler equations to HERMESHD

The first phase of development was focused on constructing a viable approach to incorporating a fluctuating hydrodynamics model within the existing structure of PERSEUS (Physics of the Extended-mhd Relaxation System using an Efficient Upwind Scheme), a numerical 3D finite-volume method (FVM) designed to model dense Z-pinch experiments generating plasmas whose energy and density scales span many order of magnitude [368]. PERSEUS formulates the extended magnetohydrodynamics (XMHD) model—closely related to two-fluid formulations based on the generalized Ohm's law (GOL) [425–427]—as a hyperbolic relaxation model to capture Hall physics and other beyond-MHD effects using time steps much larger than otherwise demanded by the high frequency dynamics. The original FVM-based PERSEUS was subsequently upgraded to a discontinuous Galerkin (DG) spatial discretization that further improved accuracy and efficiency for high-energy-density (HED) plasma applications [428] and has been applied to the study of shock structures in supersonic plasma flows past an obstacle [429].

The DG implementation solves the governing transport equations in *conservation form* (as with the original finite-volume formulation), several advantages of which include shock capturing and higher overall accuracy in calculating the conserved quantities. PERSEUS is already capable of solving the inviscid, compressible Euler equations used in MHD, but the need to model dense fluids necessitated a means to incorporate both viscous forces and thermal transport. The well-known Navier-Stokes equations differ from the Euler equations primarily by the addition of a term containing a viscous stress tensor (sometimes called the Cauchy stress tensor in continuum mechanics), the form of which is specified through a constitutive relation—namely, Newton's law of viscosity—relating the gradients of the fluid velocity to the stress components [430]. When diffusive heat transport is relevant, the

Navier-Stokes system can be supplemented with Fourier's law of heat conduction, leading to the Navier-Stokes-Fourier equations. Starting from the Euler equations, the Navier-Stokes equations are derived, followed by a description of how viscosity and heat flux (resp. momentum and heat transport) are managed by HERMESHD. Finally, a numerical "trick" of sorts—constructing a so-called *hyperbolic relaxation system*—will be used to derive the hydrodynamic equations in HERMESHD (cf. S. Jin, Pareschi, and Slemrod [431]).

6.2.1 The Euler equations

The Euler equations describe the flow of an inviscid, compressible, adiabatic fluid (i.e., an ideal fluid), and take the form of five conservation-type equations representing, respectively, the conservation of mass, vector momentum, and energy,

$$\partial_t \rho + \boldsymbol{\nabla} \cdot (\rho \mathbf{u}) = 0 \tag{6.2a}$$

$$\partial_t (\rho \mathbf{u}) + \boldsymbol{\nabla} \cdot \left(\rho \mathbf{u}\mathbf{u} + p\vec{\vec{\mathbf{1}}} \right) = 0 \tag{6.2b}$$

$$\partial_t \mathcal{E} + \boldsymbol{\nabla} \cdot (\mathbf{u}(\mathcal{E} + p)) = 0. \tag{6.2c}$$

The corresponding fluid is described in terms of five independent field quantities for the (mass) density ρ, momentum density $\rho\mathbf{u}$, and total energy density $\mathcal{E} = \rho e + \rho u^2/2$, where $\mathbf{u}$ is fluid velocity, p is the thermodynamic pressure, e is the specific internal energy (per unit mass) of the fluid, ρe is the internal energy density, and $\rho u^2/2$ is the kinetic energy density. Note that the right-hand sides of all the equations are zero—there are no source/sink terms describing additional sources of mass, momentum, or energy. Thus, the rates of change of the field quantities are identically balanced by divergences of their respective fluxes.

As dissipation is absent in an ideal fluid, the Euler equations can also be expressed using a continuity equation for the conservation of entropy density,

$$\partial_t (\rho s) + \boldsymbol{\nabla} \cdot (\rho s \mathbf{u}) = 0, \tag{6.3}$$

instead of the energy equation, Eq. 6.2c, where s is the specific entropy (per unit mass). Eq. 6.3 describes so-called *isentropic* flow, which says that the entropy density is conserved

along fluid streamlines. Though the entropy equation is not directly used in HERMESHD (at least in its current form), it is included as a point of contact to the concept of entropy balance in nonequilibrium thermodynamic theories. In particular, a nonequlibrium process would lead to source terms on the right-hand side of Eq. 6.3 corresponding to entropy *production*.‡

6.2.2 The Navier-Stokes equations

The Navier-Stokes equations can be viewed as an extension to the Euler equations that incorporates viscous effects, specifically the dissipation of momentum and energy due to shear flow, which can be respectively added to Eq. 6.2b and Eq. 6.2c as source terms to give

$$\partial_t \rho + \boldsymbol{\nabla} \cdot (\rho \mathbf{u}) = 0 \tag{6.4a}$$

$$\partial_t (\rho \mathbf{u}) + \boldsymbol{\nabla} \cdot \left(\rho \mathbf{u}\mathbf{u} + p\vec{\vec{\mathbf{1}}} \right) = \boldsymbol{\nabla} \cdot \vec{\vec{\sigma}} \tag{6.4b}$$

$$\partial_t \mathcal{E} + \boldsymbol{\nabla} \cdot (\mathbf{u}(\mathcal{E} + p)) = \boldsymbol{\nabla} \cdot (\vec{\vec{\sigma}} \cdot \mathbf{u}). \tag{6.4c}$$

While the continuity equation remains the same, the momentum and energy equations acquire additional terms on their right-hand sides that introduce a new quantity, $\vec{\vec{\sigma}}$, called the Cauchy stress tensor, the pressure tensor, or sometimes just the stress tensor [430]. The stress tensor captures all sources of stress at a given point in space. For instance, the tendency of neighboring fluid streamlines with different velocities to relax to the same velocity is due to shear stresses (forces per unit area) that are described by the off-diagonal components of the stress tensor—this "drag" on adjacent fluid layers is due to the microscopic exchange of momentum when velocity gradients are present, the result of which are dissipative stresses that depend on the material properties, i.e., microscopic molecular details, of the fluid. A full description of a fluid therefore requires the inclusion

‡It is usually demanded that the entropy production, Σ, be non-negative, which corresponds to the second law of thermodynamics and the maximization of the *global* entropy, S, in equilibrium. A central objective in contemporary theories of nonequilibrium thermodynamics is to extend the equilibrium formulation by expressing Σ in terms of quantities characterizing specific irreversible processes [409, 410].

of a constitutive equation for the *viscous* stresses,

$$\vec{\vec{\sigma}}' = \zeta(\boldsymbol{\nabla}\cdot\mathbf{u})\vec{\vec{\mathbf{1}}} + \eta\left(\boldsymbol{\nabla}\mathbf{u} + (\boldsymbol{\nabla}\mathbf{u})^{\intercal} - \frac{2}{3}(\boldsymbol{\nabla}\cdot\mathbf{u})\right), \tag{6.5}$$

where ζ and η are the bulk and (dynamic) shear viscosities, the prime on $\vec{\vec{\sigma}}'$ indicating that stresses due to the thermodynamic pressure, p, are excluded. Thus, the full stress tensor, $\vec{\vec{\sigma}}$, is the sum of the thermodynamic pressure and viscous stresses,

$$\vec{\vec{\sigma}} = -p\vec{\vec{\mathbf{1}}} + \vec{\vec{\sigma}}'. \tag{6.6}$$

It is worth noting that the bulk viscosity modulates an isotropic term, while the shear viscosity sets the scale of *deviatoric* (non-isotropic, symmetric) contributions to the stress. Alternatively, the full stress tensor is sometimes decomposed into isotropic and deviatoric components,

$$\vec{\vec{\sigma}} = \pi\vec{\vec{\mathbf{1}}} + \vec{\vec{\mathbf{s}}}, \tag{6.7}$$

where $\pi = (1/3)\,\mathrm{Tr}\,\vec{\vec{\sigma}}$ is the mean stress and $\vec{\vec{\mathbf{s}}}$ is the (traceless) stress deviator tensor; lastly, subtracting the contribution of bulk viscosity from the mean stress, π, gives

$$p = -\frac{1}{3}\,\mathrm{Tr}\left(\vec{\vec{\sigma}} - \zeta(\boldsymbol{\nabla}\cdot\mathbf{u})\vec{\vec{\mathbf{1}}}\right) = \zeta(\boldsymbol{\nabla}\cdot\mathbf{u}) - \pi, \tag{6.8}$$

which is the usual thermodynamic pressure. Note that p manifestly does *not* arise from the fluid velocity field, $\mathbf{u}$, which is an averaged quantity; the trace of the full stress tensor extracts the contributions from diagonal elements so that only thermodynamic sources of pressure remain once contributions from velocity field divergences (due to bulk viscous effects) are removed.

6.2.3 The Navier-Stokes-Fourier equations

To account for the flow of energy due to the conduction of heat, we introduce the heat flux vector,

$$\mathbf{q} = -\kappa\boldsymbol{\nabla}T. \tag{6.9}$$

Eq. 6.9 is the familiar Fourier's law, where κ is the thermal conductivity of the fluid. When thermal conduction is non-negligible, the rate at which energy can flow into (or out of) a point in space will depend on the divergence of the heat flux vector. Thus, Eq. 6.4c will acquire an additional term, i.e.,

$$\partial_t \mathcal{E} + \boldsymbol{\nabla} \cdot (\mathbf{u}(\mathcal{E} + p)) = \boldsymbol{\nabla} \cdot (\vec{\vec{\sigma}} \cdot \mathbf{u}) + \boldsymbol{\nabla} \cdot \mathbf{q}, \tag{6.10}$$

where the heat flux, q, is defined by Eq. 6.9. The mass (Eq. 6.4a), momentum (Eq. 6.4b), and energy (Eq. 6.10) equations, taken together, are the equations of motion for a compressible, viscous, thermally conducting fluid and are sometimes referred to as the *Navier-Stokes-Fourier* (NSF) equations. It is important to mention that the NSF system is only strictly valid under the assumption of local thermodynamic equilibrium or LTE—sometimes referred to as the local-equilibrium hypothesis [410].

6.2.4 Hyperbolic relaxation system

To help motivate the additional equations in the HERMESHD hydrodynamic model that go beyond Navier-Stokes, we examine an illustrative 1D hyperbolic relaxation system. As mentioned in Section 6.1, the NSF equations contain second-order spatial derivatives in the velocities and temperature (or, equivalently, the energy), which is due to the velocity and temperature gradients in the constitutive laws for viscosity and thermal conduction (resp. Eq. 6.5 and Eq. 6.9). To circumvent architectural difficulties in handling second-order spatial derivatives, HERMESHD uses a kind of numerical "trick", which involves casting the momentum and energy equations from the original (semi-parabolic) NSF system into the form of a hyperbolic relaxation system (HRS). In effect, HERMESHD solves a generalization of the following 1D HRS,

$$\partial_t u + \partial_x \gamma = 0 \tag{6.11}$$

$$\epsilon \partial_t \gamma - D \partial_x u = \gamma, \tag{6.12}$$

where u is a conserved quantity (e.g., density), γ is an auxiliary dynamical variable, D represents a diffusion coefficient, and ϵ is a (small) relaxation parameter [368]. In the fast

relaxation limit ($\epsilon \to 0$), the system reduces to a parabolic diffusion-type problem for u,

$$\partial_t u = D\partial_x^2 u. \tag{6.13}$$

An HRS as in Eq. 6.11 can thus be used to reproduce the diffusive dynamics in Eq. 6.13 by inputting small values of ϵ, which forces the simulated dynamics to relax to the equilibrium solution in the fast relaxation time limit.

The relaxation model effectively sidesteps the explicit evaluation of second-order spatial derivatives (that would otherwise be required in directly solving for the usual expression for Navier-Stokes), but requires integrating six auxiliary dynamical variables in time—one per independent stress tensor component, beyond the usual five (mass density, three momentum densities, and energy density). The hyperbolic relaxation system currently employed in the HERMESHD code contains linearized (relaxation-type) equations for the tensorial deviatoric stress (Eq. 6.15d) and vectorial heat flux (Eq. 6.16e), which are connected to the so-called Bhatnagar-Gross-Krook (BGK) [432, 433] model of the Boltzmann equation. The idea behind the BGK approach was to find solutions to the nonlinear integro-differential Boltzmann equation by replacing the nonlinear collision term, $(\partial_t f)_{\text{coll}}$, with a linear relaxation approximation to obtain

$$\left(\frac{\partial f}{\partial t}\right)_{\text{coll}} \sim -\frac{(f - f^0)}{\tau}, \tag{6.14}$$

where $f = f(\mathbf{x}, \mathbf{v}, t)$ is single-particle distribution function (from the Boltzmann equation), f^0 is the (stable) equilibrium distribution, and τ is a characteristic timescale pertaining to the relaxation of the fluid to the equilibrium distribution f^0. It is clear by inspection that Eq. 6.14 has an analogous form to Eq. 6.15d and Eq. 6.16e. Indeed, the hyperbolic relaxation perspective was not so much a numerical "trick" for reproducing the Navier-Stokes equations (in the relaxation limit where $\tau \to 0$) as was suggested at the outset, because it mirrors a number of rigorous approaches presented in the literature that have been used to derive higher-moment hydrodynamic equations, such as the use of the Knudsen number, Kn, as the expansion parameter in Chapman-Enskog theory [423, 431, 434, 435].

6.2.5 *Compressible flow with shear stresses and dissipation*

It can be seen that the Navier-Stokes equations are of the parabolic type because second-order spatial derivatives show up when dissipative stresses and thermal conduction are included—the constitutive laws for viscous stress (Eq. 6.5 and Eq. 6.6) and heat flux (Eq. 6.9) are expressed, respectively, as gradients of the velocities and temperature. HERMESHD (and PERSEUS) do *not* directly solve the Navier-Stokes equations in the form of Eq. 6.4 in large part to avoid computing second-order spatial derivatives (in velocities and temperature); both codes are designed to solve *hyperbolic* systems of equations expressed in conservation form after re-casting the Navier-Stokes or LLNS system in terms of first-order derivatives in space and time. In particular, HERMESHD uses a hyperbolic relaxation system (HRS) representation and an implicit-explicit time integration scheme to "drive" the numerical solution toward the equilibrium solution represented by the parabolic system of equations in Navier-Stokes/LLNS. The relaxation from a hyperbolic to a parabolic system is achieved in the limit as a relaxation time parameter, τ, becomes small relative to hydrodynamic "observation" timescales. The HRS approach, in effect, trades second-order spatial derivatives (and the computational expense to compute them) for an additional set of dynamical variables, each described by its own balance equation. Furthermore, the HRS representation permits the five-moment Navier-Stokes/LLNS system of equations to be naturally extended to 13-moment (or higher) hydrodynamic equations. In comparison to alternative approaches, the structure of HERMESHD is conducive to efficient simulations of dense fluids at the nanoscale, and potentially with greater physical accuracy than the conventional five-moment LLNS model.

In the absence of bulk viscous effects, the full stress tensor reduces to the shear stress tensor, $\vec{\vec{\sigma}}'$. Since $\vec{\vec{\sigma}}'$ is symmetric and traceless, the HRS must include five additional dynamical field variables: three for the off-diagonal stresses, two for the diagonal components under the traceless condition. Thus, five dynamical equations govern shear stresses in in the HRS representation—in addition to the mass, momentum, and energy equations—for a

total of 10 dynamical equations,

$$\partial_t \rho + \boldsymbol{\nabla} \cdot (\rho \mathbf{u}) = 0 \tag{6.15a}$$

$$\partial_t (\rho \mathbf{u}) + \boldsymbol{\nabla} \cdot \left(\rho \mathbf{u}\mathbf{u} + p \vec{\vec{\mathbf{1}}} \right) = 0 \tag{6.15b}$$

$$\partial_t \mathcal{E} + \boldsymbol{\nabla} \cdot (\mathbf{u}(\mathcal{E} + p)) = 0 \tag{6.15c}$$

$$\partial_t \vec{\vec{\sigma}} + \frac{\eta}{\tau_\nu} \left(\boldsymbol{\nabla}\mathbf{u} + (\boldsymbol{\nabla}\mathbf{u})^\intercal - \frac{2}{3} (\boldsymbol{\nabla} \cdot \mathbf{u}) \vec{\vec{\mathbf{1}}} \right) = -\frac{1}{\tau_\eta} \vec{\vec{\sigma}}. \tag{6.15d}$$

The tensor equation for the stress components (Eq. 6.15d) is the linearized equation from the full pressure-tensor moment equation containing the deviatoric stresses *and* the scalar energy. The linearized form of Eq. 6.15d is obtained using a closure relation that specifies the phenomenological relaxation time parameter time τ_η. While it is possible that additional physics can be extracted from the full nonlinear pressure moment equation, the present study only considers linear effects, which already include potentially important memory (non-Markovian) effects that the Navier-Stokes equations do not capture. It can also be seen that the second term in the equations for shear stress is exactly the right-hand side of Eq. 6.5 with $\zeta = 0$ and contains only first-order spatial derivatives of the velocity.[§] The relaxation parameter, τ, sets the timescale of viscous dissipation and can be associated with a microscopic collision time. It can be seen that the time derivatives of the shear stress components vanish in the small τ limit and we are left with the usual shear stress components of the Cauchy stress tensor.

A slight complication arises after the spatial domain is discretized into (rectangular) finite-volume grid cells, because the fluxes at each of the six faces of a cell will, in general, differ—the faces of a cell are not at the same point in space, so the traceless condition is not exactly satisfied after discretization. The five shear stress equations, Eq. 6.15d, must be supplemented with an additional independent variable/equation to handle the diagonal components (instead of just two equations plus the traceless condition). In total,

[§]The shear stresses, written in tensor notation in Eq. 6.15d, are not quite in conservation form. However, since only first-order spatial derivatives (of the velocities) are present and since the full system of equations (Eq. 6.15a–d) is hyperbolic, the shear stress "fluxes" can be integrated analogously to the mass, momentum, and energy fluxes in Eq. 6.15a–c. The advantage of this structure is simplicity.

HERMESHD integrates 11 dynamical equations to solve for 11 dynamical field variables: 5 for the usual mass, momentum and energy densities, 3 for the off-diagonal shear stress components, and 3 for the diagonal shear stress components.[¶]

Inclusion of bulk viscous dissipation

The bulk viscosity, ζ, captures dissipative viscous effects caused by *isotropic* changes in the velocity field, i.e., the divergence of the velocity field, $\mathbf{u}$.[||] Bulk viscous effects are relevant when compressibility effects are important; for example, bulk viscosity is responsible for the attentuation of sound waves and is important in describing the hydrodynamics of liquids containing gas bubbles [430]. In situations where bulk viscosity is non-negligible, an isotropic contribution in Eq. 6.5 manifests in the spatial derivatives for the diagonal stress components in Eq. 6.15d. In general, the quantities η and ζ are functions of temperature and pressure and ζ is usually of the same order of magnitude as η, though there are cases where ζ may exceed η considerably [366]; as such, the relaxation parameter, τ_η, may differ (or be insufficient) when bulk viscous effects are present due to the introduction of a separate timescale associated with bulk dissipation. Appendix G also provides some detail as to how bulk viscosity manifests in fluctuating hydrodynamics.

Inclusion of thermal conduction

When thermal conduction is important, we must look to Eq. 6.10 to incorporate heat flux. However, as with the stress tensor, a similar situation is encountered where second-order spatial derivatives of the temperature field are introduced upon substitution of Fourier's Law (Eq. 6.9), albeit this time (only) in the energy equation (Eq. 6.10). To solve

[¶] A thought: it might be possible to satisfy the traceless condition *identically* through some clever algorithmic means; for example, one could look at techniques in plasma codes that are used to satisfy Gauss' law for magnetic fields, $\boldsymbol{\nabla} \cdot \mathbf{B} = 0$. Such an approach would probably have to account for the manner in which numerical fluxes are calculated at each cell interface—it is not clear whether it would be desirable to do this, even if it were feasible.

[||] The bulk viscosity is alternatively referred to as the volume viscosity, expansion viscosity, or second coefficient of viscosity [430].

Eq. 6.10 using the hyperbolic relaxation approach in HERMESHD, three new dynamical field variables must be introduced—one for each vector component of the heat flux—leading to an auxiliary *vector* equation analogous to the tensor equation for shear stress (Eq. 6.15d). The full 13-moment HRS becomes

$$\partial_t \rho + \boldsymbol{\nabla} \cdot (\rho \mathbf{u}) = 0 \tag{6.16a}$$

$$\partial_t (\rho \mathbf{u}) + \boldsymbol{\nabla} \cdot \left(\rho \mathbf{u}\mathbf{u} + p \vec{\vec{\mathbf{1}}} \right) = 0 \tag{6.16b}$$

$$\partial_t \mathcal{E} + \boldsymbol{\nabla} \cdot (\mathbf{u}(\mathcal{E} + p)) = 0 \tag{6.16c}$$

$$\partial_t \vec{\vec{\sigma}} + \frac{\eta}{\tau_\eta} \left(\boldsymbol{\nabla}\mathbf{u} + (\boldsymbol{\nabla}\mathbf{u})^\intercal - \frac{2}{3} (\boldsymbol{\nabla} \cdot \mathbf{u}) \vec{\vec{\mathbf{1}}} \right) = -\frac{1}{\tau_\eta} \vec{\vec{\sigma}} \tag{6.16d}$$

$$\partial_t \mathbf{q} + \frac{\kappa}{\tau_\kappa} \boldsymbol{\nabla} T = -\frac{1}{\tau_\kappa} \mathbf{q}, \tag{6.16e}$$

where τ_κ is a phenomenological relaxation time of the same order as τ_ν and is related to the kinetic processes underlying thermal conduction. As with Eq. 6.15d, the vector equation for heat flux (Eq. 6.16e) is the linearized version of the full heat-flow moment equation, which may be obtained by using a closure assumption that determines the relaxation time, τ_κ. Eq. 6.16 thus defines the linearized 13-moment system (L13, for brevity) used in HERMESHD.

6.2.6 Fluctuations in the linearized 13-moment system

The fluctuating hydrodynamics equations in the linearized 13-moment system (L13) was achieved by incorporating stochastic fluxes in the momentum equations (Eq. 6.16b) and energy equation (Eq. 6.16c) so as to recover the LLNS equations in the limit of high collision frequency, $\tau_\eta \to 0$ and $\tau_\kappa \to 0$. The fluctuating L13 (FL13) system may then be

expressed as

$$\partial_t \rho + \boldsymbol{\nabla} \cdot (\rho \mathbf{u}) = 0 \tag{6.17a}$$

$$\partial_t (\rho \mathbf{u}) + \boldsymbol{\nabla} \cdot \left(\rho \mathbf{u}\mathbf{u} + p\vec{\vec{\mathbf{1}}} + \mathcal{S} \right) = 0 \tag{6.17b}$$

$$\partial_t \mathcal{E} + \boldsymbol{\nabla} \cdot (\mathbf{u}(\mathcal{E} + p) + \mathbf{u} \cdot \mathcal{S} + \mathcal{Q}) = 0 \tag{6.17c}$$

$$\partial_t \vec{\vec{\sigma}} + \frac{\eta}{\tau_\eta} \left(\boldsymbol{\nabla}\mathbf{u} + (\boldsymbol{\nabla}\mathbf{u})^\intercal - \frac{2}{3}(\boldsymbol{\nabla} \cdot \mathbf{u})\vec{\vec{\mathbf{1}}} \right) = -\frac{1}{\tau_\eta}\vec{\vec{\sigma}} \tag{6.17d}$$

$$\partial_t \mathbf{q} + \frac{\kappa}{\tau_\kappa}\boldsymbol{\nabla} T = -\frac{1}{\tau_\kappa}\mathbf{q}, \tag{6.17e}$$

where $\mathcal{S}$ is the stochastic stress tensor and $\mathcal{Q}$ is the stochastic heat flux vector. Both $\mathcal{S}$ and $\mathcal{Q}$ are stochastic fluxes that have zero mean and covariances (neglecting bulk viscosity) given by

$$\left\langle \mathcal{S}_{ij}(\mathbf{x}, t)\mathcal{S}_{kl}(\mathbf{x}', t') \right\rangle = 2\eta k_\mathrm{B} T \left(\delta_{il}\delta_{jk} + \delta_{ik}\delta_{jl} - \frac{2}{3}\delta_{ij}\delta_{kl} \right) \delta(\mathbf{x} - \mathbf{x}')\delta(t - t'), \tag{6.18}$$

$$\left\langle \mathcal{Q}_i(\mathbf{x}, t)\mathcal{Q}_j(\mathbf{x}', t') \right\rangle = 2\kappa k_\mathrm{B} T^2 \delta_{ij}\delta(\mathbf{x} - \mathbf{x}')\delta(t - t'), \tag{6.19}$$

and

$$\left\langle \mathcal{S}_{ij}(\mathbf{x}, t)\mathcal{Q}_k(\mathbf{x}', t') \right\rangle = 0, \tag{6.20}$$

where, as before, η is the dynamic viscosity and κ is the thermal conductivity [366, 407]. More details are given in Appendix G, which, focusing specifically on the formal structure and sampling of the stochastic stress tensor, provides an intuitive sense of how stochastic terms have been handled in HERMESHD.

6.2.7 Initial verification of HERMESHD

In order to examine the accuracy and performance of the hydrodynamic solver(s) in HERMESHD, two standard hydrodynamic test problems** were implemented: (1) 1D Sod

**Further testing is needed for rigorous verification, though the 1D Sod shock tube and 2D isentropic vortex problems sufficed for illustrative purposes during the BWGF. It is also imperative that the new models—such as the 10- and 13-moment fluctuating hydrodynamics approach in HERMESHD—eventually be validated by direct comparison with experimental data.

shock tube [436] (not shown); (2) 2D isentropic vortex [437]. The tests were performed for the linearized 10-moment (L10) equations, which neglect the effects of thermal conduction (Eq. 6.15). Fig. 6.2 shows an example of the advective evolution of the 2D isentropic vortex over 10 periods (cf. [438]) for an 80^2 grid using a linear basis with 8 Gaussian quadrature points per cell (not shown) and second-order Runge-Kutta integration. The vortex is advected horizontally (across periodic boundaries) at a uniform velocity—its shape should be preserved as it returns to its starting position. The density profile after 10 periods demonstrates low numerical diffusion and very reasonable qualitative agreement with the exact solution. As all the hydro tests have analytical flow solutions, they have been especially helpful for ensuring correctness in the process of modularizing the HERMESHD codebase.

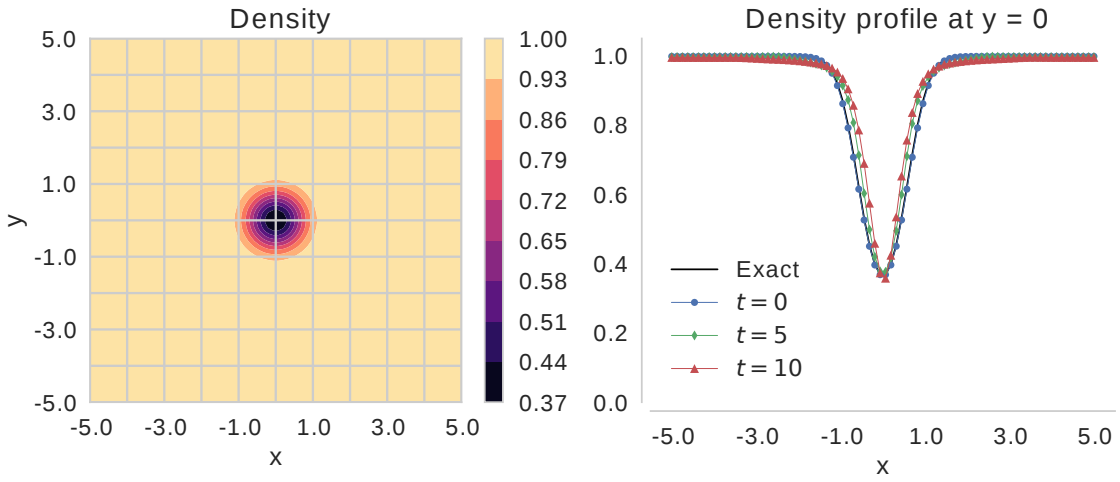

Figure 6.2: Density plots of the 2D isentropic vortex simulation for L10 system of equations (neglecting thermal conduction). Left: 2D fluid density at the initial time. Right: vortex evolution over 10 periods. The density profile at $y = 0$ along x-direction is shown at $t = 0$ (initial condition, blue circles), $t = 5$ (5 periods, green diamonds), $t = 10$ (10 periods, red triangles), with analytical solution shown as solid black line. Simulation was performed on a grid of 80×80 cells (DG basis function data within cells not shown) in dimensionless units.

Several preliminary benchmarks were also performed on the Blue Waters supercomputer, using the vortex problem for initial performance tests of the L10 system: (1) second-

and third-order Runge-Kutta integration[††], (2) spatial resolution and performance with varying combinations of grid resolution and discontinuous Galerkin (DG) basis representations, (3) parallel scaling performance of the MPI-based domain decomposition. Fig. 6.3 gives an idea of raw strong scaling performance in the 2D isentropic vortex problem for up to 8 nodes (32 MPI ranks per node) using a 160^2 grid with linear (2,8), quadratic (3,10), and cubic (4,20) DG basis representations. Overall, these results suggest that the FL10 code scales well to many processing elements in a predictable manner, with the linear and quadratic bases being very efficient as compared to simulations using cubic basis functions. In the future, formal scaling efficiency benchmarks will be carried out for HERMESHD across a range of problems, including tests of the FL13 and full, nonlinear 13-moment systems.

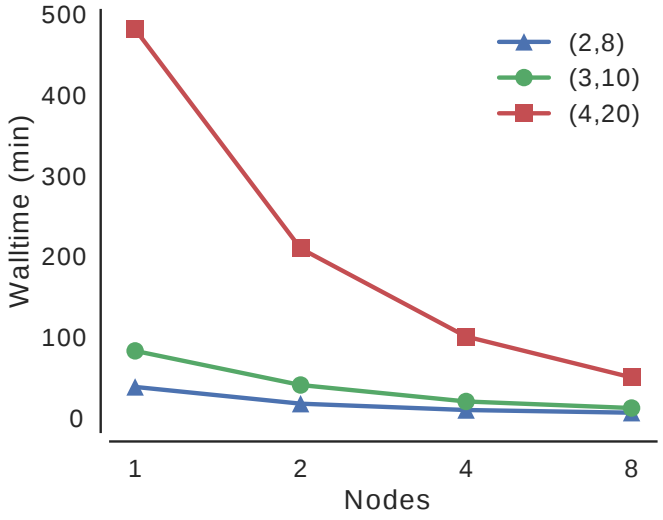

Figure 6.3: Raw strong scaling of isentropic vortex problem on Blue Waters XE nodes using 32 MPI ranks/node on a grid of 160×160 cells for linear (blue triangles), quadratic (orange circles), and cubic (green squares) basis representations with 8, 10, and 20 internal points, respectively.

As a further qualitative test of the HERMESHD fluctuating hydrodynamics model, a comparison between conventional Navier-Stokes and the fluctuating L10 (FL10) system (Eq. 6.17a–d, with $\mathcal{Q} = 0$) was carried out using a simulation of a nanoscale hydrodynamic

[††]Third-order Runge-Kutta integration is needed for variance-preserving properties to correctly recover the fluctuations in the LLNS equations [405].

jet of a viscous, compressible fluid under isothermal conditions.[‡‡] The jet was initialized in a periodic domain with small, random velocity perturbations ($\lesssim 1 k_B T$) and simulated with and without fluctuating terms (resp. top and bottom of Fig. 6.4). The velocity shear was expected to cause the perturbations to grow into a visible Kelvin-Helmholtz instability [430], with the prediction that the FHD simulation should go unstable before the conventional NS due to the action of thermal fluctuations. Indeed, jet breakup occured sooner for the FL10 simulation than the NS system, as seen around the 12 ns mark in Fig. 6.4, corroborating the importance of correctly incorporating hydrodynamic fluctuations at the nanometer observation scales where biological macromolecular dynamics are important.

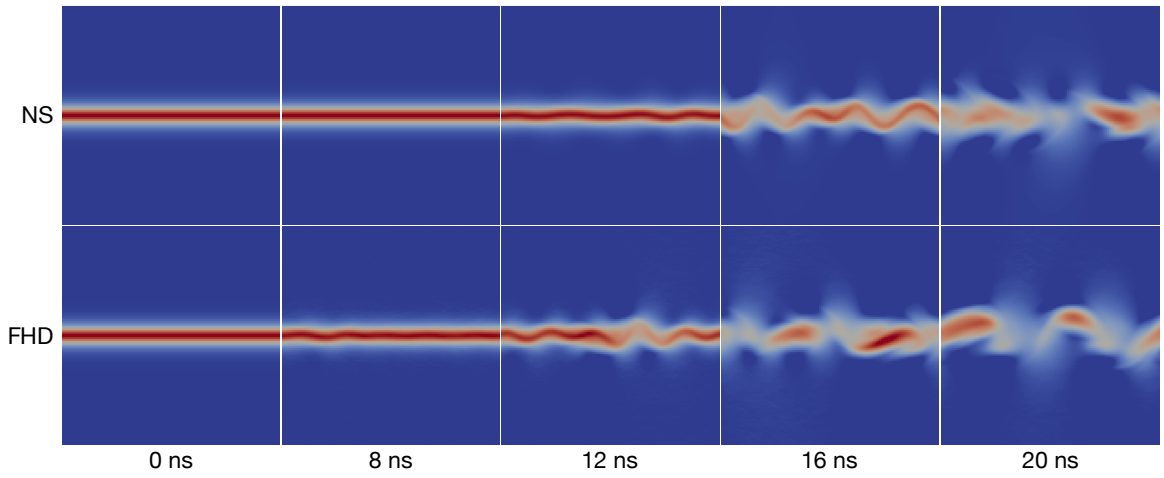

Figure 6.4: Comparison of the rightward horizontal velocities for a simple fluid around water density at STP using Navier-Stokes (NS, top) and FL10 hydrodynamics (FHD, bottom) after 8 ns, 12 ns, 16 ns, and 20 ns. The FL10 system consists of Eq. 6.17a–d with $\mathcal{Q} = 0$ (neglecting heat flux). The simulation region is a 30^2 grid measuring 300 nm per side. Blue and red correspond to small and large velocities (from 0 nm ns^{-1} to 100 nm ns^{-1}), respectively. Simulation was performed using linear quadrature with four basis elements (1,x,y,z) and RK2 integration.

6.3 Synopsis of the hybrid atomistic method

The FHD and MD components of a hybrid code must have a means of communication, though it can be difficult to determine the most appropriate approach at the outset. One

[‡‡]This simple test was performed primarily as a visual sanity check that the hydrodynamic fluctuations in the FHD model were sensible, though it was not intended as a rigorous means of code verification. One possible way in which the nanojet test could be improved would be to analytically estimate the expected time to jet break-up for the FHD and NS models and compare the analytical and numerical results.

option that was considered was to use the MD code as the master program to "call" (subroutines from) HERMESHD, though the problem with this approach was that the HAC code would have to be tied to that particular MD engine. In the opposite approach, which is a more portable alternative, the MD engine would be slave to HERMESHD acting as the master code; this was also rejected due to the fact that MD codes would have to be called from Fortran, making it somewhat difficult to maintain HERMESHD as a standalone code without the extensive modifications needed to build an interface to HERMESHD in Fortran. As such, a decision was made to design and implement the prototype HAC code using an umbrella (driver) program model, where the continuum and MD codes are accessed on an equal footing as libraries. The umbrella program was written in Python—a good language for gluing together disparate software components that may be written in a several different languages—helping to maintain the independence of the hydro and particle components. One key advantage of unifying the HAC prototype through a Python interface was the ease with which code output could be handled, which, for HAC simulations, serves to alleviate possible complications in managing the data structures representing fundamentally difference types of data (i.e., particle- and cell-based data). Paraview [439] and yt [440] are two excellent options as they can be used to visualize mesh and particle data, are open-source, and can be interfaced with Python.

HERMESHD in its current form has a basic Python interface (`hermeshd.py`) generated with the help of f90wrap [441] (based on the work of Pletzer et al. [442] and the f2py package [443]) that exposes high-level subroutines to Python. Subroutine arguments are passed as (Fortran-order) NumPy arrays that point to the underlying Fortran data in memory. Fig. 6.5 represents a functional Python script that can be used to run a hydrodynamic simulation using the Python interface to HERMESHD; in particular, one can call individual time steps, allowing, for instance, MD steps to be called in between hydro steps. As a specific example, the `step()` method advances time forward by a single step by solving for the fluid fluxes/sources and integrating the equations of motion in time; the field variables can then be directly extracted from the NumPy arrays. This is particularly useful for HAC

simulations where boundary conditions to/from HERMESHD will be passed periodically after a small number of time steps.

```
import numpy as np
from mpi4py import MPI
from hermeshd import FHDsim

# Init sim params and data arrays for fluid variables; Qio is primary data array
nx, ny, nz = 80, 80, 1              # 80 × 80 × 1 grid
nQ, nB = 11, 8                      # 11 equations (10-moment), 8 basis funcs
t  = np.array(0.00, dtype=float)    # current time, t
dt = np.array(0.01, dtype=float)    # timestep, Δt
tf = np.array(10.0, dtype=float)    # final time, t_f
dtout = np.array(0., dtype=float)   # output period (every dtout units of time), t_out
Qio = np.empty((nx,ny,nz,nQ,nB), order='F', dtype=np.float32)
Q1  = np.empty_like(Qio)  # temp array for vars w/ 2nd-order RK integrator
Q2  = np.empty_like(Qio)  # temp array for vars w/ 2nd-order RK integrator

# Setup and run HERMESHD
FHDsim.setup(Qio, t, dt, dtout, MPI.COMM_WORLD.py2f())  # fill F90 data structs
while (t < tf):
    FHDsim.step(Qio, Q1, Q2, t, dt)   # perform a single FHD timestep
    FHDsim.output(Qio, t, dt, dtout)  # generate VTK output every dtout interval
FHDsim.cleanup()
```

Figure 6.5: Example Python script (`hermes_run.py`) for performing an MPI-enabled run using `mpi4py` and the Python interface to HERMESHD. Fortran-order NumPy array data, `Qio`, is initialized in Python, sent through a Python interface to the `setup()` subroutine in HERMESHD, and updated by iterating through the main simulation loop until the final time is reached, `t>=tf`. A 16-core MPI run can be executed from the command line by calling the Python script with, for instance, `mpirun -n 16 python hermes_run.py`.

For the MD component of the HAC method, the LAMMPS MD engine was chosen for reasons of simplicity and flexibility as it: (1) is relatively easy to dynamically modify particle properties, custom potentials, and boundary conditions during a timestep using a so-called "fix", (2) supports CHARMM/Amber/OPLS force fields, (3) can be compiled as a library to be called from an external program, and (4) has a built-in Python interface. To leverage the LAMMPS Python interface for communication with the driver program, LAMMPS was built as a library with MPI support.

The prototype hybrid code was implemented by combining a minimal implementation of a Python-based driver program, a Python- and MPI-compatible build of LAMMPS as

a library, and the HERMESHD program with a custom Python interface to the Fortran code based on f90wrap.[§§] All communication was managed at the top level by the driver program, which behaved as a server that facilitated communication between the clients programs (the individual MD and hydro components); such a scheme, if designed poorly, may be detrimental to performance because data must be first moved from one client to the driver program before being sent to the other client code. On the other hand, such a consolidated approach has a clear advantage in terms of organizational and semantic clarity, since what the HAC simulation is "doing" at any given timestep can be expressed explicitly in terms of actions taken by the umbrella code (rather than hidden actions executed by either the hydro or MD codes). Fig. 6.1 illustrates the flow of the HAC method timestep, where a fixed number of MD steps were followed by the transfer of atomistic data (from the buffer region) to the hydro code, in turn followed by a single hydro time step; once the field data was updated by the hydro code, those (continuum) data in the buffer were sent to the MD code, beginning the cycle anew.

Communication between the atomistic and continuum regions was implemented as a simple one-way continuum-to-atomistic coupling (C $\rightarrow$ A) scheme where the drag force on a particle, p, was given as the difference between the particle velocity, $\mathbf{v}_p$, and the hydrodynamic velocity, $\mathbf{u}$,

$$\frac{\mathrm{d}\mathbf{x}_p}{\mathrm{d}t} = \mathbf{v}_p \tag{6.21a}$$

$$m\frac{\mathrm{d}\mathbf{v}_p}{\mathrm{d}t} = -\xi\left(\mathbf{v}_p - \mathbf{u}\right) - \boldsymbol{\nabla}_{\mathbf{x}_p} U\left(\mathbf{X}\right), \tag{6.21b}$$

where ξ is a phenomenological friction coefficient, $\mathbf{x}_p$ is the position of particle p, and $\mathbf{X}$ is the configuration space vector of particle positions; since only C $\rightarrow$ A coupling was implemented, no forces due to the particles were applied to the fluid. The advantage of Eq. 6.21 over conventional Langevin dynamics is that the particles feel a drag force with

[§§] As a starting point, f90wrap was used to generate a Fortran 90 interface to the HERMESHD source code together with a high-level Pythonic wrapper. The Python extension modules were then modified to provide the desired user-level access to subroutines and data structures in the source code.

respect to the background fluid velocity rather than a fixed reference frame; also, unlike the Langevin equation, no additional stochastic term was included in Eq. 6.21b since the (hydrodynamic) fluctuations were left to implicitly enter into the fluid velocity, **u**, through the momentum equation (Eq. 6.17b). This approach is similar in spirit to the fluctuating hydrodynamics thermostat used in Y. Wang et al. [444] and the lattice Boltzmann approach taken by Mackay, Ollila, and Denniston [445]—the implementation based on Eq. 6.21 was not intended as a realistic coupling approach, but only as a proof-of-concept to test the communication between the LAMMPS and HERMESHD codes. The test simulation was a Couette flow (shear flow) problem with fully-overlapping MD and hydro domains (data not shown), which was designed to simplify complications associated with boundary condition communication in the buffer region (i.e., region B in Fig. 6.1).

6.4 Future directions

A functional HAC prototype will demonstrate the ability to: (1) deploy HERMESHD as a library callable from Python, (2) dynamically modify MD simulation parameters and particle data through LAMMPS fixes, (3) exchange boundary condition information in the buffer region between the hydro and MD domains through the umbrella program, (4) produce a qualitatively accurate hybrid simulation of a simple fluid like liquid argon. In particular, the handling of boundary conditions (point 3) is a nontrivial task and there are many different approaches to exchanging particle and continuum information in the overlap region [446] other than the simple method mentioned above (Eq. 6.21). It is also possible that a particle-stat—a particle regulator for open boundary conditions that obeys basic thermodynamic considerations—will need to be incorporated into the HAC model; fortunately, good solutions to this problem already exist [447, 448]. Future hurdles notwithstanding, it has been demonstrated that the HERMESHD hydrodynamics model is viable and can be interfaced—using Python—with an MD engine to construct a hybrid atomistic-continuum method. Modest progress, with respect to the software and algorithmic architecture of the HAC code, has provided a clear picture of important hurdles, but has also left substantial

room to re-think various aspects of constructing a fully working model as well as the physics that underlie particle-to-continuum mapping and vice versa.

The question of how to bridge disparate length scales applies not only to our numerical models, but to our physical models. The development of consistent coarse-graining schemes that capture the most relevant aspects of atomic-resolution force fields using a smaller number of composite particles is an active area of research [449–452]. More closely related to HAC simulation are the questions surrounding the existence and nature of a proper continuum limit. Efforts on this front tend to focus on single-particle phase space distributions and the Boltzmann equation; many overlapping approaches have been used to obtain continuum (hydrodynamic) models from kinetic theory [453–455]. However, it is not know whether a consistent mathematical framework exists for mapping atomistic systems to continuum fields, though some degree of progress is being made. From the opposite end of the spectrum, contemporary theories of nonequilibrium thermodynamics are providing a top-down perspective, offering new perspectives in beyond-equilibrium thermodynamic processes.[¶¶] Despite being phenomenological in nature, generalized thermodynamics approaches may help drive the connection between hydrodynamics and nonequilibrium dissipative processes [458–461]; indeed, Grad's original 13-moment method has played a central role many of these modern thermodynamic theories. Furthermore, shedding light on these fundamental problems is central to elucidating how HAC methods and other multiphysics approaches can incorporate disparate physical models in a mathematically consistent way.

6.4.1 A teaser on the non-Markovianity of L13

The advent of experimental techniques like neutron scattering and molecular simulation methods (e.g., molecular dynamics) has led to a relatively recent re-examination of the

[¶¶] There are a number of distinct, but related frameworks, including classical irreversible thermodynamics (CIT), rational thermodynamics (RT), thermodynamics of irreversible processes (TIP), extended irreversible thermodynamics (EIT), and rational extended thermodynamics (RET) [456–459].

theory of transport coefficients. In particular, a generalized hydrodynamic theory allows a spatiotemporal dependence of the transport coefficients that permits the molecular structure of the fluid to play a role. On length scales in the vicinity of 1–10 molecular diameters and time scales on the order of 30 molecular collisions, fluids exhibit viscoelastic behavior and other effects arising from the microscopic molecular structure [388]. Progress on the FHD front has been robust, and there is promise in investigating extended hydrodynamic models that go beyond Landau-Lifschitz FHD theory and conventional Navier-Stokes. Recently, researchers used the extended thermodynamics formalism to develop a 13-moment FHD model for monatomic ideal gases [462]. Although this model is less general than HERMESHD and it is not known at the time of writing whether Arima et al. [462] have released a numerical model, it is encouraging that other researchers are taking interest in merging the formalisms of extended hydrodynamic theories and hydrodynamic fluctuations. To conclude this chapter, it seems appropriate to provide a forward-looking perspective that might inspire curiosity, starting with three interrelated points of focus:

1. Is the HAC method in any way a viable replacement for EqMD in simulating biomolecules in native-like aqueous environments?
2. Can we quantify the physical consistency of various HAC coupling methods and can a proper continuum limit (i.e., atomistic-to-continuum coarse-graining) be established?
3. What is the relevance of higher-moment FHD models to the description of (nonequilibrium) transport at the nanoscale? Furthermore, can hydrodynamic approaches offer more intuitive descriptions of biomacromolecular phenomena than MD simulations (and techniques for identifying collective variables)?

One intriguing feature of the linearized 13-moment model*** (L13) in HERMESHD is the way in which viscoelasticity emerges from the linearized stress and heat flux equations due

***In general, this is also true of the full, nonlinear G13 system, though only the L13 system is considered due to the comparative simplicity of the analyses. From the perspective of the BGK approximation [432, 433], it is seen that the relaxation term (the BGK operator), which drives relaxation of the single-particle distribution back to equilibrium in the BGK model, in some sense originates such viscoelastic and thermoacoustic memory effects in the L13 system via the stress (Eq. 6.15d) and heat flux equations (Eq. 6.16e).

to the presence of the relaxation (source) terms. Whereas NS and LLNS tacitly assume linear, *time-independent* constitutive relations connecting viscous stresses to the velocity gradients [463], and heat flux to the temperature gradient—respectively, Newton's law of viscosity and Fourier's law of heat conduction—the L13 system generates history-dependent stresses and heat flux, leading to non-Markovian behavior that emerges at small spatiotemporal scales. Certain viscoelastic effects may also be measurable in real-world and numerical experiments, possibly offering a way to verify and validate the L13 model as it pertains to dense fluids like water.

Starting with the L13 equations (Eq. 6.16). In L13, the five hydrodynamic equations describing the deviatoric stress components, σ_{ij}, which is a linearized 3$^{\text{rd}}$-moment equation, take the form

$$\frac{\partial \sigma_{ij}}{\partial t} + p_0 \dot{\epsilon}_{ij} = -\nu \sigma_{ij},$$

where p_0 is an equilibrium pressure, ν is a microscopic collision frequency (at this point, a phenomenological material parameter), and i and j can take on the values 1, 2, and 3 for each spatial component; $\dot{\epsilon}_{ij}$ (neglecting bulk viscosity, for now) is a tensorial component of the rate-of-strain tensor (from Newton's law of viscosity) [430]

$$\dot{\epsilon}_{ij} = \frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} - \frac{2}{3}\delta_{ij}\frac{\partial u_k}{\partial x_k}.$$

In the limit of large collision frequency (fast relaxation time, $\tau \ll 1$), $\partial_t \ll \nu$, the time-derivative term is dropped and σ_{ij} becomes a dependent variable,

$$\sigma_{ij} = \frac{p_0}{\nu}\Big(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} - \frac{2}{3}\delta_{ij}\frac{\partial u_k}{\partial x_k}\Big).$$

After identifying $\eta = p_0/\nu$ with the dynamic viscosity, we recover the constitutive relation for a Newtonian fluid with strain-rate tensor $\dot{\epsilon}_{ij}$, where $\sigma_{ij} = \eta \dot{\epsilon}_{ij}$ is the viscous (shear) stress tensor, which describes momentum diffusion. In other words, if the hydrodynamic timescales of observable interest are long relative to microscopic collision processes, the L13 system reduces, or *relaxes*, to the five-moment Navier-Stokes system. In the opposite limit, $\partial_t \gg \nu$, where observation timescales are much faster than microscopic collision times,

the source term, $-\nu\sigma_{ij}$, is dropped. If we then write $u_i = \partial_t \xi_i$, relating the fluid velocity components, u_i, to material displacements, ξ_i, we can eliminate the time derivatives on both sides to obtain the elastic stress tensor

$$\sigma_{ij} = p_0\Big(\frac{\partial \xi_i}{\partial x_j} + \frac{\partial \xi_j}{\partial x_i} - \frac{2}{3}\delta_{ij}\frac{\partial \xi_k}{\partial x_k}\Big).$$

The general solution of the deviatoric stress equation is

$$\sigma_{ij}(\mathbf{x}, t) = -p_0 \int_{-\infty}^{t} \mathrm{d}t' e^{\nu(t-t')} \dot{\epsilon}_{ij}(\mathbf{x}, t'),$$

which, by the above analysis, is a viscoelastic stress tensor whose non-Markovian nature is apparent. The $\dot{\epsilon}_{ij}(\mathbf{x}, t)$ are understood to depend spatially on the (gradients of) the velocity field which in turn depends explicitly on time.

Chapter 7

CONCLUSIONS

Deducing the function of a protein from information about its 3D structure is a central ambition in biophysics. This dissertation explored several aspects of computational methodology in elucidating this link between protein structure and function. The enzyme adenylate kinase (AdK) was adopted as a lens through which state-of-the-art experimental and computational techniques could be placed into a proper context. In particular, the catalytically relevant motions that AdK undergoes are relatively simple to visualize, albeit sufficiently complicated so as to: (1) illustrate the interplay between experiment and numerical simulation, (2) introduce the role of numerical path-sampling methods in surmounting the equilibrium sampling problem, (3) discuss extant gaps in our understanding of the mechanisms underlying AdK's function and enzymes in general, and (4) lay the groundwork for Path Similarity Analysis (PSA) and other computational approaches to *quantifying* and *integrating* potentially heterogeneous data.

Path-sampling methods can generate plausible conformational transition paths with relatively little computational effort, overcoming the typical timescale limitations of equilibrium MD (EqMD) simulation. Though both EqMD and path-sampling techniques have furnished some insight into the closed $\leftrightarrow$ open apo-AdK transition, contemporary sampling methodologies have yet to decisively deduce the underlying mechanism, in large part due to the difficulty of assessing the accuracy of such sampling methods. We developed the Path Similarity Analysis (PSA) framework—based on the idea of using a *path metric* to measure the similarity between two transition paths—as a step toward addressing the emerging problem of path-sampling verification; we demonstrated that PSA enables direct quantitative comparisons of large trajectory data sets while minimizing a-priori assumptions that can lead to bias. The heatmap-dendrogram approach in PSA was viable not only for a toy model system, but also realistic scenarios involving path-sampling trajectory data

obtained from apo-AdK and diphtheria toxin transitions. Despite the visual simplicity of the closed $\rightarrow$ open apo-AdK transition, there were marked differences among the methods and resultant paths that could not be fully rationized by the presented analyses. That the closed $\leftrightarrow$ open transition may be, for instance, compared to the opening and closing of a clam shell, is an oversimplification, and it is perhaps the case that AdK does not represent an ideal testbed. As hinted previously in Chapter 1 and Chapter 2, AdK (and other enzymes) seems to be hiding a much richer picture.

Due to complications caused by the lipid environment around membrane proteins, few structures have been solved in comparison to globular proteins; such proteins as transporters are thus especially challenging to study experimentally and, computationally, explicit membrane and solvent EqMD simulations are difficult and expensive. However, we were able to take initial strides toward elucidating the structure and transport mechanism of the SLC4 membrane transporter Bor1p. We succeeded in constructing viable structural models by using MD flexible fitting (MDFF) to integrate medium-resolution cryo-EM data into explicit-solvent, explicit-membrane EqMD simulations. Although serial femtosecond crystallography promises to make "movies" of macromolecules a reality, numerical simulations offer the only viable means of obtaining truly atomistic trajectories. Atomistic EqMD, path-sampling, and computational algorithms to integrate experimental data into simulation (like MDFF) are likely to become even more useful. Indeed, as the number of transporters solved in multiple conformations inevitably grows, path-sampling methods, for instance, will become an increasingly viable approach to studying transporters; combined with quantitative frameworks like PSA, it will soon be possible to generate transition paths, perform EqMD, and quantify those data with increasing confidence.

In general, nanoscale transport phenomena (e.g., transfer of mass, momentum, heat, etc.)—especially in dense fluids and aqueous solutions—are dictated by large thermal fluctuations, leading to nonequilibrium phenomena that often necessitate nonlinear descriptions of fluid flow and quasithermodynamic phenomena. The author's Blue Waters

Graduate Fellowship (BWGF) project* offered an opportunity to investigate the viability of hybrid methods and fluctuating hydrodynamics (FHD) simulation for modeling biomacromolecular dynamics. The hybrid atomistic-continuum (HAC) method—the third mode of methodological integration explored in this dissertation—is but one example of a multiphysics method that couples multiple physical models into a unified numerical method. Given that some nanoscale biophysical processes may be characteristically hydrodynamic by nature, HAC and FHD simulation are viable means of investigating such phenomena. As one example, the linearized 13-moment FHD equations used in PERSEUS were shown to predict non-Markovian behavior that leads to viscoelasticity, thermoacoustic phenomena, and other memory effects when observation scales approach the nanometer range. To wit, the analyses in Chapter 6 show that L13 contains the well known Maxwell model of viscoelastic materials, implying that a dense fluid like water behaves like a Maxwell fluid at the nanoscale [464]. There is experimental evidence that hydrodynamics may be relevant to the description of diffusion enhancement observed in enzymes upon the release of heat during catalysis [465]. Hydrodynamics may also be important in protein hydration dynamics [466–468], protein folding [469], hydrodynamic collective effects in cells [470–472], and inertial effects on slow kinetic processes of DNA [395] and other biomacromolecules. However, the particle picture of biomolecular phenomena has been reinforced by an abundance of atomistic X-ray crystal structures and the ubiquity of atomistic numerical methods like MD; perhaps a hydrodynamic perspective would not only be refreshing, but provide a new source of physical insight.

The extent to which biomolecular sampling efficiency and accuracy depends, in general, on specific aspects of the physical and numerical models is a fundamental yet unanswered question. These matters have important implications for understanding the physics involved in biomolecular dynamics and also building confidence in the accuracy of the methods and models. The analyses as presented in this dissertation are insufficient to

Developing a Hybrid Atomistic-continuum Method for Simulating Large-scale Heterogeneous Biomolecular Systems

fully elucidate how different aspects of a path-sampling method (i.e., model resolution, biasing algorithms, progress variables, collective variables, etc.) are connected to pathway that will be sampled. However, the core components of the path similarity analysis (PSA) framework—the visual heatmap-dendrogram approach based on hierarchical clustering and Hausdorff pairs analysis—were found to be useful not only for assessing pathway similarities, but more importantly in identifying and extracting molecular-structural correlates of their *differences* at the atomistic level. Given the modular nature of PSA and the ease with which it can evaluate large trajectory ensembles, it is possible to construct sophisticated analysis pipelines, and the author hopes that its core components or ideas will be integrated into new and existing quantitative analyses. It would be valuable to perform a thorough path-sampling method comparison covering a broader range of physical models, numerical algorithms, and biasing schemes. Ultimately, ascertaining the extent to which specific features of a numerical model determine sampling behavior is a necessary component of verification and validation of numerical simulations. The difficulty of such a task notwithstanding, there is an opportunity to deepen our understanding of the relationship between our physical models (i.e., how we conceptualize reality) and real natural processes (i.e., reality). Indeed, in light of the impressive number and variety of sampling methods, frameworks like PSA can quantify essential differences between physical models, allowing us to connect specific pictures we derive from our physical models to the dynamical motions and mechanisms underlying protein function.

Though this dissertation provides but a glimpse of emerging modes of scientific inquiry through the lens of protein structure-function, the importance of integration and quantification to scientific methodology is clear. It seems self-evident that, given the increasingly complexity of the physical systems being studied, the generation of data cannot be a viable long-term solution. Indeed, in a time when the volume of data and the size of data sets is growing exponentially, it is imperative that we focus on extracting *meaning* from that data, which in turn requires that we, as scientists, be mindful of the increasingly variegated ways in which both systematic and statistical error can confound our analyses. The development

of robust computational methodologies to verify and validate our methods and data is essential, and it seems inevitable that the future of scientific exploration and discovery will depend on such methodologies that emphasize deeper integration of increasingly powerful numerical simulations methods and experimental techniques.

[1] Yon, J. M., D. Perahia, and C. Ghélis (1998). Conformational dynamics and enzyme activity. *Biochimie* **80** (1), 33–42.

[2] Karplus, M., Y. Q. Gao, J. Ma, A. van der Vaart, and W. Yang (2005). Protein structural transitions and their functional role. *Philos. Trans. A Math. Phys. Eng. Sci.* **363** (1827), 331–55, discussion 355–6. DOI: 10.1098/rsta.2004.1496.

[3] Henzler-Wildman, K. and D. Kern (2007). Dynamic personalities of proteins. *Nature* **450** (7172), 964–972. DOI: 10.1038/nature06522.

[4] Seyler, S. L. and O. Beckstein (2014). Sampling large conformational transitions: adenylate kinase as a testing ground. *Mol. Simul.* **40** (10-11), 855–877. DOI: 10.1080/08927022.2014.919497.

[5] Seyler, S. L., A. Kumar, M. F. Thorpe, and O. Beckstein (2015). Path Similarity Analysis: A Method for Quantifying Macromolecular Pathways. *PLoS Comput. Biol.* **11** (10), e1004568. DOI: 10.1371/journal.pcbi.1004568.

[6] Shaw, D. E., M. M. Deneroff, R. O. Dror, J. S. Kuskin, R. H. Larson, J. K. Salmon, C. Young, B. Batson, K. J. Bowers, J. C. Chao, M. P. Eastwood, J. Gagliardo, J. P. Grossman, C. R. Ho, D. J. Ierardi, I. Kolossváry, J. L. Klepeis, T. Layman, C. McLeavey, M. A. Moraes, R. Mueller, E. C. Priest, Y. Shan, J. Spengler, M. Theobald, B. Towles, and S. C. Wang (2008). Anton, a special-purpose machine for molecular dynamics simulation. *Commun. ACM* **51** (7), 91–97. DOI: 10.1145/1364782.1364802.

[7] Coudray, N., S. L Seyler, R. Lasala, Z. Zhang, K. M. Clark, M. E. Dumont, A. Rohou, O. Beckstein, and D. L. Stokes (2017). Structure of the SLC4 transporter Bor1p in an inward-facing conformation. *Protein Sci.* **26** (1), 130–145. DOI: 10.1002/pro.3061.

[8] Forrest, L. R., R. Krämer, and C. Ziegler (2011). The structural basis of secondary active transport mechanisms. *Biochim. Biophys. Acta* **1807** (2), 167–188. DOI: 10.1016/j.bbabio.2010.10.014.

[9] Shi, Y. (2013). Common folds and transport mechanisms of secondary active transporters. *Annu. Rev. Biophys.* **42** (1), 51–72. DOI: 10.1146/annurev-biophys-083012-130429.

[10] Drew, D. and O. Boudker (2016). Shared Molecular Mechanisms of Membrane Transporters. *Annu. Rev. Biochem.* **85**, 543–572. DOI: 10.1146/annurev-biochem-060815-014520.

[11] Pisliakov, A. V., J. Cao, S. C. L. Kamerlin, and A. Warshel (2009). Enzyme millisecond conformational dynamics do not catalyze the chemical step. *Proc. Natl. Acad. Sci. U. S. A.* **106** (41), 17359–17364. DOI: 10.1073/pnas.0909150106.

[12] Kamerlin, S. C. L. and A. Warshel (2010). At the dawn of the 21st century: Is dynamics the missing link for understanding enzyme catalysis? *Proteins* **78** (6), 1339–1375. DOI: 10.1002/prot.22654.

[13] Eisenmesser, E. Z., O. Millet, W. Labeikovsky, D. M. Korzhnev, M. Wolf-Watz, D. A. Bosco, J. J. Skalicky, L. E. Kay, and D. Kern (2005). Intrinsic dynamics of an enzyme underlies catalysis. *Nature* **438** (7064), 117–121. DOI: 10.1038/nature04105.

[14] Henzler-Wildman, K. A., M. Lei, V. Thai, S. J. Kerns, M. Karplus, and D. Kern (2007a). A hierarchy of timescales in protein dynamics is linked to enzyme catalysis. *Nature* **450** (7171), 913–916. DOI: 10.1038/nature06407.

[15] Smith, A. J. T., R. Müller, M. D. Toscano, P. Kast, H. W. Hellinga, D. Hilvert, and K. N. Houk (2008). Structural reorganization and preorganization in enzyme active sites: comparisons of experimental and theoretically ideal active site geometries in the multistep serine esterase reaction cycle. *J. Am. Chem. Soc.* **130** (46), 15361–15373. DOI: 10.1021/ja803213p.

[16] Bhabha, G., J. Lee, D. C. Ekiert, J. Gam, I. A. Wilson, H. J. Dyson, S. J. Benkovic, and P. E. Wright (2011). A dynamic knockout reveals that conformational fluctuations influence the chemical step of enzyme catalysis. *Science* **332** (6026), 234–238. DOI: 10.1126/science.1198542.

[17] Kohen, A. (2015). Role of dynamics in enzyme catalysis: substantial versus semantic controversies. *Acc. Chem. Res.* **48** (2), 466–473. DOI: 10.1021/ar500322s.

[18] Hammes-Schiffer, S. and S. J. Benkovic (2006). Relating protein motion to catalysis. *Annu. Rev. Biochem.* **75**, 519–541. DOI: 10.1146/annurev.biochem.75.103004.142800.

[19] Wolf-Watz, M., V. Thai, K. Henzler-Wildman, G. Hadjipavlou, E. Z. Eisenmesser, and D. Kern (2004). Linkage between dynamics and catalysis in a thermophilic-mesophilic enzyme pair. *Nat. Struct. Mol. Biol.* **11** (10), 945–949. DOI: 10.1038/nsmb821.

[20] Watt, E. D., H. Shimada, E. L. Kovrigin, and J. P. Loria (2007). The mechanism of rate-limiting motions in enzyme function. *Proc. Natl. Acad. Sci. U. S. A.* **104** (29), 11981–11986. DOI: 10.1073/pnas.0702551104.

[21] Ma, B. and R. Nussinov (2010). Enzyme dynamics point to stepwise conformational selection in catalysis. *Curr. Opin. Chem. Biol.* **14** (5), 652–659. DOI: 10.1016/j.cbpa.2010.08.012.

[22] Sacquin-Mora, S., O. Delalande, and M. Baaden (2010). Functional modes and residue flexibility control the anisotropic response of guanylate kinase to mechanical stress. *Biophys. J.* **99** (10), 3412–3419. DOI: 10.1016/j.bpj.2010.09.026.

[23] Wang, Y. and G. Zocchi (2011). Viscoelastic transition and yield strain of the folded protein. *PLoS One* **6** (12), e28097. DOI: 10.1371/journal.pone.0028097.

[24] Tseng, C.-Y. and G. Zocchi (2013). Mechanical control of Renilla luciferase. *J. Am. Chem. Soc.* **135** (32), 11879–11886. DOI: 10.1021/ja4043565.

[25] Delalande, O., N. Férey, G. Grasseau, and M. Baaden (2009). Complex molecular assemblies at hand via interactive simulations. *J. Comput. Chem.* **30** (15), 2375–2387. DOI: 10.1002/jcc.21235.

[26] Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne (2000). The Protein Data Bank. *Nucleic Acids Res.* **28** (1), 235–242.

[27] Martin-Garcia, J. M., C. E. Conrad, J. Coe, S. Roy-Chowdhury, and P. Fromme (2016). Serial femtosecond crystallography: A revolution in structural biology. *Arch. Biochem. Biophys.* **602**, 32–47. DOI: 10.1016/j.abb.2016.03.036.

[28] Johansson, L. C., B. Stauch, A. Ishchenko, and V. Cherezov (2017). A Bright Future for Serial Femtosecond Crystallography with XFELs. *Trends Biochem. Sci.* **42** (9), 749–762. DOI: 10.1016/j.tibs.2017.06.007.

[29] Yang, L.-W. and C.-P. Chng (2008). Coarse-grained models reveal functional dynamics–I. Elastic network models–theories, comparisons and perspectives. *Bioinform. Biol. Insights* **2**, 25–45.

[30] Schlitter, J., M. Engels, and P. Krüger (1994). Targeted molecular dynamics: a new approach for searching pathways of conformational transitions. *J. Mol. Graph.* **12** (2), 84–89. DOI: 10.1016/0263-7855(94)80072-3.

[31] Atkinson, D. E. (1968). The energy charge of the adenylate pool as a regulatory parameter. Interaction with feedback modifiers. *Biochemistry* **7** (11), 4030–4034.

[32] Dzeja, P. P., R. J. Zeleznikar, and N. D. Goldberg (1998). Adenylate kinase: kinetic behavior in intact cells indicates it is integral to multiple cellular processes. *Mol. Cell. Biochem.* **184** (1-2), 169–182.

[33] Saraste, M., P. R. Sibbald, and A. Wittinghofer (1990). The P-loop–a common motif in ATP- and GTP-binding proteins. *Trends Biochem. Sci.* **15** (11), 430–434. DOI: 10.1016/0968-0004(90)90281-F.

[34] Leipe, D. D., E. V. Koonin, and L. Aravind (2003). Evolution and classification of P-loop kinases and related proteins. *J. Mol. Biol.* **333** (4), 781–815. DOI: 10.1016/j.jmb.2003.08.040.

[35] Gerstein, M., G. Schulz, and C. Chothia (1993). Domain closure in adenylate kinase. Joints on either side of two helices close like neighboring fingers. *J. Mol. Biol.* **229** (2), 494–501. DOI: 10.1006/jmbi.1993.1048.

[36] Müller, C. W., G. J. Schlauderer, J. Reinstein, and G. E. Schulz (1996). Adenylate kinase motions during catalysis: an energetic counterweight balancing substrate binding. *Structure* **4** (2), 147–156.

[37] Vonrhein, C., G. J. Schlauderer, and G. E. Schulz (1995). Movie of the structural changes during a catalytic cycle of nucleoside monophosphate kinases. *Structure* **3** (5), 483–490.

[38] Maragakis, P. and M. Karplus (2005). Large amplitude conformational change in proteins explored with a plastic network model: adenylate kinase. *J. Mol. Biol.* **352** (4), 807–822. DOI: 10.1016/j.jmb.2005.07.031.

[39] Feng, Y., L. Yang, A. Kloczkowski, and R. L. Jernigan (2009). The energy profiles of atomic conformational transition intermediates of adenylate kinase. *Proteins* **77** (3), 551–558. DOI: 10.1002/prot.22467.

[40] Beckstein, O., E. J. Denning, J. R. Perilla, and T. B. Woolf (2009). Zipping and unzipping of adenylate kinase: atomistic insights into the ensemble of openclosed transitions. *J. Mol. Biol.* **394** (1), 160–176. DOI: 10.1016/j.jmb.2009.09.009.

[41] Adén, J. and M. Wolf-Watz (2007). NMR identification of transient complexes critical to adenylate kinase catalysis. *J. Am. Chem. Soc.* **129** (45), 14003–14012. DOI: 10.1021/ja075055g.

[42] Hanson, J. A., K. Duderstadt, L. P. Watkins, S. Bhattacharyya, J. Brokaw, J.-W. Chu, and H. Yang (2007). Illuminating the mechanistic roles of enzyme conformational dynamics. *Proc. Natl. Acad. Sci. U. S. A.* **104** (46), 18055–18060. DOI: 10.1073/pnas.0708600104.

[43] Henzler-Wildman, K. A., V. Thai, M. Lei, M. Ott, M. Wolf-Watz, T. Fenn, E. Pozharski, M. A. Wilson, G. A. Petsko, M. Karplus, C. G. Hübner, and D. Kern (2007b). Intrinsic motions along an enzymatic reaction trajectory. *Nature* **450** (7171), 838–844. DOI: 10.1038/nature06410.

[44] Koshland, D. E. (1958). Application of a Theory of Enzyme Specificity to Protein Synthesis. *Proc. Natl. Acad. Sci. U. S. A.* **44** (2), 98–104.

[45] Monod, J., J. Wyman, and J. P. Changeux (1965). ON THE NATURE OF ALLOSTERIC TRANSITIONS: A PLAUSIBLE MODEL. *J. Mol. Biol.* **12**, 88–118.

[46] Hammes, G. G., Y.-C. Chang, and T. G. Oas (2009). Conformational selection or induced fit: a flux description of reaction mechanism. *Proc. Natl. Acad. Sci. U. S. A.* **106** (33), 13737–13741. DOI: 10.1073/pnas.0907195106.

[47] Csermely, P., R. Palotai, and R. Nussinov (2010). Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. *Trends Biochem. Sci.* **35** (10), 539–546. DOI: 10.1016/j.tibs.2010.04.009.

[48] Schrank, T. P., J. O. Wrabl, and V. J. Hilser (2013). "Conformational Heterogeneity Within the LID Domain Mediates Substrate Binding to Escherichia coli Adenylate Kinase: Function Follows Fluctuations". *Dynamics in Enzyme Catalysis*. Topics in Current Chemistry. Springer Berlin Heidelberg, pp. 95–121. DOI: 10.1007/128_2012_410.

[49] Tobi, D. and I. Bahar (2005). Structural changes involved in protein binding correlate with intrinsic motions of proteins in the unbound state. *Proc. Natl. Acad. Sci. U. S. A.* **102** (52), 18908–18913. DOI: 10.1073/pnas.0507603102.

[50] Cukier, R. I. (2009). Apo adenylate kinase encodes its holo form: a principal component and varimax analysis. *J. Phys. Chem. B* **113** (6), 1662–1672. DOI: 10.1021/jp8053795.

[51] Bakan, A. and I. Bahar (2009). The intrinsic dynamics of enzymes plays a dominant role in determining the structural changes induced upon inhibitor binding. *Proc. Natl. Acad. Sci. U. S. A.* **106** (34), 14349–14354. DOI: 10.1073/pnas.0904214106.

[52] Seeliger, D. and B. L. de Groot (2010). Conformational transitions upon ligand binding: holo-structure prediction from apo conformations. *PLoS Comput. Biol.* **6** (1), e1000634. DOI: 10.1371/journal.pcbi.1000634.

[53] Daily, M. D., L. Makowski, G. N. Phillips Jr, and Q. Cui (2012). Large-scale motions in the adenylate kinase solution ensemble: coarse-grained simulations and comparison with solution X-ray scattering. *Chem. Phys.* **396**, 84–91. DOI: 10.1016/j.chemphys.2011.08.015.

[54] Shapiro, Y. E. and E. Meirovitch (2006). Activation energy of catalysis-related domain motion in E. coli adenylate kinase. *J. Phys. Chem. B* **110** (23), 11519–11524. DOI: 10.1021/jp060282a.

[55] Sinev, M. A., E. V. Sineva, V. Ittah, and E. Haas (1996). Domain closure in adenylate kinase. *Biochemistry* **35** (20), 6425–6437. DOI: 10.1021/bi952687j.

[56] Matsunaga, Y., H. Fujisaki, T. Terada, T. Furuta, K. Moritsugu, and A. Kidera (2012). Minimum free energy path of ligand-induced transition in adenylate kinase. *PLoS Comput. Biol.* **8** (6), e1002555. DOI: 10.1371/journal.pcbi.1002555.

[57] Bolhuis, P. G., D. Chandler, C. Dellago, and P. L. Geissler (2002). Transition path sampling: Throwing ropes over rough mountain passes, in the dark. *Annu. Rev. Phys. Chem.* **53** (1), 291–318. DOI: 10.1146/annurev.physchem.53.182301.113146.

[58] Potoyan, D. A., P. I. Zhuravlev, and G. A. Papoian (2012). Computing free energy of a large-scale allosteric transition in adenylate kinase using all atom explicit solvent simulations. *J. Phys. Chem. B* **116** (5), 1709–1715. DOI: 10.1021/jp209980b.

[59] Lou, H. and R. I. Cukier (2006a). Molecular dynamics of apo-adenylate kinase: a principal component analysis. *J. Phys. Chem. B* **110** (25), 12796–12808. DOI: 10.1021/jp061976m.

[60] Hammes, G. G., S. J. Benkovic, and S. Hammes-Schiffer (2011). Flexibility, diversity, and cooperativity: pillars of enzyme catalysis. *Biochemistry* **50** (48), 10422–10430. DOI: 10.1021/bi201486f.

[61] Wolfenden, R. and M. J. Snider (2001). The depth of chemical time and the power of enzymes as catalysts. *Acc. Chem. Res.* **34** (12), 938–945. DOI: 10.1021/ar000058i.

[62] Garcia-Viloca, M., J. Gao, M. Karplus, and D. G. Truhlar (2004). How enzymes work: analysis by modern rate theory and computer simulations. *Science* **303** (5655), 186–195. DOI: 10.1126/science.1088172.

[63] Wolfenden, R. (2006). Degrees of difficulty of water-consuming reactions in the absence of enzymes. *Chem. Rev.* **106** (8), 3379–3396. DOI: 10.1021/cr050311y.

[64] — (1972). Analog approaches to the structure of the transition state in enzyme reactions. *Acc. Chem. Res.* **5** (1), 10–18. DOI: 10.1021/ar50049a002.

[65] Lienhard, G. E. (1973). Enzymatic catalysis and transition-state theory. *Science* **180** (4082), 149–154. DOI: 10.1126/science.180.4082.149.

[66] Pauling, L. (1946). Molecular architecture and biological reactions. *Chem. Eng. News* **24** (10), 1375–1377.

[67] Truhlar, D. G., B. C. Garrett, and S. J. Klippenstein (1996). Current Status of Transition-State Theory. *J. Phys. Chem.* **100** (31), 12771–12800. DOI: 10.1021/jp953748q.

[68] Wang, Y., J. M. Martins, and K. Lindorff-Larsen (2017). Biomolecular conformational changes and ligand binding: from kinetics to thermodynamics. *Chem. Sci.* **8** (9), 6466–6473. DOI: 10.1039/C7SC01627A.

[69] Torrie, G. M. and J. P. Valleau (1977). Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.* **23** (2), 187–199. DOI: 10.1016/0021-9991(77)90121-8.

[70] Kumar, S., J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman (1992). THE weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *J. Comput. Chem.* **13** (8), 1011–1021. DOI: 10.1002/jcc.540130812.

[71] Roux, B. (1995). The calculation of the potential of mean force using computer simulations. *Comput. Phys. Commun.* **91** (1), 275–282.

[72] Shirts, M. R. and J. D. Chodera (2008). Statistically optimal analysis of samples from multiple equilibrium states. *J. Chem. Phys.* **129** (12), 124105. DOI: 10.1063/1.2978177.

[73] Laio, A. and M. Parrinello (2002). Escaping free-energy minima. *Proc. Natl. Acad. Sci. U. S. A.* **99** (20), 12562–12566. DOI: 10.1073/pnas.202427399.

[74] Naidoo, K. J. (2011). FEARCF a multidimensional free energy method for investigating conformational landscapes and chemical reaction mechanisms. *Sci. China Chem.* **54** (12), 1962–1973. DOI: 10.1007/s11426-011-4423-7.

[75] Zhou, H.-X. (2010). Rate theories for biologists. *Q. Rev. Biophys.* **43** (2), 219–293. DOI: 10.1017/S0033583510000120.

[76] Martin, D. R., S. B. Ozkan, and D. V. Matyushov (2012). Dissipative electro-elastic network model of protein electrostatics. *Phys. Biol.* **9** (3), 036004. DOI: 10.1088/1478-3975/9/3/036004.

[77] Martin, D. R. and D. V. Matyushov (2012). Solvated dissipative electro-elastic network model of hydrated proteins. *J. Chem. Phys.* **137** (16), 165101. DOI: 10.1063/1.4759105.

[78] Tiwary, P. and M. Parrinello (2013). From metadynamics to dynamics. *Phys. Rev. Lett.* **111** (23), 230602. DOI: 10.1103/PhysRevLett.111.230602.

[79] Bohner, M. U., J. Zeman, J. Smiatek, A. Arnold, and J. Kästner (2014). Nudged-elastic band used to find reaction coordinates based on the free energy. *J. Chem. Phys.* **140** (7), 074109. DOI: 10.1063/1.4865220.

[80] Fichthorn, K. A. and S. Mubin (2015). Hyperdynamics made simple: Accelerated molecular dynamics with the Bond-Boost method. *Comput. Mater. Sci.* **100, Part B**, 104–110.

[81] Bello-Rivas, J. M. and R. Elber (2015). Exact milestoning. *J. Chem. Phys.* **142** (9), 094102. DOI: 10.1063/1.4913399.

[82] Zwier, M. C., J. L. Adelman, J. W. Kaus, A. J. Pratt, K. F. Wong, N. B. Rego, E. Suárez, S. Lettieri, D. W. Wang, M. Grabe, D. M. Zuckerman, and L. T. Chong (2015). WESTPA: an interoperable, highly scalable software package for weighted ensemble simulation and analysis. *J. Chem. Theory Comput.* **11** (2), 800–809. DOI: 10.1021/ct5010615.

[83] Valsson, O., P. Tiwary, and M. Parrinello (2016). Enhancing Important Fluctuations: Rare Events and Metadynamics from a Conceptual Viewpoint. *Annu. Rev. Phys. Chem.* **67**, 159–184. DOI: 10.1146/annurev-physchem-040215-112229.

[84] Kurkcuoglu, Z., I. Bahar, and P. Doruker (2016). ClustENM: ENM-Based Sampling of Essential Conformational Space at Full Atomic Resolution. *J. Chem. Theory Comput.* DOI: 10.1021/acs.jctc.6b00319.

[85] Yonezawa, Y. (2016). A method for predicting protein conformational pathways by using molecular dynamics simulations guided by difference distance matrices. *J. Comput. Chem.* DOI: 10.1002/jcc.24296.

[86] Shao, Q. (2016). Enhanced conformational sampling technique provides an energy landscape view of large-scale protein conformational transitions. *Phys. Chem. Chem. Phys.* **18** (42), 29170–29182. DOI: 10.1039/c6cp05634b.

[87] Chandrasekaran, S. N., J. Das, N. V. Dokholyan, and C. W. Carter Jr (2016). A modified PATH algorithm rapidly generates transition states comparable to those found by other well established algorithms. *Structural Dynamics* **3** (1), 012101. DOI: 10.1063/1.4941599.

[88] Koehl, P. (2016). Minimum action transition paths connecting minima on an energy surface. *J. Chem. Phys.* **145** (18), 184111. DOI: 10.1063/1.4966974.

[89] Zeller, F. and M. Zacharias (2015). Substrate Binding Specifically Modulates Domain Arrangements in Adenylate Kinase. *Biophys. J.* **109** (9), 1978–1985. DOI: 10.1016/j.bpj.2015.08.049.

[90] Li, D., M. S. Liu, and B. Ji (2015). Mapping the Dynamics Landscape of Conformational Transitions in Enzyme: The Adenylate Kinase Case. *Biophys. J.* **109** (3), 647–660. DOI: 10.1016/j.bpj.2015.06.059.

[91] Formoso, E., V. Limongelli, and M. Parrinello (2015). Energetics and structural characterization of the large-scale functional motion of adenylate kinase. *Sci. Rep.* **5**, 8425. DOI: 10.1038/srep08425.

[92] Lee, J., K. Joo, B. R. Brooks, and J. Lee (2015). The Atomistic Mechanism of Conformational Transition of Adenylate Kinase Investigated by Lorentzian Structure-Based Potential. *J. Chem. Theory Comput.* **11** (7), 3211–3224. DOI: 10.1021/acs.jctc.5b00268.

[93] Unan, H., A. Yildirim, and M. Tekpinar (2015). Opening mechanism of adenylate kinase can vary according to selected molecular dynamics force field. *J. Comput. Aided Mol. Des.* **29** (7), 655–665. DOI: 10.1007/s10822-015-9849-0.

[94] Kerns, S. J., R. V. Agafonov, Y.-J. Cho, F. Pontiggia, R. Otten, D. V. Pachov, S. Kutter, L. A. Phung, P. N. Murphy, V. Thai, T. Alber, M. F. Hagan, and D. Kern (2015). The energy landscape of adenylate kinase during catalysis. *Nat. Struct. Mol. Biol.* **22** (2), 124–131. DOI: 10.1038/nsmb.2941.

[95] Kurkcuoglu, Z. and P. Doruker (2016). Ligand Docking to Intermediate and Close-To-Bound Conformers Generated by an Elastic Network Model Based Algorithm

for Highly Flexible Proteins. *PLoS One* **11** (6), e0158063. DOI: 10.1371/journal.pone.0158063.

[96] Matsunaga, Y., Y. Komuro, C. Kobayashi, J. Jung, T. Mori, and Y. Sugita (2016). Dimensionality of Collective Variables for Describing Conformational Changes of a Multi-Domain Protein. *J. Phys. Chem. Lett.* **7** (8), 1446–1451. DOI: 10.1021/acs.jpclett.6b00317.

[97] Halder, R., R. N. Manna, S. Chakraborty, and B. Jana (2017). Modulation of the Conformational Dynamics of Apo-Adenylate Kinase through a π-Cation Interaction. *J. Phys. Chem. B* **121** (23), 5699–5708. DOI: 10.1021/acs.jpcb.7b01736.

[98] Allen, M. P. and D. J. Tildesley (1989). *Computer Simulation of Liquids*. Revised ed. Oxford Science Publications. Clarendon Press.

[99] Leach, A. R. (2001). *Molecular Modelling: Principles and Applications*. Second. Prentice Hall.

[100] Frenkel, D. and B. Smit (2002). *Understanding Molecular Simulation: From Algorithms to Applications*. 2nd ed. San Diego: Academic Press.

[101] Tuckerman, M. E. (2010). *Statistical Mechanics: Theory and Molecular Simulation*. Oxford Graduate Texts. Great Clarendon Street, Oxford OX2 6DP: Oxford University Press.

[102] Lee, E. H., J. Hsin, M. Sotomayor, G. Comellas, and K. Schulten (2009). Discovery through the computational microscope. *Structure* **17** (10), 1295–1306. DOI: 10.1016/j.str.2009.09.001.

[103] Dror, R. O., R. M. Dirks, J. P. Grossman, H. Xu, and D. E. Shaw (2012). Biomolecular simulation: a computational microscope for molecular biology. *Annu. Rev. Biophys.* **41**, 429–452. DOI: 10.1146/annurev-biophys-042910-155245.

[104] Adcock, S. A. and J. A. McCammon (2006). Molecular dynamics: survey of methods for simulating the activity of proteins. *Chem. Rev.* **106** (5), 1589–1615. DOI: 10.1021/cr040426m.

[105] Kubitzki, M. B., B. L. de Groot, and D. Seeliger (2009). "Protein Dynamics: From Structure to Function". *From Protein Structure to Function with Bioinformatics*. Ed. by D. J. Rigden. Dordrecht: Springer Netherlands, pp. 217–249. DOI: 10.1007/978-1-4020-9058-5_9.

[106] Orozco, M. (2014). A theoretical view of protein dynamics. *Chem. Soc. Rev.* **43** (14), 5051–5066. DOI: 10.1039/c3cs60474h.

[107] Predescu, C., R. A. Lippert, M. P. Eastwood, D. Ierardi, H. Xu, M. Ø. Jensen, K. J. Bowers, J. Gullingsrud, C. A. Rendleman, R. O. Dror, and D. E. Shaw (2012). Com-

putationally efficient molecular dynamics integrators with improved sampling accuracy. *Mol. Phys.* **110** (9-10), 967–983. DOI: 10.1080/00268976.2012.681311.

[108] Donnelly, D. and E. Rogers (2005). Symplectic integrators: An introduction. *Am. J. Phys.* **73** (10), 938–945. DOI: 10.1119/1.2034523.

[109] Sanz-Serna, J. M. and M. P. Calvo (1994). *Numerical Hamiltonian problems*. Chapman & Hall.

[110] Gans, J. and D. Shalloway (2000). Shadow mass and the relationship between velocity and momentum in symplectic numerical integration. *Phys. Rev. E Stat. Phys. Plasmas Fluids Relat. Interdiscip. Topics* **61** (4 Pt B), 4587–4592. DOI: 10.1103/PhysRevE.61.4587.

[111] Engle, R. D., R. D. Skeel, and M. Drees (2005). Monitoring energy drift with shadow Hamiltonians. *J. Comput. Phys.* **206** (2), 432–452. DOI: 10.1016/j.jcp.2004.12.009.

[112] MacKerell, A. D., D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiórkiewicz-Kuczera, D. Yin, and M. Karplus (1998). All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **102** (18), 3586–3616. DOI: 10.1021/jp973084f.

[113] Cornell, W. D., P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman (1995). A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **117** (19), 5179–5197. DOI: 10.1021/ja00124a002.

[114] Scott, W. R. P., P. H. Hünenberger, I. G. Tironi, A. E. Mark, S. R. Billeter, J. Fennen, A. E. Torda, T. Huber, P. Krüger, and W. F. van Gunsteren (1999). The GROMOS Biomolecular Simulation Program Package. *J. Phys. Chem. A* **103** (19), 3596–3607. DOI: 10.1021/jp984217f.

[115] Jorgensen, W. L. and J. Tirado-Rives (1988). The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *J. Am. Chem. Soc.* **110** (6), 1657–1666. DOI: 10.1021/ja00214a001.

[116] Ponder, J. W. and D. A. Case (2003). Force fields for protein simulations. *Adv. Protein Chem.* **66**, 27–85.

[117] Mackerell Jr, A. D. (2004). Empirical force fields for biological macromolecules: overview and issues. *J. Comput. Chem.* **25** (13), 1584–1604. DOI: 10.1002/jcc.20082.

[118] Abraham, M. J., D. van der Spoel, E. Lindahl, B. Hess, and the GROMACS development team (2017). *GROMACS User Manual version 2016.3*. DOI: www.gromacs.org.

[119] Piana, S., K. Lindorff-Larsen, R. M. Dirks, J. K. Salmon, R. O. Dror, and D. E. Shaw (2012). Evaluating the effects of cutoffs and treatment of long-range electrostatics in protein folding simulations. *PLoS One* **7** (6), e39918. DOI: 10.1371/journal.pone.0039918.

[120] Luty, B. A., M. E. Davis, I. G. Tironi, and W. F. Van Gunsteren (1994). A Comparison of Particle-Particle, Particle-Mesh and Ewald Methods for Calculating Electrostatic Interactions in Periodic Molecular Systems. *Mol. Simul.* **14** (1), 11–20. DOI: 10.1080/08927029408022004.

[121] Deserno, M. and C. Holm (1998a). How to mesh up Ewald sums. I. A theoretical and numerical comparison of various particle mesh routines. *J. Chem. Phys.* **109** (18), 7678–7693. DOI: 10.1063/1.477414.

[122] — (1998b). How to mesh up Ewald sums. II. An accurate error estimate for the particle–particle–particle-mesh algorithm. *J. Chem. Phys.* **109** (18), 7694–7701. DOI: 10.1063/1.477415.

[123] Shan, Y., J. L. Klepeis, M. P. Eastwood, R. O. Dror, and D. E. Shaw (2005). Gaussian split Ewald: A fast Ewald mesh method for molecular simulation. *J. Chem. Phys.* **122** (5), 54101. DOI: 10.1063/1.1839571.

[124] Hardy, D. J., Z. Wu, J. C. Phillips, J. E. Stone, R. D. Skeel, and K. Schulten (2015). Multilevel summation method for electrostatic force evaluation. *J. Chem. Theory Comput.* **11** (2), 766–779. DOI: 10.1021/ct5009075.

[125] Sagui, C. and T. A. Darden (1999). Molecular dynamics simulations of biomolecules: long-range electrostatic effects. *Annu. Rev. Biophys. Biomol. Struct.* **28** (1), 155–179. DOI: 10.1146/annurev.biophys.28.1.155.

[126] Fukuda, I. and H. Nakamura (2012). Non-Ewald methods: theory and applications to molecular systems. *Biophys. Rev.* **4** (3), 161–170. DOI: 10.1007/s12551-012-0089-4.

[127] Cisneros, G. A., M. Karttunen, P. Ren, and C. Sagui (2014). Classical electrostatics for biomolecular simulations. *Chem. Rev.* **114** (1), 779–814. DOI: 10.1021/cr300461d.

[128] Lindorff-Larsen, K., P. Maragakis, S. Piana, M. P. Eastwood, R. O. Dror, and D. E. Shaw (2012). Systematic validation of protein force fields against experimental data. *PLoS One* **7** (2), e32131. DOI: 10.1371/journal.pone.0032131.

[129] Car, R. and M. Parrinello (1985). Unified approach for molecular dynamics and density-functional theory. *Phys. Rev. Lett.* **55** (22), 2471–2474. DOI: 10.1103/PhysRevLett.55.2471.

[130] Tuckerman, M. E., P. J. Ungar, T. von Rosenvinge, and M. L. Klein (1996). Ab Initio Molecular Dynamics Simulations. *J. Phys. Chem.* **100** (31), 12878–12887. DOI: 10.1021/jp960480+.

[131] Tuckerman, M. E. and G. J. Martyna (2000). Understanding Modern Molecular Dynamics: Techniques and Applications. *J. Phys. Chem. B* **104** (2), 159–178. DOI: 10.1021/jp992433y.

[132] Tuckerman, M. E. (2002). Ab initio molecular dynamics: basic concepts, current trends and novel applications. *J. Phys. Condens. Matter* **14** (50), R1297.

[133] Freddolino, P. L., S. Park, B. Roux, and K. Schulten (2009). Force field bias in protein folding simulations. *Biophys. J.* **96** (9), 3772–3780. DOI: 10.1016/j.bpj.2009.02.033.

[134] Piana, S., K. Lindorff-Larsen, and D. E. Shaw (2011). How robust are protein folding simulations with respect to force field parameterization? *Biophys. J.* **100** (9), L47–9. DOI: 10.1016/j.bpj.2011.03.051.

[135] Petrov, D. and B. Zagrovic (2014). Are current atomistic force fields accurate enough to study proteins in crowded environments? *PLoS Comput. Biol.* **10** (5), e1003638. DOI: 10.1371/journal.pcbi.1003638.

[136] Palazzesi, F., M. K. Prakash, M. Bonomi, and A. Barducci (2015). Accuracy of current all-atom force-fields in modeling protein disordered states. *J. Chem. Theory Comput.* **11** (1), 2–7. DOI: 10.1021/ct500718s.

[137] Piana, S., A. G. Donchev, P. Robustelli, and D. E. Shaw (2015). Water Dispersion Interactions Strongly Influence Simulated Structural Properties of Disordered Protein States. *J. Phys. Chem. B*, DOI: 10.1021/jp508971m.

[138] Schappals, M., A. Mecklenfeld, L. Kröger, V. Botan, A. Köster, S. Stephan, E. J. García, G. Rutkai, G. Raabe, P. Klein, K. Leonhard, C. W. Glass, J. Lenhard, J. Vrabec, and H. Hasse (2017). Round Robin Study: Molecular Simulation of Thermodynamic Properties from Models with Internal Degrees of Freedom. *J. Chem. Theory Comput.* **13** (9), 4270–4280. DOI: 10.1021/acs.jctc.7b00489.

[139] Mackerell Jr, A. D., M. Feig, and C. L. Brooks 3rd (2004). Extending the treatment of backbone energetics in protein force fields: limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J. Comput. Chem.* **25** (11), 1400–1415. DOI: 10.1002/jcc.20065.

[140] Lindorff-Larsen, K., S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, and D. E. Shaw (2010). Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins* **78** (8), 1950–1958. DOI: 10.1002/prot.22711.

[141] Huang, J. and A. D. MacKerell Jr (2013). CHARMM36 all-atom additive protein force field: validation based on comparison to NMR data. *J. Comput. Chem.* **34** (25), 2135–2145. DOI: 10.1002/jcc.23354.

[142] Robertson, M. J., J. Tirado-Rives, and W. L. Jorgensen (2015). Improved Peptide and Protein Torsional Energetics with the OPLSAA Force Field. *J. Chem. Theory Comput.* **11** (7), 3499–3509. DOI: 10.1021/acs.jctc.5b00356.

[143] Maier, J. A., C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, and C. Simmerling (2015). ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **11** (8), 3696–3713. DOI: 10.1021/acs.jctc.5b00255.

[144] Lindorff-Larsen, K., S. Piana, R. O. Dror, and D. E. Shaw (2011). How fast-folding proteins fold. *Science* **334** (6055), 517–520. DOI: 10.1126/science.1208351.

[145] Piana, S., J. L. Klepeis, and D. E. Shaw (2014). Assessing the accuracy of physical models used in protein-folding simulations: quantitative evidence from long molecular dynamics simulations. *Curr. Opin. Struct. Biol.* **24**, 98–105. DOI: 10.1016/j.sbi.2013.12.006.

[146] Medders, G. R. and F. Paesani (2015). Infrared and Raman Spectroscopy of Liquid Water through "First-Principles" Many-Body Molecular Dynamics. *J. Chem. Theory Comput.* **11** (3), 1145–1154. DOI: 10.1021/ct501131j.

[147] Ponder, J. W., C. Wu, P. Ren, V. S. Pande, J. D. Chodera, M. J. Schnieders, I. Haque, D. L. Mobley, D. S. Lambrecht, R. A. DiStasio Jr, M. Head-Gordon, G. N. I. Clark, M. E. Johnson, and T. Head-Gordon (2010). Current status of the AMOEBA polarizable force field. *J. Phys. Chem. B* **114** (8), 2549–2564. DOI: 10.1021/jp910674d.

[148] Jiang, W., D. J. Hardy, J. C. Phillips, A. D. Mackerell Jr, K. Schulten, and B. Roux (2011). High-performance scalable molecular dynamics simulations of a polarizable force field based on classical Drude oscillators in NAMD. *J. Phys. Chem. Lett.* **2** (2), 87–92. DOI: 10.1021/jz101461d.

[149] Lopes, P. E. M., J. Huang, J. Shim, Y. Luo, H. Li, B. Roux, and A. D. Mackerell Jr (2013). Force Field for Peptides and Proteins based on the Classical Drude Oscillator. *J. Chem. Theory Comput.* **9** (12), 5430–5449. DOI: 10.1021/ct400781b.

[150] Baptista, A. M., V. H. Teixeira, and C. M. Soares (2002). Constant-pH molecular dynamics using stochastic titration. *J. Chem. Phys.* **117** (9), 4184–4200. DOI: 10.1063/1.1497164.

[151] Huang, Y., W. Chen, J. A. Wallace, and J. Shen (2016). All-Atom Continuous Constant pH Molecular Dynamics With Particle Mesh Ewald and Titratable Water. *J. Chem. Theory Comput.* **12** (11), 5411–5421. DOI: 10.1021/acs.jctc.6b00552.

[152] Donnini, S., R. T. Ullmann, G. Groenhof, and H. Grubmüller (2016). Charge-Neutral Constant pH Molecular Dynamics Simulations Using a Parsimonious Proton Buffer. *J. Chem. Theory Comput.* **12** (3), 1040–1051. DOI: 10.1021/acs.jctc.5b01160.

[153] Krasnoshchekov, S. V. and N. F. Stepanov (2008). Anharmonic Force Fields and Perturbation Theory in the Interpretation of Vibrational Spectra of Polyatomic Molecules. *Russ. J. Phys. Chem.* **82** (4), 592–602. DOI: 10.1134/S0036024408040158.

[154] Császár, A. G. (2012). Anharmonic molecular force fields. *WIREs Comput Mol Sci* **2** (2), 273–289. DOI: 10.1002/wcms.75.

[155] Hagler, A. T. (2015). Quantum Derivative Fitting and Biomolecular Force Fields: Functional Form, Coupling Terms, Charge Flux, Nonbond Anharmonicity, and Individual Dihedral Potentials. *J. Chem. Theory Comput.* **11** (12), 5555–5572. DOI: 10.1021/acs.jctc.5b00666.

[156] Yin, J., A. T. Fenley, N. M. Henriksen, and M. K. Gilson (2015). Toward Improved Force-Field Accuracy through Sensitivity Analysis of Host-Guest Binding Thermodynamics. *J. Phys. Chem. B* **119** (32), 10145–10155. DOI: 10.1021/acs.jpcb.5b04262.

[157] Glielmo, A., P. Sollich, and A. De Vita (2017). Accurate interatomic force fields via machine learning with covariant kernels. *Phys. Rev. B Condens. Matter* **95** (21), 214302. DOI: 10.1103/PhysRevB.95.214302.

[158] Huan, T. D., R. Batra, J. Chapman, S. Krishnan, L. Chen, and R. Ramprasad (2017). A universal strategy for the creation of machine learning-based atomistic force fields. *npj Computational Materials* **3** (1), 37. DOI: 10.1038/s41524-017-0042-y.

[159] Vashisth, H. and C. L. Brooks 3rd (2012). Conformational Sampling of Maltose-transporter Components in Cartesian Collective Variables is Governed by the Low-frequency Normal Modes. *J. Phys. Chem. Lett.* **3** (22), 3379–3384. DOI: 10.1021/jz301650q.

[160] Stone, J. E., D. J. Hardy, I. S. Ufimtsev, and K. Schulten (2010). GPU-accelerated molecular modeling coming of age. *J. Mol. Graph. Model.* **29** (2), 116–125. DOI: 10.1016/j.jmgm.2010.06.010.

[161] Götz, A. W., M. J. Williamson, D. Xu, D. Poole, S. Le Grand, and R. C. Walker (2012). Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 1. Generalized Born. *J. Chem. Theory Comput.* **8** (5), 1542–1555. DOI: 10.1021/ct200909j.

[162] Salomon-Ferrer, R., A. W. Götz, D. Poole, S. Le Grand, and R. C. Walker (2013). Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. *J. Chem. Theory Comput.* **9** (9), 3878–3888. DOI: 10.1021/ct400314y.

[163] Kutzner, C., S. Páll, M. Fechner, A. Esztermann, B. L. de Groot, and H. Grubmüller (2015). Best bang for your buck: GPU nodes for GROMACS biomolecular simulations. *J. Comput. Chem.* **36** (26), 1990–2008. DOI: 10.1002/jcc.24030.

[164] Eastman, P., J. Swails, J. D. Chodera, R. T. McGibbon, Y. Zhao, K. A. Beauchamp, L.-P. Wang, A. C. Simmonett, M. P. Harrigan, C. D. Stern, R. P. Wiewiora, B. R. Brooks, and V. S. Pande (2017). OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comput. Biol.* **13** (7), e1005659. DOI: 10.1371/journal.pcbi.1005659.

[165] Lyman, E. and D. M. Zuckerman (2007). On the structural convergence of biomolecular simulations by determination of the effective sample size. *J. Phys. Chem. B* **111** (44), 12876–12882. DOI: 10.1021/jp073061t.

[166] Zuckerman, D. M. (2011). Equilibrium sampling in biomolecular simulations. *Annu. Rev. Biophys.* **40**, 41–62. DOI: 10.1146/annurev-biophys-042910-155255.

[167] Schwartz, S. D. and V. L. Schramm (2009). Enzymatic transition states and dynamic motion in barrier crossing. *Nat. Chem. Biol.* **5** (8), 551–558. DOI: 10.1038/nchembio.202.

[168] Lei, H. and Y. Duan (2007). Improved sampling methods for molecular simulation. *Curr. Opin. Struct. Biol.* **17** (2), 187–191. DOI: 10.1016/j.sbi.2007.03.003.

[169] Chng, C.-P. and L.-W. Yang (2008). Coarse-grained models reveal functional dynamics–II. Molecular dynamics simulation at the coarse-grained level–theories and biological applications. *Bioinform. Biol. Insights* **2**, 171–185.

[170] Christen, M. and W. F. van Gunsteren (2008). On searching in, sampling of, and dynamically moving through conformational space of biomolecular systems: A review. *J. Comput. Chem.* **29** (2), 157–166. DOI: 10.1002/jcc.20725.

[171] Maximova, T., R. Moffatt, B. Ma, R. Nussinov, and A. Shehu (2016). Principles and Overview of Sampling Methods for Modeling Macromolecular Structure and Dynamics. *PLoS Comput. Biol.* **12** (4), e1004619. DOI: 10.1371/journal.pcbi.1004619.

[172] Woolf, T. B. (1998). Path corrected functionals of stochastic trajectories: towards relative free energy and reaction coordinate calculations. *Chem. Phys. Lett.* **289** (5), 433–441. DOI: 10.1016/s0009-2614(98)00427-8.

[173] Zuckerman, D. M. and T. B. Woolf (1999). Dynamic reaction paths and rates through importance-sampled stochastic dynamics. *J. Chem. Phys.* **111** (21), 9475–9484. DOI: 10.1063/1.480278.

[174] Perilla, J. R., O. Beckstein, E. J. Denning, and T. B. Woolf (2011). Computing ensembles of transitions from stable states: Dynamic importance sampling. *J. Comput. Chem.* **32** (2), 196–209. DOI: 10.1002/jcc.21564.

[175] Flores, S., N. Echols, D. Milburn, B. Hespenheide, K. Keating, J. Lu, S. Wells, E. Z. Yu, M. Thorpe, and M. Gerstein (2006). The Database of Macromolecular Motions: new

features added at the decade mark. *Nucleic Acids Res.* **34** (Database issue), D296–301. DOI: 10.1093/nar/gkj046.

[176] Krebs, W. G. and M. Gerstein (2000). The morph server: a standardized system for analyzing and visualizing macromolecular motions in a database framework. *Nucleic Acids Res.* **28** (8), 1665–1675.

[177] Weiss, D. R. and M. Levitt (2009). Can morphing methods predict intermediate structures? *J. Mol. Biol.* **385** (2), 665–674. DOI: 10.1016/j.jmb.2008.10.064.

[178] Dellago, C., P. G. Bolhuis, F. S. Csajka, and D. Chandler (1998). Transition path sampling and the calculation of rate constants. *J. Chem. Phys.* **108** (5), 1964–1977. DOI: 10.1063/1.475562.

[179] Taketomi, H., Y. Ueda, and N. Gō (1975). Studies on protein folding, unfolding and fluctuations by computer simulation. I. The effect of specific amino acid sequence represented by specific inter-unit interactions. *Int. J. Pept. Protein Res.* **7** (6), 445–459.

[180] Farrell, D. W., K. Speranskiy, and M. F. Thorpe (2010). Generating stereochemically acceptable protein pathways. *Proteins* **78** (14), 2908–2921. DOI: 10.1002/prot.22810.

[181] Voter, A. F. (1997). Hyperdynamics: Accelerated Molecular Dynamics of Infrequent Events. *Phys. Rev. Lett.* **78** (20), 3908–3911. DOI: 10.1103/PhysRevLett.78.3908.

[182] Sugita, Y. and Y. Okamoto (1999). Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **314** (1), 141–151. DOI: 10.1016/S0009-2614(99)01123-9.

[183] Hamelberg, D., J. Mongan, and J. A. McCammon (2004). Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules. *J. Chem. Phys.* **120** (24), 11919–11929. DOI: 10.1063/1.1755656.

[184] Kubitzki, M. B. and B. L. de Groot (2008). The atomistic mechanism of conformational transition in adenylate kinase: a TEE-REX molecular dynamics study. *Structure* **16** (8), 1175–1182. DOI: 10.1016/j.str.2008.04.013.

[185] Barnett, C. B. and K. J. Naidoo (2009). Free Energies from Adaptive Reaction Coordinate Forces (FEARCF): an application to ring puckering. *Mol. Phys.* **107** (8-12), 1243–1250. DOI: 10.1080/00268970902852608.

[186] Abrams, C. and G. Bussi (2013). Enhanced Sampling in Molecular Dynamics Using Metadynamics, Replica-Exchange, and Temperature-Acceleration. *Entropy* **16** (1), 163–199. DOI: 10.3390/e16010163.

[187] Snow, C., G. Qi, and S. Hayward (2007). Essential dynamics sampling study of adenylate kinase: comparison to citrate synthase and implication for the hinge and

shear mechanisms of domain motions. *Proteins* **67** (2), 325–337. DOI: 10.1002/prot.21280.

[188] Zuckerman, D. M. and T. B. Woolf (2001). Efficient dynamic importance sampling of rare events in one dimension. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **63** (1 Pt 2), 016702. DOI: 10.1103/PhysRevE.63.016702.

[189] Marchi, M. and P. Ballone (1999). Adiabatic bias molecular dynamics: A method to navigate the conformational space of complex molecular systems. *J. Chem. Phys.* **110** (8), 3697–3702. DOI: 10.1063/1.478259.

[190] Tiana, G. and C. Camilloni (2012). Ratcheted molecular-dynamics simulations identify efficiently the transition state of protein folding. *J. Chem. Phys.* **137** (23), 235101. DOI: 10.1063/1.4769085.

[191] Ferrara, P., J. Apostolakis, and A. Caflisch (2000a). Targeted Molecular Dynamics Simulations of Protein Unfolding. *J. Phys. Chem. B* **104** (18), 4511–4518. DOI: 10.1021/jp9943878.

[192] Vaart, A. van der and M. Karplus (2005). Simulation of conformational transitions by the restricted perturbation-targeted molecular dynamics method. *J. Chem. Phys.* **122** (11), 114903. DOI: 10.1063/1.1861885.

[193] — (2007). Minimum free energy pathways and free energy profiles for conformational transitions based on atomistic molecular dynamics simulations. *J. Chem. Phys.* **126** (16), 164106. DOI: 10.1063/1.2719697.

[194] Darve, E., D. Rodríguez-Gómez, and A. Pohorille (2008). Adaptive biasing force method for scalar and vector free energy calculations. *J. Chem. Phys.* **128** (14), 144120. DOI: 10.1063/1.2829861.

[195] Hénin, J., G. Fiorin, C. Chipot, and M. L. Klein (2010). Exploring Multidimensional Free Energy Landscapes Using Time-Dependent Biases on Collective Variables. *J. Chem. Theory Comput.* **6** (1), 35–47. DOI: 10.1021/ct9004432.

[196] Barducci, A., G. Bussi, and M. Parrinello (2008). Well-tempered metadynamics: a smoothly converging and tunable free-energy method. *Phys. Rev. Lett.* **100** (2), 020603. DOI: 10.1103/PhysRevLett.100.020603.

[197] Maragliano, L. and E. Vanden-Eijnden (2006). A temperature accelerated method for sampling free energy and determining reaction pathways in rare events simulations. *Chem. Phys. Lett.* **426** (1), 168–175. DOI: 10.1016/j.cplett.2006.05.062.

[198] Abrams, C. F. and E. Vanden-Eijnden (2010). Large-scale conformational sampling of proteins using temperature-accelerated molecular dynamics. *Proc. Natl. Acad. Sci. U. S. A.* **107** (11), 4961–4966. DOI: 10.1073/pnas.0914540107.

[199] Noid, W. G. (2013). Perspective: Coarse-grained models for biomolecular systems. *J. Chem. Phys.* **139** (9), 090901. DOI: 10.1063/1.4818908.

[200] Whitford, P. C., K. Y. Sanbonmatsu, and J. N. Onuchic (2012). Biomolecular dynamics: order-disorder transitions and energy landscapes. *Rep. Prog. Phys.* **75** (7), 076601. DOI: 10.1088/0034-4885/75/7/076601.

[201] Whitford, P. C., O. Miyashita, Y. Levy, and J. N. Onuchic (2007). Conformational transitions of adenylate kinase: switching by cracking. *J. Mol. Biol.* **366** (5), 1661–1671. DOI: 10.1016/j.jmb.2006.11.085.

[202] Whitford, P. C., J. N. Onuchic, and P. G. Wolynes (2008). Energy landscape along an enzymatic reaction trajectory: hinges or cracks? *HFSP J.* **2** (2), 61–64. DOI: 10.2976/1.2894846.

[203] Lu, Q. and J. Wang (2008). Single molecule conformational dynamics of adenylate kinase: energy landscape, structural correlations, and transition state ensembles. *J. Am. Chem. Soc.* **130** (14), 4772–4783. DOI: 10.1021/ja0780481.

[204] Daily, M. D., G. N. Phillips Jr, and Q. Cui (2010). Many local motions cooperate to produce the adenylate kinase conformational transition. *J. Mol. Biol.* **400** (3), 618–631. DOI: 10.1016/j.jmb.2010.05.015.

[205] Tirion, M. M. (1996). Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis. *Phys. Rev. Lett.* **77** (9), 1905–1908. DOI: 10.1103/PhysRevLett.77.1905.

[206] Bahar, I., A. R. Atilgan, and B. Erman (1997). Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Fold. Des.* **2** (3), 173–181. DOI: 10.1016/S1359-0278(97)00024-2.

[207] Atilgan, A. R., S. R. Durell, R. L. Jernigan, M. C. Demirel, O. Keskin, and I. Bahar (2001). Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys. J.* **80** (1), 505–515. DOI: 10.1016/S0006-3495(01)76033-X.

[208] Zheng, W., B. R. Brooks, and G. Hummer (2007). Protein conformational transitions explored by mixed elastic network models. *Proteins* **69** (1), 43–57. DOI: 10.1002/prot.21465.

[209] Tekpinar, M. and W. Zheng (2010). Predicting order of conformational changes during protein conformational transitions using an interpolated elastic network model. *Proteins* **78** (11), 2469–2481. DOI: 10.1002/prot.22755.

[210] Bahar, I., T. R. Lezon, L.-W. Yang, and E. Eyal (2010a). Global dynamics of proteins: bridging between structure and function. *Annu. Rev. Biophys.* **39**, 23–42. DOI: 10.1146/annurev.biophys.093008.131258.

[211] Bahar, I., T. R. Lezon, A. Bakan, and I. H. Shrivastava (2010b). Normal mode analysis of biomolecular structures: functional mechanisms of membrane proteins. *Chem. Rev.* **110** (3), 1463–1497. DOI: 10.1021/cr900095e.

[212] Sanejouand, Y.-H. (2013). Elastic network models: theoretical and empirical foundations. *Methods Mol. Biol.* **924**, 601–616. DOI: 10.1007/978-1-62703-017-5_23.

[213] Hinsen, K. (1998). Analysis of domain motions by approximate normal mode calculations. *Proteins* **33** (3), 417–429. DOI: 10.1002/(SICI)1097-0134(19981115)33:3<417::AID-PROT10>3.0.CO;2-8.

[214] Hinsen, K. and G. R. Kneller (1999). A simplified force field for describing vibrational protein dynamics over the whole frequency range. *J. Chem. Phys.* **111** (24), 10766–10769. DOI: 10.1063/1.480441.

[215] Peng, C. and L. Zhang (2009). "Assessing Iterative Normal Modes on Representing Protein Conformational Transitions". *2009 3rd International Conference on Bioinformatics and Biomedical Engineering*, pp. 1–5. DOI: 10.1109/ICBBE.2009.5162411.

[216] Jimenez-Roldan, J. E., R. B. Freedman, R. A. Römer, and S. A. Wells (2012). Rapid simulation of protein motion: merging flexibility, rigidity and normal mode analyses. *Phys. Biol.* **9** (1), 016008. DOI: 10.1088/1478-3975/9/1/016008.

[217] López-Blanco, J. R., J. I. Aliaga, E. S. Quintana-Ortí, and P. Chacón (2014). iMODS: internal coordinates normal mode analysis server. *Nucleic Acids Res.* **42** (Web Server issue), W271–6. DOI: 10.1093/nar/gku339.

[218] Franklin, J., P. Koehl, S. Doniach, and M. Delarue (2007). MinActionPath: maximum likelihood trajectory for large-scale structural transitions in a coarse-grained locally harmonic energy landscape. *Nucleic Acids Res.* **35** (suppl_2), W477–82. DOI: 10.1093/nar/gkm342.

[219] Chu, J.-W. and G. A. Voth (2007). Coarse-grained free energy functions for studying protein conformational changes: a double-well network model. *Biophys. J.* **93** (11), 3860–3871. DOI: 10.1529/biophysj.107.112060.

[220] Kantarci-Carsibasi, N., T. Haliloglu, and P. Doruker (2008). Conformational transition pathways explored by Monte Carlo simulation integrated with collective modes. *Biophys. J.* **95** (12), 5862–5873. DOI: 10.1529/biophysj.107.128447.

[221] Ahmed, A., F. Rippmann, G. Barnickel, and H. Gohlke (2011). A normal mode-based geometric simulation approach for exploring biologically relevant conformational transitions in proteins. *J. Chem. Inf. Model.* **51** (7), 1604–1622. DOI: 10.1021/ci100461k.

[222] Gur, M., J. D. Madura, and I. Bahar (2013). Global transitions of proteins explored by a multiscale hybrid methodology: application to adenylate kinase. *Biophys. J.* **105** (7), 1643–1652. DOI: 10.1016/j.bpj.2013.07.058.

[223] Sfriso, P., A. Hospital, A. Emperador, and M. Orozco (2013). Exploration of conformational transition pathways from coarse-grained simulations. *Bioinformatics* **29** (16), 1980–1986. DOI: 10.1093/bioinformatics/btt324.

[224] Kirillova, S., J. Cortés, A. Stefaniu, and T. Siméon (2008). An NMA-guided path planning approach for computing large-amplitude conformational changes in proteins. *Proteins* **70** (1), 131–143. DOI: 10.1002/prot.21570.

[225] Raveh, B., A. Enosh, O. Schueler-Furman, and D. Halperin (2009). Rapid sampling of molecular motions with prior information constraints. *PLoS Comput. Biol.* **5** (2), e1000295. DOI: 10.1371/journal.pcbi.1000295.

[226] Kavraki, L. E., P. Svestka, J. C. Latombe, and M. H. Overmars (1996). Probabilistic roadmaps for path planning in high-dimensional configuration spaces. *IEEE Trans. Rob. Autom.* **12** (4), 566–580. DOI: 10.1109/70.508439.

[227] Apaydin, M. S., A. P. Singh, D. L. Brutlag, and J. C. Latombe (2001). "Capturing molecular energy landscapes with probabilistic conformational roadmaps". *Proceedings 2001 ICRA. IEEE International Conference on Robotics and Automation (Cat. No.01CH37164)*. Vol. 1, 932–939 vol.1. DOI: 10.1109/ROBOT.2001.932670.

[228] Head, M. S., J. A. Given, and M. K. Gilson (1997). "Mining Minima": Direct Computation of Conformational Free Energy. *J. Phys. Chem. A* **101** (8), 1609–1618. DOI: 10.1021/jp963817g.

[229] Seeliger, D., J. Haas, and B. L. de Groot (2007). Geometry-based sampling of conformational transitions in proteins. *Structure* **15** (11), 1482–1492. DOI: 10.1016/j.str.2007.09.017.

[230] Cortés, J., T. Siméon, V. Ruiz de Angulo, D. Guieysse, M. Remaud-Siméon, and V. Tran (2005). A path planning approach for computing large-amplitude motions of flexible molecules. *Bioinformatics* **21 Suppl 1**, i116–25. DOI: 10.1093/bioinformatics/bti1017.

[231] Erp, T. S. van, D. Moroni, and P. G. Bolhuis (2003). A novel path sampling method for the calculation of rate constants. *J. Chem. Phys.* **118** (17), 7762–7774. DOI: 10.1063/1.1562614.

[232] Erp, T. S. van and P. G. Bolhuis (2005). Elaborating transition interface sampling methods. *J. Comput. Phys.* **205** (1), 157–181.

[233] Allen, R. J., D. Frenkel, and P. R. ten Wolde (2006a). Simulating rare events in equilibrium or nonequilibrium stochastic systems. *J. Chem. Phys.* **124** (2), 024102. DOI: 10.1063/1.2140273.

[234] — (2006b). Forward flux sampling-type schemes for simulating rare events: efficiency analysis. *J. Chem. Phys.* **124** (19), 194111. DOI: 10.1063/1.2198827.

[235] Allen, R. J., C. Valeriani, and P. Rein Ten Wolde (2009). Forward flux sampling for rare event simulations. *J. Phys. Condens. Matter* **21** (46), 463102. DOI: 10.1088/0953-8984/21/46/463102.

[236] Huber, G. A. and S. Kim (1996). Weighted-ensemble Brownian dynamics simulations for protein association reactions. *Biophys. J.* **70** (1), 97–110. DOI: 10.1016/S0006-3495(96)79552-8.

[237] Bhatt, D., B. W. Zhang, and D. M. Zuckerman (2010). Steady-state simulations using weighted ensemble path sampling. *J. Chem. Phys.* **133** (1), 014110. DOI: 10.1063/1.3456985.

[238] Zhang, B. W., D. Jasnow, and D. M. Zuckerman (2010). The "weighted ensemble" path sampling method is statistically exact for a broad class of stochastic processes and binning procedures. *J. Chem. Phys.* **132** (5), 054107. DOI: 10.1063/1.3306345.

[239] Vanden-Eijnden, E. and M. Venturoli (2009a). Markovian milestoning with Voronoi tessellations. *J. Chem. Phys.* **130** (19), 194101. DOI: 10.1063/1.3129843.

[240] E, W., W. Ren, and E. Vanden-Eijnden (2005a). Finite temperature string method for the study of rare events. *J. Phys. Chem. B* **109** (14), 6688–6693. DOI: 10.1021/jp0455430.

[241] — (2005b). Transition pathways in complex systems: Reaction coordinates, isocommittor surfaces, and transition tubes. *Chem. Phys. Lett.* **413** (1), 242–247. DOI: 10.1016/j.cplett.2005.07.084.

[242] Maragliano, L., A. Fischer, E. Vanden-Eijnden, and G. Ciccotti (2006). String method in collective variables: minimum free energy paths and isocommittor surfaces. *J. Chem. Phys.* **125** (2), 24106. DOI: 10.1063/1.2212942.

[243] Maragliano, L. and E. Vanden-Eijnden (2007). On-the-fly string method for minimum free energy paths calculation. *Chem. Phys. Lett.* **446** (1), 182–190. DOI: 10.1016/j.cplett.2007.08.017.

[244] Pan, A. C., D. Sezer, and B. Roux (2008). Finding transition pathways using the string method with swarms of trajectories. *J. Phys. Chem. B* **112** (11), 3432–3440. DOI: 10.1021/jp0777059.

[245] E, W., W. Ren, and E. Vanden-Eijnden (2002). String method for the study of rare events. *Phys. Rev. B Condens. Matter* **66** (5), 052301. DOI: 10.1103/PhysRevB.66.052301.

[246] Jónsson, H., G. Mills, and K. W. Jacobsen (1998). "Nudged elastic band method for finding minimum energy paths of transitions". *Classical and Quantum Dynamics in Condensed Phase Simulations*. Ed. by B. J. Berne, G. Ciccoti, and D. F. Coker. World Scientific. Chap. 16, pp. 385–394.

[247] Henkelman, G. and H. Jónsson (2000). Improved tangent estimate in the nudged elastic band method for finding minimum energy paths and saddle points. *J. Chem. Phys.* **113** (22), 9978–9985. DOI: 10.1063/1.1323224.

[248] Henkelman, G., B. P. Uberuaga, and H. Jónsson (2000). A climbing image nudged elastic band method for finding saddle points and minimum energy paths. *J. Chem. Phys.* **113** (22), 9901–9904. DOI: 10.1063/1.1329672.

[249] Fischer, S. and M. Karplus (1992). Conjugate peak refinement: an algorithm for finding reaction paths and accurate transition states in systems with many degrees of freedom. *Chem. Phys. Lett.* **194** (3), 252–261. DOI: 10.1016/0009-2614(92)85543-j.

[250] Sugita, Y. and Y. Okamoto (2000). Replica-exchange multicanonical algorithm and multicanonical replica-exchange method for simulating systems with rough energy landscape. *Chem. Phys. Lett.* **329** (3), 261–270.

[251] Steiner, M. M., P.-A. Genilloud, and J. W. Wilkins (1998). Simple bias potential for boosting molecular dynamics with the hyperdynamics scheme. *Phys. Rev. B Condens. Matter* **57**, 10236–10239. DOI: 10.1103/PhysRevB.57.10236.

[252] Sørensen, M. R. and A. F. Voter (2000). Temperature-accelerated dynamics for simulation of infrequent events. *J. Chem. Phys.* **112** (21), 9599–9606. DOI: 10.1063/1.481576.

[253] Montalenti, F. and A. F. Voter (2002). Exploiting past visits or minimum-barrier knowledge to gain further boost in the temperature-accelerated dynamics method. *J. Chem. Phys.* **116** (12), 4819–4828. DOI: 10.1063/1.1449865.

[254] Lou, H. and R. I. Cukier (2006b). Molecular dynamics of apo-adenylate kinase: a distance replica exchange method for the free energy of conformational fluctuations. *J. Phys. Chem. B* **110** (47), 24121–24137. DOI: 10.1021/jp064303c.

[255] Bhatt, D. and D. M. Zuckerman (2010). Heterogeneous path ensembles for conformational transitions in semi-atomistic models of adenylate kinase. *J. Chem. Theory Comput.* **6** (11), 3527–3539. DOI: 10.1021/ct100406t.

[256] Arora, K. and C. L. Brooks 3rd (2007). Large-scale allosteric conformational transitions of adenylate kinase appear to involve a population-shift mechanism. *Proc. Natl. Acad. Sci. U. S. A.* **104** (47), 18496–18501. DOI: 10.1073/pnas.0706443104.

[257] Song, H. D. and F. Zhu (2013). Conformational dynamics of a ligand-free adenylate kinase. *PLoS One* **8** (7), e68023. DOI: 10.1371/journal.pone.0068023.

[258] Brokaw, J. B. and J.-W. Chu (2010). On the roles of substrate binding and hinge unfolding in conformational changes of adenylate kinase. *Biophys. J.* **99** (10), 3420–3429. DOI: 10.1016/j.bpj.2010.09.040.

[259] Jana, B., B. V. Adkar, R. Biswas, and B. Bagchi (2011). Dynamic coupling between the LID and NMP domain motions in the catalytic conversion of ATP and AMP to ADP by adenylate kinase. *J. Chem. Phys.* **134** (3), 035101. DOI: 10.1063/1.3516588.

[260] Wang, Y., L. Gan, E. Wang, and J. Wang (2013a). Exploring the Dynamic Functional Landscape of Adenylate Kinase Modulated by Substrates. *J. Chem. Theory Comput.* **9** (1), 84–95. DOI: 10.1021/ct300720s.

[261] Peng, C., L. Zhang, and T. Head-Gordon (2010). Instantaneous normal modes as an unforced reaction coordinate for protein conformational transitions. *Biophys. J.* **98** (10), 2356–2364. DOI: 10.1016/j.bpj.2010.01.044.

[262] Whitford, P. C., S. Gosavi, and J. N. Onuchic (2008). Conformational transitions in adenylate kinase. Allosteric communication reduces misligation. *J. Biol. Chem.* **283** (4), 2042–2048. DOI: 10.1074/jbc.M707632200.

[263] Adkar, B. V., B. Jana, and B. Bagchi (2011). Role of water in the enzymatic catalysis: study of ATP + AMP $\rightarrow$ 2ADP conversion by adenylate kinase. *J. Phys. Chem. A* **115** (16), 3691–3697. DOI: 10.1021/jp104787s.

[264] Pontiggia, F., A. Zen, and C. Micheletti (2008). Small- and large-scale conformational changes of adenylate kinase: a molecular dynamics study of the subdomain motion and mechanics. *Biophys. J.* **95** (12), 5901–5912. DOI: 10.1529/biophysj.108.135467.

[265] Delalande, O., S. Sacquin-Mora, and M. Baaden (2011). Enzyme closure and nucleotide binding structurally lock guanylate kinase. *Biophys. J.* **101** (6), 1440–1449. DOI: 10.1016/j.bpj.2011.07.048.

[266] Shan, Y., M. A. Seeliger, M. P. Eastwood, F. Frank, H. Xu, M. Ø. Jensen, R. O. Dror, J. Kuriyan, and D. E. Shaw (2009). A conserved protonation-dependent switch controls drug binding in the Abl kinase. *Proc. Natl. Acad. Sci. U. S. A.* **106** (1), 139–144. DOI: 10.1073/pnas.0811223106.

[267] Bae, E. and G. N. Phillips Jr (2005). Identifying and engineering ion pairs in adenylate kinases. Insights from molecular dynamics simulations of thermophilic and mesophilic homologues. *J. Biol. Chem.* **280** (35), 30943–30948. DOI: 10.1074/jbc.M504216200.

[268] Kundu, S. and D. Roy (2009). Comparative structural studies of psychrophilic and mesophilic protein homologues by molecular dynamics simulation. *J. Mol. Graph. Model.* **27** (8), 871–880. DOI: 10.1016/j.jmgm.2009.01.004.

[269] Korkut, A. and W. A. Hendrickson (2009). Computation of conformational transitions in proteins by virtual atom molecular mechanics as validated in application to adenylate kinase. *Proc. Natl. Acad. Sci. U. S. A.* **106** (37), 15673–15678. DOI: 10.1073/pnas.0907684106.

[270] Pan, A. C., T. M. Weinreich, Y. Shan, D. P. Scarpazza, and D. E. Shaw (2014). Assessing the Accuracy of Two Enhanced Sampling Methods Using EGFR Kinase Transition Pathways: The Influence of Collective Variable Choice. *J. Chem. Theory Comput.* **10** (7), 2860–2865. DOI: 10.1021/ct500223p.

[271] Skeel, R. D., R. Zhao, and C. B. Post (2017). A minimization principle for transition paths of maximum flux for collective variables. *Theor. Chem. Acc.* **136** (1), 14. DOI: 10.1007/s00214-016-2041-3.

[272] Michaud-Agrawal, N., E. J. Denning, T. B. Woolf, and O. Beckstein (2011). MDAnalysis: a toolkit for the analysis of molecular dynamics simulations. *J. Comput. Chem.* **32** (10), 2319–2327. DOI: 10.1002/jcc.21787.

[273] Gowers, R. J., M. Linke, J. Barnoud, T. J. E. Reddy, M. N. Melo, S. L. Seyler, D. L. Dotson, J. Domanski, S. Buchoux, I. M. Kenney, et al. (2016). "MDAnalysis: a Python package for the rapid analysis of molecular dynamics simulations". *Proceedings of the 15th Python in Science Conference, Austin, TX*.

[274] Zuckerman, D. M. (2017). *What I have against (most) PMF calculations*. URL: http://statisticalbiophysicsblog.org/.

[275] Maas, U. and S. B. Pope (1992). Simplifying chemical kinetics: Intrinsic low-dimensional manifolds in composition space. *Combust. Flame* **88** (3), 239–264. DOI: 10.1016/0010-2180(92)90034-M.

[276] Teodoro, M. L., G. N. Phillips Jr, and L. E. Kavraki (2003). Understanding protein flexibility through dimensionality reduction. *J. Comput. Biol.* **10** (3-4), 617–634. DOI: 10.1089/10665270360688228.

[277] Mesentean, S., S. Fischer, and J. C. Smith (2006). Analyzing large-scale structural change in proteins: comparison of principal component projection and Sammon mapping. *Proteins* **64** (1), 210–218. DOI: 10.1002/prot.20981.

[278] Balsera, M. A., W. Wriggers, Y. Oono, and K. Schulten (1996). Principal Component Analysis and Long Time Protein Dynamics. *J. Phys. Chem.* **100** (7), 2567–2572. DOI: 10.1021/jp9536920.

[279] Kitao, A. and N. Go (1999). Investigating protein dynamics in collective coordinate space. *Curr. Opin. Struct. Biol.* **9** (2), 164–169. DOI: 10.1016/S0959-440X(99)80023-2.

[280] Best, R. B., G. Hummer, and W. A. Eaton (2013). Native contacts determine protein folding mechanisms in atomistic simulations. *Proc. Natl. Acad. Sci. U. S. A.* **110** (44), 17874–17879. DOI: 10.1073/pnas.1311599110.

[281] Huttenlocher, D. P., G. A. Klanderman, and W. J. Rucklidge (1993). Comparing images using the Hausdorff distance. *IEEE Trans. Pattern Anal. Mach. Intell.* **15** (9), 850–863. DOI: 10.1109/34.232073.

[282] Alt, H., B. Behrends, and J. Blömer (1995). Approximate matching of polygonal shapes. *Ann. Math. Artif. Intell.* **13** (3-4), 251–265. DOI: 10.1007/BF01530830.

[283] Alt, H. and L. Scharf (2008). COMPUTING THE HAUSDORFF DISTANCE BETWEEN CURVED OBJECTS. *Int. J. Comput. Geom. Appl.* **18** (04), 307–320. DOI: 10.1142/S0218195908002647.

[284] Fréchet, M. (1906). Sur quelques points du calcul fonctionnel. *Rend. Circ. Mat. Palermo* **22** (1), 1–72.

[285] Alt, H. and M. Godau (1995). COMPUTING THE FRÉCHET DISTANCE BETWEEN TWO POLYGONAL CURVES. *Int. J. Comput. Geom. Appl.* **05** (01n02), 75–91. DOI: 10.1142/S0218195995000064.

[286] Jin, X. and J. Han (2011). "Partitional Clustering". *Encyclopedia of Machine Learning*. Ed. by C. Sammut and G. I. Webb. Springer US, pp. 766–766. DOI: 10.1007/978-0-387-30164-8_631.

[287] Baraty, S., D. A. Simovici, and C. Zara (2011). "The Impact of Triangular Inequality Violations on Medoid-Based Clustering". *Foundations of Intelligent Systems*. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 280–289. DOI: 10.1007/978-3-642-21916-0_31.

[288] Taha, A. A. and A. Hanbury (2015). An efficient algorithm for calculating the exact Hausdorff distance. *IEEE Trans. Pattern Anal. Mach. Intell.* **37** (11), 2153–2163. DOI: 10.1109/TPAMI.2015.2408351.

[289] Jones, E., T. Oliphant, P. Peterson, et al. (2001–). *SciPy: Open source scientific tools for Python*. URL: http://www.scipy.org/.

[290] Driemel, A., S. Har-Peled, and C. Wenk (2012). Approximating the Fréchet Distance for Realistic Curves in Near Linear Time. *Discrete Comput. Geom.* **48** (1), 94–127. DOI: 10.1007/s00454-012-9402-z.

[291] Har-Peled, S. and B. Raichel (2014). The fréchet distance revisited and extended. *ACM Trans. Algorithms* **10** (1), 3. DOI: 10.1145/2532646.

[292] Eiter, T. and H. Mannila (1994). *Computing Discrete Fréchet Distance*. Tech. rep. Wien: Christian Doppler Laboratory for Expert Systems, Technische Universität Wien.

[293] Alt, H., C. Knauer, and C. Wenk (2001). "Bounding the Fréchet distance by the Hausdorff distance". *In Proceedings of the Seventeenth European Workshop on Computational Geometry*, pp. 166–169.

[294] Buchin, K., M. Buchin, and C. Wenk (2008). Computing the Fréchet distance between simple polygons. *Comput. Geom.* **41** (1), 2–20. DOI: 10.1016/j.comgeo.2007.08.003.

[295] Crippen, G. M. (2003). Series approximation of protein structure and constructing conformation space. *Polymer* **44** (15), 4373–4379. DOI: 10.1016/S0032-3861(03)00131-9.

[296] Lindorff-Larsen, K. and J. Ferkinghoff-Borg (2009). Similarity measures for protein ensembles. *PLoS One* **4** (1), e4203. DOI: 10.1371/journal.pone.0004203.

[297] Tiberti, M., E. Papaleo, T. Bengtsen, W. Boomsma, and K. Lindorff-Larsen (2015). ENCORE: Software for Quantitative Ensemble Comparison. *PLoS Comput. Biol.* **11** (10), e1004415. DOI: 10.1371/journal.pcbi.1004415.

[298] Sfriso, P., A. Emperador, L. Orellana, A. Hospital, J. L. Gelpí, and M. Orozco (2012). Finding Conformational Transition Pathways from Discrete Molecular Dynamics Simulations. *J. Chem. Theory Comput.* **8** (11), 4707–4718. DOI: 10.1021/ct300494q.

[299] Das, A., M. Gur, M. H. Cheng, S. Jo, I. Bahar, and B. Roux (2014). Exploring the conformational transitions of biomolecular systems using a simple two-state anisotropic network model. *PLoS Comput. Biol.* **10** (4), e1003521. DOI: 10.1371/journal.pcbi.1003521.

[300] Shakhnovich, E., G. Farztdinov, A. M. Gutin, and M. Karplus (1991). Protein folding bottlenecks: A lattice Monte Carlo simulation. *Phys. Rev. Lett.* **67** (12), 1665–1668. DOI: 10.1103/PhysRevLett.67.1665.

[301] Emperador, A., T. Meyer, and M. Orozco (2010). Protein flexibility from discrete molecular dynamics simulations using quasi-physical potentials. *Proteins* **78** (1), 83–94. DOI: 10.1002/prot.22563.

[302] Husic, B. E. and V. S. Pande (2017). Ward Clustering Improves Cross-Validated Markov State Models of Protein Folding. *J. Chem. Theory Comput.* **13** (3), 963–967. DOI: 10.1021/acs.jctc.6b01238.

[303] Vanden-Eijnden, E. and M. Venturoli (2009b). Revisiting the finite temperature string method for the calculation of reaction tubes and free energies. *J. Chem. Phys.* **130** (19), 194103. DOI: 10.1063/1.3130083.

[304] Bonomi, M., D. Branduardi, G. Bussi, C. Camilloni, D. Provasi, P. Raiteri, D. Donadio, F. Marinelli, F. Pietrucci, R. A. Broglia, and M. Parrinello (2009). PLUMED: A portable plugin for free-energy calculations with molecular dynamics. *Comput. Phys. Commun.* **180** (10), 1961–1972. DOI: 10.1016/j.cpc.2009.05.011.

[305] Tribello, G. A., M. Bonomi, D. Branduardi, C. Camilloni, and G. Bussi (2014). PLUMED 2: New feathers for an old bird. *Comput. Phys. Commun.* **185** (2), 604–613. DOI: 10.1016/j.cpc.2013.09.018.

[306] Pappenheimer Jr, A. M. (1977). Diphtheria toxin. *Annu. Rev. Biochem.* **46**, 69–94. DOI: 10.1146/annurev.bi.46.070177.000441.

[307] Bell, C. E. and D. Eisenberg (1996). Crystal structure of diphtheria toxin bound to nicotinamide adenine dinucleotide. *Biochemistry* **35** (4), 1137–1149. DOI: 10.1021/bi9520848.

[308] Bennett, M. J. and D. Eisenberg (1994). Refined structure of monomeric diphtheria toxin at 2.3 A resolution. *Protein Sci.* **3** (9), 1464–1475. DOI: 10.1002/pro.5560030912.

[309] Bennett, M. J., S. Choe, and D. Eisenberg (1994). Refined structure of dimeric diphtheria toxin at 2.0 A resolution. *Protein Sci.* **3** (9), 1444–1463. DOI: 10.1002/pro.5560030911.

[310] Carroll, S. F., J. T. Barbieri, and R. J. Collier (1986). Dimeric form of diphtheria toxin: purification and characterization. *Biochemistry* **25** (9), 2425–2430. DOI: 10.1021/bi00357a019.

[311] Shaw, D. E., J. P. Grossman, J. A. Bank, B. Batson, J. A. Butts, J. C. Chao, M. M. Deneroff, R. O. Dror, A. Even, C. H. Fenton, A. Forte, J. Gagliardo, G. Gill, B. Greskamp, C. R. Ho, D. J. Ierardi, L. Iserovich, J. S. Kuskin, R. H. Larson, T. Layman, L.-S. Lee, A. K. Lerer, C. Li, D. Killebrew, K. M. Mackenzie, S. Y.-H. Mok, M. A. Moraes, R. Mueller, L. J. Nociolo, J. L. Peticolas, T. Quan, D. Ramot, J. K. Salmon, D. P. Scarpazza, U. Ben Schafer, N. Siddique, C. W. Snyder, J. Spengler, P. T. P. Tang, M. Theobald, H. Toma, B. Towles, B. Vitale, S. C. Wang, and C. Young (2014). "Anton 2: Raising the Bar for Performance and Programmability in a Special-purpose Molecular Dynamics Supercomputer". *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. SC '14. Piscataway, NJ, USA: IEEE Press, pp. 41–53. DOI: 10.1109/SC.2014.9.

[312] Vastermark, A., S. Wollwage, M. E. Houle, R. Rio, and M. H. Saier Jr (2014). Expansion of the APC superfamily of secondary carriers. *Proteins* **82** (10), 2797–2811. DOI: 10.1002/prot.24643.

[313] Romero, M. F., A.-P. Chen, M. D. Parker, and W. F. Boron (2013). The SLC4 family of bicarbonate (HCO3-) transporters. *Mol. Aspects Med.* **34** (2), 159–182.

[314] Alper, S. L. (2009). Molecular physiology and genetics of Na+-independent SLC4 anion exchangers. *J. Exp. Biol.* **212** (Pt 11), 1672–1683. DOI: 10.1242/jeb.029454.

[315] Jennings, M. L., T. R. Howren, J. Cui, M. Winters, and R. Hannigan (2007). Transport and regulatory characteristics of the yeast bicarbonate transporter homolog Bor1p. *Am. J. Physiol. Cell Physiol.* **293** (1), C468–76. DOI: 10.1152/ajpcell.00286.2005.

[316] Nozawa, A., J. Takano, M. Kobayashi, N. von Wirén, and T. Fujiwara (2006). Roles of BOR1, DUR3, and FPS1 in boron transport and tolerance in Saccharomyces cerevisiae. *FEMS Microbiol. Lett.* **262** (2), 216–222. DOI: 10.1111/j.1574-6968.2006.00395.x.

[317] Park, M., Q. Li, N. Shcheynikov, S. Muallem, and W. Zeng (2005). Borate transport and cell growth and proliferation. Not only in plants. *Cell Cycle* **4** (1), 24–26. DOI: 10.4161/cc.4.1.1394.

[318] Xu, F., H. E. Goldbach, P. H. Brown, R. W. Bell, T. Fujiwara, C. D. Hunt, S. Goldberg, and L. Shi, eds. (2007). *Advances in Plant and Animal Boron Nutrition: Proceedings of the 3rd International Symposium on all Aspects of Plant and Animal Boron Nutrition*. Springer Netherlands.

[319] Arakawa, T., T. Kobayashi-Yurugi, Y. Alguel, H. Iwanari, H. Hatae, M. Iwata, Y. Abe, T. Hino, C. Ikeda-Suno, H. Kuma, D. Kang, T. Murata, T. Hamakubo, A. D. Cameron, T. Kobayashi, N. Hamasaki, and S. Iwata (2015). Crystal structure of the anion exchanger domain of human erythrocyte band 3. *Science* **350** (6261), 680–684. DOI: 10.1126/science.aaa4335.

[320] Lu, F., S. Li, Y. Jiang, J. Jiang, H. Fan, G. Lu, D. Deng, S. Dang, X. Zhang, J. Wang, and N. Yan (2011). Structure and mechanism of the uracil transporter UraA. *Nature* **472** (7342), 243–246. DOI: 10.1038/nature09885.

[321] Alguel, Y., S. Amillis, J. Leung, G. Lambrinidis, S. Capaldi, N. J. Scull, G. Craven, S. Iwata, A. Armstrong, E. Mikros, G. Diallinas, A. D. Cameron, and B. Byrne (2016). Structure of eukaryotic purine/H(+) symporter UapA suggests a role for homodimerization in transport activity. *Nat. Commun.* **7**, 11336. DOI: 10.1038/ncomms11336.

[322] Kalli, A. C., M. S. P. Sansom, and R. A. F. Reithmeier (2015). Molecular dynamics simulations of the bacterial UraA H+-uracil symporter in lipid bilayers reveal a closed state and a selective interaction with cardiolipin. *PLoS Comput. Biol.* **11** (3), e1004123. DOI: 10.1371/journal.pcbi.1004123.

[323] Coincon, M., P. Uzdavinys, E. Nji, D. L. Dotson, I. Winkelmann, S. Abdul-Hussein, A. D. Cameron, O. Beckstein, and D. Drew (2016). Crystal structures reveal the molecular basis of ion translocation in sodium/proton antiporters. *Nat. Struct. Mol. Biol.* **23** (3), 248–255. DOI: 10.1038/nsmb.3164.

[324] Zhou, X., E. J. Levin, Y. Pan, J. G. McCoy, R. Sharma, B. Kloss, R. Bruni, M. Quick, and M. Zhou (2014). Structural basis of the alternating-access mechanism in a bile acid transporter. *Nature* **505** (7484), 569–573. DOI: 10.1038/nature12811.

[325] Wöhlert, D., M. J. Grötzinger, W. Kühlbrandt, and Ö. Yildiz (2015). Mechanism of Na(+)-dependent citrate transport from the structure of an asymmetrical CitS dimer. *Elife* **4**, e09375. DOI: 10.7554/eLife.09375.

[326] Reyes, N., C. Ginter, and O. Boudker (2009). Transport mechanism of a bacterial homologue of glutamate transporters. *Nature* **462** (7275), 880–885. DOI: 10.1038/nature08616.

[327] Fernandez-Leiro, R. and S. H. W. Scheres (2016). Unravelling biological macromolecules with cryo-electron microscopy. *Nature* **537** (7620), 339–346. DOI: 10.1038/nature19948.

[328] Milne, J. L. S., M. J. Borgnia, A. Bartesaghi, E. E. H. Tran, L. A. Earl, D. M. Schauder, J. Lengyel, J. Pierson, A. Patwardhan, and S. Subramaniam (2013). Cryo-electron microscopy–a primer for the non-microscopist. *FEBS J.* **280** (1), 28–45. DOI: 10.1111/febs.12078.

[329] Dubochet, J., J.-J. Chang, R. Freeman, J. Lepault, and A. W. McDowall (1982). Frozen aqueous suspensions. *Ultramicroscopy* **10** (1), 55–61.

[330] Reichow, S. L. and T. Gonen (2009). Lipid-protein interactions probed by electron crystallography. *Curr. Opin. Struct. Biol.* **19** (5), 560–565. DOI: 10.1016/j.sbi.2009.07.012.

[331] Abe, K. and Y. Fujiyoshi (2016). Cryo-electron microscopy for structure analyses of membrane proteins in the lipid bilayer. *Curr. Opin. Struct. Biol.* **39**, 71–78. DOI: 10.1016/j.sbi.2016.06.001.

[332] Efremov, R. G., C. Gatsogiannis, and S. Raunser (2017). "Chapter One - Lipid Nanodiscs as a Tool for High-Resolution Structure Determination of Membrane Proteins by Single-Particle Cryo-EM". *Methods in Enzymology*. Ed. by C. Ziegler. Vol. 594. Academic Press, pp. 1–30. DOI: 10.1016/bs.mie.2017.05.007.

[333] Bai, X.-C., G. McMullan, and S. H. W. Scheres (2015). How cryo-EM is revolutionizing structural biology. *Trends Biochem. Sci.* **40** (1), 49–57. DOI: 10.1016/j.tibs.2014.10.005.

[334] Cheng, Y. (2015). Single-Particle Cryo-EM at Crystallographic Resolution. *Cell* **161** (3), 450–457. DOI: 10.1016/j.cell.2015.03.049.

[335] Zhang, X., L. Jin, Q. Fang, W. H. Hui, and Z. H. Zhou (2010). 3.3 A cryo-EM structure of a nonenveloped virus reveals a priming mechanism for cell entry. *Cell* **141** (3), 472–482. DOI: 10.1016/j.cell.2010.03.041.

[336] Trabuco, L. G., E. Villa, K. Mitra, J. Frank, and K. Schulten (2008). Flexible fitting of atomic structures into electron microscopy maps using molecular dynamics. *Structure* **16** (5), 673–683. DOI: 10.1016/j.str.2008.03.005.

[337] Jolley, C. C., S. A. Wells, P. Fromme, and M. F. Thorpe (2008). Fitting low-resolution cryo-EM maps of proteins using constrained geometric simulations. *Biophys. J.* **94** (5), 1613–1621. DOI: 10.1529/biophysj.107.115949.

[338] Topf, M., K. Lasker, B. Webb, H. Wolfson, W. Chiu, and A. Sali (2008). Protein structure fitting and refinement guided by cryo-EM density. *Structure* **16** (2), 295–307. DOI: 10.1016/j.str.2007.11.016.

[339] Zheng, W. (2011). Accurate flexible fitting of high-resolution protein structures into cryo-electron microscopy maps using coarse-grained pseudo-energy minimization. *Biophys. J.* **100** (2), 478–488. DOI: 10.1016/j.bpj.2010.12.3680.

[340] Lopéz-Blanco, J. R. and P. Chacón (2013). iMODFIT: efficient and robust flexible fitting based on vibrational analysis in internal coordinates. *J. Struct. Biol.* **184** (2), 261–270. DOI: 10.1016/j.jsb.2013.08.010.

[341] Kirmizialtin, S., J. Loerke, E. Behrmann, C. M. T. Spahn, and K. Y. Sanbonmatsu (2015). "Using Molecular Simulation to Model High-Resolution Cryo-EM Reconstructions". *Methods in Enzymology Structures of Large RNA Molecules and Their Complexes*. Ed. by S. A. Woodson and F. H. T. Allain. Vol. 558. Methods in Enzymology. Academic Press, pp. 497–514.

[342] Noel, J. K., M. Levi, M. Raghunathan, H. Lammert, R. L. Hayes, J. N. Onuchic, and P. C. Whitford (2016). SMOG 2: A Versatile Software Package for Generating Structure-Based Models. *PLoS Comput. Biol.* **12** (3), e1004794. DOI: 10.1371/journal.pcbi.1004794.

[343] Singharoy, A., I. Teo, R. McGreevy, J. E. Stone, J. Zhao, and K. Schulten (2016). Molecular dynamics-based refinement and validation for sub-5 Å cryo-electron microscopy maps. *Elife* **5**, DOI: 10.7554/eLife.16105.

[344] Ahmed, A., P. C. Whitford, K. Y. Sanbonmatsu, and F. Tama (2012). Consensus among flexible fitting approaches improves the interpretation of cryo-EM data. *J. Struct. Biol.* **177** (2), 561–570. DOI: 10.1016/j.jsb.2011.10.002.

[345] Ahmed, A. and F. Tama (2013). Consensus among multiple approaches as a reliability measure for flexible fitting into cryo-EM data. *J. Struct. Biol.* **182** (2), 67–77. DOI: 10.1016/j.jsb.2013.02.002.

[346] Webb, B. and A. Sali (2014). Protein structure modeling with MODELLER. *Methods Mol. Biol.* **1137**, 1–15. DOI: 10.1007/978-1-4939-0366-5_1.

[347] Alva, V., S.-Z. Nam, J. Söding, and A. N. Lupas (2016). The MPI bioinformatics Toolkit as an integrative platform for advanced protein sequence and structure analysis. *Nucleic Acids Res.* **44** (W1), W410–5. DOI: 10.1093/nar/gkw348.

[348] Rohou, A. and N. Grigorieff (2014). Frealix: model-based refinement of helical filament structures from electron micrographs. *J. Struct. Biol.* **186** (2), 234–244. DOI: 10.1016/j.jsb.2014.03.012.

[349] Kucukelbir, A., F. J. Sigworth, and H. D. Tagare (2014). Quantifying the local resolution of cryo-EM density maps. *Nat. Methods* **11** (1), 63–65. DOI: 10.1038/nmeth.2727.

[350] Wriggers, W., R. A. Milligan, and J. A. McCammon (1999). Situs: A package for docking crystal structures into low-resolution maps from electron microscopy. *J. Struct. Biol.* **125** (2-3), 185–195. DOI: 10.1006/jsbi.1998.4080.

[351] Humphrey, W., A. Dalke, and K. Schulten (1996). VMD: visual molecular dynamics. *J. Mol. Graph.* **14** (1), 33–8, 27–8.

[352] Phillips, J. C., R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kalé, and K. Schulten (2005). Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **26** (16), 1781–1802. DOI: 10.1002/jcc.20289.

[353] Jo, S., J. B. Lim, J. B. Klauda, and W. Im (2009). CHARMM-GUI Membrane Builder for mixed bilayers and its application to yeast membranes. *Biophys. J.* **97** (1), 50–58. DOI: 10.1016/j.bpj.2009.04.013.

[354] Lee, J., X. Cheng, J. M. Swails, M. S. Yeom, P. K. Eastman, J. A. Lemkul, S. Wei, J. Buckner, J. C. Jeong, Y. Qi, S. Jo, V. S. Pande, D. A. Case, C. L. Brooks 3rd, A. D. MacKerell Jr, J. B. Klauda, and W. Im (2016). CHARMM-GUI Input Generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM Simulations Using the CHARMM36 Additive Force Field. *J. Chem. Theory Comput.* **12** (1), 405–413. DOI: 10.1021/acs.jctc.5b00935.

[355] Abraham, M. J., T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, and E. Lindahl (2015). GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1–2**, 19–25. DOI: 10.1016/j.softx.2015.06.001.

[356] Klauda, J. B., R. M. Venable, J. A. Freites, J. W. O'Connor, D. J. Tobias, C. Mondragon-Ramirez, I. Vorobyov, A. D. MacKerell Jr, and R. W. Pastor (2010). Update of the CHARMM all-atom additive force field for lipids: validation on six lipid types. *J. Phys. Chem. B* **114** (23), 7830–7843. DOI: 10.1021/jp101759q.

[357] Bussi, G., D. Donadio, and M. Parrinello (2007). Canonical sampling through velocity rescaling. *J. Chem. Phys.* **126** (1), 014101. DOI: 10.1063/1.2408420.

[358] Parrinello, M. and A. Rahman (1981). Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.* **52** (12), 7182–7190. DOI: 10.1063/1.328693.

[359] Essmann, U., L. Perera, M. L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen (1995). A smooth particle mesh Ewald method. *J. Chem. Phys.* **103** (19), 8577–8593. DOI: 10.1063/1.470117.

[360] Kabsch, W. and C. Sander (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22** (12), 2577–2637. DOI: 10.1002/bip.360221211.

[361] Smart, O. S., J. M. Goodfellow, and B. A. Wallace (1993). The pore dimensions of gramicidin A. *Biophys. J.* **65** (6), 2455–2460. DOI: 10.1016/S0006-3495(93)81293-1.

[362] Thurtle-Schmidt, B. H. and R. M. Stroud (2016). Structure of Bor1 supports an elevator transport mechanism for SLC4 anion exchangers. *Proc. Natl. Acad. Sci. U. S. A.* **113** (38), 10542–10546. DOI: 10.1073/pnas.1612603113.

[363] Mohamed, K. M. and A. A. Mohamad (2010). A review of the development of hybrid atomistic–continuum methods for dense fluids. *Microfluid. Nanofluidics* **8** (3), 283–302. DOI: 10.1007/s10404-009-0529-z.

[364] O'Connell, S. T. and P. A. Thompson (1995). Molecular dynamics-continuum hybrid computations: A tool for studying complex fluid flows. *Phys. Rev. E Stat. Phys. Plasmas Fluids Relat. Interdiscip. Topics* **52** (6), R5792–R5795.

[365] Wijesinghe, H. S. and N. G. Hadjiconstantinou (2004). Discussion of Hybrid Atomistic-Continuum Methods for Multiscale Hydrodynamics. *International Journal for Multiscale Computational Engineering* **2** (2), 189–202. DOI: 10.1615/IntJMultCompEng.v2.i2.20.

[366] Landau, L. D. and E. M. Lifschitz (1966). *Fluid Mechanics*. third. Vol. 6.

[367] Plimpton, S. (1995). Fast Parallel Algorithms for Short-Range Molecular Dynamics. *J. Comput. Phys.* **117** (1), 1–19. DOI: 10.1006/jcph.1995.1039.

[368] Seyler, C. E. and M. R. Martin (2011). Relaxation model for extended magnetohydrodynamics: Comparison to magnetohydrodynamics for dense Z-pinches. *Phys. Plasmas* **18** (1), 012703. DOI: 10.1063/1.3543799.

[369] Chen, J. and C. L. Brooks 3rd (2008). Implicit modeling of nonpolar solvation for simulating protein folding and conformational transitions. *Phys. Chem. Chem. Phys.* **10** (4), 471–481. DOI: 10.1039/b714141f.

[370] Vega, C. and J. L. F. Abascal (2011). Simulating water with rigid non-polarizable models: a general perspective. *Phys. Chem. Chem. Phys.* **13** (44), 19663–19688. DOI: 10.1039/c1cp22168j.

[371] Huggins, D. J. (2012). Correlations in liquid water for the TIP3P-Ewald, TIP4P-2005, TIP5P-Ewald, and SWM4-NDP models. *J. Chem. Phys.* **136** (6), 064518. DOI: 10.1063/1.3683447.

[372] Heckmann, L. and B. Drossel (2013). Common features of simple water models. *J. Chem. Phys.* **138** (23), 234503. DOI: 10.1063/1.4810875.

[373] Fogolari, F., A. Brigo, and H. Molinari (2002). The Poisson-Boltzmann equation for biomolecular electrostatics: a tool for structural biology. *J. Mol. Recognit.* **15** (6), 377–392. DOI: 10.1002/jmr.577.

[374] Feig, M., A. Onufriev, M. S. Lee, W. Im, D. A. Case, and C. L. Brooks 3rd (2004). Performance comparison of generalized born and Poisson methods in the calculation of electrostatic solvation energies for protein structures. *J. Comput. Chem.* **25** (2), 265–284. DOI: 10.1002/jcc.10378.

[375] Lee, M. S., F. R. Salsbury Jr, and M. A. Olson (2004). An efficient hybrid explicit/implicit solvent method for biomolecular simulations. *J. Comput. Chem.* **25** (16), 1967–1978. DOI: 10.1002/jcc.20119.

[376] Fennell, C. J. and K. A. Dill (2011). Physical Modeling of Aqueous Solvation. *J. Stat. Phys.* **145** (2), 209–226. DOI: 10.1007/s10955-011-0232-9.

[377] Kleinjung, J. and F. Fraternali (2014). Design and application of implicit solvent models in biomolecular simulations. *Curr. Opin. Struct. Biol.* **25**, 126–134. DOI: 10.1016/j.sbi.2014.04.003.

[378] Anandakrishnan, R., A. Drozdetski, R. C. Walker, and A. V. Onufriev (2015). Speed of conformational change: comparing explicit and implicit solvent molecular dynamics simulations. *Biophys. J.* **108** (5), 1153–1164. DOI: 10.1016/j.bpj.2014.12.047.

[379] Bussi, G. and M. Parrinello (2007). Accurate sampling using Langevin dynamics. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **75** (5 Pt 2), 056707. DOI: 10.1103/PhysRevE.75.056707.

[380] Mor, A., G. Ziv, and Y. Levy (2008). Simulations of proteins with inhomogeneous degrees of freedom: The effect of thermostats. *J. Comput. Chem.* **29** (12), 1992–1998. DOI: 10.1002/jcc.20951.

[381] Widmalm, G. and R. W. Pastor (1992). Comparison of Langevin and molecular dynamics simulations. Equilibrium and dynamics of ethylene glycol in water. *J. Chem. Soc. Faraday Trans.* **88** (13), 1747–1754. DOI: 10.1039/FT9928801747.

[382] Basconi, J. E. and M. R. Shirts (2013). Effects of Temperature Control Algorithms on Transport Properties and Kinetics in Molecular Dynamics Simulations. *J. Chem. Theory Comput.* **9** (7), 2887–2899. DOI: 10.1021/ct400109a.

[383] David, L., R. Luo, and M. K. Gilson (2000). Comparison of generalized born and poisson models: Energetics and dynamics of HIV protease. *J. Comput. Chem.* **21** (4), 295–309. DOI: 10.1002/(SICI)1096-987X(200003)21:4<295::AID-JCC5>3.0.CO;2-8.

[384] Stultz, C. M. (2004). An Assessment of Potential of Mean Force Calculations with Implicit Solvent Models. *J. Phys. Chem. B* **108** (42), 16525–16532. DOI: 10.1021/jp047126t.

[385] Geney, R., M. Layten, R. Gomperts, V. Hornak, and C. Simmerling (2006). Investigation of Salt Bridge Stability in a Generalized Born Solvent Model. *J. Chem. Theory Comput.* **2** (1), 115–127. DOI: 10.1021/ct050183l.

[386] Roe, D. R., A. Okur, L. Wickstrom, V. Hornak, and C. Simmerling (2007). Secondary structure bias in generalized Born solvent models: comparison of conformational ensembles and free energy of solvent polarization from explicit and implicit solvation. *J. Phys. Chem. B* **111** (7), 1846–1857. DOI: 10.1021/jp066831u.

[387] Paul, G. L. and P. N. Pusey (1981). Observation of a long-time tail in Brownian motion. *J. Phys. A Math. Gen.* **14** (12), 3301–3327. DOI: 10.1088/0305-4470/14/12/025.

[388] Alder, B. J. and W. E. Alley (1984). Generalized hydrodynamics. *Phys. Today* **37** (1), 56–63. DOI: 10.1063/1.2916048.

[389] Kheifets, S., A. Simha, K. Melin, T. Li, and M. G. Raizen (2014). Observation of Brownian motion in liquids at short times: instantaneous velocity and memory loss. *Science* **343** (6178), 1493–1496.

[390] Hinsen, K. and G. R. Kneller (2008). Solvent effects in the slow dynamics of proteins. *Proteins* **70** (4), 1235–1242. DOI: 10.1002/prot.21655.

[391] Boussinesq, J. (1885). Sur la résistance qu'oppose un fluide indéfini au repos, sans pesanteur, au mouvement varié d'une sphère solide qu'il mouille sur toute sa surface, quand les vitesses restent bien continues et assez faibles pour que leurs carrés et produits soient néglige. *C. R. Hebd. Seances Acad. Sci.* **100** (935).

[392] Basset, A. B. (1888). III. On the motion of a sphere in a viscous liquid. *Philos. Trans. R. Soc. Lond. A* **179**, 43–63. DOI: 10.1098/rsta.1888.0003.

[393] Michaelides, E. E. (2003). Hydrodynamic Force and Heat/Mass Transfer From Particles, Bubbles, and Drops—The Freeman Scholar Lecture. *J. Fluids Eng.* **125** (2), 209–238. DOI: 10.1115/1.1537258.

[394] Candelier, F., J. R. Angilella, and M. Souhar (2004). On the effect of the Boussinesq–Basset force on the radial migration of a Stokes particle in a vortex. *Phys. Fluids* **16** (5), 1765–1776. DOI: 10.1063/1.1689970.

[395] Beard, D. A. and T. Schlick (2000). Inertial stochastic dynamics. II. Influence of inertia on slow kinetic processes of supercoiled DNA. *J. Chem. Phys.* **112** (17), 7323–7338. DOI: 10.1063/1.481371.

[396] Delgado-Buscalioni, R. and G. De Fabritiis (2007). Embedding molecular dynamics within fluctuating hydrodynamics in multiscale simulations of liquids. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **76** (3 Pt 2), 036709. DOI: 10.1103/PhysRevE.76.036709.

[397] Ritos, K., M. K. Borg, D. A. Lockerby, D. R. Emerson, and J. M. Reese (2015). Hybrid molecular-continuum simulations of water flow through carbon nanotube membranes of realistic thickness. *Microfluid. Nanofluidics* **19** (5), 997–1010. DOI: 10.1007/s10404-015-1617-x.

[398] Van Huyen, V., B. Trouette, Q. D. To, and E. Chénier (2016). Multi-scale modelling and hybrid atomistic-continuum simulation of non-isothermal flows in microchannels. *Microfluid. Nanofluidics* **20** (2), 43. DOI: 10.1007/s10404-016-1709-2.

[399] Cao, Q., C. Zuo, and L. Li (2012). Hybrid Particle–Continuum Simulations of Polymer Brushes in Shear Flow. *J. Macromol. Sci. Pt. B: Phys.* **51** (4), 707–719. DOI: 10.1080/00222348.2011.609793.

[400] De Nicola, A., T. Kawakatsu, and G. Milano (2014). Generation of Well-Relaxed All-Atom Models of Large Molecular Weight Polymer Melts: A Hybrid Particle-Continuum Approach Based on Particle-Field Molecular Dynamics Simulations. *J. Chem. Theory Comput.* **10** (12), 5651–5667. DOI: 10.1021/ct500492h.

[401] Giupponi, G., G. De Fabritiis, and P. V. Coveney (2007). Hybrid method coupling fluctuating hydrodynamics and molecular dynamics for the simulation of macromolecules. *J. Chem. Phys.* **126** (15), 154903. DOI: 10.1063/1.2720385.

[402] Donev, A., J. Bell, A. Garcia, and B. Alder (2010). A Hybrid Particle-Continuum Method for Hydrodynamics of Complex Fluids. *Multiscale Model. Simul.* **8** (3), 871–911. DOI: 10.1137/090774501.

[403] Shang, B. Z. (2013). "Multiscale Simulations: From Enzyme Kinetics to Fluctuating Hydrodynamics". PhD thesis. University of California, Berkeley.

[404] Müser, M. H., G. Sutmann, and R. G. Winkler, eds. (2013). *Hybrid Particle-Continuum Methods in Computational Materials Physics*. Vol. 46. NIC Series. Forschungszentrum Jülich GmbH Zentralbibliothek.

[405] Bell, J. B., A. L. Garcia, and S. A. Williams (2007). Numerical methods for the stochastic Landau-Lifshitz Navier-Stokes equations. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **76** (1 Pt 2), 016708. DOI: 10.1103/PhysRevE.76.016708.

[406] — (2010). Computational fluctuating fluid dynamics. *Esaim Math. Model. Numer. Anal.* **44** (5), 1085–1105. DOI: 10.1051/m2an/2010053.

[407] Fox, R. F. and G. E. Uhlenbeck (1970). Contributions to Nonequilibrium Thermodynamics. II. Fluctuation Theory for the Boltzmann Equation. *The Physics of Fluids* **13** (12), 2881–2890. DOI: 10.1063/1.1692878.

[408] Inoue, A. and T. Funaki (1979). On a new derivation of the Navier-Stokes equation. *Commun.Math. Phys.* **65** (1), 83–90. DOI: 10.1007/BF01940961.

[409] Müller, I. and T. Ruggeri (1998). *Rational Extended Thermodynamics*. Ed. by C. Truesdell. second. Vol. 37. Spring Tracts in Natural Philosophy. New York: Springer-Verlag.

[410] Jou, D., J. Casas-Vázquez, and G. Lebon (2010). *Extended Irreversible Thermodynamics*. Fourth. New York: Springer, pp. 41–74. DOI: 10.1007/978-90-481-3074-0.

[411] Karniadakis, G., A. Beskok, and N. Aluru (2005). *Microflows and Nanoflows: Fundamentals and Simulation*. Ed. by S. S. Antman, J. E. Marsden, and L. Sirovich. Vol. 29. Interdisciplinary Applied Mathematics. New York: Springer.

[412] Hadjiconstantinou, N. G. (2006). The limits of Navier-Stokes theory and kinetic extensions for describing small-scale gaseous hydrodynamics. *Phys. Fluids* **18** (11), 111301. DOI: 10.1063/1.2393436.

[413] Torrilhon, M. (2011). *Regularization of Grad's 13-moment-equations in kinetic gas theory*. Tech. rep. RTO-EN-AVT-194-10. RWTH AACHEN UNIV (GERMANY). DOI: 10.14339/RTO-EN-AVT-194.

[414] Howard, J. (2002). "Kinetic Theory". *Introduction to Plasma Physics C17 Lecture Notes*. Ed. by J. Howard. Chap. 2, pp. 29–64.

[415] Levermore, D. C. and N. Masmoudi (2010). From the Boltzmann Equation to an Incompressible Navier–Stokes–Fourier System. *Arch. Ration. Mech. Anal.* **196** (3), 753–809. DOI: 10.1007/s00205-009-0254-5.

[416] Tong, D. (2012). "Kinetic Theory".

[417] Grad, H. (1949). On the kinetic theory of rarefied gases. *Communications on pure and applied mathematics* **2** (4), 331–407. DOI: 10.1002/cpa.3160020403.

[418] Reinecke, S. and G. M. Kremer (1990). Method of moments of Grad. *Phys. Rev. A* **42** (2), 815–820.

[419] Torrilhon, M. (2000). Characteristic waves and dissipation in the 13-moment-case. *Continuum Mech. Thermodyn.* **12** (5), 289–301. DOI: 10.1007/s001610050138.

[420] Brini, F. (2001). Hyperbolicity region in extended thermodynamics with 14 moments. *Continuum Mech. Thermodyn.* **13** (1), 1–8. DOI: 10.1007/s001610100036.

[421] Torrilhon, M. (2009). Hyperbolic moment equations in kinetic gas theory based on multi-variate Pearson-IV-distributions. *CiCP*, DOI: 10.4208/cicp.2009.09.049.

[422] Karlin, I. V. and A. N. Gorban (2002). Hydrodynamics from Grad's equations: What can we learn from exact solutions? *Ann. Phys.* **11** (10-11), 783–833. DOI: 10.1002/1521-3889(200211)11:10/11<783::AID-ANDP783>3.0.CO;2-V.

[423] Pekker, L., O. Pekker, and V. Timchenko (2010). A Complete Set of Grad's Thirteen Regularized Moment Equations.

[424] Kremer, G. M. (2011). *The Methods of Chapman-Enskog and Grad and Applications*. Tech. rep. RTO-EN-AVT-194. Universidade Federal do Paraná.

[425] Szabó, J. and I. Abonyi (1965). Generalized Ohm's Law in a Magnetic Plasma. *Beitr. Plasmaphys.* **5** (1-2), 9–12. DOI: 10.1002/ctpp.19650050103.

[426] "The Generalized Ohm's Law in Plasma" (2007). *Plasma Astrophysics*. Vol. 341. Astrophysics and Space Science Library. New York, NY: Springer New York, pp. 193–204. DOI: 10.1007/978-0-387-68894-7_12.

[427] Gourdain, P.-A. (2017). The impact of the Hall term on tokamak plasmas.

[428] Zhao, X., Y. Yang, and C. E. Seyler (2014). A positivity-preserving semi-implicit discontinuous Galerkin scheme for solving extended magnetohydrodynamics equations. *J. Comput. Phys.* **278** (0), 400–415. DOI: 10.1016/j.jcp.2014.08.044.

[429] Zhao, X. and C. E. Seyler (2015). Computational extended magneto-hydrodynamical study of shock structure generated by flows past an obstacle. *Phys. Plasmas* **22** (7), 072102. DOI: 10.1063/1.4923426.

[430] Batchelor, G. K. (2000). *An Introduction to Fluid Dynamics*. Cambridge University Press.

[431] Jin, S., L. Pareschi, and M. Slemrod (2002). A Relaxation Scheme for Solving the Boltzmann Equation Based on the Chapman-Enskog Expansion. *Acta Math. Appl. Sin.* **18** (1), 37–62. DOI: 10.1007/s102550200003.

[432] Bhatnagar, P. L., E. P. Gross, and M. Krook (1954). A Model for Collision Processes in Gases. I. Small Amplitude Processes in Charged and Neutral One-Component Systems. *Phys. Rev.* **94** (3), 511–525. DOI: 10.1103/PhysRev.94.511.

[433] Nassios, J. (2008). Kinetic Theory an the BGK equation: Gas Dynamics for the Nanoscale. *University of Melbourne Mathematiics Department*.

[434] Struchtrup, H. and M. Torrilhon (2003). Regularization of Grad's 13 moment equations: Derivation and linear analysis. *Phys. Fluids* **15** (9), 2668–2680. DOI: 10.1063/1.1597472.

[435] Torrilhon, M. (2015). Convergence Study of Moment Approximations for Boundary Value Problems of the Boltzmann-BGK Equation. *Commun. Comput. Phys.* **18** (3), 529–557. DOI: 10.4208/cicp.061013.160215a.

[436] Sod, G. A. (1978). A survey of several finite difference methods for systems of nonlinear hyperbolic conservation laws. *J. Comput. Phys.* **27** (1), 1–31. DOI: 10.1016/0021-9991(78)90023-2.

[437] Shu, C.-W. (1998). *Essentially non-oscillatory and weighted essentially non-oscillatory schemes for hyperbolic conservation laws*. Tech. rep. Brown University.

[438] Zhou, Y. C. and G. W. Wei (2001). High resolution conjugate filters for the simulation of flows.

[439] Ahrens, J., B. Geveci, and C. Law (2005). *ParaView: An End-User Tool for Large-Data Visualization*. Ed. by C. D. Hansen and C. R. Johnson. Burlington: Elsevier, pp. 717–731. DOI: 10.1016/B978-012387582-2/50038-1.

[440] Turk, M. J., B. D. Smith, J. S. Oishi, S. Skory, S. W. Skillman, T. Abel, and M. L. Norman (2010). yt: A MULTI-CODE ANALYSIS TOOLKIT FOR ASTROPHYSICAL SIMULATION DATA. *ApJS* **192** (1), 9. DOI: 10.1088/0067-0049/192/1/9.

[441] Kermode, J. (2011). *f90wrap*. URL: https://github.com/charlespwd/project-title.

[442] Pletzer, A., D. McCune, S. Maszala, S. Vadlamani, and S. Kruger (2008). Exposing Fortran Derived Types to C and Other Languages. *Comput. Sci. Eng.* **10** (4), 86–92. DOI: 10.1109/MCSE.2008.94.

[443] Peterson, P. (2009). F2PY: a tool for connecting Fortran and Python programs. *Int. J. Comput. Sci. Eng.* **4** (4), 296–305. DOI: 10.1504/IJCSE.2009.029165.

[444] Wang, Y., J. K. Sigurdsson, E. Brandt, and P. J. Atzberger (2013b). Dynamic implicit-solvent coarse-grained models of lipid bilayer membranes: fluctuating hydrodynamics thermostat. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **88** (2), 023301. DOI: 10.1103/PhysRevE.88.023301.

[445] Mackay, F. E., S. T. T. Ollila, and C. Denniston (2013). Hydrodynamic forces implemented into LAMMPS through a lattice-Boltzmann fluid. *Comput. Phys. Commun.* **184** (8), 2021–2031. DOI: 10.1016/j.cpc.2013.03.024.

[446] Flekkøy, E. G., R. Delgado-Buscalioni, and P. V. Coveney (2005). Flux boundary conditions in particle simulations. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **72** (2 Pt 2), 026703. DOI: 10.1103/PhysRevE.72.026703.

[447] Delgado-Buscalioni, R. and P. V. Coveney (2003). USHER: An algorithm for particle insertion in dense fluids. *J. Chem. Phys.* **119** (2), 978–987. DOI: 10.1063/1.1579475.

[448] De Fabritiis, G., R. Delgado-Buscalioni, and P. V. Coveney (2004). Energy controlled insertion of polar molecules in dense fluids. *J. Chem. Phys.* **121** (24), 12139–12142. DOI: 10.1063/1.1835957.

[449] Riniker, S., J. R. Allison, and W. F. van Gunsteren (2012). On developing coarse-grained models for biomolecular simulation: a review. *Phys. Chem. Chem. Phys.* **14** (36), 12423–12430. DOI: 10.1039/c2cp40934h.

[450] Wassenaar, T. A., K. Pluhackova, R. A. Böckmann, S. J. Marrink, and D. P. Tieleman (2014). Going Backward: A Flexible Geometric Approach to Reverse Transformation

from Coarse Grained to Atomistic Models. *J. Chem. Theory Comput.* **10** (2), 676–690. DOI: 10.1021/ct400617g.

[451] Davtyan, A., J. F. Dama, G. A. Voth, and H. C. Andersen (2015). Dynamic force matching: A method for constructing dynamical coarse-grained models with realistic time dependence. *J. Chem. Phys.* **142** (15), 154104. DOI: 10.1063/1.4917454.

[452] Na, H., R. L. Jernigan, and G. Song (2015). Bridging between NMA and Elastic Network Models: Preserving All-Atom Accuracy in Coarse-Grained Models. *PLoS Comput. Biol.* **11** (10), e1004542. DOI: 10.1371/journal.pcbi.1004542.

[453] Golse, F. (2010). From the Kinetic Theory of Gases to Continuum Mechanics.

[454] Klimontovich, Y. L. (1995). "Langevin Method in Kinetic Theory of Fluctuations". *Statistical Theory of Open Systems*. Fundamental Theories of Physics. Springer, Dordrecht, pp. 159–171. DOI: 10.1007/978-94-011-0175-2_10.

[455] Belyi, V. V. (2005). Klimontovich–Langevin approach to the fluctuation-dissipation theorem for a nonlocal plasma. *J. Phys. Conf. Ser.* **11** (1), 61. DOI: 10.1088/1742-6596/11/1/006.

[456] Jou, D. and J. Casas-Vázquez (2001). Extended irreversible thermodynamics and its relation with other continuum approaches. *J. Non-Newtonian Fluid Mech.* **96** (1), 77–104. DOI: 10.1016/S0377-0257(00)00138-5.

[457] Pokrovski, V. N. (2005). Extended thermodynamics in a discrete-system approach. *Eur. J. Phys.* **26** (5), 769.

[458] Cimmelli, V., D. Jou, T. Ruggeri, and P. Ván (2014). Entropy Principle and Recent Results in Non-Equilibrium Theories. *Entropy* **16** (3), 1756–1807. DOI: 10.3390/e16031756.

[459] Carrisi, M. C., R. E. Tchame, M. Obounou, and S. Pennisi (2015a). Extended Thermodynamics for Dense Gases up to Whatever Order and with Only Some Symmetries. *Entropy* **17** (10), 7052–7075. DOI: 10.3390/e17107052.

[460] Ruggeri, T. and M. Sugiyama (2014). Recent Developments in Extended Thermodynamics of Dense and Rarefied Polyatomic Gases. *Acta Appl. Math.* **132** (1), 527–548. DOI: 10.1007/s10440-014-9923-y.

[461] Carrisi, M. C., S. Pennisi, T. Ruggeri, and M. Sugiyama (2015b). Extended thermodynamics of dense gases in the presence of dynamic pressure. *Ric. Mat.* **64** (2), 403–419. DOI: 10.1007/s11587-015-0247-7.

[462] Arima, T., S. Taniguchi, T. Ruggeri, and M. Sugiyama (2011). Extended thermodynamics of dense gases. *Continuum Mech. Thermodyn.* **24** (4-6), 271–292. DOI: 10.1007/s00161-011-0213-x.

[463] Morozov, A. and S. E. Spagnolie (2014). "Introduction to Complex Fluids". *Complex Fluids in Biological Systems: Experiment, Theory, and Computation*. Ed. by S. E. Spagnolie. Biological and Medical Physics, Biomedical Engineering. New York: Springer. Chap. 1, pp. 3–52. DOI: 10.1007/978-1-4939-2065-5_1.

[464] Roylance, D. (2001). *ENGINEERING VISCOELASTICITY*. Tech. rep. Massachusetts Institute of Technology.

[465] Riedel, C., R. Gabizon, C. A. M. Wilson, K. Hamadani, K. Tsekouras, S. Marqusee, S. Pressé, and C. Bustamante (2015). The heat released during catalytic turnover enhances the diffusion of an enzyme. *Nature* **517** (7533), 227–230. DOI: 10.1038/nature14043.

[466] Halle, B. and M. Davidovic (2003). Biomolecular hydration: from water dynamics to hydrodynamics. *Proc. Natl. Acad. Sci. U. S. A.* **100** (21), 12135–12140. DOI: 10.1073/pnas.2033320100.

[467] Rai, N., M. Nöllmann, B. Spotorno, G. Tassara, O. Byron, and M. Rocco (2005). SOMO (SOlution MOdeler) differences between X-Ray- and NMR-derived bead models suggest a role for side chain flexibility in protein hydrodynamics. *Structure* **13** (5), 723–734. DOI: 10.1016/j.str.2005.02.012.

[468] Lau, E. Y. and V. V. Krishnan (2007). Temperature dependence of protein-hydration hydrodynamics by molecular dynamics simulations. *Biophys. Chem.* **130** (1-2), 55–64. DOI: 10.1016/j.bpc.2007.07.004.

[469] Frembgen-Kesner, T. and A. H. Elcock (2009). Striking Effects of Hydrodynamic Interactions on the Simulated Diffusion and Folding of Proteins. *J. Chem. Theory Comput.* **5** (2), 242–256. DOI: 10.1021/ct800499p.

[470] Jönsson, P. and B. Jönsson (2015). Hydrodynamic Forces on Macromolecules Protruding from Lipid Bilayers Due to External Liquid Flows. *Langmuir* **31** (46), 12708–12718. DOI: 10.1021/acs.langmuir.5b03421.

[471] Mikhailov, A. S. and R. Kapral (2015). Hydrodynamic collective effects of active protein machines in solution and lipid bilayers. *Proc. Natl. Acad. Sci. U. S. A.* **112** (28), E3639–44. DOI: 10.1073/pnas.1506825112.

[472] Kapral, R. and A. S. Mikhailov (2016). Stirring a fluid at low Reynolds numbers: Hydrodynamic collective effects of active proteins in biological cells. *Physica D* **318-319** (Supplement C), 100–104. DOI: 10.1016/j.physd.2015.10.024.

[473] Einstein, A. (1956). *Investigations on the Theory of the Brownian Movement*. Ed. by D. Fürth and A. D. Cowper. Dover Books on Physics Series. Dover Publications.

[474] Peskir, G. (2003). On the Diffusion Coefficient: The Einstein Relation and Beyond. *Stoch. Models* **19** (3), 383–405. DOI: 10.1081/STM-120023566.

[475] Cerutti, D. S., R. Duke, P. L. Freddolino, H. Fan, and T. P. Lybrand (2008). Vulnerability in Popular Molecular Dynamics Packages Concerning Langevin and Andersen Dynamics. *J. Chem. Theory Comput.* **4** (10), 1669–1680. DOI: 10.1021/ct8002173.

[476] Loncharich, R. J., B. R. Brooks, and R. W. Pastor (1992). Langevin dynamics of peptides: the frictional dependence of isomerization rates of N-acetylalanyl-N'-methylamide. *Biopolymers* **32** (5), 523–535. DOI: 10.1002/bip.360320508.

[477] Pastort, R. W. and M. Karplus (1988). Parametrizatlon of the Friction Constant for Stochastic Simulations of Polymers. *J. Phys. Chem.* **92** (9), 2636–2641. DOI: 10.1021/j100320a047.

[478] Romo, T. D. and A. Grossfield (2011). Validating and improving elastic network models with molecular dynamics simulations. *Proteins* **79** (1), 23–34. DOI: 10.1002/prot.22855.

[479] Leioatts, N., T. D. Romo, and A. Grossfield (2012). Elastic Network Models are Robust to Variations in Formalism. *J. Chem. Theory Comput.* **8** (7), 2424–2434. DOI: 10.1021/ct3000316.

[480] Ferrara, P., J. Apostolakis, and A. Caflisch (2000b). Computer simulations of protein folding by targeted molecular dynamics. *Proteins* **39** (3), 252–260. DOI: 10.1002/(SICI)1097-0134(20000515)39:3<252::AID-PROT80>3.0.CO;2-3.

[481] Paci, E. and M. Karplus (1999). Forced unfolding of fibronectin type 3 modules: an analysis by biased molecular dynamics simulations. *J. Mol. Biol.* **288** (3), 441–459. DOI: 10.1006/jmbi.1999.2670.

[482] Theobald, D. L. (2005). Rapid calculation of RMSDs using a quaternion-based characteristic polynomial. *Acta Crystallogr. A* **61** (Pt 4), 478–480. DOI: 10.1107/S0108767305015266.

[483] Liu, P., D. K. Agrafiotis, and D. L. Theobald (2010). Fast determination of the optimal rotational matrix for macromolecular superpositions. *J. Comput. Chem.* **31** (7), 1561–1563. DOI: 10.1002/jcc.21439.

[484] Xu, R. and D. Wunsch (2008). *Clustering*. IEEE Press Series on Computational Intelligence. John Wiley & Sons.

[485] *Statistics and Machine Learning Toolbox: User's Guide* (2017). Natick, Massachusetts: The MathWorks, Inc.

[486] Dickson, B. M., H. Huang, and C. B. Post (2012). Unrestrained computation of free energy along a path. *J. Phys. Chem. B* **116** (36), 11046–11055. DOI: 10.1021/jp304720m.

[487] Buchin, M. (2007). "On the Computability of the Fréchet Distance Between Triangulated Surfaces". PhD thesis. Institut für Informatik Freie Universität Berlin.

[488] Moseler, M. and U. Landman (2000). Formation, stability, and breakup of nanojets. *Science* **289** (5482), 1165–1170. DOI: 10.1126/science.289.5482.1165.

[489] Eggers, J. (2002). Dynamics of liquid nanojets. *Phys. Rev. Lett.* **89** (8), 084502. DOI: 10.1103/PhysRevLett.89.084502.

[490] Astumian, R. D. and P. Hänggi (2002). Brownian Motors. *Phys. Today* **55** (11), 33–39. DOI: 10.1063/1.1535005.

[491] Brangwynne, C. P., G. H. Koenderink, F. C. MacKintosh, and D. A. Weitz (2008). Cytoplasmic diffusion: molecular motors mix it up. *J. Cell Biol.* **183** (4), 583–587. DOI: 10.1083/jcb.200806149.

[492] Woodhouse, F. G. and R. E. Goldstein (2012). Spontaneous circulation of confined active suspensions. *Phys. Rev. Lett.* **109** (16), 168105. DOI: 10.1103/PhysRevLett.109.168105.

[493] Wu, M., G. Ahlers, and D. S. Cannell (1995). Thermally Induced Fluctuations below the Onset of Rayleigh-Bénard Convection. *Phys. Rev. Lett.* **75** (9), 1743–1746. DOI: 10.1103/PhysRevLett.75.1743.

[494] Kadau, K., T. C. Germann, N. G. Hadjiconstantinou, P. S. Lomdahl, G. Dimonte, B. L. Holian, and B. J. Alder (2004). Nanohydrodynamics simulations: an atomistic view of the Rayleigh-Taylor instability. *Proc. Natl. Acad. Sci. U. S. A.* **101** (16), 5851–5855. DOI: 10.1073/pnas.0401228101.

[495] Kadau, K., C. Rosenblatt, J. L. Barber, T. C. Germann, Z. Huang, P. Carlès, and B. J. Alder (2007). The importance of fluctuations in fluid mixing. *Proc. Natl. Acad. Sci. U. S. A.* **104** (19), 7741–7745. DOI: 10.1073/pnas.0702871104.

[496] Español, P. (1998). Stochastic differential equations for non-linear hydrodynamics. *Physica A: Statistical Mechanics and its Applications* **248** (1), 77–96. DOI: 10.1016/S0378-4371(97)00461-5.

[497] Español, P., J. G. Anero, and I. Zúñiga (2009). Microscopic derivation of discrete hydrodynamics. *J. Chem. Phys.* **131** (24), 244117. DOI: 10.1063/1.3274222.

[498] Balboa Usabiaga, F., J. Bell, R. Delgado-Buscalioni, A. Donev, T. Fai, B. Griffith, and C. Peskin (2012). Staggered Schemes for Fluctuating Hydrodynamics. *Multiscale Model. Simul.* **10** (4), 1369–1408. DOI: 10.1137/120864520.

[499] Scukins, A. (2014). "Bridging large and small scales of water models using hybrid Molecular Dynamics/Fluctuating Hydrodynamics framework". PhD thesis. Aston University.

APPENDIX A

STATEMENT OF CO-AUTHOR PERMISSIONS

All co-authors have granted their permissions to use data and figures from the following articles:

- Sean L. Seyler and Oliver Beckstein (2014). *Sampling large conformational transitions: adenylate kinase as a testing ground.* Molecular Simulation, 40: 855–877. [4] for Chapters 1 and 2.
- Sean L. Seyler, Avishek Kumar, Michael F. Thorpe, and Oliver Beckstein (2015). *Path Similarity Analysis: A Method for Quantifying Macromolecular Pathways.* PLoS Comput Biol 11(10): e1004568. [5] for Chapters 3 and 4.
- Nicolas Coudray, Sean L. Seyler, Ralph Lasala, Zhening Zhang, Kathy M. Clark, Mark E. Dumont, Alexis Rohou, Oliver Beckstein, and David L. Stokes (2017). *Structure of the SLC4 transporter Bor1p in an inward-facing conformation.* Protein Science, 26: 130–145. [7] for Chapter 5.

Permission from Rafael Delgado Buscalioni was obtained to use the "Hybrid MD setup" schematic (labeled FH–B–MD) embedded in Fig. 6.1.

APPENDIX B

MATHEMATICAL MODEL FOR THE DOUBLE-SLIDE TOY SYSTEM

Mathematical details are provided for the simulation of the double-slide model. The system assumes overdamped Langevin dynamics (Brownian motion) and numerical integration was performed using a first-order scheme in time. The construction of the model permits consistent coarse-graining with respect to the number of particles in a cluster, effectively allowing tuning of the number of degrees of freedom, or the dimensionality of the configuration space.

B.1 Simulating dynamics in the double-slide model

In the double-slide model, we consider the dynamics of clusters of particles in an external potential energy landscape. A cluster of size N (an N-cluster) is a collection of N identical, interconnected particles of mass m, where each particle interacts with the other $N-1$ particles through springs with a force constant k. To produce a simple model, we assume that (1) particles in a cluster only interact through their springs, (2) particles in a cluster are identical smooth spheres moving in the Stokes' flow regime of a Newtonian fluid of uniform viscosity, and (3) each particle obeys a Langevin equation in the overdamped regime (Brownian motion).

We use assumption (1) to neglect complicated particle-particle interactions, such as hard-sphere potentials. Using assumption (2), individual particles are subject to identical viscous Stokes' flow drag forces via immersion in a uniform solvent bath. Under all three assumptions, all particles in a cluster are then subject to the same (Brownian) dynamics, where each particle interacts identically with a solvent acting as a thermal bath, and where inter-particle interactions enter only through the external force term in Langevin equation. All particles in an N-cluster thus have identical collision frequencies and, therefore, friction/diffusion coefficients. Furthermore, assumption (3) implies that the dynamics are Markovian, so that particles do not perturb the velocity of the surrounding fluid, and are thus not subject to memory effects due to the motion of nearby particles—the model ignores interactions between particles through the solvent bath.

B.1.1 Particle dynamics

Individual particles within a cluster were subject to the equation of motion for Brownian dynamics [118],

$$\dot{\mathbf{r}}_i = \frac{1}{\zeta}\boldsymbol{\nabla}_i U(\mathbf{r}(t)) + \hat{\mathbf{W}}_i(t), \tag{B.1}$$

where $\mathbf{r} = (\mathbf{r}_1, \ldots, \mathbf{r}_N)$ is a $3N$-dimensional vector of all particle positions, $\mathbf{r}_i$ is the position of particle i, $-\boldsymbol{\nabla}_i U(\mathbf{r}(t)) = -\partial U(\mathbf{r})/\partial \mathbf{r}_i$ is the non-thermal force on particle i as a function of all particle positions at time t, ζ is the viscous damping constant, and $\hat{\mathbf{W}}(t)$ is a stochastic process. The stochastic term, $\hat{\mathbf{W}}(t)$, is a delta-correlated, zero-mean, stationary Gaussian process,

$$\left\langle \hat{W}_\alpha(t)\, \hat{W}_\beta(\tau) \right\rangle = 2D\delta_{\alpha\beta}\delta(t-\tau), \tag{B.2}$$

where $D = k_B T/\zeta$ is the diffusion coefficient, and α and β run over all spatial dimensions.

B.1.2 Numerical simulation

The trajectory of particle i in a cluster was generated using a first-order integration scheme (cf. Ch. 3.9 of the 2016.3 GROMACS Manual Abraham et al. [118]),

$$\mathbf{r}_i^{n+1} = \mathbf{r}_i^n + \frac{\Delta t}{\zeta}\boldsymbol{\nabla}_i U(\mathbf{r}^n) + \sqrt{2D\Delta t}\,\hat{\mathbf{G}}_i^n, \tag{B.3}$$

where a superscript n denotes a time $t = n\Delta t$, and $-\boldsymbol{\nabla}_i U(\mathbf{r}^n)$ is the non-thermal force on particle i at time step n; $\hat{\mathbf{G}}_i^n$ is a random variable representing the n^{th} realization of the stochastic process—for each spatial component of particle i at each time step n, it was generated with a Gaussian random number generator having zero mean and unit variance:

$$\left\langle \hat{W}_{i,\alpha}^n\, \hat{W}_{i,\beta}^m \right\rangle = 2D\Delta t \left\langle \hat{G}_{i,\alpha}^n\, \hat{G}_{i,\beta}^m \right\rangle = 2D\Delta t\, \delta_{\alpha\beta}\delta_{nm}, \tag{B.4}$$

where α and β run over all spatial dimensions, and n and m are the n^{th} and m^{th} time steps, respectively.

B.2 Double-slide potential energy landscape

It is prudent to reiterate that the double-slide model's purpose is to generate "large-scale" transitions that are induced by an external ramp potential, particularly so that certain aspects of path similarity analysis (PSA) can be examined. The total potential energy in Eq. B.3 consists of an external potential and particle interaction terms: $U = U^{(\text{ext})} + U^{(\text{int})}$, respectively. The external potential is broken into a ramp potential and double-slide potential, where, for particle i (in a cluster), located at position $\mathbf{r}_i \doteq (x_i, y_i, z_i)$,

$$U^{(\text{ext})}(\mathbf{r}_i) = U_{\text{ramp}}(\mathbf{r}_i) + U_{\text{slide}}(\mathbf{r}_i) \tag{B.5}$$

$$U^{(\text{ramp})} = F\, z_i, \tag{B.6}$$

$$U^{(\text{slide})} = A x_i^2 + B\, y_i^2 \left(y_i^2 - 2C^2\right), \tag{B.7}$$

where A controls the strength of confinement in the x-direction, B and C control the shape of the slides with $y = \pm C$ defining the slide minima in nanometers, and F dictates the ramp steepness or transition force. The inter-particle potential is modeled by Hookean springs:

$$U^{(\text{int})}(\mathbf{r}) = \frac{1}{2}\frac{k_s}{2}\sum_{j\neq i}^{N}\left(\mathbf{r}_i - \mathbf{r}_j\right)^2, \tag{B.8}$$

k_s is the inter-particle spring constant.

The ramp potential generates directed progress along the z-coordinate that replicates a large-scale transition. (Note that although these transitions were produced by forcing center-of-mass translation in the positive z-direction, one should keep in mind that real transitions do not necessarily correspond to center-of-mass translations of the system, though they may be present.) The double-slide potential operates as a means to produce two distinct pathways at sufficiently low temperatures. The inter-particle springs are independent of the coarse-graining level and serve to confine particles into compact clusters.

B.3 Conditions and parameter selection for consistent coarse-graining

As PSA can take advantage of the full configuration space, the double-slide model was designed in part to test PSA when a system is coarse-grained to reduce the degrees of freedom. Transitions generated by the double-slide model should therefore be consistent in some sense across clusters of varying size. To achieve a reasonable coarse-graining scheme, (1) the transition rate and (2) net diffusivity of a cluster should be independent of its size. In particular, condition (1) demands that if all particles in a cluster are initialized at precisely the same location under zero-temperature conditions—so they all move identically under the external potential—they must transition at a rate independent of the number of particles in the cluster.

As discussed above, we assume all particles in an N-cluster are identical spheres subject to Stokes' flow in the same solvent bath. The Stokes' (viscous) drag force on a particle of radius r in a fluid of dynamic viscosity η is

$$F_{\text{drag}} = 6\pi\eta r U_\infty, \tag{B.9}$$

where U_∞ is the uniform (far-field) velocity of the surrounding fluid [430], which we treat as unperturbed by neighboring particles. For the above fluid at temperature T, the Einstein relation gives the diffusion coefficient,

$$D = \frac{k_B T}{m\gamma} = \frac{k_B T}{\zeta} = \frac{k_B T}{6\pi\eta r}, \tag{B.10}$$

in terms of, γ, the collision frequency, and $\zeta = 6\pi\eta r$, the viscous damping coefficient [473, 474]. The last equality in Eq. B.10 follows directly from Eq. B.9. In the analysis that follows, we consider an N-cluster, where each particle, i, is subject to the dynamics in Eq. B.1 in the cases of zero and finite temperature.

B.3.1 Constraints on the friction and potential from zero-temperature dynamics

We first examine conditions for constancy of the transition rate at zero temperature given identical initial conditions of all particles in an N-cluster (i.e., $\mathbf{r}_i(t) = \mathbf{r}^0$ for $i = 1, \ldots, N$). Under these conditions, the stochastic term vanishes and Eq. B.1 reduces to

$$\dot{\mathbf{r}}_i = \frac{1}{\zeta}\boldsymbol{\nabla}_i U(\mathbf{r}^0). \tag{B.11}$$

Since a transition takes place by traversing a set length of the slides along the z-coordinate, the transition rate depends only on the z-component of the velocity in Eq. B.11. Furthermore, the velocity will depend only on the ratio of the magnitude of the gradient of the external potential and the friction coefficient, ζ. Thus, under coarse-graining, we are permitted to change ζ and U for the CG particle so long as they are scaled by the same factor.

One may appeal to the necessity of conserving mass under coarse-graining so that the CG particle mass is the sum of the masses in the N-cluster. In this case, the friction coefficient of the CG particle should be N-times larger than those of the individual N-cluster particles. However, the diffusion coefficient for the CG particle will *decrease* by a factor of

N, which will change the dynamics at finite temperature when the stochastic term is non-vanishing. We keep the general analyses above in view before deciding on an appropriate scaling factor for ζ (or, equivalently, D) as we proceed to the finite-temperature case.

B.3.2 Diffusion constraints from finite-temperature dynamics

To derive more general CG constraints at finite temperature, we further require the net diffusive behavior of the CG particle to match its N-cluster. Indeed, if the diffusion coefficient in Eq. B.3, which is proportional to the average squared displacement, is preserved, then the average velocity and, thus, average transition rate will also be preserved.

We consider replacing an N-cluster of total mass M by a CG particle positioned at the center of mass of the cluster, where

$$\bar{\mathbf{r}}(t) = \frac{1}{M}\sum_{i=1}^{N} m_i \mathbf{r}_i(t) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{r}_i(t), \tag{B.12}$$

is the center of mass, and

$$\dot{\bar{\mathbf{r}}}(t) = \frac{1}{N}\sum_{i=1}^{N} \dot{\mathbf{r}}_i(t) \tag{B.13}$$

is the center-of-mass velocity. We formulate our constraint on the transition rate to be equivalent to requiring that the diffusion of the CG particle be identical to the net diffusion of center of mass of the N-cluster it replaces.

To see how coarse-graining modifies the dynamics, we take Eq. B.1, sum over all particles in the cluster, and divide by N, to obtain the center-of-mass equation of motion for an N-cluster:

$$\dot{\bar{\mathbf{r}}}(t) = \frac{1}{\zeta}\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\nabla}_i U(\mathbf{r}(t)) + \frac{1}{N}\sum_{i=1}^{N} \sqrt{2D}\,\hat{\mathbf{G}}_i(t), \tag{B.14}$$

where ζ and D are the friction and diffusion coefficients of each particle, and $\sqrt{2D}\,\hat{\mathbf{G}}_i(t) = \hat{\mathbf{W}}_i(t)$. We can simplify the equation of motion by working individually with each term on the right-hand side.

Potential term. The external potential, $U^{(\mathrm{ext})}$, is a linear function of the positions and takes the same form for each particle, but the inter-particle potential, $U^{(\mathrm{int})}$, includes nonlinear cross terms. To simplify the analysis, we begin by assuming the particles in the N-cluster are initialized at a single point and are rigidly connected. Then $\mathbf{R}(t) = \mathbf{r}_1(t) = \cdots = \mathbf{r}_N(t)$ for all t. We furthermore allow the spring potential to remain constant in the limit as $k_s \to \infty$ and $(\mathbf{r}_i - \mathbf{r}_j)^2 \to 0$ for all i and j. Under these conditions, U and its gradient become linear functions of the positions, so that the force on each particle i,

$$\boldsymbol{\nabla}_i U(\mathbf{r}) = \mathbf{F}_i(\bar{\mathbf{r}}), \tag{B.15}$$

depends only on the center-of-mass coordinate; the potential term in Eq. (B.14) then becomes

$$\frac{1}{\zeta}\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\nabla}_i U(\mathbf{r}) = \frac{1}{\zeta}\frac{1}{N}\sum_{i=1}^{N} \mathbf{F}_i(\bar{\mathbf{r}}) = \frac{1}{\zeta}\,\bar{\mathbf{F}}(\bar{\mathbf{r}}), \tag{B.16}$$

where $\bar{\mathbf{F}}(\bar{\mathbf{r}})$ is the average of the individual forces acting on the particles, which acts at the center of mass of the N-cluster. We consider scenarios where the particles have arbitrary locations and nonzero inter-particle forces in the subsequent section.

Thermal noise term. The second term on the right hand side of Eq. B.14 can be viewed as the average of N realizations of the stochastic process $\hat{\mathbf{G}}_i$. Alternatively, it is the average of the partial sum of N independent, identically distributed (iid) random variables, $\hat{\mathbf{S}}_N = \sum_{i=1}^{N} \hat{\mathbf{G}}_i$. The mean and variance of $\hat{\mathbf{S}}_N$ is the sum of the means and sum of the variances, respectively, of the $\hat{\mathbf{G}}_i$. Furthermore, since the $\hat{\mathbf{G}}_i$ are iid Gaussian random variables, $\hat{\mathbf{S}}_N$ and $\hat{\mathbf{M}}_N = \hat{\mathbf{S}}_N/N$ are also a Gaussian random variables, with zero mean and respective variances N and $1/N$. Using the central limit theorem, we replace the stochastic term in Eq. B.14 with a new stochastic process,

$$\hat{\mathbf{M}}(t) = \frac{1}{N} \sum_{i=1}^{N} \hat{\mathbf{G}}_i(t), \tag{B.17}$$

where

$$\left\langle \hat{M}_\alpha(t)\, \hat{M}_\beta(\tau) \right\rangle = \frac{1}{N} \left\langle \hat{G}_\alpha(t)\, \hat{G}_\beta(\tau) \right\rangle = \frac{2D}{N}\, \delta_{\alpha\beta}\delta(t-\tau). \tag{B.18}$$

Therefore, the center-of-mass motion of an N-cluster due to N independent, identical stochastic processes acting on the particles is statistically equivalent to a single stochastic process with a rescaled diffusion coefficient, $D^{(\mathrm{com})} = D/N$, acting on the center of mass of the particles.

Coarse-grained equation of motion. By Eq. B.18, we can write the stochastic term in terms of $\hat{\mathbf{G}}_i(t)$ with a rescaled diffusion coefficient. Combining this with Eq. B.16, the expression for the center-of-mass dynamics of a rigid N-cluster becomes

$$\dot{\bar{\mathbf{r}}}(t) = \frac{1}{\zeta}\, \bar{\mathbf{F}}(\bar{\mathbf{r}}(t)) + \sqrt{2\left(\frac{D}{N}\right)}\, \hat{\mathbf{G}}_i(t). \tag{B.19}$$

We see that the center-of-mass diffusion is smaller than the individual particle diffusion by a factor of N. If one were to replace the cluster with a CG particle at the center of mass of an N-cluster using the same diffusion coefficient D, as for each of the particles, the CG particle would diffuse too quickly. The CG particle's diffusion coefficient should thus be scaled by N so that $D^{(\mathrm{CG})} = D^{(\mathrm{com})} = D/N$ to preserve the rate of diffusion and, thus, the transition rate.

To maintain the proper relationship between the friction and diffusion coefficients through Eq. B.10, the new friction coefficient must also be increased by a factor of N: $\zeta^{(\mathrm{CG})} = N\zeta$. However, the external force must in turn be increased by a factor of N to ensure the external force on the CG particle is the same as the net force on the corresponding N-cluster. Thus, if we begin with an N-cluster under forces $F_i(\mathbf{r})$, with diffusion coefficient D and friction ζ, the equation of motion for the CG particle should be

$$\dot{\mathbf{R}}(t) = \frac{1}{\zeta^{(\mathrm{CG})}} \mathbf{F}^{(\mathrm{CG})}(\mathbf{R}(t)) + \sqrt{2D^{(\mathrm{CG})}}\, \hat{\mathbf{G}}_1(t), \tag{B.20}$$

where $\mathbf{R}(t)$ is the position of the CG particle at time t, $D^{(\mathrm{CG})} = D/N = k_B T/\zeta^{(\mathrm{CG})}$, and $\mathbf{F}^{(\mathrm{CG})}(\mathbf{R}) = \mathbf{F}_1(\mathbf{R})$ (where we picked $i = 1$ to denote terms identical to quantities for any

single particle in the cluster). The new friction coefficient, $\zeta^{(\mathrm{CG})}$, is additionally consistent with the previous consideration of conserving particle mass under coarse-graining:

$$\zeta^{(\mathrm{CG})} = N\zeta = (Nm)\gamma = m^{(\mathrm{CG})}\gamma, \tag{B.21}$$

where the mass of the CG particle, $m^{(\mathrm{CG})} = Nm = M$, is the sum of the masses of the individual (identical) particles. Furthermore, we see that the collision frequency, γ, can be held fixed across simulations at all levels of coarse-graining.

B.3.3 Deviations from rigid-cluster coarse-grained behavior

It is clear that $\mathbf{F}^{(\mathrm{CG})}(\mathbf{R}) = \mathbf{F}_1(\mathbf{R}) = \bar{\mathbf{F}}(\mathbf{R})$ (averaged over all particles i) in the case where all particles are constrained to the same point for all time. This assumption guarantees the linearity of the potential energy function. Our simulation system on the other hand has finite inter-particle spring constants and an N-cluster will tend to have its constituent particles wandering to some degree. If the single-point, rigid-cluster assumption is relaxed, we can then ask under what conditions the force on the center of mass is well-approximated by the average of the individual particle forces.

If all particles are located at different positions, then they will see different parts of the external potential and will also have nonzero inter-particle forces between them. However, due to Newton's third law these inter-particle forces sum to zero and do not contribute to a force acting on the center of mass. On the other hand, the sum of the external forces on all particles will *not* generally be equal to the gradient of the external potential evaluated at their center of mass. Now consider replacing an N-cluster with a CG particle that feels N times the force of a regular particle at the same point: $F^{(\mathrm{CG})}(\mathbf{r}^*) = NF_1(\mathbf{r}^*)$. In this case, the net force on the N-cluster (at its center of mass) is, in general, different than the force on the CG particle due to the external potential (at the center of mass):

$$\mathbf{F}^{(\mathrm{CG})}(\mathbf{R}) = -N\boldsymbol{\nabla}_{\mathbf{R}}U^{\mathrm{ext}}(\mathbf{R}) \neq \sum_{i=1}^{N}\mathbf{F}_i(\mathbf{r}) = -\sum_{i=1}^{N}\boldsymbol{\nabla}_i U^{\mathrm{ext}}(\mathbf{r}_i) \tag{B.22}$$

In particular, Eq. B.22 will occur when the external potential U^{ext} has non-vanishing second (or higher) derivatives, i.e., the external force is not a constant function of position.

These statements can be made more precise by first transforming to coordinates relative to the center of mass and writing the positions as

$$\mathbf{r}_i = \Delta\mathbf{r}_i + \mathbf{R}, \tag{B.23}$$

where $\Delta\mathbf{r}_i$ is the displacement vector of the i^{th} particle relative to the center of mass. The total external force,

$$\mathbf{F}^{(\mathrm{total})} = \sum_{i=1}^{N}\mathbf{F}_i(\mathbf{r}_i) = \sum_{i=1}^{N}\mathbf{F}_i(\Delta\mathbf{r}_i + \mathbf{R}), \tag{B.24}$$

can then be expressed, for small displacements in $\Delta\mathbf{r}_i$, as a Taylor series expansion about the center of mass,

$$\sum_{i=1}^{N}\mathbf{F}_i(\Delta\mathbf{r}_i + \mathbf{R}) = \sum_{i=1}^{N}\left[\mathbf{F}_i(\mathbf{R}) + \frac{\partial\mathbf{F}_i}{\partial\mathbf{r}_i}\cdot\Delta\mathbf{r}_i + \mathrm{O}\left(\Delta\mathbf{r}_i^2\right)\right]. \tag{B.25}$$

For sufficiently small displacements, the product $\frac{\partial \mathbf{F}_i}{\partial \mathbf{r}_i} \cdot \Delta \mathbf{r}_i$ can be dropped, giving

$$\mathbf{F}^{(\text{total})} \sim \sum_{i=1}^{N} \mathbf{F}_i(\mathbf{R}). \tag{B.26}$$

In the case that the (external) forces are derivable from a potential function, Eq. B.26 holds when the second derivatives of the potential are small.

If we assume the particles are sufficiently proximate at time t such that $\boldsymbol{\nabla}_i U(\mathbf{r}(t))$ can be treated as approximately constant across the space occupied by the N-cluster, then coarse-graining according to the above prescription will approximately preserve net diffusion to first order in the displacements, $\Delta \mathbf{r}_i$, of the particles. Under this condition, the total force on the cluster can be approximated by the external force on the CG particle. For the form of the potential in Eq. (B.7), we expect diffusive behavior orthogonal to the z-coordinate for an N-cluster to deviate from that of the CG particle since the potential is nonlinear along those directions. However, as our transition rate is determined by the constant-gradient potential in the z-direction in Eq. (B.6), the transition rate will be statistically identical for CG particles if the same identical initial conditions as the N-clusters are used. This is the key result that assures the average transition rate is unchanging with respect to N.

B.4 Double-slide simulation parameters

We emphasize that the double-slide system was designed to produce noisy, non-trivial trajectories to test PSA and the path metrics. The parameter values used for the double-slide simulations, while guided by realistic physical systems, were determined in part by our coarse-graining constraints and practical considerations (i.e., numerical stability of the integrator, simulation time, number of simulation steps, etc.).

Table B.1 summarizes the parameters used for simulations of single- and eight-particle molecules. The potential was constructed to produce potential energy changes on the order of (or less than) $10\,\mathrm{kJ/(mol\,\AA)}$ with a time step then chosen as large as allowed by the numerical stability of the integrator. The friction coefficient, γ, was heuristically chosen to reflect typical values using water as a solvent, which can range from 0.1–100 ps to depending on the solute of interest [381, 475–477]. Spring constants, which approximate bond strength, are usually on the order of 0.1–10 $\mathrm{kJ/(mol\,\AA^2)}$ [205, 478, 479] and were set to within an order of magnitude of values used in the literature.

Table B.1: Double-slide model parameters for one- and eight-particle molecules.

N	Potential landscape parameters					Dynamical parameters		
	A^* (kJ/(mol nm^2))	B^* (kJ/(mol nm^4))	C (nm)	F^* (kJ/(mol nm))	k_s (kJ/(mol nm^2))	$\gamma^\dagger$ (ps^{-1})	m^* (Da)	$\Delta t^\ddagger$ (fs)
1	30.0	12.18	0.8	7.5	—	50.0	10.0	0.5-2.5
8	3.75	1.523	0.8	0.9375	41.84	50.0	1.25	0.5-2.5

Simulations were performed in 50 K steps from 0 K to 500 K, and also at 600 K. Values for $N = 8$ are either unchanged from the $N = 1$ case or scaled down by a factor of eight to produce (zero-temperature) center-of-mass dynamics corresponding to single-particle dynamics. A controls x-direction confinement, B and C control the slide shape, with $y = \pm C$ being the slide minima in nanometers, and F is the ramp steepness. The collision frequency, γ, was the same for all systems, but particle masses (and friction coefficient, $\zeta = m\gamma$) were scaled to keep total cluster mass constant.

* Value of parameter was scaled inversely proportionally to N.

† Constant collision frequency; friction coefficient, $\zeta = m\gamma$, scales proportionally to the particle mass, m, and particle number, N.

‡ Time steps were chosen to mitigate simulation time while maintaining integrator stability for a given temperature. Values for $N = 8$ are either unchanged from the $N = 1$ case or scaled down (by a factor of eight) to match the (zero-temperature) center-of-mass dynamics corresponding to one-particle systems. Simulations were performed in 50 K steps from 0 K to 500 K, and also at 600 K. Time steps were chosen heuristically at each temperature to balance simulation length and stability of the integrator.

APPENDIX C

PATH-SAMPLING METHODS USED IN THE ADK COMPARISON

A summary of the transition path generating algorithms, used in the comparison of sampling methods, is provided for convenience. We summarize the key aspects of each of the physical models and path generating algorithms to help lay the groundwork for connecting algorithmic/model differences to differences between the respective transition paths that were produced.

C.1 Overview

The sampling methods used to generate paths were selected primarily among those available on publicly accessible servers. DIMS-MD [172, 174] and FRODA [180] trajectories were generated in-house to produce large AdK and DT path ensembles, which was not feasible for the other methods through their web server interfaces. Table C.1 summarizes the primary references and web server locations for each method. It is urged that readers interested in further details about the methods refer to the original published articles for complete descriptions.

Table C.1: Primary references and public web servers (if available) for methods employed in path-sampling comparison.

Name	Reference	Web server*	Adjusted parameter†
DIMS	Perilla et al. [174]	–	re-run
FRODA	Farrell, Speranskiy, and Thorpe [180]	http://pathways.asu.edu	re-run
MDdMD	Sfriso et al. [298]	http://mmb.irbbarcelona.org/MDdMD/	re-run
GOdMD	Sfriso et al. [223]	http://mmb.irbbarcelona.org/GOdMD/	relax: 20,50,80 ps
rTMD	Ferrara, Apostolakis, and Caflisch [480]	–	k, pull spd, re-run
ANMP	Das et al. [299]	http://anmpathway.lcrc.anl.gov/anmpathway.cgi/	cutoff: 12,15,18 Å
MAP	Franklin et al. [218]	http://lorentz.dynstr.pasteur.fr/joel	cutoff: 7,10,13 Å
MENM	Zheng, B. R. Brooks, and Hummer [208]	http://enm.lobos.nih.gov/start_path.html	cutoff: 8,10,12 Å
iENM	Tekpinar and Zheng [209]	http://enm.lobos.nih.gov/start_ienm.html	cutoff: 8,10,12 Å
Morph	Krebs and Gerstein [176]	http://molmovdb.org/cgi-bin/submit.cgi	fit/min, fit, none
LinInt	–	–	–

The methods used to generate transition paths are given with their primary references and corresponding URL for submitting jobs. Three different transitions were generated per method (except LinInt) by adjusting a single parameter to three different values; Morph transitions were produced by varying input structure superimposition and energy minimization settings. The URL of the server for FRODA is shown in red to indicate that the site was down at the time of writing.

* Web server: URL of job submission page

† Adjusted parameter: re-run, initial conditions automatically randomized; relaxation, relaxation window in picoseconds; k, spring constant; pull spd, speed of equilibrium position of moving restraint; cutoff, spring cutoff distance in Å; fit, superimposition of input structures; min, energy minimization (adiabatic mapping).

DIMS, MDdMD, and GOdMD are all non-deterministic dynamical algorithms that implement importance sampling in the form of a soft-ratcheting Metropolis condition, although MDdMD and GOdMD are hybrid methods that also incorporate additional biasing based on normal mode information derived from coarse-grained potentials. The FRODA algorithm is dynamical and deterministic, but can be made stochastic by turning on a random motion setting that performs a random displacement/rotation of each rigid subunit prior to each step. The five ENM-based methods (counting MENM-SD and MENM-

SP separately) are all based on double-well anisotropic network models, but use different approaches to generating a single (mixed) potential, and how they generate a minimum energy path. MAP is unique in that it solves for the path that minimizes the Onsager-Machlup action subject to Brownian dynamics boundary conditions, whereas the other ENMs find steepest descent or saddle point paths.

C.2 Transition path methods details

DIMS MD [172, 174] employs an informational criterion to bias a transition toward a target structure. Here we use DIMS MD with the soft ratcheting algorithm together with Langevin dynamics and the Analytical Continuum Electrostatics (ACE) model to generate stochastic MD steps in an implicit solvent bath, as described previously [40]. A 1 fs time step was used with a collision frequency (damping coefficient) of $25\,\mathrm{ps}^{-1}$. Steps toward the target (measured using heavy-atom rmsd-to-target as the progress variable) are always accepted, while steps increasingly farther away from the target are accepted with decreasing probability according to a Boltzmann distribution of the square of the order parameter. Soft ratcheting has emerged as a robust approach to ensure overall progress toward the target structure while retaining the capability to back out of energetic dead ends [40, 174].

FRODA [180] advances proteins toward a target structure by incrementally reducing the rmsd while enforcing steric constraints. Random paths are produced by introducing random perturbations to the orientation of one side chain per simulation step. Contact constraints shared by the initial and final conformations were preserved throughout the simulations, while non-common contacts were allowed to be broken and/or formed. Contacts were defined as atom pairs falling within an 8 Å cutoff radius.

Both **MDdMD** [298] and **GOdMD** [223] employ deterministic discrete MD to quickly sample large regions of configuration space; for each simulation a sub-sequence of configurations is dynamically constructed by selecting conformer snapshots on the fly according to a Metropolis-like Monte Carlo (MC) procedure combined with informational criteria. The sub-sequences represent physically plausible transition paths. In both approaches, progress is driven toward a final state using an soft ratcheting-like algorithm. MDdMD further selects pre-accepted snapshots based on the degree of eigenvector overlap with the essential transition vector. The essential transition vector is computed (via NMA of a Gō-like potential) using primarily the initial state. GOdMD differs in that it employs an ENM-metadynamics method to bias proteins away from the initial energetic well.

The MDdMD method employs simplified or multi-step square potentials for bonded interactions at the atomic level. Solvent, van der Waals, and electrostatic forces are modeled using two-step square wells for attractive interactions and soft barriers for repulsion. GOdMD is based on a C_α representation. Square wells define the physical chemistry while a multi-well Gō-like potential describes non-bonded interactions. For both methods the initial velocities are randomized so that separate runs generate distinct trajectories. The three GOdMD trajectories were produced, however, by varying the relaxation window: 20 ps, 50 ps, and 100 ps.

The **rTMD** [480] (restrained TMD) method resembles the original TMD algorithm which was based on a time-dependent holonomic constraint that moves linearly toward a target structure; in the rTMD variation, a moving harmonic restraint is used instead. As with

DIMS, the rTMD method requires a progress variable in the form of an atom selection for biasing. rTMD simulations were set up to mimic DIMS simulations as closely as possible: the heavy-atom rmsd-to-target progress variable was used for biasing and rmsd fitting; we used Langevin dynamics with an identical time step and collision frequency, and the Generalized Born implicit solvent as implemented in the NAMD simulation code [352] with a matching time step, collision frequency. During each time step, the equilibrium position of the harmonic restraint is moved a fixed amount by decreasing the rmsd to the target conformation. We generated six transitions in total, three for each of the following settings (using velocities randomly sampled from the Maxwell-Boltzmann distribution): three *fast*-pulling simulations with $k = 4.184 \times 10^5\,\mathrm{kJ\,mol^{-1}\,\AA^{-1}}$ and $\sim 1\,\mathrm{\AA\,ps^{-1}}$ pulling speed; three *slow*-pulling simulations with $k = 4.184 \times 10^3\,\mathrm{kJ\,mol^{-1}\,\AA^{-1}}$ and $\sim 0.01\,\mathrm{\AA\,ps^{-1}}$ pulling speed.

Using the C_α positions of an initial and final structure, the **CG-EN** models examined here construct two energetic potentials around the respective states. The separate potentials are then combined (e.g. through a heuristic mixing rule) to form a double-well potential that models the energy landscape of the configuration space spanning the two states. A path is generated from the double well potential using, for instance, a minimum energy path (MEP) approach. The **ANMP** [299] algorithm constructs two anisotropic network models (ANM) about their respective end states, iteratively searches for transition state defined as the minimum energy structure lying on the cusp hypersurface of a combined two-state potential, and performs two steepest descent (SD) minimizations from the transition state to generate a sequence of conformers that represent a conformational path. Three transitions were generated using the following spring cutoff distances: 12 Å, 15 Å, and 18 Å. Similarly, **MAP** [218] generates a separate ANM potential around each end state. The MAP transition is generated by first minimizing the Onsager-Machlup action assuming overdamped Langevin dynamics, then the resulting deterministic equations are solved analytically for positions and velocities using appropriate boundary conditions. Spring cutoff distances were set to 7 Å, 10 Å, and 13 Å.

The AD-ENM Web Server [208, 209] offers two approaches to generating ENM transitions. Like the ANMP and MAP models, both approaches are based on ANM representations of the initial and final protein configurations. The **MENM** of the PATH-ENM server generates a double-well potential from a predefined mixing function. The energetic minima and saddle points (SPs) are computed and transitions are generated in one of two ways: (1) The steepest descent (SD) paths about each SP; (2) The path tracing the minima and SPs. The **iENM** server exploits the independence of the equation for the minima and SPs on the detailed form of the mixing function. An iterative procedure is used to solve for the universal minimum-energy path (MEP) through the SPs of an arbitrary double-well potential that includes a predefined steric collision energy. Three distinct transitions for each method were produced using an 8 Å, 10 Å, and 12 Å spring cutoff distance.

The **Morph** [176] servers generate transitions using adiabatic mapping, which combines simple linear interpolation (between two structures) and CHARMM-based energy minimization to reduce severe steric clashing. Two transitions were generated using stepwise energy minimization: one transition used the 'superimpose before morphing' option, while the other relied on our CORE alignment procedure (described in S4 Text), allowing us to

compare possible path differences arising from structural pre-alignment. A third transition without both energy minimization and superimposition was also performed.

We also generated a zeroth-order transition path using naive linear interpolation (**LinInt**). The path is defined by a sequence of evenly spaced conformers lying along the line formed by the separation vector between the C_α representations of the initial and final states. We used LinInt as a convenient reference path for the comparison of the higher-order methods.

C.3 Similarities between DIMS-MD and ABMD

There is an MD-based method, going by several names, that uses a ratchet-and-pawl approach, where a one-sided harmonic potential guides a simulation toward a target structure, much like SRA in DIMS; it has been referred to as adiabatic bias molecular dynamics (ABMD) [189], biased MD (BMD) [481], and ratcheted MD [190]. There are several subtle differences between the ABMD algorithm and SRA, the most important one being that ABMD introduces an explicit bias force into the molecular potential along the progress variable with a harmonic form

$$V_{\text{rat}}(\rho(t)) = \begin{cases} 0, & \rho(t) \leq \rho_m(t) \\ (\rho(t) - \rho_m(t))^2, & \rho(t) > \rho_m(t), \end{cases} \tag{C.1}$$

where $\rho(t) = d^2_{\text{RMS}}(\mathbf{X}|\mathbf{X}^F)$ is the mean-squared distance to the target (similar to Eq. 2.2) and $\rho_m(t) = \min_{0\leq\tau\leq t}\{\rho(\tau)\}$ is the equilibrium value of the harmonic potential; $\rho_m(t)$ takes on a value that corresponds to the smallest value of $\rho(t)$ that occurred over the entire simulation (up to the current time), which is to say that the potential moves (all the way) up to the point at which the simulation was nearest the target. ABMD thus also differs from DIMS in that the potential bias V_{rat} moves up to and remains at the point of maximum progress (the pawl aspect prevents V_{rat} from moving backward), while the SRA algorithm "follows" the simulation along both forward and backward steps. However, both methods are similar in the following way: V_{rat} will sample backward steps when $\rho(t) > \rho_m(t)$ with the Boltzmann-weighted probability $\exp(-\rho^2/k_{\text{B}}T)$, which matches the form of p_{acc} (when $\Delta\phi > 0$). In the adiabatic limit (and ignoring the pawl aspect of ABMD for the moment to match DIMS, so that $\rho_m^{n+1} = \rho^n$, where n is the current step), the bias potential V_{rat} in ABMD generates a distribution identical to p_{acc} when we set $(\Delta\phi_0)^2 = k_{\text{B}}T$, with ABMD generating an purely enthalpic force and DIMS an purely entropic force.

APPENDIX D

ALIGNMENT PROCEDURE FOR PROTEINS USED IN PATH SIMILARITY ANALYSIS

In this addendum, the best-fit RMSD is discussed as it pertains to the application of the path metrics used in this dissertation. Theoretical considerations are discussed, motivating the means by which the structural alignment of the AdK and DT systems were carried out prior to applying PSA.

D.1 Best-fit rmsd as a point metric

In general, a simulation snapshot will be a (protein) structure having an arbitrary orientation and center-of-mass translation. Since the rmsd between two conformations of the same structure will depend on their relative orientation and separation, it is necessary to employ an alignment procedure to ensure a unique rmsd will be computed for the pair. The best-fit rmsd between a structure pair is computed by first aligning the centers of mass and finding the rotation (matrix) that minimizes the rmsd. Though this approach is common, we discuss several reasons why the best-fit rmsd may not be the best choice of point metric for PSA.

To perform PSA of a collection of transition paths, we must compute path similarities for all unique pairs of paths. For each pair of paths P and Q, there are p and q snapshots, respectively. Using the rmsd as the point metric, the Hausdorff and Fréchet distance between P and Q both require pq rmsd calculations. If the best-fit rmsd is to be used, then pq optimizations must be performed, corresponding to pairwise alignment of all unique pairs of structures between P and Q.

While it may be thought that computing the best-fit rmsd be computed on a pairwise basis, there are two drawbacks to this approach: The first issue has to do with computational cost, as rmsd optimization, though relatively fast using the Quaternion Characteristic Polynomial (QCP) algorithm [482, 483], can quickly become computationally expensive. Given N transition paths, each having s snapshots, to be compared with PSA, there are $N(N-1)/2$ unique path comparisons. For each comparison (of two paths), there are s^2 unique pairs of conformers, one conformer each, between them. The pairwise best-fit rmsd approach would then necessitate $N(N-1)s^2/2$ optimizations: the cost grows proportionally to the product of square of the number of paths and the product of the numbers of time steps in each pair of paths. Comparing an ensemble of hundreds of trajectories, each composed of hundreds of conformer snapshots, can easily demand upwards of a billion best-fit rmsd optimizations. The second problem is that the pairwise-minimal rmsd will not generally preserve the triangle inequality and thus will not behave as a proper metric on configuration space [295]. Intuitively, it can be seen that, since a triplet of trajectories will have three unique rotational alignments between each of the three combinations of unique trajectory pairs, and since each rotation depends only on the trajectories in the pair, pairwise-minimal rmsd measurements need not obey the transitive property and can thus violate the triangle inequality.

The Hausdorff and Fréchet metrics are only proper metrics provided that they are defined in terms of a (proper) point metric. Thus, Hausdorff and Fréchet calculations will not preserve metric properties if the best-fit rmsd is used. One may argue that the metric requirement may be relaxed for the path metrics without affecting the quality of results. It is also possible that there are benefits to using the best-fit rmsd to maximize consistency with its use in the literature. We were nevertheless able to obtain sensible and consistent results

using heuristic alignment schemes (discussed next) that preserve all the properties of a metric. Although the relationship of structural similarity measures and PSA is interesting and likely worth further examination, it is out of the scope of this paper.

D.2 Alignment procedures used in study

To mitigate computational costs, our current implementation of PSA utilizes a pre-alignment procedure: for each path, the rotation matrix R_i that minimizes the rmsd between frame i and a single reference structure is computed, then used to rotate frame i. Thus, the alignment of all conformer snapshots in a path P to a pre-defined reference structure scales linearly with respect to the number of snapshots, p, in a path, whereas pairwise alignment scales quadratically. Given a path ensemble, a single reference structure common to the entire ensemble is used as the basis for aligning each conformer in each path. Although a poor choice of reference structure may be suboptimal, we found that our procedure produces sensible results at much reduced computational cost and complexity. Furthermore, the use of a single reference structure for the structural superposition preserves the triangle inequality for the point metric and thus imbues δ_H and δ_F with the qualities of a proper metric for paths.

D.2.1 AdK trajectory alignment

In the special case of AdK, since the conformational motion is known to be primarily confined to the NMP and AMP domains and the hinges connecting them to the CORE domain, we chose to align the CORE domain of each intermediate conformation to the average of the aligned CORE C_α coordinates (using the best-fit rmsd) of the 1AKE:A and 4AKE:A structures with the center of mass of the averaged CORE at the origin. We chose to use the average CORE C_α coordinates as a putative reference structure to reduce the alignment bias of intermediate snapshots residing closer to one of the boundary conformations than the other. The alignment procedure, in the context of the comparison of transition path methods, is demonstrated in the example Python script `psa\_full.py` in the PSA tutorial, which is available as open source at github.com/Becksteinlab/PSAnalysisTutorial under the GNU General Public License 3.

To align a given AdK conformer snapshot, it was translated so that the center of mass of its C_α CORE atoms coincided with the origin. All of the atoms of the conformer were then rigidly rotated according to the rotation matrix that generated the best-fit rmsd between the conformer's C_α CORE and the (C_α-CORE) reference structure, using the QCP algorithm [482, 483] implemented in MDAnalysis [272]. The entire structure of each snapshot in each path, for all paths in a transition path comparison, was translated and rotated so as to align each C_α-CORE region to the reference C_α-CORE coordinates. Path metric calculations were then performed on the set of aligned paths, where for each structural comparison, the C_α rmsd was directly computed without any further rotation/alignment. In the hypothetical case where the C_α CORE domains of two intermediate conformers have coordinates identical to the reference coordinates, the C_α rmsd between the (entire) intermediate structures will be solely due to C_α deviations in the LID and NMP domains. Therefore, this alignment procedure produces rmsds reflecting residue displacements in the

mobile LID and NMP domains, and not the CORE domains. In the case of the path-sampling methods comparison, structural rmsd measurements were made once CORE domains were aligned. The comparison between DIMS and FRODA used the same alignment protocol outlined above with the added step of translating the center of mass of all conformer snapshots to the origin (instead of just the CORE). This was done in part to lower the rmsd and path distance measurements further to see if the effectiveness of the Hausdorff pairs comparison would be reduced; our results suggest that Hausdorff pairs analysis is still viable.

D.2.2 DT trajectory alignment

We found that a satisfactory alignment procedure for DT transition paths was achieved by using the average C_α coordinates of the full 1MDT:A and 1DDT:A structures to generate a reference structure. The DT path ensembles were otherwise aligned in an identical manner to the AdK ensembles: each conformer snapshot for each path was aligned to the reference prior to calculating path similarities; after aligning each path, the rmsd between conformers in different paths without any further rotations/alignments was used as the point metric.

Although DT is conceptualized as having a mobile Receptor-binding (R) domain that moves relative the Catalytic (C) and Translocation (T) domains, it was not amenable to the alignment procedure we used for AdK (see Fig. 4.5 in the main text). If we were to carry out an analogous procedure, we must align the C_α atoms corresponding to the C and T domains in the 1MDT:A and 1DDT:A end structures. The average coordinates of the 1MDT:A C/T C_α atoms and 1DDT:A C/T C_α atoms would then be then be used as the reference structure to which individual conformers would be aligned. We found, however, that the procedure produced larger Hausdorff and Fréchet distances than when simply aligning the C_α atoms of entire conformers to either of the end structures.

The reason that C/R domain alignment performs poorly is likely due to a confluence of several factors. First, the C and T domains were seen to fluctuate greatly throughout DIMS and FRODA simulations and were not static to the same degree as the AdK CORE. Second, the overall shape of the C and T domains taken together is somewhat cylindrical with an approximate axis running through both, while the R domain is displaced mostly orthogonally from this axis. A structure aligned to this region may be rotated relative to another in the sense that its R domain has been swung around the cylindrical R/C domains during alignment. R domains will therefore tend to be displaced if only the C and T domains are used to produce a reference. After alignment, the rmsd becomes too sensitive to the orientation of the R domain about the T and C domain. In the case of using the entire structure to produce reference coordinates, alignment is sufficiently constrained about the T/C "axis" to prevent erroneous R domain displacements.

APPENDIX E

ON SELECTING AND VALIDATING HIERARCHICAL CLUSTERING LINKAGES

In this addendum we mention qualitative and quantitative considerations in selecting the Ward linkage criterion for hierarchical clustering, in the context of other linkage criteria. Some comments on potential data interpretation pitfalls when performing general cluster analyses are provided with a view toward viable approaches to PSA cluster and data validation.

E.1 Hierarchical clustering and linkage criteria

Since we use PSA to generate a distance (proximity) matrix (of all path pairs), cluster analysis is directly amenable as a mode of exploratory data visualization. One approach that requires little external input is agglomerative hierarchical clustering, which generates a binary tree (dendrogram) depicting the relationships between objects and clusters.

The similarity between one object (a singleton cluster) and another is specified directly by the distance matrix, while a *linkage* defines a general inter-cluster distance as a function of the pairwise distances of the individual objects (in the clusters). A given linkage will emphasize certain features of the distribution(s) underlying the objects and choosing a good linkage may depend on a confluence of factors. In general, a linkage should generate a clustering that corresponds well with the original distance matrix while highlighting potentially relevant patterns in the data. We summarize below—in the context of transition paths—several linkage algorithms that we applied. The linkages were implemented using the SciPy clustering package and are also described in the SciPy Reference Guide online [289]. R. Xu and Wunsch [484] also provides alternative definitions and descriptions of the linkages, as well as a general overview of hierarchical clustering.

E.1.1 Ward linkage

Ward's (minimum variance) method is motivated by the statistical analysis of variance between distributions. A candidate object will be grouped with the cluster whose sum of squared errors increases the least upon the object's inclusion [484]. The inter-cluster distance for Ward linkage between clusters u and v, where u is constructed from two sub-clusters, s and t, is

$$d(u,v) = \sqrt{\frac{n(s)+n(v)}{N}d^2(s,v) + \frac{n(t)+n(v)}{N}d^2(t,v) + \frac{n(v)}{N}d^2(s,t)}, \tag{E.1}$$

where d is the inter-cluster distance, $n(*)$ is the cardinality of (numbers of paths in) cluster $*$, and $N = n(u) + n(v) = n(s) + n(t) + n(v)$ is the total number of paths. The definition is recursive in that the distance between two clusters is defined in terms of distance between their constituent clusters. The first two terms under the square root give the fractional contributions of paths in sub-clusters s and t to the overall distance. The third term is a distance "penalty" that is proportional to the relative size of v, $n(v)/N$, and the "spread" of u, $\delta^2(s,t)$, which is the squared distance between its sub-clusters. Under Ward linkage, the clustering procedure, which during each pass seeks the two clusters with the smallest distance, will tend to agglomerate small, compact clusters [484, 485].

E.1.2 Other linkages

Single linkage. Also called the Nearest Point Algorithm, single linkage defines the distance between two clusters of paths as

$$d(u,v) = \min\left(\delta(u_i, v_j)\right), \tag{E.2}$$

u and v, and u_i and v_j are the i^{th} and j^{th} paths in clusters u and v, respectively.

Complete linkage. Similar to single linkage, complete linkage (Farthest Point Algorithm) defines inter-cluster distance as

$$d(u,v) = \max\left(\delta(u_i, v_j)\right). \tag{E.3}$$

Average linkage. Also called the Unweighted Pair Group Method Average (UPGMA), average linkage defines the distance between clusters as

$$d(u,v) = \frac{1}{n(u)\cdot n(v)} \sum_{i=1}^{n(u)} \sum_{j=1}^{n(v)} \delta(u_i, v_j), \tag{E.4}$$

In other words, the distance between u and v is the average distance between a path in u and a path in v. Average linkage can also be defined recursively in terms of two sub-clusters, s and t, comprising u:

$$d(u,v) = \frac{1}{n(s)+n(t)}\left[n(s)d(s,v) + n(t)d(t,v)\right]. \tag{E.5}$$

The contributions of s and t to the distance, $d(u,v)$, between u and v are proportional to their relative sizes so that each path contributes equally to the cluster distance, d.

Weighted average linkage. The Weighted Pair Group Method Average (WPGMA) or *weighted average* linkage is defined recursively as

$$d(u,v) = \frac{1}{2}\left[d(s,v) + d(t,v)\right]. \tag{E.6}$$

Weighted average linkage gives equal importance to sub-clusters s and t, regardless of the number of paths in each. Thus, if s contains half the number of paths in t, a path in s will have twice the weight as a path in t in contributing to the inter-cluster distance, $d(u,v)$.

E.2 Examining cluster quality

Although analyzing in depth the consequences of using different linkages was out of the scope of this study, we produced four additional Fréchet heat map dendrograms (see Fig. S5) for the path-sampling methods analysis using the single, complete, average, and weighted average linkages to perform a basic comparison. The structure of each dendrogram is a function of the linkage algorithm, although it can be difficult (and tricky) to draw strong conclusions about the underlying data.

E.2.1 Quantitative measures

The *cophenetic distance* is a measure of correlation between the original distance matrix and the distances between objects according to the inter-cluster distances assigned by a given linkage. To obtain an approximate measure of the quality of cluster divisions, we examined the *inconsistency coefficients* for clusters containing more than two children (i.e., ignoring singleton clusters and clusters with vanishing inconsistency coefficients). We also computed the maximum inconsistency coefficients for each linkage, where the coefficient for a given cluster is the largest of the coefficients between itself and its children [289, 485].

We computed cophenetic distances for the clusterings produced by each linkage. To concisely examine the inconsistency coefficient data, we averaged the (nonzero) values for each linkage; we likewise computed the average of the maximum inconsistencies for each linkage. There were eleven clusters with inconsistency coefficients of zero among the non-singleton clusters, which were excluded from both averages. The three values computed for each linkage are summarized in Table E.1.

Table E.1: Summary of the computed clustering-quality measures for each linkage for the methods comparison of Fréchet distances.

Linkage	Cophenetic Distance	Inconsistency Coefficient Statistics	
		Average*	Maxima average†
Ward	0.78	0.86	0.93
single	0.85	0.75	0.80
complete	0.82	0.88	0.92
average	0.85	0.85	0.91
weighted	0.86	0.83	0.91

A cophenetic distance near zero indicates that the overall clustering poorly reflects the actual pairwise distances between the paths, while values close to unity indicate a viable—but not necessarily good—clustering. All linkages produced clusterings with adequate correspondence. A large inconsistency coefficient indicates that the height of the link corresponding to a cluster is large compared to the average link heights of its children, which in turn indicates that the two child clusters joined at this level are dissimilar. Clusters (links) having larger inconsistency coefficients have more distinct child clusters.

* Average of the nonzero inconsistency coefficients computed for non-singleton clusters.

† For each non-singleton cluster along with its descendents, find the maximum inconsistency coefficient, take the average of the nonzero maxima.

E.2.2 Selecting a linkage criterion

Inter-cluster distances will tend to be smaller for a single linkage than a complete linkage, since the former defines inter-cluster distance as the minimum (rather than maximum) inter-point distance. As a result, fluctuations due to measurement uncertainty will be relatively larger for single linkage, making it more susceptible to noise and a "chaining" effect that can produce stretched clusters [484]; the single linkage (see Fig. S5A) in the methods comparison produced a large (yellow) cluster resembling this effect. The other four (Ward,

average, weighted average, complete), on the other hand, generated qualitatively viable clusterings. We note that the average and weighted average linkages merged MDdMD with the Morph/MAP cluster, then merged the result with DIMS, then finally again with FRODA, while the other ENM methods still formed their own cluster. The complete linkage, however, while having placed MDdMD and DIMS in their own cluster, merged FRODA with the ENM methods! It is difficult at this point to determine why FRODA transitions are subject to such different clusterings, although It appears that FRODA paths, being somewhat uniformly different from all others, are subject to being clustered by their dissimilarity, in some sense, rather than their similarity to other paths.

To decide which linkage is "best" at this stage corrupts the purpose of exploratory data analysis. A more productive view is that each linkage emphasizes different aspects of the data and can be used to tease out specific types of information. A sound general approach is likely to involve the integration of results from a variety of linkages as a means to identify common features and patterns. If a particular cluster, say, Morph/MAP/LinInt, were to appear as a motif across several linkages, it would be a more robust indication that its constituent paths were truly similar (relative to other paths). This approach can be expanded to include other clustering algorithms, such as partitional clustering approaches (e.g., the K-means algorithm).

In light of the previous discussion and the quantitative results in Table E.1, the Ward, complete, average, and weighted average linkages perform adequately for the purposes of this study. Both Ward and complete linkage produced particularly well-defined clusters, which was reflected by their overall large inconsistency coefficients. Arguments can easily made in favor of the average-type linkages as well. To keep the focus of the study on the presentation of PSA, we elected to concentrate on one—the Ward linkage–for clarity of presentation.

APPENDIX F

EXPLORING DISCRETE AVERAGE FRÉCHET AND AVERAGE HAUSDORFF DISTANCE FUNCTIONS

This addendum explores variations of the Hausdorff and discrete Fréchet metrics that are based on averages rather than maxima. Straightforward definitions are provided for these average-type path *distance functions*. We walk through an example where they violate the triangle inequality to show that they do not define proper metrics. Both average-type path distance functions are applied in the context of the path-sampling methods comparison (using Ward hierarchical clustering) and their behavior is discussed. Future studies may serve to explore other possible definitions and the extents of their application.

F.1 Definitions and implementation

The Hausdorff and discrete Fréchet metrics are both sensitive to path outliers—even when the majority of points in two paths are spatially proximate, a single point that deviates substantially can generate large path distance measurements. Indeed, Hausdorff and discrete Fréchet may not be suitable for calculating an "overall" similarity along the entire lengths of two paths. There exist many possible definitions for path distance functions based on measures of central tendency, such as variations based on average or median distances; we consider two variations of the Hausdorff and discrete Fréchet distance functions based on averages rather than maxima.

Recall that a metric must satisfy the following properties:

$$\delta(A, B) \geq 0 \tag{F.1a}$$
$$\delta(A, B) = 0 \iff A = B \tag{F.1b}$$
$$\delta(A, B) = \delta(B, A) \tag{F.1c}$$
$$\delta(A, C) \leq \delta(A, B) + \delta(B, C). \tag{F.1d}$$

A *distance function* may be said to satisfy the first three properties, whereas the triangle inequality (Eq. F.1d) may be violated. Below, we provide explicit definitions for average-type Hausdorff and discrete Fréchet distance functions and show that they are *not* metrics using an example where the triangle inequality is not satisfied.

F.1.1 Average Hausdorff distance

We define the average Hausdorff distance between two paths to be the sum of the nearest neighbor distances for all points (on both paths) divided by the total number of points. We define the one-sided summed Hausdorff distance from path P to path Q as

$$\delta_h^{\text{sum}}(P \mid Q) = \sum_{p \in P} \min_{q \in Q} d(p, q), \tag{F.2}$$

so that the (symmetric) average Hausdorff distance is the total sum normalized by the total number of points:

$$\delta_H^{\text{avg}}(P, Q) = \frac{1}{|P| + |Q|} \left[\delta_h(P \mid Q) + \delta_h(Q \mid P)\right], \tag{F.3}$$

where $|P|$ and $|Q|$ are the cardinalities of (the number of points comprising) P and Q, respectively. A similar definition was examined by Eiter and Mannila [292], the only

difference being that they normalized by the number of paths ($\frac{1}{2}$) rather than the total number of points in the paths ($\frac{1}{|P|+|Q|}$). A sensible alternative is a weighted average Hausdorff distance,

$$\delta_H^{\text{wavg}}(P,Q) = \frac{1}{2}\left[\frac{1}{|P|}\delta_h(P \mid Q) + \frac{1}{|Q|}\delta_h(Q \mid P)\right], \tag{F.4}$$

which normalizes the contribution of each one-sided sum so that each path contributes equally irrespective of the number of constituent points. This prevents paths with relatively many points from "diluting" the overall average when compared to those with fewer points. Weighted average Hausdorff should thus minimize sensitivity to the number of points used to parameterize a path, whereas the average Hausdorff distance will tend to discount distance contributions from paths with relatively few points.

In Eq. F.3, the average is over $|P| + |Q|$ total points, while in Eq. F.4, separate averages over $|P|$ and $|Q|$ points are averaged together with equal weights. These definitions are consistent with the usage of "average" and "weighted average" for hierarchical clustering linkages (see S2 Text). For brevity, we focus on the weighted average Hausdorff distance (Eq. F.4) in this text.

F.1.2 Discrete average Fréchet distance

As with average-type Hausdorff distances, several definitions for a discrete average Fréchet distance function are possible. We use a definition identical to that employed by Dickson, H. Huang, and Post [486] where the coupling distance is defined as the average link length (in the coupling), which is defined in the subsequent discussion.

Following the description of the conventional discrete Fréchet distance in the main paper, we consider two polygonal curves P and Q, each with n and m ordered points (respectively), in a metric space (V, d) for some metric d. The sequence of line segments of P and Q are respectively defined as $\sigma(P) = (p_1, \ldots, p_n)$ and $\sigma(Q) = (q_1, \ldots, q_m)$. The coupling (in the product space $\sigma(Q,P) \equiv \sigma(P) \times \sigma(Q)$) between P and Q is

$$C(P,Q) \equiv (p_{a_1}, q_{b_1}), (p_{a_2}, q_{b_2}), \ldots, (p_{a_L}, q_{b_L}), \tag{F.5}$$

of L unique pairs of points (i.e., number of links) and satisfies the following conditions: (1) The first/last pairs correspond to the first/last points of the respective paths ($a_1 = b_1 = 1$, $a_L = n$ and $b_L = m$); (2) at least one point on either of the paths must be advanced to its successive point, i.e., ($a_{i+1} = a_i$ and $b_{i+1} = b_i + 1$) or ($a_{i+1} = a_i + 1$ and $b_{i+1} = b_i$) or ($a_{i+1} = a_i + 1$ and $b_{i+1} = b_i + 1$) for all $i = 1, \ldots, L$.

The definitions up to this point are identical to conventional Fréchet; however, the coupling distance, C, is now defined as an average distance over all pairs of points in a coupling:

$$\|C\| \equiv \frac{1}{L}\sum_{i=1}^{L} d(p_{a_i}, q_{b_i}). \tag{F.6}$$

We now consider, as usual, the set of all possible couplings between P and Q, $\Gamma_{P,Q}$, and take the *discrete average Fréchet distance* between P and Q to be the minimum coupling distance:

$$\delta_{dF}^{\text{avg}}(P,Q) = \min_{C \in \Gamma_{P,Q}} \|C\|. \tag{F.7}$$

Since the average is taken over the number of links in the coupling, the normalization factor for two paths depends not only on the numbers of points in the paths, but the optimal coupling between them. We can generate a coupling between P and Q with the possible fewest links by jumping simultaneously along both paths ($a_{i+1} = a_i + 1$ and $b_{i+1} = b_i + 1$) until an end point on one path is reached, then stepping along the rest of the points on the other path. Quantitatively, we have

$$L_{\min} = \min\{|P|, |Q|\} + |(|P| - |Q|)| = \max\{|P|, |Q|\}. \tag{F.8}$$

On the other hand, the maximum number of links is generated when a step is taken on only one path at a time, i.e., no simultaneous jumps occur ($a_{i+1} = a_i$, $b_{i+1} = b_i + 1$ or $a_{i+1} = a_i + 1$ and $b_{i+1} = b_i$):

$$L_{\max} = |P| + |Q| - 1. \tag{F.9}$$

In general, the normalization factor for discrete average Fréchet will be smaller than that for average Hausdorff, and greater than or equal to the number of points in the larger of two paths:

$$\max\{|P|, |Q|\} \leq L \leq |P| + |Q| - 1. \tag{F.10}$$

F.2 Are such average-type distance functions also metrics?

Motivated by Buchin's examination of summed and average Fréchet distances [487], we construct three polygonal paths whose mutual average Hausdorff and discrete average Fréchet distances violate the triangle inequality. We consider three paths P, Q, and R whose direction of traversal along each path is from left to right

F.2.1 Average Hausdorff violates the triangle inequality

In Fig. F.1, we consider the nearest neighbor distances for each pair of paths among P, Q, and R. As with the usual Hausdorff distance, the average Hausdorff distance, δ_H^{wavg}, will be invariant to the ordering of points. Following the illustration in Fig. F.1A, the distance between P and Q, $\delta_H^{\text{wavg}}(P, Q)$, is computed by considering the nearest neighbors in Q for all points in P and those in P for all points in Q.

Procedure:

1. For each point in P, locate its nearest neighbor (the nearest point) in Q and record the distance.
2. Compute the average nearest neighbor distance over all points in P, normalizing by the number of points in P (five).
3. Repeat the process for Q (same as in previous step by symmetry).
4. Average the two average nearest neighbor distances for P and Q to compute the (weighted) average Hausdorff distance.

Summing the distances explicitly, we have

$$\delta_H^{\text{wavg}}(P, Q) = \frac{1}{2}\left[\frac{1}{5}(2l + l + 2l + l + 0) + \frac{1}{5}(2l + l + 2l + l + 0)\right] = \frac{6}{5}l. \tag{F.11}$$

The distance between P and R is calculated analogously from Fig. F.1B:

$$\delta_H^{\text{wavg}}(P,R) = \frac{1}{2}\left[\frac{1}{5}(l+0+l+0+0) + \frac{1}{6}(0+0+\frac{l}{2}+\frac{l}{2}+0+l)\right] = \frac{11}{30}l. \tag{F.12}$$

By symmetry, we also have $\delta_H^{\text{wavg}}(P,R) = \delta_H^{\text{wavg}}(Q,R)$, so we have the relationship

$$\frac{11}{30}l + \frac{11}{30}l < \frac{6}{5}l \tag{F.13}$$

$$\delta_H^{\text{wavg}}(P,R) + \delta_H^{\text{wavg}}(Q,R) < \delta_H^{\text{wavg}}(P,Q), \tag{F.14}$$

which does not satisfy the triangle inequality. It can be easily shown that the (unweighted) average Hausdorff distance (Eq. F.3) also violates the triangle inequality for the paths in Fig. F.1.

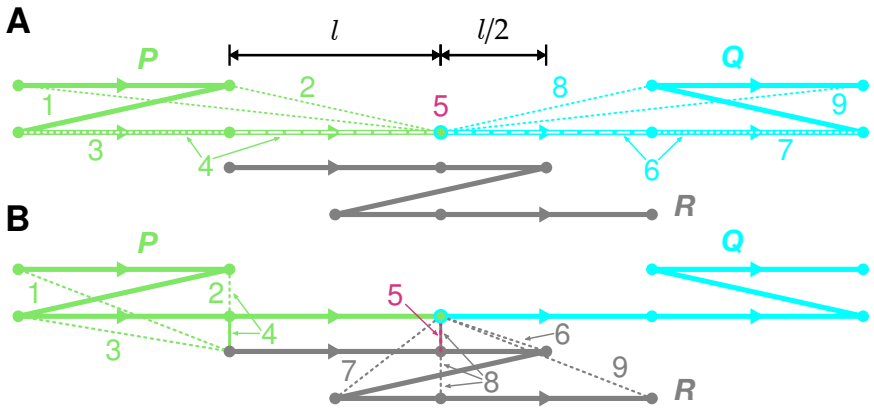

Figure F.1: Computing the average Hausdorff distance between three discretized paths P, Q, and R. The vertical direction is expanded for the purpose of illustration and does not represent actual separation; each curve is imagined to lie on the same horizontal axis. Numbered points have a corresponding dashed line representing their nearest neighbor (on the other path). The average Hausdorff distance is computed by averaging the horizontal lengths of dashed lines of a given color and then averaging the two averages for a given pair of paths. (A) Nearest neighbors depicted for P (green) and Q (cyan). (B) Nearest neighbors depicted for P and R (gray).

F.2.2 Discrete average Fréchet violates the triangle inequality

The schematic in Fig. F.2 explicitly shows the links comprising the optimal couplings between path pairs among P, Q, and R. As with the usual Fréchet distance, the discrete average Fréchet distance, δ_F^{avg}, is sensitive to the ordering of points; arrows on the paths indicate directionality. The illustration in Fig. F.2A depicts the sequence of links (i.e., coupling) with the minimal average link length; the distance between P and Q, $\delta_F^{\text{avg}}(P,Q)$, is computed from the depicted coupling.

Procedure:

1. Begin at the starting (leftmost) points in P and Q connected by the first link.
2. Step along P only (staying at the initial point on Q) until the fifth link (shown in magenta, which has zero length) is reached.

3. Movement along P is completed; step along Q to its last point.
4. Compute the average link length—sum the lengths of all links (in the coupling) and divide by the total number of links—to compute the discrete average Fréchet distance.

Averaging the link lengths explicitly, we have

$$\delta_F^{\text{avg}}(P,Q) = \frac{1}{9}\left(2l + l + 2l + l + 0 + l + 2l + l + 2l\right) = \frac{12}{9}l. \tag{F.15}$$

The distance between P and R is calculated from the coupling shown in Fig. F.2B; we remark that the second, fourth, fifth, and eighth (vertical) links each have zero length since P, Q, and R all lie on the same horizontal line:

$$\delta_F^{\text{avg}}(P,R) = \frac{1}{9}\left(l + 0 + l + 0 + 0 + \frac{l}{2} + \frac{l}{2} + 0 + l\right) = \frac{4}{9}l. \tag{F.16}$$

Again by symmetry, $\delta_F^{\text{avg}}(P,R) = \delta_F^{\text{avg}}(Q,R)$, so we have the relationship

$$\frac{4}{9}l + \frac{4}{9}l < \frac{12}{9}l \tag{F.17}$$

$$\delta_F^{\text{avg}}(P,R) + \delta_F^{\text{avg}}(Q,R) < \delta_F^{\text{avg}}(P,Q), \tag{F.18}$$

which violates the triangle inequality.

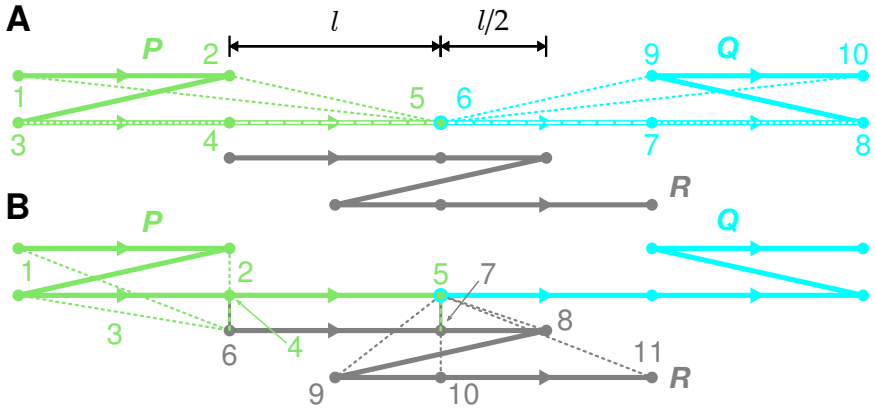

Figure F.2: Computing the discrete average Fréchet distance between three discretized paths P, Q, and R. The vertical direction is expanded for the purpose of illustration and does not represent actual separation; each curve is imagined to lie on the same horizontal axis. The (optimal) couplings producing the minimal average link length (i.e., the discrete average Fréchet distance) are shown for P and Q (A), and P and R (B). Links are represented by dashed lines, colored according to the path along which a step is taken, and numbered sequentially for the given coupling; at step 5 in panel (A) and (B), the link length is zero, progress along P is completed, and the remaining movement is solely along the other path.

F.2.3 Summary and additional considerations

Our analyses demonstrate that the average-type Hausdorff distance functions (Eq. F.3 and Eq. F.4), and the discrete average Fréchet distance (eqs. (F.5) to (F.7)), do not generally

satisfy the triangle inequality and are therefore *not* proper path metrics. We note that there may be some problems where the *relaxed* triangle inequality,

$$\delta(A, C) \leq \kappa \left[\delta(A, B) + \delta(B, C)\right], \tag{F.19}$$

which scales the upper bound on $\delta(A, C)$, normally set by the full triangle inequality, by a finite constant κ, may be sufficient [487]. However, by modifying the paths in Fig. F.1 and Fig. F.2 (by continually increasing the number of "zig-zags" in P, Q, and R), it can be shown that κ becomes arbitrarily large (see Ch. 6 in [487])—the relaxed triangle inequality is also violated by the path distance functions.

While neither PSA nor hierarchical clustering require the use of true metrics, the triangle inequality is a useful property in that it is an intuitive extension of the transitive property. That is, when two objects, A and B, in some metric space are close to a third object, C, in the same space, then A can be considered close to B in the sense that their maximal separation is bounded from above by the triangle inequality: $d(A, B) \leq d(A, C) + d(B, C)$. In order to preserve commonsense intuition about the pairwise relationships between paths, the main text discusses PSA exclusively in the context of metrics.

F.3 Methods comparison using average-type distance functions

Although it was shown above that average Hausdorff and discrete average Fréchet are not metrics, they may still prove to be useful distance functions. We generated heat map dendrograms for the path-sampling methods comparison (using the definitions in Eq. F.4 and Eq. F.7) to get a feel for their behavior under familiar circumstances.

F.3.1 Weighted average Hausdorff

Average Hausdorff distances substantially smaller than conventional Hausdorff (Fig. F.3A), with the largest average Hausdorff distance (2.84 Å) being about 1.5 Å smaller than the largest Hausdorff distance (4.67 Å). In terms of clustering, the primary differences were that both GOdMD and ANMP clustered with the rest of the dynamical methods. The MENM methods formed their own cluster, which was in turn grouped with iENM. DIMS, rTMD-S, MDdMD, and FRODA clustered very similarly to Hausdorff; this was also the case with the rTMD-F, Morph, MAP and LinInt cluster, although Morph groups with MAP instead of rTMD-F in the average Hausdorff heat map. In Fig. F.3B, points fall noticeably below the diagonal (of unity slope), indicating that the magnitudes of average Hausdorff distances are both bounded from above by the conventional Hausdorff distance. The Pearson correlation was also weaker (0.868 versus 0.977) and the average Hausdorff distance distribution, though qualitatively similar to that of conventional Hausdorff, is skewed toward smaller values.

F.3.2 Discrete average Fréchet

On the other hand, discrete average Fréchet generates a different clustering (Fig. F.4A) than the other distance functions and, on the whole, exhibits much smaller distances. GOdMD clustered with the rest of the dynamical methods (with the exception of TMD-F),

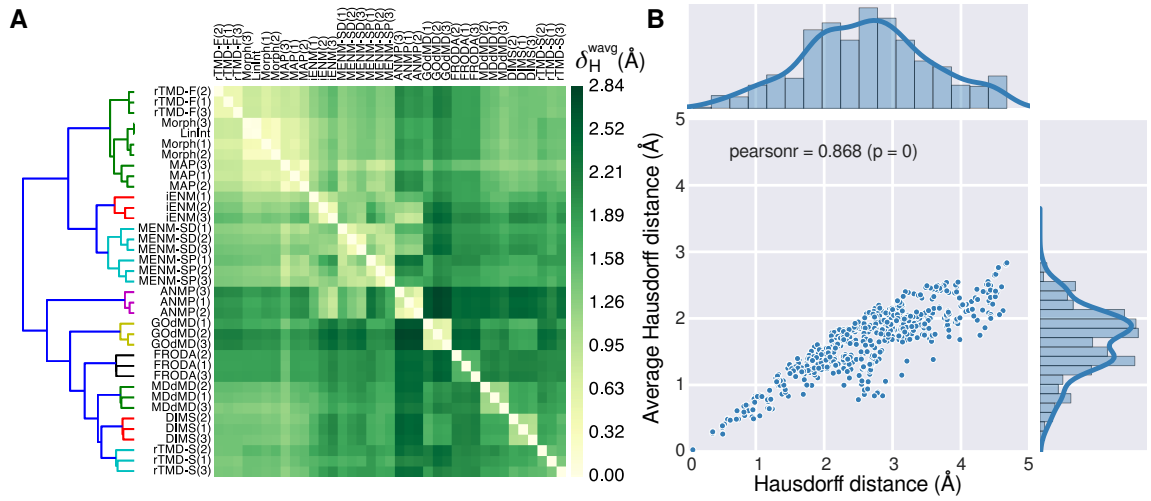

Figure F.3: (A) Path-sampling methods comparison for AdK closed → open transition of (weighted) average Hausdorff distances using Ward linkage. (B) Correlation and joint distributions between Hausdorff and (weighted) average Hausdorff distances (in Å RMSD) for the AdK closed → open methods comparison. Reasonably strong linear correlation indicated by the scatter plot, with a Pearson correlation coefficient close to unity.

while all of the elastic network models were grouped along with the cluster of TMD-F, Morph, and LinInt; this is the reverse of what was produced by both Hausdorff distance functions and conventional Fréchet, where GOdMD ended up with the ENMs and the TMD-F, Morph, and LinInt cluster was grouped with the dynamical methods. The majority of points in Fig. F.4B fall substantially below the diagonal and are more scattered compared with Fig. F.3B; the marginal distribution of discrete average Fréchet closely resembles the distribution produced by average Hausdorff, though the Pearson correlation was slightly stronger (0.892 versus 0.868).

F.3.3 Discussion

From a qualitative standpoint, Fig. F.3A and Fig. F.4A are satisfactory in distinguishing between the most obvious patterns among the path-sampling methods. In particular, both clusterings generate one group containing DIMS, FRODA, MDdMD, and TMD-S, another group with MENM-SD/SP, iENM, and ANMP, and a third group with MAP, Morph, TMD-F, and LinInt. The clustering within each of these groups are also quite similar, with the second and third MENM and iENM paths being closer than the first paths. Interestingly, ANMP clusters with the methods based on dynamical algorithms when using average Hausdorff, whereas average Fréchet places it with the other ENM-based approaches. Meanwhile, the average Hausdorff and discrete average Fréchet distances from the GOdMD paths and MAP(3), to paths from other methods, are substantially smaller than those generated by conventional Hausdorff and Fréchet; the average Hausdorff distances of MAP(3) also correspond to a relatively light band, although it is less pronounced than in the average Fréchet heat map.

In conclusion, the path distance functions based on simple averages (rather than max-

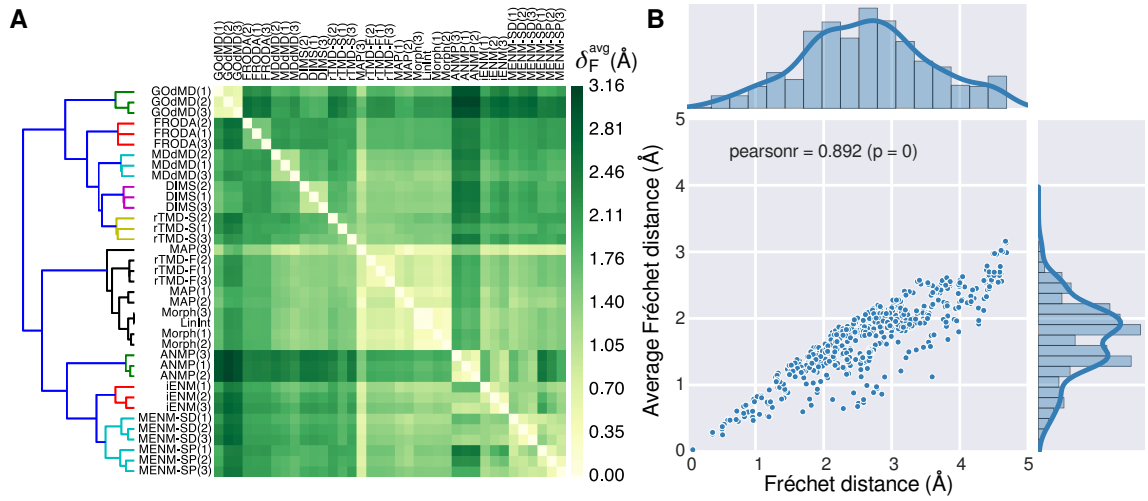

Figure F.4: (A) Path-sampling methods comparison for AdK closed → open transition of discrete average Fréchet distances using Ward linkage. (B) Correlation and joint distributions between Fréchet and (weighted) average Fréchet distances (in Å RMSD) for the AdK closed → open methods comparison. Noticeably weaker linear correlation is indicated by the Pearson correlation and the scatter points show that discrete average Fréchet produces distances substantially smaller than conventional discrete Fréchet.

ima) considered in this supplement are not true metrics as they violate the triangle inequality. We acknowledge that there may be other path metrics based on measures of central tendency, or at least more robust against outlier points, that would be worth exploring in the future. Furthermore, we did not examine whether the triangle inequality was (or would likely to be) satisfied for a typical path comparison where one may be dealing with a restricted class of curves. Situations may arise where average-type path distance functions behave as metrics for a specific problem. PSA is also not limited to the use of proper path (and point) metrics and it was seen in Fig. F.3A and Fig. F.4A that our average-type Hausdorff and Fréchet distance functions both generated qualitatively acceptable distance measurements. In light of these results, our main study focuses exclusively on the conventional Hausdorff and (discrete) Fréchet metrics because they are satisfactory measures of path similarity that also respect common intuitions about notions of closeness or dissimilarity.

APPENDIX G

STOCHASTIC ELEMENTS IN HERMESHD

Accounting for microscopic thermal fluctuations in numerical simulations of fluids necessitates the inclusion of stochastic forcing terms when modeling microscale or nanoscale flows. The Landau-Lifschitz Navier-Stokes equations, a well known extension to the Navier-Stokes equations, incorporate stochastic flux terms in the form of a stochastic stress tensor (in the momentum and energy equations) and a stochastic heat flux vector (in the energy equation). In this appendix, a brief overview of the stochastic stress tensor provides an intuitive feel for how hydrodynamic fluctuations are incorporated in HERMESHD, the FHD simulation code developed by the author for the Blue Waters Graduate Fellowship project.

G.1 Hydrodynamic fluctuations

The study of fluctuations in microscale and nanoscale fluids is important for describing such processes as droplet breakup in liquid nanojets [488, 489], molecular motors [490, 491] and diffusion enhancement in cells [471, 492], hydrodynamic instabilities [493–495], and many other nanoscale flows [405, 406]. The macroscopic Navier-Stokes equations describe ordinary linear hydrodynamics by supplementing the conservation equations for mass, momentum, and energy with empirical constitutive relations connecting the gradients of the conserved quantities with linear responses in their corresponding fluxes; the linear transport coefficients are the proportionality constants defined by these relations. The macroscopic hydrodynamic equations provide a surprisingly effective model of fluid systems down to about ten molecular diameters and only an order of magnitude above the kinetic time scale of molecular collisions (about 30 collision times) [388]. Although the macroscopic, linear Navier-Stokes equations model quite well average flows down to, perhaps surprisingly, the nanoscale, the modeling of a dense fluid (i.e. liquids) at truly molecular distance scales requires properly accounting for the fluctuations that naturally arise when taking averages over small numbers of particles.

G.2 Stochastic shear stresses

To give a general sense of how hydrodynamic fluctuations enter the picture, it suffices to illustrate the case of thermal fluctuations in the fluid stress (i.e., the fluid momentum density).* These fluctuations take the form of spatiotemporally delta-correlated stochastic stresses with zero mean, which are elements of a random stress tensor, $\mathcal{S}$. For small-amplitude fluctuations, one can use linear fluctuation-dissipation theory to derive the correlations of the components of the random stress tensor [366, 407]. Including the bulk viscosity, the covariances are given by

$$\left\langle \mathcal{S}_{ij}(\mathbf{x},t)\mathcal{S}_{kl}(\mathbf{x}',t')\right\rangle = 2k_{\mathrm{B}}T\left[\eta\left(\delta_{il}\delta_{jk}+\delta_{ik}\delta_{jl}\right)+\left(\zeta-\frac{2}{3}\eta\right)\delta_{ij}\delta_{kl}\right]\delta(\mathbf{x}-\mathbf{x}')\delta(t-t'), \quad \text{(G.1)}$$

where ζ is the bulk viscosity, η is the dynamic shear viscosity, and the indices i, j, k, and l run independently over the three spatial directions. It is clear from the delta functions,

*For a complete discussion of both the stochastic stresses and heat fluxes, see Landau and Lifschitz [366]; a succinct description of the 1D LLNS equations is also provided by J. B. Bell, Garcia, and Williams [405].

$\delta(\mathbf{x}-\mathbf{x}')$ and $\delta(t-t')$, that $\mathcal{S}$ is spatiotemporally delta-correlated—in a numerical FHD simulation, $\mathcal{S}$ should be independently sampled at every timestep in each grid cell.

The structure of $\mathcal{S}$ can be seen more clearly by extracting from Eq. G.1 explicit representations of the correlations between various combinations of the random stress tensor components (assuming a given time, t, and point in space, $\mathbf{x}$). In particular, each individual matrix element, $\mathcal{S}_{ij}$, may be written in terms of a zero-mean (stationary) Gaussian process, which conceptually simplifies numerical sampling. First, the off-diagonal components in, say, the upper triangle of the random stress matrix, are mutually uncorrelated, so

$$\langle \mathcal{S}_{12}\mathcal{S}_{13}\rangle = \langle \mathcal{S}_{13}\mathcal{S}_{23}\rangle = \langle \mathcal{S}_{23}\mathcal{S}_{12}\rangle = 0. \tag{G.2}$$

The same conditions hold for the lower triangle. Second, the off-diagonal components are correlated with themselves and their symmetric partners,

$$\left\langle \mathcal{S}_{ij}\mathcal{S}_{ij}\right\rangle = \left\langle \mathcal{S}_{ij}\mathcal{S}_{ji}\right\rangle = 2\eta k_{\mathrm{B}}T, \tag{G.3}$$

for $i \neq j$. Furthermore, given that the diagonal components are uncorrelated with respect to the off-diagonal stresses, a total of three stochastic samples are needed to completely capture off-diagonal contributions. Finally, for the deviatoric stresses (neglecting the diagonal contribution from the bulk viscosity), it can be seen that there are non-vanishing correlations among the diagonal components,

$$\left\langle \mathcal{S}_{ii}\mathcal{S}_{jj}\right\rangle = \begin{cases} +\frac{8}{3}\eta k_{\mathrm{B}}T & i=j, \\ -\frac{4}{3}\eta k_{\mathrm{B}}T & i\neq j, \end{cases} \tag{G.4}$$

which, together, effectively contribute two independent stochastic terms. Combined with the three stochastic terms needed for the off-diagonal elements, a minimum of five independent stochastic samples would be needed, in principle, to generate the random stress components at each point in space and time; however, to satisfy the traceless condition, six samples are taken per timestep per grid cell as shown in the subsequent section.

G.3 Numerical sampling of stochastic terms

In order to sample the random stresses at each point in space at a given timestep, six Gaussian samples are taken in HERMESHD. Say we have a delta-correlated stationary Gaussian process with zero mean so that

$$\langle R(t)\rangle = 0 \tag{G.5}$$

$$\left\langle R(t)R(t')\right\rangle = \delta(t-t'). \tag{G.6}$$

We seek a form for the random stress tensor such that its components, $\mathcal{S}_{ij}$, may be written in terms of $R(t)$; such a construction would allow one to sample $\mathcal{S}$ in the numerical code by only taking several samples of a zero-mean Gaussian distribution, $R(t)$.

Let us first generalize $R(t)$ so that we have a 3×3 matrix, $R_{ij}(t)$, of nine independent Gaussian-distributed random variables, where $R_{ij}(t) = R(t)$ is an independent and identically distributed delta-correlated Gaussian process with zero mean specified by Eq. G.5.

For a given timestep, Δt, and (Cartesian) spatial discretization, $\Delta V = \Delta x \Delta y \Delta z$, the random stress components may then be written as

$$\mathcal{S}_{ij}(t) = \sqrt{\frac{2k_\mathrm{B}T}{\Delta t \Delta V}} \left(\sqrt{2\eta}\, G_{ij}(t) + \sqrt{\frac{\zeta}{D}} R_{kk}(t) \right), \tag{G.7}$$

where $R_{kk}(t)$ is the trace of $G_{ij}(t)$ is a traceless symmetric matrix constructed from $R_{ij}(t)$,

$$G_{ij}(t) = \frac{1}{2} \left(R_{ij}(t) + R_{ji}(t) \right) - \frac{1}{D} \delta_{ij} R_{kk}(t), \tag{G.8}$$

and D is the number of spatial dimensions [496–499]. Note that the factor of $\sqrt{2}$ in Eq. G.7 compensates for the reduction in variance due to the averaging in Eq. G.8 of $R_{ij}(t)$ and $R_{ji}(t)$ [498]. It can be easily verified that, at each time t, $G_{ij}(t)$ is identically traceless and symmetric:

$$G_{ij}(t) = G_{ji}(t) \tag{G.9}$$

$$G_{ii}(t) = 0. \tag{G.10}$$

Although Eq. G.7 may be used to construct the stochastic stress tensor with the correct correlations (Eq. G.1) using Eq. G.8, nine pseudorandom numbers must be generated. To reduce the number of generated pseudorandom numbers per grid cell per timestep to six, it is possible to generate only three pseudorandom numbers for the off-diagonal components in the upper (lower) triangle and exploit the symmetry of $G_{ij}(t)$ to get the corresponding components in the lower (upper) triangle; then, without the averaging, the factor of $\sqrt{2}$ can be dropped and the diagonal components—including the contribution from bulk viscosity—can be generated separately with three additional pseudorandom numbers.