THE UNIVERSITY OF CHICAGO

ENHANCED SAMPLING AND DYNAMICAL ANALYSES OF MULTIPATHWAY REACTIONS: APPLICATIONS TO INSULIN AND KAIB

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF CHEMISTRY

BY
ADAM ANTOSZEWSKI

CHICAGO, ILLINOIS
AUGUST 2022

Dedicated to JoAnn and Paul Antoszewski

"If we were to name the most powerful assumption of all, which leads one on and on in an attempt to understand life, it is that all things are made of atoms, and that everything that living things do can be understood in terms of the jigglings and wigglings of atoms."

-Richard Feynman

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

How does one start an acknowledgment for a PhD? It's not like it was a single thing, a single achievement that a few people were partially responsible for. It was just under one fifth of my life to date, and over half of my "adult life." The group of people to thank is too huge, and the list of reasons why too long, to do anyone justice in a few paragraphs. What follows is a shallow attempt at doing exactly that.

First, I would have quit long ago if it wasn't for my family. My parents, JoAnn and Paul Antoszewski, have always been there to listen. From the first year in college, existential panic, *what-in-the-world-do-people-even-do-with-their-lives* call, to the fifth year in grad school, existential panic, *what-in-the-world-do-people-even-do-with-their-lives* call, they've seen or heard it all. But they were also there for everything in between - for the simple joy of describing my day on the walk home, for the inane prattling about about how I've gotten really into F1 during the pandemic, or for the cooking advice when I couldn't remember how long to boil sausage. They were there when I came out as gay during my first year at UChicago (although I think they knew long before), and they were there when I wanted to drop my PhD altogether and move back home. For being there during all of this and more, I am grateful beyond measure. I love you.

I'm also deeply thankful for my siblings, Graham and Trinity Antoszewski. No matter the time, or how busy any of us were, I always knew I had you two to turn to. Also, I know I must be simply the *worst* texter that exists, so thanks for not strangling me when I didn't respond for weeks at a time. You are both, in different ways, everything I aspire to be. I love you, and I'm so, so excited for what comes next for both of you.

Of course, I owe a great debt to my advisor. How to describe Aaron? Well, to borrow an Aaron-ism, he is "not the worst advisor in the entire world." Advisor-advisee is not quite the relationship that we have, but there's not one snappy word that summarizes what he has meant to me during my PhD. Aaron for some reason decided to let me join his group, despite

me not knowing anything about molecules, biology, simulation techniques, or really anything that would be of practical use in my research. Truly, for someone with an appointment in the Institute for Biophysical Dynamics, accepting someone who literally had never taken a real biology or biochemistry class was a bold move. I hesitate to put words in his mouth, but he perhaps had his first moment of regret when, immediately after joining his group, we had a two-hour meeting where I steadfastly refused to grasp what we could possibly use molecular dynamics for. What it all really meant. In some sense, that has been the story of my PhD. Me, struggling to understand what any of this data means or why we might care, and Aaron, constantly trying to poke and prod me in the right direction, but never forcing me into any conclusion. For me, Aaron has been the ideal advisor (far beyond what I reasonably could have expected), both scientifically and personally. Surely, I wasn't the easiest person to work with over my time in graduate school. To Aaron's eternal chagrin, my research was almost always paired with a revolving door of departmental commitments - PSD Social Committee, Tigger Talks, DoGSI, and the ever-dreaded Recruitment Committee. It is an immense credit to him that he didn't kick me out of the group the exact moment he found out that I'd be helping with recruitment for a third year in a row. But Aaron respected that these roles were important to me. He also was more flexible than any advisor should have to be, when I had to take a leave of absence during my first summer in grad school, or when I decided to abandon him in the middle of writing a paper for some internship. Despite all of this, Aaron has always remained in my corner, and I will forever be grateful.

And where would I be without the denizens of GCIS E145C? Or, for that matter, the attendees at Sampling Subgroup? I could imagine no better group of people to waste time bullshitting with on a Friday morning. When I first joined, Erik Thiede and Lu Hong were the elder statesmen who were responsible for teaching me, well, everything (but primarily math and molecular dynamics). Erik is perhaps the most patient teacher I've ever had. Patient would not be a word to describe Lu, but he was an amazing mentor nonetheless.

John Strahan and Chatipat Lorpaiboon have both been invaluable resources when it comes to TPT, DGA, or anything else related to math. John also has taught me a lot about his favorite place, New York City, and his favorite topic, machine learning. Chatipat, amongst his seemingly-endless projects, consistently finds time to be the Dinner Group's "fixer," always volunteering to help when things break. A huge thank you to you both. Spencer Guo has also been an invaluable resource in the last two years, as one of the few remaining lab members who really cares about proteins. He is also a good friend, even if he did take the title of Dinner Group's Best Dressed away from me.

I'm also thankful for the Fun Times Party Peeps Happy Squad, comprised of Cat Triandafillou, Bodhi Vani, and Elizabeth White, for being amazing collaborators and even better friends. Cat taught me how to truly be excited about science, and Elizabeth taught me about the limits of human circadian rhythms. And what did Bo not teach me? Of course, she gave me The Insulin Project, which turned into Chapter 2 of this thesis. But more than that, she was an electric presence in the office. She was the ideal stretching partner, the perfect person to "rubber duck" to, a great cook, and above all else, a real human connection. Bo heard more about my personal life than anyone else in the office, and for that I am thankful. I would be remiss if I didn't mention the two people who joined me in weekly Insulin Parties for well over two years, Luis Busto de Moner and Chi-Jui Feng. Both were wonderful collaborators, and were relentlessly helpful. There was never one request I had, or one question I asked, that they did not immediately tackle head-on. Thanks to you both, even if Luis did insist on scheduling these "parties" at 8 AM on Mondays.

A huge thanks also goes to my co-advisor, Jon Weare, whose wit and general presence at Sampling Subgroup has always been a joy. His willingness to provide excellent mathematical advice helps me to forgive him for briefly taking my name off of his website. I'm also very appreciative for my incredible colleagues at Merck during my internship (namely Sebastian Schneider, Melissa Ford, Hannah Bruce Macdonald, and Michael Altman) for being a huge

help during my job search process. I owe an immense debt of gratitude to Vera Dragisich, Laura Luburich, Melinda Moore, and Brenda Thomas for patiently putting up with me and my departmental shenanigans for 5 years. Speaking of departmental shenanigans, I also deeply thank Hannah Yi, Josh Portner, Jake Higgins, and Mark Levin for all of the help with the Recruitment Behemoth. This section is already getting prohibitively long, so a general thank you goes to Steven Redford, Chris Chi, ChuHui Fu, Tegan Marianchuk, Cal Floyd, Yuqing Qiu, Noah Gamble, Sammy Allaw, Rob Webber, Justin Finkle, Irena Hsu, Sam Greene, Huan Zhang, Charlotte Cheng, and Sherry Li for making the Dinner and Weare Groups what they are. A particular thank you is due to Tegan, whose work on the KaiB project eventually turned into Chapter 4 of this thesis. Along with Tegan, I thank Tobin Sosnick, Xiangda Peng, Andy LiWang, Nanhao Chen, Ning Zhang, Supratim Dey, Lee-Ping Wang, and Damini Sood for data and helpful conversations regarding KaiB.

I thank Biman Bagchi and Michael Weiss for helpful discussions, along with Albert Pan, Bryan Jackson, and coworkers at D.E. Shaw for providing their trajectories for comparison in Chapter 2. This work was supported by National Institutes of Health awards R01GM109455-01, 5R01GM118774-02, and R35 GM136381. Computations were performed on resources provided by the University of Chicago Research Computing Center, the GM4 cluster supported by the NSF Major Research Instrumentation award DMR-1828629, the Beagle3 cluster supported by the NIH Shared Instrumentation Grant 1S10OD028655-01, and the Extreme Science and Engineering Discovery Environment [6] (NSF Grant ACI-1548562) Bridges (PSC) computing nodes through allocation TG-MCB180007.

The last batch of thanks, perhaps among some of the most important, go to the many friends I've made in the last 5 years. To be brief, you all are what have made grad school such a pleasure. I can't possibly list everyone here, so I will keep the list short. To Ben Soloway, Nathaniel Durfee, Adam Weiss, Kendall Bryant, Kate Henn, and Rudy G., you all somehow made every outing feel a little bit special. To Matt Zajac and Phil Gemmel, I didn't expect

to find such amazing friends so late in grad school, but here we are. I wouldn't want to have been DoGSIs with anyone else. To my former roommates Isaac and Ella Hirsch-Hock, living with you for four(!) years was one of the highlights of my time in grad school. From the infamous Wolf Party to the two years of quarantine to celebrating all of our defenses, we've been through it all. To my current roommate Sarah Willson and my honorary roommate Brad Studnitzer, it's impossible to put into words what you both mean to me. I won't try. I love you both.

ABSTRACT

Simulation of biological processes is an important complement to traditional *in vitro* and *in vivo* experimentation, but such simulation often requires enormous computational effort. This is because many biological processes are so-called "rare events", meaning that their timescales are long compared to the timescale of molecular simulation (typically on the order of μs). Indeed, using traditional molecular dynamics, the simulation of statistically meaningful numbers of biological rare events is generally completely intractable. Thus, much work has been done to bridge the separation of computational and experimental timescales. Two broad (and non-exhaustive) approaches are (1) enhanced sampling, where the simulations themselves are biased, accelerated, or manipulated to encourage rare events to occur, and (2) dynamical analysis, where existing simulations are statistically recombined to gain information about long-time dynamical statistics from relatively short-time simulations. This thesis aims to apply these approaches to study two proteins: insulin, the hormone that regulates cellular glucose uptake, and KaiB, an element of the core oscillator of the cyanobacterial circadian clock. For all investigated systems, we found an ensemble of pathways that were either energetically or dynamically accessible, highlighting the complexity of biological processes and the folly of assuming *a priori* that a process of interest follows a single mechanism.

The first part of this thesis focuses on the enhanced sampling approach, using a technique named Replica Exchange Umbrella Sampling to characterize the energetic and structural features of the insulin dimer dissociation. The dimer dissociation is an essential prerequisite for cellular insulin binding and ultimate biological function. We discovered a large set of dissociation pathways with comparable free energy barriers, ranging from extremes of conformational selection to induced fit. These results reconciled previous experimental and simulation results, seemingly in disagreement, as part of a broader ensemble of coupled (un)folding and (un)binding. We for the first time explicitly characterized monomeric unfolding during the dissociation, which involved the detachment of insulin's B chain C-terminus

from the hydrophobic core of the monomer. We also identified key interfacial rotations, and showed using computational infrared spectroscopy that limiting pathways could be experimentally distinguished based on differences in these rotations and associated backbone amide solvations.

The second part of this thesis focuses on the dynamical analysis approach, using a recently-developed method called the Dynamical Galerkin Approximation to explore the dynamics of phenol release from the insulin hexamer. By investigating the mechanisms of phenol release and how they might be altered by targeted insulin mutations, we aimed to understand how one might design new diabetes therapeutics that either encourage or discourage phenol dissociation. We identified and quantitatively characterized six phenol binding/unbinding pathways for wildtype, Ile A10 Val, and Glu B13 Gln mutant insulins. A number of these pathways involved large-scale opening of the primary escape channel, suggesting that the hexamer is much more dynamic than previously appreciated. We show that phenol unbinding is a multipathway process, with no single pathway representing more than 50% of the reactive current and all pathways representing at least 10%. We also showed how the contributions of specific pathways can be manipulated by mutating residues both in and out of the phenolic binding pocket.

The final part of the thesis combines dynamical analysis with near-atomic molecular dynamics simulations to investigate mechanisms of fold switching for the cyanobacterial circadian protein KaiB. KaiB can reversibly switch between two stable folds, the so-called "ground state" (gs) and "fold switched state" (fs). We used a combination of near-atomic simulation and dynamical analysis to explore the mechanisms of this fold switch. We compared computational predictions of hydrogen-deuterium exchange to experiments to validate simulation parameters, and added proline isomerization into the model. We further discovered that three prolines (P63, P70, and P71) preferentially occupy the *cis* state in fs KaiB and the *trans* state in gs KaiB. The two folds have nearly identical free energies. The mech-

anisms of fold switching are quite complex. Secondary structure elements in the C-terminal fold switching domain can fold and refold in almost any order, with very little unfolding observed in the otherwise-stable N-terminal domain. The primary free energy barrier is correlated with the breaking of β sheets in the fold switched state. The isomerization of P63, P70, and P71 largely occurs before this barrier. Overall, the secondary structure elements in the C-terminal fold switching domain act as foldons, tending to fold and unfold as units independent of one another. This work is the first statistical treatment of the mechanisms of fold switching of a metamorphic protein and underscores the inherently multipathway nature of fold switching.

CHAPTER 1

INTRODUCTION

The fields of chemistry and chemical biology are broadly concerned with chemical reactions and/or the interactions between biological components. In many instances, however, important interactions occur in ways that traditional experiments cannot directly probe. For example, it is common in drug discovery to crystallize a drug target, often a cellular receptor, both in its apo form and when bound to a series of potential drugs. These crystal structures only provide static snapshots of proteins, and may miss conformational dynamics essential to designing effective drugs. For example, the binding pocket of mollusk acetylcholine binding protein (AChBP) is compact and relatively dry when bound to small molecule ligands. However, when bound to larger ligands, this pocket is much wider and much more exposed to solvent [7]. To better understand how AChBP can dramatically change its structure to accommodate several different classes of ligands, we need a more detailed understanding of the molecular motions accessible to AChBP. Specifically, we would want residue-level resolution of side chain motions in and near the binding pocket. This would allow us to, for example, design a drug that binds to pockets in AChBP that only open transiently, and are not found in X-ray crystal structures. Molecular simulation is one way to investigate the motions of these regions. All-atom molecular dynamics can take as an input X-ray crystallographic structures with atomic-level precision, and it can propagate these coordinates forward in time by integrating Newton's equations of motion. Doing this allows scientists to "see" these structures move in time, providing invaluable insight to complement experiment.

While simulation and experiment can certainly be a powerful tool when combined, one of the most persistent challenges in current biophysical computation is the so-called "separation of timescales;" that is, the timescale of individual simulation steps, often on the order of femtoseconds, is far removed from the timescale of the process itself, which can range from milliseconds to even days. Even for the most specialized and powerful supercomputers for

molecular dynamics (i.e., Anton 3), a 512-node simulation of a million atoms (a reasonable size, for example, for an all-atom simulation involving the cellular insulin receptor surrounded by solvent) yields approximately 100 μs of simulation time per day [8]. While certainly an achievement in terms of computational power, this benchmark is put into context when one considers that the proposed timescale for insulin unbinding from the insulin receptor is on the order of hundreds of seconds [9]. Thus, if one wanted to investigate insulin unbinding from its receptor using traditional molecular dynamics, one would on average need millions of days worth of simulation time to simulate just one unbinding event. Compounding this problem, drawing conclusions from simulations requires a statistically meaningful number of transitions. Clearly, long equilibrium simulations of biomolecules, while a powerful tool for many smaller systems [10], can be of limited use for larger, more complex systems.

To gain meaningful insight from computation and to compare to experiment, much work has been done to bridge this separation of timescales. While a complete review on this topic is beyond the scope of this work, two broad (and non-exhaustive) approaches are (1) enhanced sampling, where the simulations themselves are biased, accelerated, or manipulated to encourage rare events to occur, and (2) dynamical analysis, where existing simulations are statistically recombined to gain information about long-time dynamical statistics from relatively short-time simulations. The former category is quite mature, with biased techniques like umbrella sampling [11, 12] and metadynamics [13] being common for decades, and splitting-based techniques like weighted ensemble [14] and forward flux sampling [15] also seeing widespread use. Splitting-based approaches, in general, systematically clone and prune a large number of parallel simulations, encouraging simulations to evolve to a target state and keeping track of their individual trajectory weights. In comparison, biased approaches like umbrella sampling and metadynamics systematically apply biases (often harmonic or Gaussian) to encourage the system to explore high-free-energy areas of collective variable (CV) space, and then account for these biases when calculating averages or building

a potential of mean force. Provided that these CVs are chosen so as to capture the essential motions of the system of interest, this free energy surface can be used to deduce energetically preferred pathways and structural features of these putative pathways. Chapter 2 of this thesis uses a version of umbrella sampling to understand the dissociation of the insulin dimer.

Dynamical analysis is arguably less mature, only seeing widespread use relatively recently. Markov State Models (MSMs) extract long-time dynamical statistics by constructing and analyzing a transition rate matrix, modeling kinetics as a memoryless jump process between states [16–18]. MSMs and their biological applications have been the subject of multiple recent reviews [19–21]. MSMs and their variants have been used to characterize the binding of barnase and barstar [22], to describe the coupled folding and binding of the inhibitor peptide PMI to an oncoprotein [23], and to discover an array of cryptic binding pockets across the SARS-CoV-2 proteome [24]. The last work was enabled by Folding@Home, a massively distributed supercomputing project for molecular dynamics that enabled the simulation of 0.1 s of data for various SARS-CoV-2 systems. Using these data to build MSMs allowed for the discovery of a wide variety of potentially druggable hidden binding pockets, the opening of which occurs on timescales much longer than 0.1 s. Clearly, MSMs and related techniques show great promise for bridging the separation of timescales in biophysical simulation.

Dynamical Galerkin Approximation (DGA [25]) is a generalization of Markov State Modeling. Instead of solving for dynamical variables by building and manipulating a coarse-grained kinetic model, DGA solves for these dynamical variables directly by casting them as solutions to operator equations involving the stopped transition operator. These solutions are determined through a basis expansion. One clear advantage of DGA is that by solving equations of the stopped transition operator (instead of just the transition operator used in MSMs), DGA correctly applies absorption boundary conditions. This in turn allows for more reliable dynamical statistics. This technique was recently used to characterize the

mechanism of folding for the trp-cage miniprotein [26], and it is further explored in Chapters 3 and 4 of this thesis.

Protein-related processes range from protein-protein/protein-ligand interactions to protein folding. These types of problems are attractive for enhanced sampling, as these processes are often both physiologically relevant and inherently rare on simulation timescales. In this thesis, we focus on two broad classes of protein-related processes: (1) coupled (un)folding and (un)binding in protein-protein and protein-ligand systems, and (2) reversible fold switching of metamorphic proteins. The first class arises when two proteins (or a protein and a ligand) encounter each other and bind. The traditional treatment of these systems uses terminology taken from the literature surrounding enzyme-substrate binding: when neither species undergoes much restructuring to allow the binding, the process corresponds to a lock-and-key model of molecular recognition. To fit together, however, two species can rearrange or adopt conformations/folds that are only stabilized by the interspecies interaction. The binding process can thereby be coupled to the folding of one or more of the interacting species [27–29]. Traditionally, coupled (un)folding and (un)binding processes can range between two extremes: conformational selection and induced fit. In the former, the two species, fluctuating in their native state ensembles, first adopt the conformations seen in the bound complex, and then bind. In the latter, the two species form an initial encounter complex, and then rearrange/refold to form the ultimately stable bound structure [30]. Very few studies have investigated to what extent these two extremes might coexist, either in a single pathway or in an ensemble of pathways accessible to a single biological process.

In contrast, the fold switching of metamorphic proteins consists of the reversible unfolding/refolding of a single protein into two or more stable structures [31]. While fold switching is a relatively nascent field [32], the more general concept of protein folding, where a disordered sequence adopts a single stable tertiary structure, has been extensively studied through both theory and experiment [33–38]. One prominent perspective on protein folding

is the Energy Landscape (or Folding Funnel) Hypothesis. Here, the protein navigates a wide diversity of folding pathways, primarily consisting of the folding/unfolding of various native substructures that eventually combine to form the native structure [39–41]. Another theory is the diffusion-collision model, where small "microdomains" can form transiently stable secondary structures, and these microdomains slowly diffuse and collide with each other. Once collided, these microdomains can form higher-order aggregates, which eventually adopt the native protein structure [42, 43]. Yet another model is the nucleation and growth model (or nucleation-collapse model), where near-native folded regions spontaneously nucleate and spread across the protein [44–46]. Although different in specifics, all of these models conclude that proteins can fold through an ensemble of pathways.

In contrast, a recently-developed hypothesis called the Foldon Hypothesis asserts that proteins instead fold through one single, ordered pathway. This hypothesis relies on the sequential folding of small, cooperative folding units, called foldons [47, 48]. The requirement of folding via one pathway is relaxed in the Foldon Funnel Model, which tries to bridge the previously-described multipathway models and the Foldon Hypothesis. In this model, a protein folds through the formation of foldons, but these foldons can fold/refold in almost any order, allowing for a diversity of pathways [49, 50]. This instead places the single-pathway Foldon Hypothesis as a subset of the more multipathway Energy Landscape view. Despite the rich literature on different models of protein folding, it is currently unclear how these concepts relate to the fold switching seen in metamorphic proteins. Since fold switching is a specific type of protein folding with two stable folds instead of one, it is reasonable to assume that metamorphic proteins might exhibit multipathway fold switching.

In their essence, both fold switching and coupled folding and binding are specific instances of the same underlying process. Namely, a sequence of amino acids is undergoing an order-to-disorder transition (or vice versa), which is the essence of the protein folding problem. In coupled folding and binding, this order-to-disorder transition is coupled to the interaction

with another species. In fold switching, it is instead paired with an external trigger that leads to refolding. As discussed above, molecular simulations have played a central role in the development of the theory of protein folding and in the interpretation of experiments. There is just as much potential for simulations to aid the field's understanding of both coupled folding and binding and fold switching. Regarding the former, many questions still remain as to how folding specifically couples to binding. How might two proteins with intrinsic disorder dissociate from one another? To what extent can induced fit and conformational selection coexist in a single mechanism? Could both extremes be experimentally accessible? Regarding the latter, an even wider set of questions exist. How do various models of protein folding apply to fold switching proteins? Is folding dominated by the formation of foldons? To what extent can multiple pathways coexist? Answering these questions in the general sense will require extensive experimental and theoretical collaboration; indeed, the protein folding problem itself has not yet been definitively solved, despite decades of dedicated research. This thesis uses enhanced sampling to answer some of these questions for three specific biological systems: the insulin dimer, the insulin hexamer, and KaiB. For these systems, at least, we find reactions that are broadly multipathway. Beyond providing evidence that can be used to further explore the above general questions, the results in this work are of biological importance. Insulin and KaiB are both proteins with key biological functions, and in the following chapters we characterize some of their most biologically-relevant equilibria.

Chapter 2 focuses on the insulin dimer. We use umbrella sampling to investigate the putative pathways for the dissociation of the dimer. Insulin can exist in a wide variety of monomeric and polymeric forms, including the hexamer and the dimer. Insulin is typically a hexamer in therapeutic formulations and a dimer when circulating in the blood [51, 52]. The dimer must dissociate for insulin to regulate blood glucose, as insulin binds to its cellular receptor as a monomer [53]. While dimer dissociation is thus a key step in the biological function of insulin, very little was known about the mechanism of dissociation. Much was

known, however, about the structural ensemble of the dimer and monomer individually. The dimer is well structured and primarily comprised of two interfaces: the β interface, made of two antiparallel β sheets, and the α interface, made of two adjacently-packed α helices [54, 55]. The monomeric state, in contrast, is thought to contain significant disorder [54, 56–59]. The extent to which this monomeric unfolding coupled to the overall dissociation was unknown.

To explore the dissociation, we developed and introduced a computational pipeline that combined the string method, adiabatic biased molecular dynamics, umbrella sampling with replica exchange, and computational infrared spectroscopy. We discovered a set of collective variables (CVs) that best capture the dissociation. These include distances and pseudodihedral angles that describe how the α and β interfaces are oriented relative to one another. Using these CVs, we computed two-dimensional (2D) potentials of mean force (PMFs) and associated equilibrium averages. These data revealed an ensemble of putative dissociation pathways that ranged from induced fit (the so-called "α path", where the α interface breaks first) to conformational selection (the β path, where the β interface breaks first). Along the α path, we observed significant monomeric unfolding that was coupled to the unbinding, in the form of C-terminal detachment of insulin's B-chain. We found no such disorder along the β path. Importantly, we also discovered many intermediate paths that blend these two extremes, all with very similar maximum free energy barriers. We concluded that the dissociation proceeds through an ensemble of pathways, and we used computational IR spectroscopy to show that these pathways could be experimentally distinguished based on different orderings of key backbone amide solvations.

Chapter 3 focuses on the insulin hexamer and its interaction with phenol. Insulin is a common treatment for type I diabetes, and most therapeutic formulations include phenol. Phenol prolongs shelf life by shifting the population of insulin hexamer from T_6, which has a lifetime of hours, to R_6, which has a lifetime of days [60]. This shift to the R_6 form

presumably suppresses hexamer dissociation and in turn off-pathway equilibria that lead to fibrillation [61–63] or on-pathway equilibria that lead to binding [52]. The possibility of further stabilizing the insulin hexamer is attractive for future generations of slow-acting diabetes therapeutics. Indeed, much effort is currently directed toward modulating insulin's equilibria to achieve therapeutics with desired properties [64]. For example, insulin aspart and lispro, widely used fast-acting diabetes therapeutics, both introduce mutations that destabilize the insulin dimer interface [65, 66]. In contrast, slow-acting basal insulin analogs like glargine [67] and detemir [68, 69] function by either decreasing insulin solubility or adsorption in the body [70]. Longer-acting insulin analogs are also being developed [71, 72], but such formulations are often expensive and complicated. The ability to discourage phenol release through targeted protein mutation would open an avenue to creating newer, simpler long-lasting insulin analogs. To enable design of improved insulin mutants and analogs, we sought to better understand the mechanism(s) of phenol release from the insulin hexamer.

We thus used DGA to expand on an existing simulation study by Vashith and Abrams [73]. They used Steered Molecular Dynamics to discover three phenol unbinding pathways for WT human insulin. We found these three pathways and also discovered three new pathways, two of which involved large-scale opening of the primary escape channel. This reveals that the hexamer is much more dynamic than previously appreciated. The reorganization of the hexamer during multiple dissociation pathways is an example of disorder introduced during unbinding, much like the coupled unfolding and unbinding found in the insulin dimer dissociation. All six of our phenol-release pathways contained at least 10% of the reactive current, underscoring how this process is also multipathway. We repeated all of these calculations for two singly-mutated insulins: Ile A10 Val and Glu B13 Gln. We demonstrated how these mutations affected the contributions of specific pathways, showing that the kinetics of this multipathway process could be manipulated by mutating residues both in and out of the phenol binding pocket. This work reveals the potential of designing insulin analogs with

multiple mutations, chosen as to simultaneously affect multiple phenol escape pathways.

Chapter 4 considers the fold switching of KaiB. KaiA, KaiB, and KaiC together form the central circadian oscillator in cyanobacteria. KaiB is also a metamorphic protein that can switch between two stable folds. One fold is the so-called "ground state" (gs), which tends to exist as a tetramer and is largely inactive in the broader circadian oscillator. The other fold is the "fold switched state" (fs), which adopts a thioredoxin-like fold and binds to KaiC and sequesters KaiA as part of the cycle. Crystal structures exist for both gs and fs KaiB. The primary structural differences between the two folds manifest in the fold switching C-terminal domain, a 57-amino acid long sequence that adopts the thioredoxin-like secondary structure pattern $\alpha\beta\beta\alpha$ in fs KaiB. In gs KaiB, this pattern inverts to $\beta\alpha\alpha\beta$. In both gs and fs KaiB, the 50-amino acid N-terminal domain remains largely unchanged [74, 75]. Because the mechanism of fold switching was unknown, we implemented a combination of near-atomic molecular dynamics, hydrogen-deuterium exchange experiments, and DGA to investigate this mechanism.

We compared computational predictions of hydrogen-deuterium exchange to experiments to validate simulation parameters and discovered that a fs-stabilized KaiB mutant exposes its amide backbone primarily through subglobal unfolding events. Additionally, both fs and gs KaiB have near-identical free energies. During the fold switch, the secondary structure elements in the C-terminal fold switching domain fold and refold independently of one other in various orders. Thus, we observe a diversity of paths involving the melting/formation of individual foldons, consistent with multiple existing models of protein folding [39, 40, 50]. By examining the reactive current, we observed very little unfolding in the otherwise-stable N-terminal domain, corresponding to the partial melting of a single β sheet or α helix. The primary free energy barrier is correlated with the breaking of β sheets in fs KaiB, and not the isomerization of prolines in the C-terminal fold switching domain. Between 40 and 50% of structures in the transition state ensemble contain the melted C-terminal fs KaiB α helix.

This ensemble, however, is quite heterogeneous, containing a broad mix of melted/folded secondary structures. We further discovered that three prolines (P63, P70, and P71) prefer certain isomerization states depending on KaiB's fold. All three preferentially populate the *cis* isomer in fs KaiB and the *trans* isomer in gs KaiB. For P71 this preference is particularly strong. This is in contrast to the crystal structure of the fs KaiB mutant, which resolves P71 as *trans* and P72 as *cis* [75]. This work is the first statistical treatment of the mechanisms of fold switching of a metamorphic protein. Much like what has been found in investigations of proteins folding to a single stable structure, the fold switching of KaiB tends to proceed through a diversity of pathways.

CHAPTER 2

INSULIN DISSOCIATES BY DIVERSE MECHANISMS OF COUPLED UNFOLDING AND UNBINDING

This chapter was published under this title as Adam Antoszewski, Chi-Jui Feng, Bodhi P. Vani, Erik H. Thiede, Lu Hong, Jonathan Weare, Andrei Tokmakoff, and Aaron R. Dinner, *J. Phys. Chem. B*, 124(27), 5571–5587, 2020 [76].

Abstract

The protein hormone insulin exists in various oligomeric forms, and a key step in binding its cellular receptor is dissociation of the dimer. This dissociation process and its corresponding association process have come to serve as paradigms of coupled (un)folding and (un)binding more generally. Despite its fundamental and practical importance, the mechanism of insulin dimer dissociation remains poorly understood. Here, we use molecular dynamics simulations, leveraging recent developments in umbrella sampling, to characterize the energetic and structural features of dissociation in unprecedented detail. We find that the dissociation is inherently multipathway with limiting behaviors corresponding to conformational selection and induced fit, the two prototypical mechanisms of coupled folding and binding. Along one limiting path, the dissociation leads to detachment of the C-terminal segment of the insulin B chain from the protein core, a feature believed to be essential for receptor binding. We simulate IR spectroscopy experiments to aid in interpreting current experiments and identify sites where isotopic labeling can be most effective for distinguishing the contributions of the limiting mechanisms.

2.1 Introduction

Protein-protein association and dissociation are key to many cellular processes, ranging from transmembrane signaling [77–79] to endocytosis [80]. While some protein complexes may involve little (re)structuring of the participating components and thus conform to a lock-and-key model of molecular recognition, it is now clear that often (un)folding and (un)binding are coupled [27–29]. Coupled folding and binding can be described by two limiting mechanisms: induced fit, in which nonnative subunits form an initial encounter complex that then rearranges to a stable bound structure, and conformational selection, in which individual subunits first rearrange to conformations similar to those in the associated state and then bind [30]. Detailed characterizations of protein-protein association/dissociation simulations [22, 23] suggest that coupled folding and binding is often multipathway, combining elements of both of these limiting mechanisms. This makes both experimental and computational study of coupled folding and binding challenging.

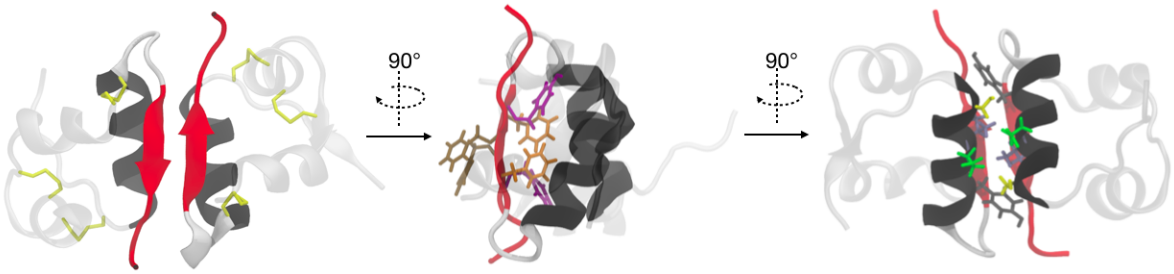


Figure 2.1: Three views of the insulin dimer. The A chain and residues Phe^{B1}-Gly^{B8} of each monomer are shown in translucent silver, while interfacial residues are opaque. The interfacial α helices are shown in black, the β turn is shown in white, and the β sheet is shown in red. In the left panel, cysteine bonds are shown in yellow. In the middle panel, side chains for residues Phe^{B24} (orange), Phe^{B25} (brown), and Tyr^{B26} (purple) are shown. In the right panel, side chains for residues Ser^{B9} (yellow), Val^{B12} (blue), Glu^{B13} (green), and Tyr^{B16} (gray) are shown.

The protein hormone insulin has come to serve as a model for studying coupled folding and binding owing to its small size and the therapeutic importance of its equilibrium between different oligomeric states [63, 81–83]. One such equilibrium is the one between dimer (Figure

2.1) and monomer. Each insulin monomer is 51 amino acids, organized into two polypeptide chains (A and B) joined by disulfide bonds (yellow in the left view). The 21-residue A chain forms two α helices (translucent), while the 30-residue B chain consists of an α helix (residues Ser^{B9}-Cys^{B19}, black) with a β turn (Gly^{B20}-Gly^{B23}, white) that leads to a C-terminal β sheet (Phe^{B24}-Ala^{B30}, red) in the dimer. Both experimental alanine scanning mutagenesis data [84] and free energy simulations [85, 86] point to the importance of specific interfacial residues for stabilizing the dimer interface. These residues include the aromatic triplet of Phe^{B24}-Phe^{B25}-Tyr^{B26} on the interfacial β sheet [83], Tyr^{B16} on the interfacial α helix, and both Gly^{B23} and Pro^{B28} on the β turn and the C-terminal segment of the B chain, respectively.

Dimer dissociation, which is a prerequisite for insulin to bind to its cellular receptor [51], is thought to be an example of coupled unfolding and unbinding. While the dimer is well structured [54, 55], the monomeric state is thought to contain significant disorder. Specifically, experimental and computational studies indicate that Phe^{B24}-Ala^{B30} can detach from the B-chain α helix and become at least partially disordered in the monomeric state [54, 56–58, 87–89]. This detachment is thought to be important for insulin to bind its receptor [79, 82, 83] based on structures of insulin in complex with fragments of the receptor [53, 90]. An outstanding question is how Phe^{B24}-Ala^{B30} detachment is coupled to dissociation of the dimer. More generally, the pathways of dimer dissociation remain poorly characterized. For example, it is unclear what role, if any, the interfacial α helices play in dissociation, and whether there are partially solvated or unfolded intermediates.

There is some experimental evidence suggesting the dimer dissociation could couple unfolding to unbinding. In particular, temperature jump two-dimensional (2D) amide-I infrared (IR) spectroscopy measurements suggest that, during dissociation, there is conformational rearrangement within the monomers on the timescale of 5 to 150 μs, prior to loss of the β sheet at the dimer interface between 250 and 1000 μs [57, 58]. Time-resolved X-ray scattering

data also suggest an intermediate with conserved secondary structure on the timescale of 900 ns [91]. These experiments, although mechanistically suggestive, provide limited structural information; complementary simulations are needed to microscopically interpret these data.

Recently, Bagchi and coworkers used metadynamics to compute the free energy as a function of the monomer-monomer center-of-mass distance and the number of intermolecular contacts, subject to restraints on the radii of gyration of the monomers [2, 3, 92]. They identified a single major pathway of dissociation in which the number of intermolecular contacts was first observed to markedly decrease before the center-of mass distance increased. Through additional collective variables, they also characterized the protein-protein and protein-solvent interactions of $\mathrm{Phe}^{\mathrm{B24}}$ and $\mathrm{Tyr}^{\mathrm{B26}}$, indicating that conformational rearrangement and intramonomeric unfolding are both coupled to the dissociation. Shaw and coworkers recently characterized the association of the insulin dimer through both unbiased simulation and tempered binding, an enhanced sampling technique that scales the protein interaction energies to encourage binding [10]. In contrast to the simulations described immediately above, they found that successful association events consisted of insulin monomers adopting conformations similar to those found in the dimer before binding, and observed very little intramonomeric unfolding. The extent to which these two binding/unbinding pathways, one which involves monomeric unfolding and one which does not, can coexist is currently unknown.

In this work, we use a computational pipeline that combines multiple methods for enhanced sampling of rare events in molecular dynamics simulations to investigate coupled unfolding and unbinding during insulin dimer dissociation. In particular, we identify collective variables that fully resolve the possible pathways for the dissociation, and we show how an error estimator that we recently introduced [93, 94] can be used to quantitatively monitor convergence and allocate computational resources efficiently. The error estimator that we employ both provides quantitative evidence as to the convergence of our simulations, and al-

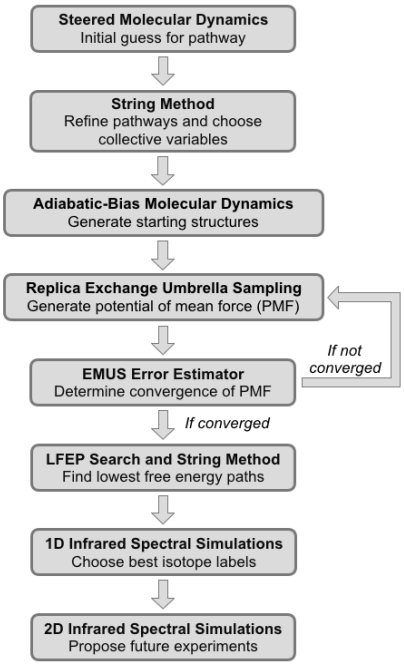


Figure 2.2: An overview of the computational pipeline. Each panel shows the method used and the information it yields. See the Methods section for further details.

lows us to meaningfully compare the free energy profiles of competing pathways by explicitly quantifying asymptotic errors. The computational pipeline enables us to show that there are multiple competing pathways for dimer dissociation, and we characterize these pathways in detail through additional collective variables that describe intra- and intermonomeric rearrangements. The pipeline is summarized in Figure 2.2 and at the start of the Results and Discussion section, so that readers interested primarily in the results can skip the Methods section without loss of continuity.

The limiting behaviors observed correspond to induced fit and conformational selection mechanisms. Our simulations thus provide a unified perspective on the binding/unbinding paths observed in previous simulations. We go on to propose a set of experiments to investigate the relative contributions of our limiting paths. Specifically, we simulate IR spectroscopy experiments for a variety of isotope-labeled insulins and identify two labels which, when measured via T-jump IR spectroscopy, could experimentally distinguish the contributions from

the limiting pathways to dimer dissociation. These simulated IR spectra also provide references to which future measurements can be compared, facilitating the interpretation of both equilibrium and T-jump IR spectra for the insulin dimer.

2.2 Methods

With a view toward providing a quantitative interpretation of experimental observations, we model insulin in solution at atomic resolution (System Setup and Equilibration), such that dimer dissociation occurs on timescales that are long compared with the molecular dynamics timestep. Consequently, both efficient sampling and informative analysis rely on identifying collective variables (CVs) that capture the slowest relaxing degrees of freedom involved in dimer dissociation. To this end, we tested many combinations of CVs for their ability to enable us to harvest reactive events (String Method and Collective Variable Selection). We found that CVs based on selected intermolecular contacts in the dimer enabled us to harvest reactive events without the addition of restraints to prevent monomer unfolding, and we improved the contact definition over the course of the study, as we gained understanding of the system (Definition of Contacts). Care was taken to converge the potential of mean force (free energy) as a function of those CVs (Adiabatic-Bias Molecular Dynamics; Replica Exchange Umbrella Sampling; Eigenvector Method for Umbrella Sampling and Adaptive Sampling). We were able to trace multiple minimum free energy paths with comparable barriers on that surface, which we validated as stable through further simulations (Finding and Confirming Energetically Favorable Paths). Finally, we computed simulated infrared spectra to guide the design of further experiments (FTIR and 2DIR Simulation). These steps are summarized in Figure 2.2, and we describe each in detail below in the parenthetically indicated sections.

System Setup and Equilibration. The system was modeled with the CHARMM36m force field [95–97]. All simulations were performed using GROMACS 5.1.4 [98], and the system was prepared using CHARMM-GUI 2.1 [99, 100]. Unless otherwise noted, simulations were carried out in the isochoric isothermal (NVT) ensemble at 303.15 K using a Langevin thermostat [101] with a 2 fs timestep and a friction constant of 0.5 ps^{-1} applied to all atoms. All bonds to hydrogen atoms were constrained using the LINCS algorithm [102]. Periodic boundary conditions were employed and the particle-mesh Ewald method [103] was used to calculate electrostatic forces with a cutoff distance of 1.2 nm. The Lennard-Jones interactions were smoothly switched off from 1.0 to 1.2 nm through the built-in GROMACS force-switch function. All molecular visualizations were done in VMD [104], and residue interaction energies were calculated using its NAMDenergy plug-in [105].

The dimer structure was based on the human insulin crystal structure (PDB ID 3W7Y) [106]. To fully equilibrate the system at the desired temperature and pressure, the protein was solvated, equilibrated with restraints in both the NVT and isobaric isothermal (NPT) ensembles, and then equilibrated restraint-free in the NVT ensemble. Specifically, hydrogens were added to the PDB structure, and it was solvated in a cubic box of size $(8\ nm)^3$ using TIP3P water [107]; 48 K^+ and 44 Cl^- ions were added to neutralize the system and bring it to a concentration of 150 mM KCl [108]. There was a total of 48,260 atoms. The system was energetically minimized using the steepest descent method, until the maximum force felt by the system was below 1000 kJ/mol nm. The system was then equilibrated for 100 ps in the NVT ensemble with a 1 fs timestep, followed by 10 ns in the NPT ensemble at 1 bar using the Parrinello-Rahman barostat [109], with a 2 fs timestep and time constant of 5.0 ps. For the energy minimization and equilibration above, harmonic restraints were used to stabilize the positions of all non-hydrogen protein atoms. The system was equilibrated further for 1 ns in the NPT ensemble without position restraints, and the average box size was determined to be $(7.82\ nm)^3$. This box size was used for all further simulations. The

system was equilibrated once more without position restraints for 1 ns in the NVT ensemble. The resulting equilibrated structure, with a root-mean-square deviation (RMSD) of 2.02 Å from the 3WY7 crystal structure, was used to initialize further simulations as described below.

Definition of Contacts Throughout the simulations in this work, relevant inter-residue α carbon distances were transformed by contact functions that smoothly vary between a small range of values. This was done to improve computational control in various methods, providing a consistent scale for biasing variables as distances varied between small and large values. The contact functions were tuned to each method, and as we learned more about the structural features of the dissociation. Specifically, we used the following three contact definitions to transform the distance between α carbons of residues i and j (d_{ij}):

$$s_{ij} = \left[1 - \left(\frac{d_{ij}}{\mathbb{E}[d_{ij}]}\right)^6\right] \Bigg/ \left[1 - \left(\frac{d_{ij}}{\mathbb{E}[d_{ij}]}\right)^{12}\right] \tag{2.1}$$

$$s_{ij} = \begin{cases} 1 & \text{if } d_{ij} < \mathbb{E}[d_{ij}] \\ \exp\left(\frac{-(d_{ij}-\mathbb{E}[d_{ij}])^2}{2r_0^2}\right) & \text{otherwise} \end{cases} \tag{2.2}$$

$$s_{ij} = \begin{cases} 1 & \text{if } d_{ij} < \mathbb{E}[d_{ij}] \\ 1 - \tanh\left(\frac{d_{ij}-\mathbb{E}[d_{ij}]}{\gamma d_{\text{mon}}}\right) & \text{otherwise} \end{cases} \tag{2.3}$$

In the above equations, $\mathbb{E}$ denotes an equilibrium average, and the average distance $\mathbb{E}[d_{ij}]$ was measured for each contact pair from a 5 ns simulation, initialized from the equilibrated dimer structure. The definition in Equation 2.1 was used for the initial driving and string method simulations used to discover collective variables. The definition in Equation 2.2 was used for the umbrella sampling calculations, as it provided a gentler bias near the dimer state. In this definition, r_0 is a parameter that sets the location of the inflection point

of the Gaussian transformation. This parameter was set to 0.6 nm to ensure that there were at least three layers of water between all monomeric residues at dissociation, based on visual inspection. Finally, the definition in Equation 2.3 was used for the final string calculations used to verify the stability of our observed low-energy paths, as it provided better resolution near the dimer state. In this definition, d_{mon} is the average residue pair distance that corresponds to the monomeric state, again chosen so that at least three layers of water separate the residues of interest. This was thus set to be 2.2 nm for the α contacts and 2.0 nm for the β contacts (see Results). The parameter choice $\gamma = 0.65$ tunes the sizes of the monomeric and dimeric states in the 2D contact space.

String Method and Collective Variable Selection. Umbrella sampling is predicated on finding a small number of collective variables (CVs) that capture the slowest relaxing degrees of freedom relevant to the process of interest. To determine reasonable CVs for insulin dimer dissociation, we tested various combinations of CVs for their ability to drive dissociation in steered molecular dynamics simulations (SMD) [110] and then selected CVs that preserved the ability to distinguish refined dissociation paths obtained from the string method [111, 112], discussed in further detail below. The CVs explored were based on interacting pairs of residues with high differential solvent accessible surface area (SASA) between dimer and monomer states (Supplemental Table 2.1, where the apostrophe differentiates residues on one monomer from residues on the other). These included the aromatic triplet in the interfacial β sheet, which was previously identified as important for dimer stability [54, 86, 113].

Distances between the C_α atoms of these residue pairs were computed and transformed using Equation 2.1 as described above. Constant velocity SMD simulations, in which harmonic restraints were used to advance random subsets of s_{ij} from 0.5 to 0.0, were used to drive the system from the equilibrated dimer structure to the dissociated state. By visual analysis, we selected dissociation paths that both led to complete dissociation and did not

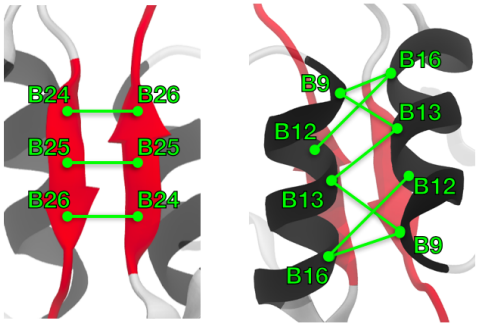


Figure 2.3: Schematic showing the β sheet contact pairs (left) and the α helix contact pairs (right). These correspond to the similarly labeled rows of Supplemental Table 2.1.

involve significant unfolding of the monomers since there is limited experimental evidence for extensive loss of secondary structure [58, 91]. We also only selected paths that had maximum free energies within the range of previous simulations [2, 54]. These were used to initialize string method simulations in the 22-dimensional space of intermonomeric contacts described above (Supplemental Table 2.1). Convergence of strings during the simulations was computed by the Hausdorff distance metric between the current string iteration and the initial string, and simulations were run until this distance metric did not change significantly [114].

To choose a small number of CVs sufficient to describe the dissociation, we clustered the initial and final strings in the space of the first few coordinates obtained from applying the diffusion maps method [115, 116] to their images. We then sought physically interpretable CVs that preserved the clusters, which led to selection of two average contact functions (averaging performed after the transformation), one for three residue pairs at the β sheet interface ($\overline{\beta_c}$) and one for seven residues at the α helical interface ($\overline{\alpha_c}$). These are detailed in Supplemental Table 2.1 and Figure 2.3. These residue pairs are consistent with important interfacial interactions identified in a recent steered molecular dynamics study [86]. We denote the average of the raw distances associated with $\overline{\beta_c}$ and $\overline{\alpha_c}$, used for visualization, by $\overline{\beta}$ and $\overline{\alpha}$, respectively.

Adiabatic-Bias Molecular Dynamics (ABMD). To initialize sampling, 49 independent ABMD [117] molecular dynamics simulations (using the PLUMED 2.3 wrapper for GROMACS [118–120]) were used to drive the system from the dimer to a 7×7 grid of points evenly covering the 2D CV space of $\overline{\beta_c}$ and $\overline{\alpha_c}$. ABMD is similar to SMD but ratchets the system to its target following unbiased fluctuations along the CVs; we came to prefer it to SMD because we found that SMD but not ABMD resulted in melting of the interfacial α helices. That said, because ABMD relies on unbiased fluctuations, we found that the initial 49 simulations did not adequately sample the space close to the $\overline{\beta_c} = 0$ and $\overline{\alpha_c} = 0$ axes. We thus performed 13 extra simulations to drive the system to supplementary points near where either one or both of $\overline{\beta_c}$ or $\overline{\alpha_c}$ went to zero. This driving, whose bias was applied on the 10 individual distances associated with $\overline{\beta_c}$ and $\overline{\alpha_c}$, was repeated with force constants of 1000, 3000, and 5000 kJ/(mol nm), generating a database of trajectories that covered all of the relevant average contact space.

Replica Exchange Umbrella Sampling (REUS). The window centers for the umbrella sampling were distributed on a logarithmically spaced grid in the space of ($\overline{\beta_c}$, $\overline{\alpha_c}$) with the expectation that the initial steps of the dissociation would involve larger changes in free energy, as seen in Figure 2.4A. The force constants, k, for the harmonic biases associated with each window were described by the following equation, adapted from an expression derived by Im and coworkers [121, 122]:

$$\sqrt{k} d_{\max} = 0.8643 \sqrt{2 k_B T}. \tag{2.4}$$

Here, as the windows are unevenly spaced, $d_{\max}$ refers to the maximum distance between adjacent window centers. Additional weak upper walls (half harmonic potentials with $k = 50$ kJ/(mol nm), turned on at 3 nm for each distance) were placed on each distance to prevent artificial interactions across periodic boundaries. To initialize each window, the ABMD-

generated structure that was nearest to each minimum was selected and equilibrated for 100 ps using the harmonic restraint described by equation 2.4. A 2D replica exchange procedure, in which there were exchanges between windows [123], was implemented by taking advantage of the built-in functionality of GROMACS. Each of the 784 windows were simulated for 100 ps, with exchanges attempted at 1 ps intervals only between adjacent windows in the same row. The windows were then simulated for an additional 100 ps, with exchanges attempted every 1 ps between adjacent windows in the same column. This procedure was repeated for a total of 5 ns of simulation time per window, with structures saved every 5 ps. In this way, replicas were exchanged via all nearest neighbors across the entire lattice of windows. Exchange probabilities between 10-50% were achieved depending on the specific window pairs being swapped.

Eigenvector Method for Umbrella Sampling (EMUS) and Adaptive Sampling. The 5 ns of sampling per window was combined to generate a potential of mean force (PMF) by using the Eigenvector Method for Umbrella Sampling (EMUS) [93]. Once the PMF was created, we wanted to investigate whether the PMF was converged and, if not, add additional sampling selectively where it would be most effective. To do this, EMUS was used to estimate the asymptotic variance of replica exchange umbrella sampling simulations. However, because of the replica exchange, assumption VII.3 of ref. 93, namely that sampling in each window is independent, does not hold for our study. Nonetheless, one can still apply Lemma VII.2 in ref. 93 to derive a central limit theorem for EMUS with replica exchange by casting sampling over all windows as a Markov chain; in this case, the asymptotic covariance matrix, Σ, is not block diagonal. This leads to a definition of the asymptotic variance for arbitrary averages that one can approximate by an expectation of integrated autocovariances over the sampled data. For details, see the Supplemental Information.

The contact space PMF is shown in Figure 2.4B, with its associated asymptotic variance in Figure 2.4C. The area of highest asymptotic variance in the PMF was identified, and

is marked by a red box in Figure 2.4C. Using the process described in ref. 93, the per-window error contributions to this region were determined; although these include only the error we would observe if the off-diagonal blocks of Σ were zero, we believe they are sufficient to diagnose the behavior of the umbrella sampling scheme. These contributions are shown in Figure 2.4D, and reveal a J-shaped region of windows which contribute the most to the asymptotic variance of the region marked in Figure 2.4C. These windows were then identified as areas to add additional sampling. This additional sampling used a similar procedure as described above for the initial replica exchange simulations, with the sampling and proposed exchanges restricted to the bottom-most five rows and right-most 5 columns in Figure 2.4D. At 1 ns intervals, this additional sampling was independently processed with EMUS as follows. These data were used to compute a new PMF and its associated asymptotic variance per bin, $\sigma^2_{ij,\text{supp}}$. We then combined $\sigma^2_{ij,\text{supp}}$ with the initial asymptotic variance, $\sigma^2_{ij,\text{init}}$, weighted by the squared ratio of simulation lengths between supplemental and total sampling, f^2_{supp}:

$$\sigma^2_{\text{tot}} = f^2_{\text{supp}}\sigma^2_{ij,\text{supp}} + (1 - f_{\text{supp}})^2\sigma^2_{ij,\text{init}} \tag{2.5}$$

Equation 2.5 assumes the initial 5 ns and supplemental 5 ns of sampling are independent, which given that the autocorrelation time of the quantities needed for EMUS was 6 ps on average across the windows, is a reasonable assumption.

As seen in Figure 2.4E, the peak variance decreased from 0.25 to 0.12 $(\text{kcal/mol})^2$ upon the addition of 5 ns of additional sampling per window in the outlined J-shaped region. This region would not obviously be chosen in the absence of a quantitative procedure, though it can be rationalized in hindsight as corresponding to a major dissociation pathway that we characterize in detail in Results and Discussion. This illustrates how EMUS allows users to monitor the convergence of US simulations as sampling proceeds, and to adaptively identify regions of state space that are most in need of additional sampling, despite the neglect of

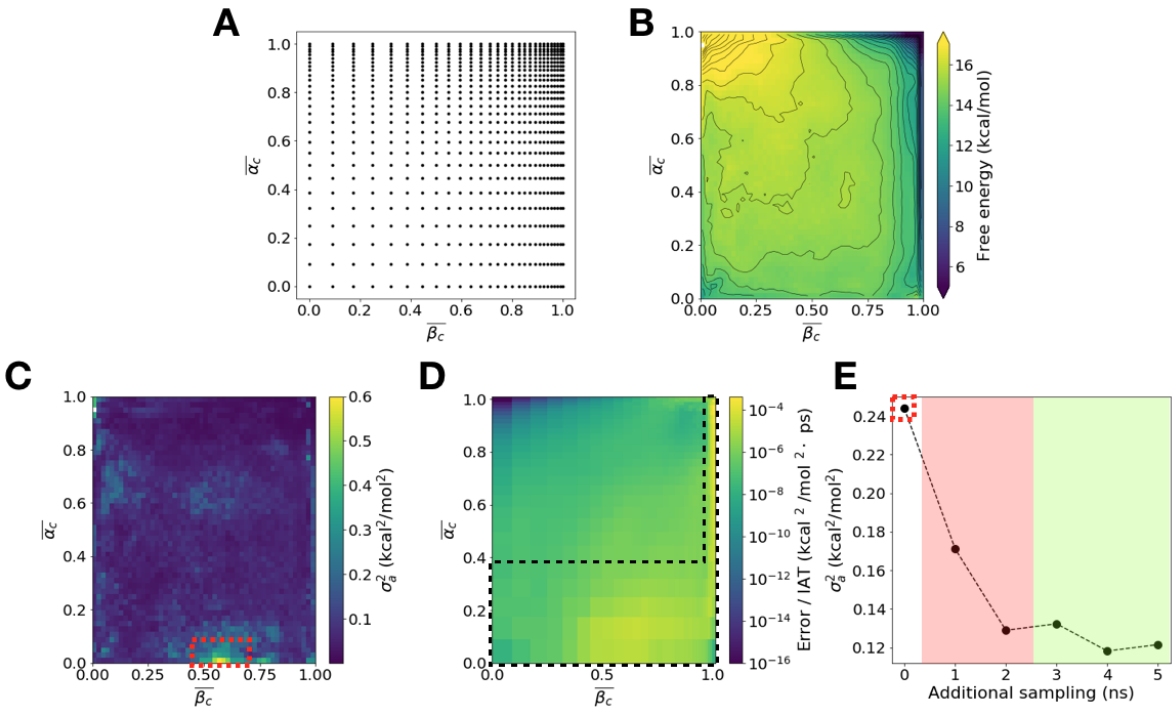


Figure 2.4: Umbrella sampling. (A) The location of the window centers used for the REUS procedure, shown in the space of $\overline{\beta_c}$ and $\overline{\alpha_c}$. These are logarithmically spaced to place more density near the dimer (upper right corner). (B) Free energy as a function of the average numbers of β and α contacts at the insulin dimer interface (contour spacing 0.5 kcal/mol). 5 ns of sampling was gathered per window (784 windows). (C) The asymptotic variance associated with the free energy in (B). The region of highest variance, with average 0.25 and maximum 0.59 kcal^2/mol^2, is marked by the red box. (D) The per-window error contributions to the marked variance in (C), assuming that the matrix Σ is diagonal. 5 ns of additional sampling was added to only the boxed black area of large error contributions. (E) How the average asymptotic variance of the marked region in (C) decreased as 5 more ns of sampling was added per selected window. The red shaded region represents the area where the additional sampling is shorter than 10 times the autocorrelation time for EMUS quantities. The asymptotic variance data in this region is thus unreliable. Reliable asymptotic variances are obtained in the green shaded region.

correlations discussed above. Furthermore, EMUS allows for the calculation of PMFs in arbitrary CV spaces, not just the space in which the biasing was done, without the need for additional sampling. Examples of these PMFs are seen in Results and Discussion.

Finding and Confirming Energetically Favorable Paths. Our analysis of dissociation is based on minimum free energy paths. Initially, seven such paths were drawn on the

2D PMF by using the lfep search algorithm [1]. To ensure that these were stable in the 10-dimensional space of all of the contacts associated with the averages $\overline{\alpha_c}$ and $\overline{\beta_c}$, another iteration of the string method was run in this space, this time initialized from structures drawn from the REUS database along these 2D minimum free energy paths. The specific contact definition used is in Equation 2.3, and the strings were run until converged as measured by the Hausdorff distance, as discussed previously. The starting and ending positions of these strings are shown in Supplemental Figure 2.11. In particular, the α path shows almost no variation, and the β path shifts only minimally, and this shift does not change any of the molecular trends discussed in the Results. The others paths also exhibit minimal variation. This provides evidence that most of the pathways we identify and the limiting pathways in particular are indeed stable in a broader space.

FTIR and 2DIR Simulation. Simulated IR spectra were calculated from the Fourier transform of a vibrational transition dipole time correlation function using a mixed quantum-classical model described in refs. 124 and 125. Briefly, electrostatic collective variables can be used to translate molecular-dynamics sampling of protein structure into (i) a time-dependent Hamiltonian and (ii) a transition dipole moment that describes the amide I vibrations of protein backbones; these quantities in turn can be used to calculate the dipole time correlation functions that are needed to create simulated FTIR and 2D IR spectra. Furthermore, these spectra can be calculated for both native proteins and isotope-labeled proteins [126]. Here, we aimed to generate FTIR spectra for 50 points equally spaced along both the α and the β paths for a variety of isotope-labeled insulins. 2DIR spectra were then calculated for specific states for two isotope-labeled insulins. These spectra were then used to propose possible experiments to validate our results.

First, each path was divided into a series of 50 points, referred to as image centers. By comparing these image centers to the REUS database, 20 structures were selected to be associated with each image center. These structures were randomly drawn from the sampling

that was within both 0.4 nm of each image center and the Voronoi tessellation associated with each image center. Each of these structures served as the starting point for an additional short molecular dynamics simulation consisting of 100 ps of equilibration followed by 100 ps of sampling every 20 fs. These simulations were then used to generate both the FTIR and 2DIR spectra by using the procedure outlined below.

As IR spectra correspond to manifestly quantum-mechanical vibrational transitions, our classical molecular dynamics trajectories had to be translated into a time-dependent Hamiltonian and transition dipole trajectories. To this end, we associated each amide I vibration with a site, defined by the atomic positions of the backbone amide groups (C, O, N, and H atoms). The frequency of each site was calculated using an empirical electrostatic frequency map optimized against experimental spectra of isotope-edited NuG2b protein, which evaluates the electrostatic potential value at the C, O, N, and H positions [127]. This potential-based map (4PN-150) has an estimated frequency uncertainty of 2.25 cm^{-1}. When applying the map, we used modified glycine charges described previously [127]. Additionally, we considered coupling between sites, including both through-bond mechanical coupling and through-space electrostatic coupling. Through-bond coupling between adjacent sites was generated using a density functional theory (DFT)-based nearest-neighbor coupling map, while through-space coupling was computed by a transition charge coupling map [128]. We did not account for vibrations from protein side-chain and terminal groups. The transition dipole of each site was assigned using the zero-field values from a DFT-based electrostatic map [129]. When generating simulated 2DIR spectra, there are signal contributions from excited state absorption (ESA), which correspond to vibrational transitions between states with one quantum of excitation energy and those with two quanta of excitation energy. To deal with this, the corresponding two-quantum Hamiltonian and transition dipole moments were constructed using a weak anharmonic model [124, 130].

The time-dependent Hamiltonian and transition dipole trajectories were converted to sim-

ulated FTIR spectra and 2D IR spectra using a dynamic wavefunction-propagation scheme with a Trotter expansion to reduce computation time [131, 132]. The window time for calculating dipole time correlation functions was set to 2.5 ps. The anharmonicity of the amide I oscillator was set to 16 cm^{-1} [130]. The amide I vibrational lifetime was modeled by an ad hoc single exponential decay, with a time constant of 1.0 ps determined by transient absorption experiments of Ala–Ala [133]. The isotope frequency shift introduced by a $^{13}C^{18}O$ label was set to 65 cm^{-1} [124]. The spectra for the structures that were selected from the REUS database were uniformly averaged within each of the 50 images across both paths. This created a simulated spectrum representative of the location of each image in CV space. The isotope labeled FTIR spectra along each path, with the corresponding simulated unlabeled spectra subtracted to create difference spectra, are shown in Supplemental Figure 2.12. Based on these results, the data were regrouped and reaveraged as described in the Supplemental Information to create the 2DIR spectra shown in a future section.

2.3 Results and Discussion

Our goal was to investigate the molecular changes in intra- and intermolecular structure during insulin dimer dissociation. To this end, as summarized in Figure 2.2 and detailed in Methods, we first used steered molecular dynamics to generate multiple dissociation events, naively biasing the simulations to force the monomers apart. We then refined the resulting paths with the string method, which relaxes these paths to local minimum free energy paths. Based on these simulations, we identified a small number of distances that provided good control over sampling (specifically, replica exchange umbrella sampling): 7 between C_α atoms in the interfacial α helices and 3 between C_α atoms in the interfacial β sheet (Figure 2.3). The α and β distances were separately averaged to define collective variables $\overline{\alpha}$ and $\overline{\beta}$, respectively. We used replica exchange umbrella sampling together with an error estimator that we recently introduced [93] to ensure good sampling of configurations consistent

with each combination of these variables. From these data, we constructed the PMF as a function of the average interfacial distances. Below, we describe this PMF, followed by additional statistical averages that provide further insights into specific intra- and intermolecular structural features. We conclude by showing how simulated vibrational spectra can serve as references for the design and interpretation of experiments.

Dissociation is multipathway, with two limiting cases. The PMF as a function of $\overline{\alpha}$ and $\overline{\beta}$ is shown in Figure 2.5, surrounded by representative structures. The minimum of the dimeric basin is marked by the circle at $(\overline{\beta}, \overline{\alpha}) = (0.53, 0.63)$ nm, and we set it to be the zero of free energy. The dimer is flanked by a trough along each axis, corresponding to breaking the β contacts while maintaining the α contacts and vice versa. There is a shoulder at $\overline{\beta} \approx 0.75$ nm; calculations described further below show that it coincides with solvent penetration of the β sheet. The remainder of the PMF is relatively flat. We take the monomeric state to be $\overline{\beta} > 2.0$ nm and $\overline{\alpha} > 2.2$ nm (marked by the dotted white box in Figure 2.5), which ensures that there are at least three layers of water between interfacial residues (see Methods). The free energy in this region ranges from 13 to 15.5 kcal/mol, within the range of previous estimates of the stability of the dimer [2, 85, 86, 134]. The plateau surrounding the monomeric state is between 1-3 k_BT higher in free energy.

Consistent with the diversity of paths that we obtained in our steered molecular dynamics and string method simulations (see Methods), many minimum free energy paths can be drawn on the PMF (Supplemental Figure 2.11). These paths are stable not only in this 2D average distance space, but also the full 10-dimensional space of all individual distances, as indicated by the string method results in Supplemental Figure 2.11. Despite their similar maximum free energies, these paths imply dramatically different mechanisms of dissociation. For clarity, we focus on two limiting cases that initially follow the aforementioned troughs in free energy flanking the dimer. Along the α path (black in Figure 2.5), the interfacial α helices separate prior to the strands of the β sheet; along the β path (red in Figure 2.5), the

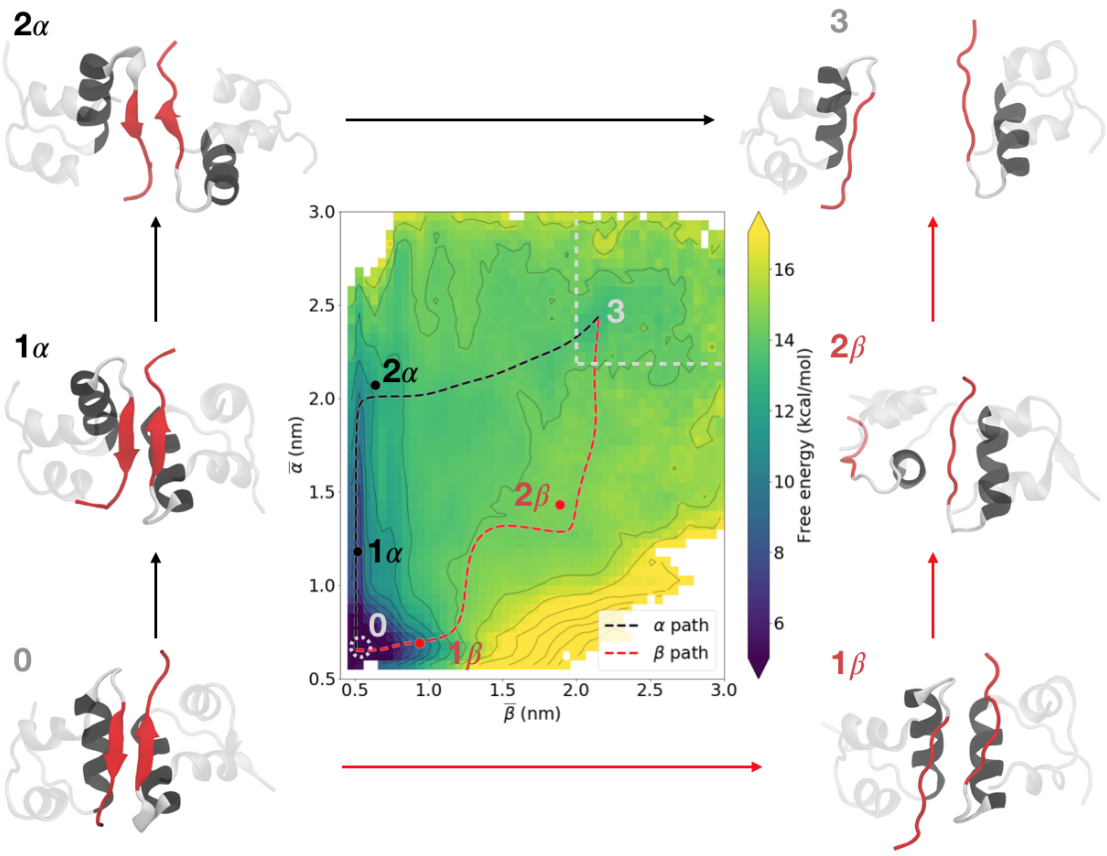


Figure 2.5: Potential of mean force (PMF) as a function of $\overline{\alpha}$ and $\overline{\beta}$. Limiting mean free energy paths in which the interfacial α or β contacts break first are indicated by black and red dashed lines, respectively. Representative structures corresponding to the marked points along the paths are labeled and shown adjacent to the PMF. These structures are referenced throughout the paper and are available in the supplemental material. The dimer is marked by a dotted white circle, and the monomeric state is marked by a dotted white box. Contour lines are every 2 k_BT. The color scale is capped at both the upper and lower ends to more clearly show the variation in the partially-dissociated regime.

order is reversed.

The free energy profile of the α path is consistent with the minimum free energy path obtained by Bagchi and coworkers (compare the black lines in Supplemental Figure 2.13 with Paths 1 and 4 in Figure 3 of ref. 2): there is an initial rise to an intermediate of 7.7 kcal/mol (5.3 kcal/mol in ref. 2), followed by a shoulder ∼2.4 kcal/mol higher in free energy and then a barrier of ∼4.0 kcal/mol. The free energy profile of the β path exhibits no comparable shoulders, but its maximum is comparable to that of the α path (14.7 kcal/mol and 14.1

kcal/mol, respectively). Considering the typical asymptotic variance of our PMF is on the order of 0.2 kcal^2/mol^2, we thus expect both of these limiting paths, as well as the many paths that fall between them (Supplemental Figure 2.11), to contribute to dissociation.

The monomers rotate relative to each other in opposite ways along the two limiting paths. Previous studies [2, 3] reported rotation of the interfacial β strands relative to each other, as characterized by a pseudodihedral angle (here denoted Φ_β) defined by the C_α atoms of Tyr^{B26}, Phe^{B24}, Tyr$^{B'26}$, and Phe$^{B'24}$, where the primes distinguish one monomer from the other. In addition to Φ_β, we calculate the pseudodihedral angle (Φ_α) between the geometric centers of the backbone atoms of the residues that define the dimeric interfacial alpha helices: Ser^{B9}- Leu^{B11}, Leu^{B17}-Cys^{B19}, Leu$^{B'17}$-Cys$^{B'19}$, and Ser$^{B'9}$-Leu$^{B'11}$. Explicit illustrations of these angles are shown in Supplemental Figure 2.14. By examining both Φ_β and Φ_α, we can better understand whether the rotation is restricted to the β strands, or if the entire dimer interface moves together. Since molecular dynamics simulations allow any collective variable to be calculated within machine precision, we create PMFs in four 2D spaces that measure interfacial rotations as both $\overline{\alpha}$ and $\overline{\beta}$ change (Figure 2.6A); the plots are restricted to ranges of $\overline{\alpha}$ and $\overline{\beta}$ that correspond to the initial steps of dissociation because the interfacial pseudodihedrals become poorly defined when the average distances are large. All of the PMFs, both in Figure 2.5 and Figure 2.6, are generated from the same dataset, as EMUS allows for the calculation of PMFs in arbitrary collective variable spaces without the need for additional sampling.

There is a deep free energy basin corresponding to the dimer at $(\overline{\beta}, \overline{\alpha}) = (0.58, 0.63)$ nm and $(\Phi_\beta, \Phi_\alpha) = (-15^\circ, 125^\circ)$. This reflects the fact that in the dimer state, in agreement with available crystal structures, there is a slight rotation from parallel between the β sheet residues, and a more pronounced rotation between α helices, consistent with well-known characterizations of α helix packing [135–137]. The PMFs are dominated by the troughs flanking the dimer in Figure 2.5. The trough along the α path (black dotted arrows) is readily

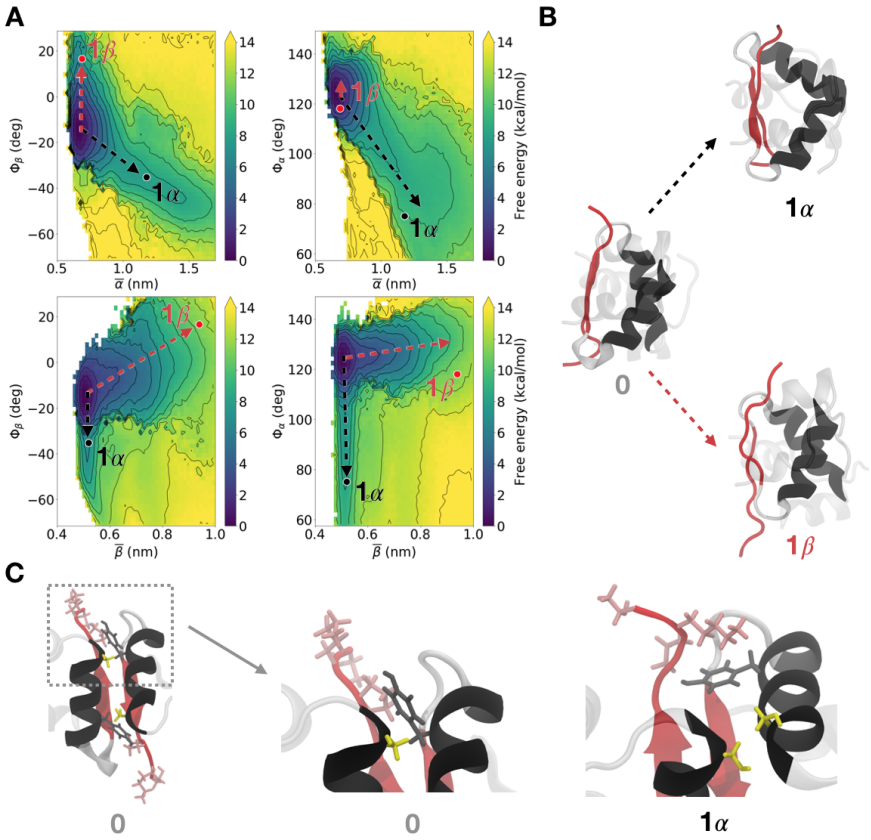


Figure 2.6: The monomers rotate relative to each other during dissociation. (A) PMFs characterizing the rotations as pairwise functions of $\overline{\alpha}$ (top) or $\overline{\beta}$ (bottom), and Φ_β (left) or Φ_α (right). Superimposed arrows show the negative rotations associated with the α path (black) and the positive rotations associated with the β path (red). Intermediates are marked on the PMF, and are as labeled in Figure 2.5. Structures were chosen to show the rotation of Φ_β; for this reason, the arrows in the left plots terminate at the dots but those on the right plots do not. Contour lines are every 2 k_BT. The color scale was capped at 14 kcal/mol. (B) Representative structures for the rotations along the α and β paths, represented by the black and red arrows, respectively. These structures, labeled in (A), are the same as those labeled in Figure 2.5. (C) The dimer with the interfacial α helices in front, showing the side chains for Ser^{B9} (yellow), Tyr^{B16} (gray) and Pro^{B28}-Ala^{B30} (pink). Zooming in (middle), one can see the native contact of Ser^{B9}-$\text{Tyr}^{\text{B}'16}$, with Pro^{B28}-Ala^{B30} behind. Along the α path (right), $\text{Tyr}^{\text{B}'16}$ has rotated away from Ser^{B9}, and is instead in contact with Pro^{B28}-Ala^{B30}. Furthermore, this rotation brings Ser^{B9} and $\text{Ser}^{\text{B}'9}$ together.

visible in both sets of plots and shows that increases in $\overline{\alpha}$ are coupled to negative rotations of both Φ_α and Φ_β (by $-35°$ and $-45°$, respectively). The trough along the β path (red dotted arrows) is more readily visible in the bottom plots and shows that increases in $\overline{\beta}$ are coupled to positive rotations of Φ_β; there is little change in Φ_α, suggesting the interfacial α helical

contacts are maintained. Side views (similar to the middle panel of Figure 2.1) of the dimer and rotated species show the structural consequences of these interfacial rotations (Figure 2.6B). Namely, comparing these side views with the corresponding front views (Figure 2.5) suggests that the initial dissociation along the limiting β path involves the breaking of the interfacial β sheet and the positive rotation of Φ_β. In contrast, the initial dissociation along the α path comes as the α helices twist away from each other, coupled with negative rotations of both Φ_α and Φ_β.

The negative rotations along the α path enable formation of nonnative interactions (Figure 2.6C). The serine side chains of Ser^{B9} and $\text{Ser}^{\text{B}'9}$ (yellow in Figure 2.6C) form a hydrogen bond at $\overline{\alpha} \in [1.0, 1.3]$ nm. Then, as the β strands rotate relative to each other along the α path, Pro^{B28}-Ala^{B30} (pink) breaks its native contacts and instead forms a contact with tyrosine $\text{Tyr}^{\text{B}'16}$ (gray), in the α helix of the opposite monomer. This tyrosine at Tyr^{B16} contacts Ser^{B9}, Val^{B12}, and $\text{Tyr}^{\text{B}'26}$ in the dimeric state (see Supplemental Table 2.1), but these interactions are broken as the α helices separate. Averages of side chain contacts that quantitatively show these trends are seen in Supplemental Figure 2.15. Furthermore, the necessity of breaking the native contacts between Pro^{B28}-Ala^{B30} and $\text{Gly}^{\text{B}'20}$-$\text{Gly}^{\text{B}'23}$ as the dissociation progresses is consistent with the mutations that yield fast-acting insulin analogs, discussed further in the Supplemental Information (Supplemental Figure 2.16). No comparable nonnative interactions are observed along the β path.

These projections further allow us to compare with Bagchi and coworkers' results. In particular, Figure 8 of ref. 3 indicates a path which begins with a slight increase in Φ_β coupled to an increase in $\overline{\beta}$ from approximately 0.6 to 1.1 nm. This initial step is followed by a 30° decrease in Φ_β coupled to a return to a dimer-like $\overline{\beta}$ (near 0.6 nm) as the dissociation progresses. The β strands only separate at the final step of their described mechanism. We interpret this to correspond to taking an initial step along the β path, then collapsing back to a near-dimer like β interface before proceeding along the α path. That said, the

projections of the α and β paths onto Bagchi and coworkers' coordinates do not fall precisely on top of their minimum free energy path (Supplemental Figure 2.17). By explicitly probing the multipathway nature of the dissociation, our results reveal the homogeneity of rotation profiles depending on dissociation path.

Water solvates key interfacial residues as dissociation progresses. Solvent plays a key role in protein association/dissociation processes. Moreover, time resolved X-ray scattering data suggest at least one intermediate in the insulin dimer dissociation that involves quick solvent uptake by a species with dimer-like secondary structure, coincident with a slight increase in molecular volume [91]. To investigate the possible presence of a similar feature in our simulations, we defined three collective variables; in order of increasing specificity, they are (i) the total molecular volume, (ii) the solvent accessible surface area (SASA) of eight residues that make up the hydrophobic core of the interface (Val^{B12}, Tyr^{B16} , Phe^{B24}, Tyr^{B26} on each monomer), and (iii) the number of native interfacial hydrogen bonds, which only form between Phe^{B24} and Tyr^{B26}. The total molecular volume, probed by the X-ray scattering, can reflect solvent uptake, but it can also correspond to large-scale conformational change. The core SASA reflects the solvation of the interface in general, while the hydrogen bonding between interfacial residues probes the loss of dimer-like secondary structure. Averages of these variables are seen in Figure 2.7A.

The red and red black dots in Figure 2.7A mark the positions of the structures in Figure 2.7B, which show characteristic solvations of the β and α interfaces, respectively. These structures represent low free energy states relative to the barriers along the β or α paths. Moving from the dimeric state to either of these positions, the total molecular volume increases by 0.3-0.4 nm^3. This small increase in molecular volume allows some solvation of the interface, increasing core SASA by between 0.8-1.2 nm^2. Along the α path, this solvation is at the α interface, while, along the β path, this solvation is at the β interface (Figure 2.7B). The former does not result in a loss of secondary structure as measured by STRIDE [138].

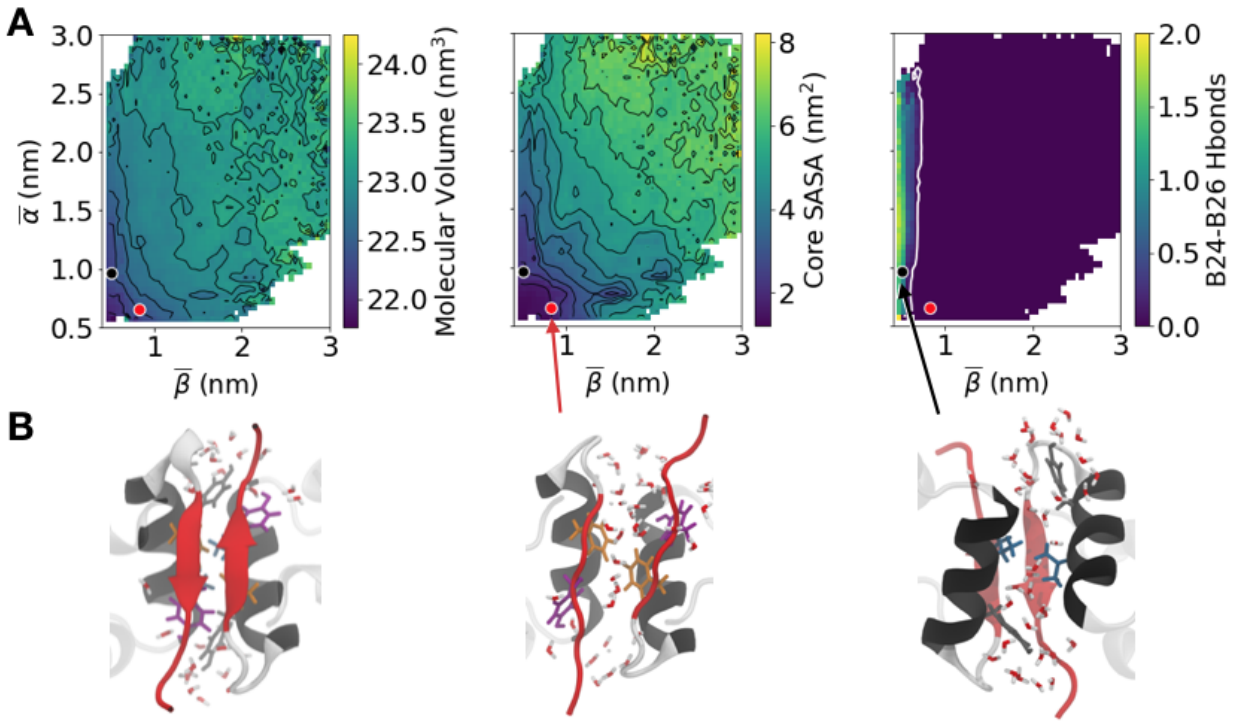


Figure 2.7: Characterizing solvation. (A) Averages of total molecular volume (left), core SASA (middle), and number of interfacial Phe^{B24}-Tyr^{B26} hydrogen bonds (right) as a function of $\overline{\alpha}$ and $\overline{\beta}$. Contours are every 0.2 nm^3 and 0.5 nm^2 for the molecular volume and SASA plots, respectively. The white contour on the right plot indicates where the number of hydrogen bonds drops to 2% of the average in the dimer. (B) Insulin structures showing the unsolvated dimer interface (left), the solvation of the β interface (middle), and the solvation of the α interface (right). The locations of these structures are marked in (A).

By contrast, we do see a distinct loss of β sheet content along the β path. Specifically, as $\overline{\beta}$ increases, the Phe^{B24} and Tyr^{B26} hydrogen bonds across the interface are replaced with ones to solvent (white contour in the right panel in Figure 2.7A and Supplemental Figure 2.18), signaling the loss of the interfacial β sheet.

One would expect that the loss of the interfacial β sheet and the solvation of the hydrophobic core are highly correlated because the separation of the β strands would allow water to penetrate between the monomeric units. We observe this to be the case when the α helices are already partly separated ($\overline{\alpha} > 1.0$ nm). The breaking of the interfacial hydrogen bonds, represented by the white contour in the right panel of Figure 2.7A, occurs in the same area of the collective variable space as the rapid solvation of the hydrophobic core,

represented by the tight black contours in the middle panel of Figure 2.7A. This area, where $\overline{\beta} \approx 0.75$ nm, is also co-located with the shoulder in the PMF mentioned earlier (Figure 2.5), suggesting that the loss of β sheet content and concomitant solvation gives rise to a slight decrease in free energy. The stabilization that we observe is consistent with previous simulations of mutants of the insulin dimer which indicate that water can mediate β sheet interactions by forming hydrogen bonds that bridge between residues [134]. The existence of this shoulder is also consistent with states involved in the dewetting transition seen in ref. 92, in which the center-of-mass separation of the monomers is 2 nm and there are a few water molecules at the interface.

However, when the α helices are in a near-native distance ($\overline{\alpha} < 1.0$ nm), the solvation of the hydrophobic core occurs after the loss of the interfacial β sheet. In this case, the protein-protein hydrogen bonds are broken when $\overline{\beta} \approx 0.65$ nm (white contour in the right panel of Figure 2.7A) but the hydrophobic core residues do not become significantly solvated until the β strands are even further separated, at approximately $\overline{\beta} = 0.9$ nm (middle panel of Figure 2.7A). This occurs within a low-free-energy trough of the PMF; in other words, the PMF only rises sharply as the core is solvated, not with the loss of hydrogen bonds. Evidence of these dynamics are further seen in the simulated IR results presented later, in which the peak absorption/emission doublet is red shifted with only a minimal loss in intensity.

As mentioned in the introduction, Chen and coworkers found evidence for a partially-solvated intermediate with dimer-like secondary structure, implying conserved interfacial β sheet character [91]. Although we find no distinct free energy basin that obviously corresponds to this intermediate, the initial partially solvated structures we observe along the α path are consistent with these data, as the interfacial β sheet is conserved. However, even with the loss of the β sheet along the β path, our simulations do not rule out the possibility of structures along the β path also being consistent with the X-ray scattering data. Specifically, the experimental data were collected at 316 K and 0.27 M DCl (pH $\approx$ 0) in a solution

of ethanol and water, while our simulations are for 303.15 K and pH 7 with only water as the solvent. Previous computational work suggests that ethanol appears to facilitate partial solvation of the β interface [3], so further simulations are needed to definitively interpret these T-jump X-ray scattering experiments.

The B-chain C-terminal segment detaches along the α path but not the β path. As noted above, there is extensive evidence that the B-chain C-terminal segment must detach from the B-chain α helix for insulin to bind its receptor.[53, 83, 90, 139, 140]. Detachment and partial unfolding have also been invoked to explain the diagonal elongation of features in equilibrium 2D infrared spectra of insulin at elevated temperatures [57]. Whether detachment and partial unfolding occurs during dissociation remains an open question.

These considerations, combined with the known partial disorder of the β turn and the B-chain C-terminal segment in the insulin monomer [57, 87–89], motivated our definition of two average angles to study intramonomeric unfolding during dimer dissociation: $\overline{\Psi_d}$, which measures detachment of the B-chain C-terminal segment from the B-chain α helix, and $\overline{\Psi_t}$, which characterizes the disorder of the β turn. Results for $\overline{\Psi_t}$ are discussed in the Supplemental Information (Supplemental Figure 2.19); here we focus on $\overline{\Psi_d}$. $\overline{\Psi_d}$ is the angle between the C_α atoms of Arg^{B22}, Phe^{B24}, and Tyr^{B26} (Supplemental Figure 2.20), measured for each monomeric unit and then averaged. $\overline{\Psi_d} = 180^\circ$ indicates an attached structure, because a flat β strand tucks against the B-chain α helix. In contrast, $\overline{\Psi_d} = 90^\circ$ indicates a structure that is almost completely detached, with the B-chain C-terminal segment bent away from the α helix. This detachment is also coupled to the solvation of Gly^{A1}-Val^{A3}, which are involved in binding to the insulin receptor [55] (Supplemental Figure 2.21). Examples of monomeric structures with attached and detached B-chain C-terminal segments are shown in Figure 2.8A. Figure 2.8B shows the full dimer view of the same detached intermediate with $\Psi_d = 140^\circ$ to illustrate how the detachment of the C-terminal segment allows for the nonnative interaction of Pro^{B28}-Ala^{B30} and $Tyr^{B'16}$, the same nonnative interaction shown

in Figure 2.6C.

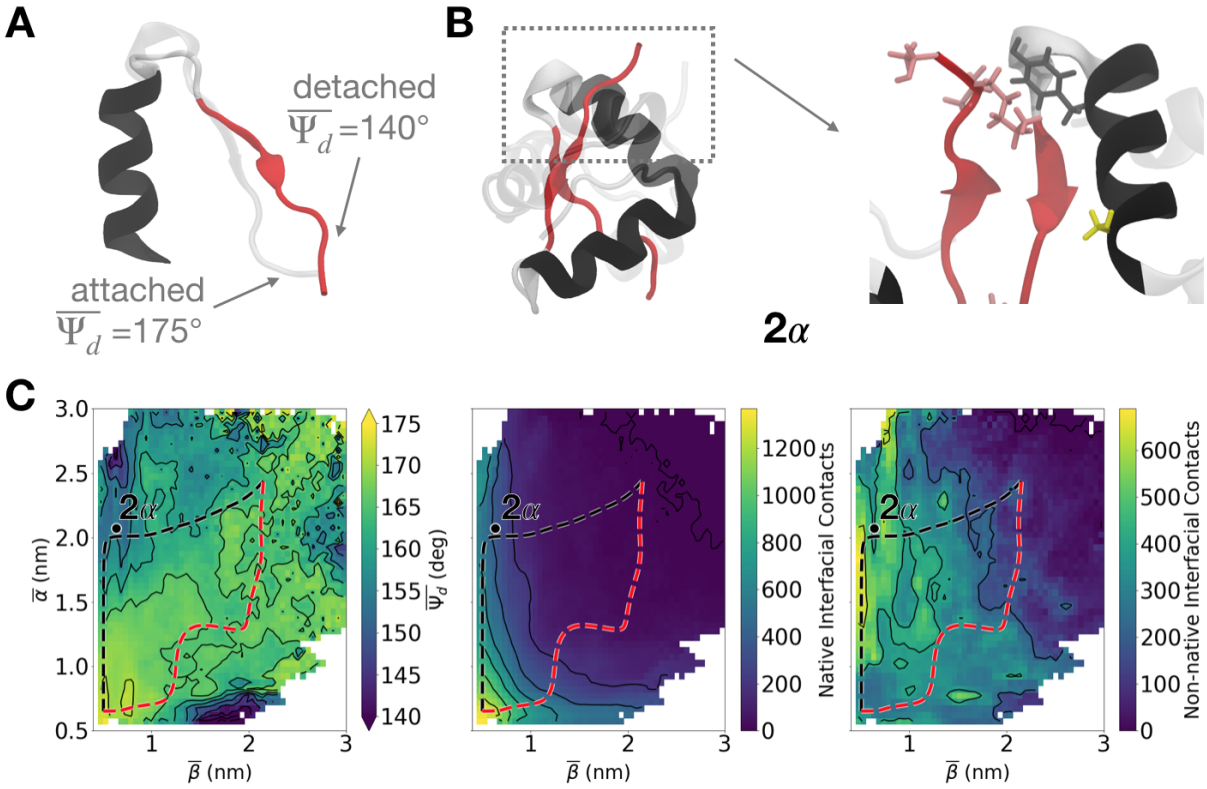


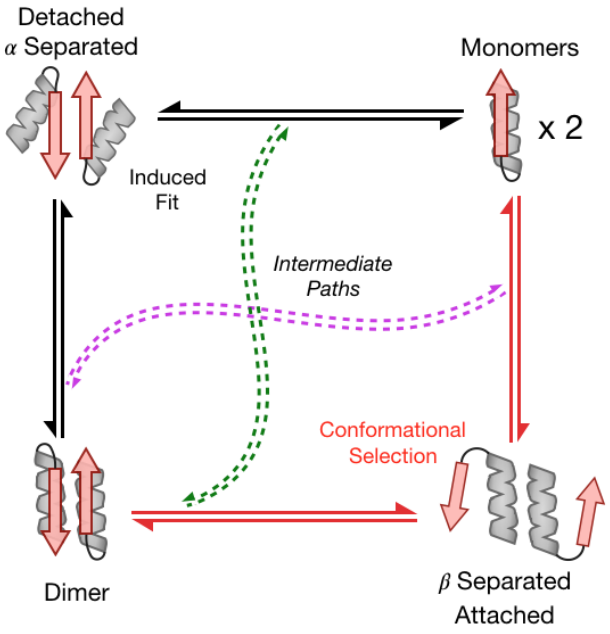
Figure 2.8: Characterizing detachment. (A) A representative monomeric structure contrasting attached and detached B-chain C-terminal segments. (B) Structural depiction of how the detachment of the B-chain C-terminal segment allows for continued nonnative interactions between Pro^{B28}-Ala^{B30} and $\text{Tyr}^{\text{B'16}}$. (C) (Left) Average of $\overline{\Psi_d}$ as a function of $\overline{\alpha}$ and $\overline{\beta}$ with black contour lines shown every 5°. (Middle) Number of native non-hydrogen atom native interfacial contacts and (right) non-hydrogen atom nonnative interfacial contacts (cutoff 7 Å), with contour lines shown every 200 contacts. On all graphs, the α (black) and β (red) paths are shown, as is the location of structure 2α shown in (B).

The average value of Ψ_d as a function of $\overline{\alpha}$ and $\overline{\beta}$ is plotted in the left panel of Figure 2.8C. The dimer state corresponds to $\overline{\Psi_d} = 172°$, consistent with the β strand being attached. $\overline{\Psi_d}$ decreases significantly along the α path; the latter half of the path ($\overline{\alpha} > 1.5$nm) consists mainly of structures in which the B-chain C-terminal segment is detached ($\overline{\Psi_d} < 165°$). This behavior contrasts with the β path, in which the minimum value of $\overline{\Psi_d}$ is $\overline{\Psi_d} = 167°$ (we neglect the low values of $\overline{\Psi_d}$ in the lower right corner of Figure 2.8C because that region is very high in free energy; see Figure 2.5). Overall, while we observe some detachment along

the β path, it is much more pronounced along the α path. It is also worth noting that in the monomeric region identified earlier ($\overline{\beta} > 2.0$ nm and $\overline{\alpha} > 2.2$ nm) the detachment is less pronounced than at intermediate stages of the α path. However, the monomeric region has more variability in detachment angle than in the dimeric region, consistent with previous results showing a limited amount of disorder in the C-terminal segment of the B chain [57, 87–89].

Along the β path, instead of unfolding, the β sheets separate and the monomers drift away from one another, forming a diverse set of non-specific, nonnative interfacial contacts, as in structure 2β in Figure 2.5. The numbers of native and nonnative interfacial contacts are plotted in the center and right panels of Figure 2.8C, respectively. Along the β path (red), the native contacts are almost completely broken as $\overline{\beta} > 1.0$ nm, although a limited number of nonnative contacts persist as the dissociation proceeds further. Similarly, along the α path (black), we also see the formation of nonnative contacts coupled to the breaking of native contacts, consistent with the side-chain interactions discussed previously (Figure 2.8B).

Finally, we note that T-jump 2D amide-I IR spectroscopy experiments in 20% ethanol indicated two contributions to the dimer dissociation process: melting of the dimer β sheet observed between 250-1000 μs, and a 5-150 μs process that was assigned to α helix disordering [58]. These timescales thus suggest that unfolding occurs before the loss of the interfacial β sheet. In our (aqueous) simulations, we do not observe any loss of α helix content, although it may be possible that helix rotation could give rise to such a signal. Instead, the unfolding that we observe is restricted to detachment of the B-chain C-terminal β strand and disorder of the β turn (see Supplemental Figure 2.19), which primarily occurs along the α path. Furthermore, the detachment along the α path (following the black path in the left panel of Figure 2.8C) starts to occur before the loss of the interfacial β sheet (right panel of Figure 2.7A). The α path, which exhibits monomeric unfolding in the form of C-terminal detachment



l

Figure 2.9: A schematic representation of the pathways of insulin dimer dissociation/association, oriented as in Figure 2.5, and labeled to describe the limiting behaviors of coupled folding and binding. The α path is depicted by the black solid double arrows, and the β path is depicted by the red solid double arrows. Intermediate paths, shown by the dashed double arrows, are colored as in Supplemental Figure 2.11.

while maintaining dimer-like secondary structure, is thus consistent with the T-jump data gathered by Tokmakoff and coworkers [58]. Moreover, this detached state also corresponds to an increase in molecular volume of between 0.6 and 0.8 nm^3, which is consistent with the evidence from Chen and coworkers for a second intermediate that corresponds to a large increase in molecular volume while maintaining secondary structure [91].

The α and β path correspond to induced fit and conformational selection. In this section, we connect the intramonomeric detachment and core solvation with the intermonomeric rotations discussed earlier to characterize the various mechanisms of the insulin dimer dissociation, making explicit comparisons to the induced fit and conformational selection models of coupled folding and binding. A summary is shown in Figure 2.9.

The α path, as mentioned before, initially consists of rotations at both the α and β interfaces (Figure 2.6A); the alpha helices twist away from each other, forming nonnative

contacts between Ser^{B9}:$Ser^{B'9}$ and Pro^{B28}-Ala^{B30}:$Tyr^{B'16}$ (Figure 2.6C). This initial α helix separation and rotation increases the SASA of the hydrophobic core by $\sim$1.5 nm^2, while further α helix separation (between $\overline{\alpha}$ = 1.0 nm and $\overline{\alpha}$ = 1.35 nm) leads to very little additional core solvation (Figure 2.7A). This solvation correlates with the free energy along the α path, which also sharply increases until $\overline{\alpha}$ = 1.0 nm, where it levels off around 8 kcal/mol until $\overline{\alpha}$ = 1.35 nm (Figure 2.5). There is then another free energy barrier of 2 kcal/mol when $\overline{\alpha} \in [1.35, 2.0]$ nm, which correlates well with the β sheet starting to detach from the core (Figure 2.8C). This detachment, while partially maintaining the nonnative interactions between Pro^{B28}-Ala^{B30}:$Tyr^{B'16}$, sacrifices native contacts and exposes Gly^{A1}-Val^{A3} to the solvent (Supplemental Figure 2.21). The last barrier of 4 kcal/mol correlates to the breaking of the interfacial β sheet hydrogen bonds and the subsequent separation of the β strands. Structurally this final step is heterogeneous, involving varying amounts of detachment, rotation, and sliding of the β strands as $\overline{\beta}$ increases.

In terms of the association process, the α path corresponds to the induced fit model of coupled folding and binding. Following the black trace in Figure 2.8C from monomer to dimer (top right to bottom left), association is initiated by the B-chain C-terminal strands detaching and encountering one another. The β interface becomes structured; then, the α helices are recruited into the interface, leading to a dimer-like structure. The monomers only adopt their structures in the dimer upon association.

The β path first couples the initial separation of the β strands to their rotation (Figure 2.6A). This correlates to the sudden increase in free energy from 0 to 14 kcal/mol as $\overline{\beta}$ increases from 0.5 to 1.2 nm. The subsequent steps of the dissociation correspond to the two monomers, with conformations similar to the ones found in the dimer, drifting away from one another. Specifically, as the dissociation proceeds, a diverse set of nonnative interfaces are formed that are structurally close to the native interface, but involve minimal detachment of the B-chain C-terminal segment (Figure 2.8C). This is consistent with the relatively flat

free energy trace when $\overline{\beta} > 1.2$ nm (Supplemental Figure 2.13), as all of these near-native, folded structures are similar in terms of solvation and protein-protein interactions.

In terms of the association process, the β path corresponds to the conformational selection model of coupled folding and binding. Following the red trace in Figure 2.8C from monomer to dimer, the monomers first adopt structures like those in the dimer, with attached B-chain C-terminal segments; they then approach one another, forming a variety of nonnative interfacial contacts that eventually collapse to the dimer-like set of native contacts. We note that the β path is consistent with the unbiased association trajectories simulated by Shaw and coworkers, who found that insulin dimer association is characterized by monomers, largely folded into their conformations in the dimer, forming near-native interfaces that eventually collapse to the native dimer interface [10]. Additionally, to remind the reader, the free energy profile, interfacial rotations, and monomeric unfolding observed along the α path are all consistent with the free energy minimum path discussed by Bagchi and coworkers [2].

As mentioned previously, the α and β paths are limiting mechanisms from a broader ensemble of pathways (Supplemental Figure 2.11), a few of which are shown schematically in Figure 2.9. A variety of intermediate pathways are thus possible, and our data suggest that these pathways combine the behaviors observed along the limiting paths. For example, the green path in Figure 2.9 initially exhibits a partial separation and solvation of the β interface, but then otherwise exhibits behaviors similar to the those found along the α path. Specifically, the α helices twist away from one another, the B chain C-terminal strands detach from the nearby α helices and only then fully separate. All this occurs with a partially solvated β interface. As mentioned in a previous section, both partially solvated and partially unfolded species along this path are consistent with previous experimental studies [58, 91]. In contrast, the pink path in Figure 2.9 initially follows the α path, then switches to exhibit behaviors more characteristic of the β path. The α helices partially twist away from one another, and then the B chain C-terminal strands separate while still attached to the α

helices. This intermediate path is itself consistent with the biased association trajectories from Shaw and coworkers [10]. These trajectories primarily exhibit interfacial rotations and nonnative interactions (namely, $\mathrm{Ser}^{\mathrm{B9}}$:$\mathrm{Ser}^{\mathrm{B'9}}$) similar to the α path, but notably lack any substantial B-chain C-terminal detachment. Thus, the collective variables we identified enabled us to obtain an ensemble of paths that encompass the diversity seen in previous studies [2, 10, 58, 91].

Simulations enable determination of optimal isotopic labeling sites for infrared spectroscopy. Even though the α and β paths are only limiting cases, most of the paths in Supplemental Figure 2.11 initially follow one of these limiting paths. As discussed previously, the initial steps along the α and β paths correspond to solvation of the α and β interfaces, respectively. In particular, the α path consists first of the solvation of the α interface, followed by the solvation of the β interface; the β path reverses this ordering. Thus, these simulation results can be further investigated by experimental techniques that can resolve residue-level solvation.

Fourier-transform (FT) and two-dimensional (2D) amide-I infrared (IR) spectroscopies are useful for studying protein secondary structure and solvation because they are sensitive probes of the hydrogen bonding of the carbonyl groups in protein backbones. In particular, one typical hydrogen bond to an amide carbonyl causes a redshift of about 16 cm^{-1} [124]. This means the location of the carbonyl stretch is sensitive to the number and strength of hydrogen bonds made by the backbone carbonyls, which includes hydrogen bonds associated with both secondary structure and protein-solvent interactions. Moreover, one can isotopically label specific amide carbonyl groups with $^{13}\mathrm{C}^{18}\mathrm{O}$, redshifting their vibrations by 65 cm^{-1} to isolate them from the other amide vibrations. Both 2D and 1D IR spectra can be generated through molecular modeling, effectively "mapping" the classical variables from molecular trajectories into a quantum-mechanical Hamiltonian (see ref. 124 for further review of 2DIR methods and their simulation).

These simulated spectra have been used to interpret both equilibrium and T-jump measurements of protein folding [141]. Furthermore, recent spectral simulation work from Meuwly and coworkers has shown that isotope labeled spectra for both the insulin dimer and monomer are qualitatively sensitive to the number of waters hydrating the labeled backbone carbonyl group [142]. They note, however, that no one structural feature is particularly strongly correlated to spectroscopic behavior, which is instead sensitive to the rapidly fluctuating environment around each backbone oscillator. When considering how best to experimentally probe the dynamics of the dissociation, it is thus difficult to *a priori* suggest the best sites to label. In previous sections, we discussed our results in the context of previous equilibrium [57] and T-jump [58] experiments of unlabeled insulin. Below, we combine our umbrella sampling results with additional simulations of IR spectra to propose new sites for isotopic labeling, with a view toward achieving residue-specific characterization of the dimer dissociation mechanism.

We expected the most promising sites for isotopic labeling to be at the dimer interface, as the participating residues exhibit large changes in SASA upon dissociation (Figure 2.7). Owing to the computational cost of 2DIR simulations, we first simulated FTIR spectra along both limiting paths for all possible constructs with a single interfacial residue isotopically labeled (between Ser^{B9} and Ala^{B30}). This process, summarized in the Supplemental Information (see Supplemental Figure 2.12), revealed that the isotopic labels that produce simulated spectra most sensitive to solvation are those on the residues in the previously identified hydrophobic core. For clarity, we focus on two specific residues: Phe^{B24}, at the β interface, and Glu^{B13}, at the α interface. A construct with the former label was studied in refs. 142 and 126, while, to the best of our knowledge, a construct with the latter label was not previously synthesized. For each label, we identify and simulate 2DIR for three states: the dimeric state, the monomeric state, and the solvated state (Figure 2.10). The solvated states for the Phe^{B24} and Glu^{B13} labels represent partially dissociated species with solvated

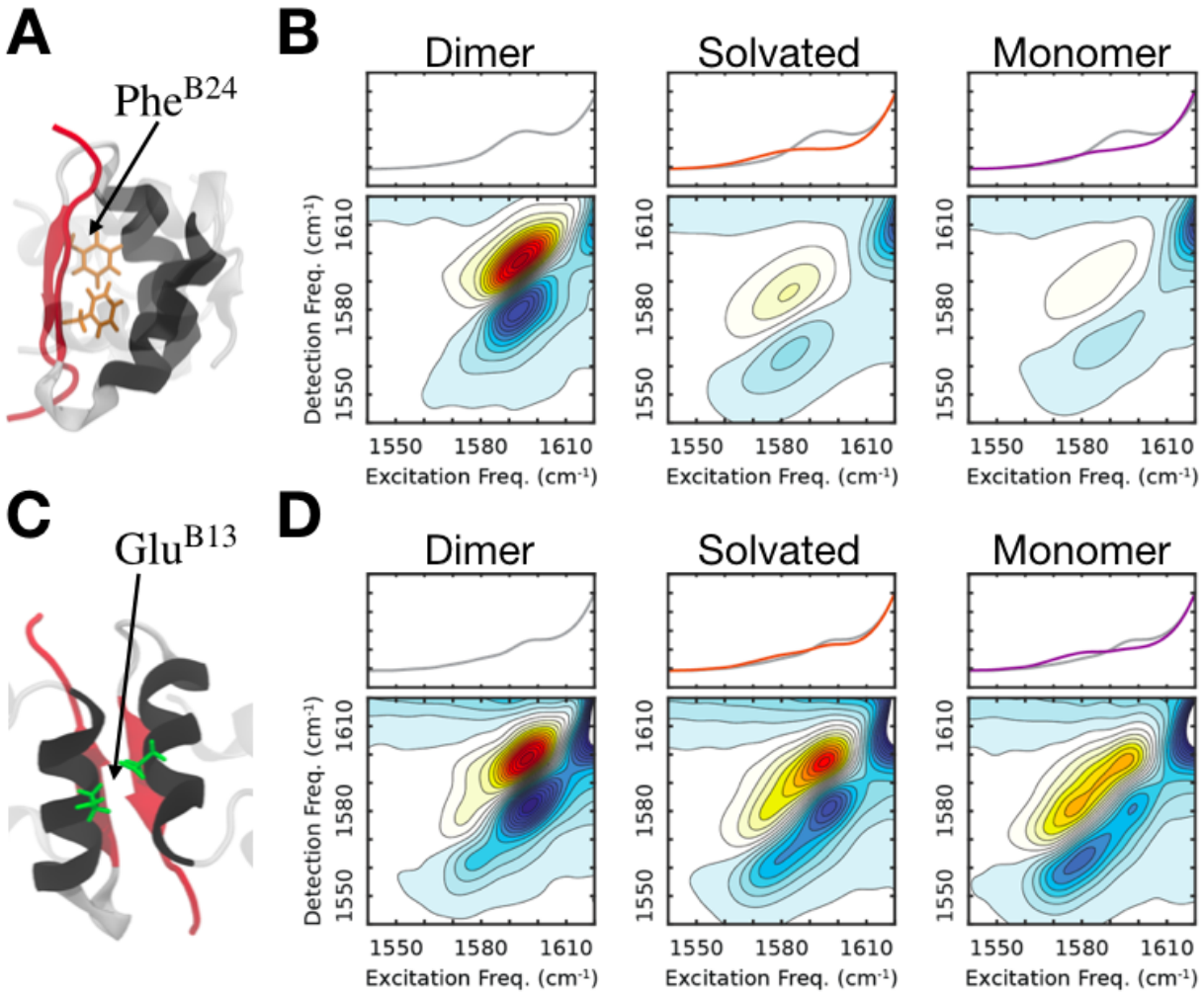


Figure 2.10: Simulated IR spectra for selected isotopically labeled constructs. (A) Dimeric structure showing the Phe^{B24} side chain, which was isotopically labeled on its backbone carbonyl. (B) Simulated 2DIR spectra of the Phe^{B24}-labeled dimer (left), solvated species (middle), and monomer (right). Intensities are normalized using the peak intensity of the dimer spectrum, with the contours spaced by 7.5%. (C & D) Similar structures/spectra, but for the Glu^{B13} -labeled insulin. In both cases, the spectra for the solvated species were generated from structures along both the α and β paths.

β and α interfaces, respectively. The process of defining these solvated states is described in the Supplemental Information (Supplemental Figure 2.12).

For insulin labeled at Phe^{B24} (at the β interface, shown in Figure 2.10A), there is a strong peak at 1595 cm^{-1} in the dimer spectrum. This feature decreases in intensity, becomes redshifted to 1582 cm^{-1}, and broadens significantly along the diagonal as the carbonyl

group of Phe^{B24} becomes solvated (Figure 2.10B), consistent with previous measurements on the monomer [126]. Microscopically, we interpret the changes to reflect the hydrogen bond between Phe^{B24} and $Tyr^{B'26}$ breaking and the backbone becoming solvated. This behavior was observed once $\overline{\beta} \gtrsim 0.9$ nm along both the α and β paths (Supplemental Figure 2.12), suggesting that 2DIR experiments using this label should serve as a probe for the solvation of the β interface regardless of the dissociation mechanism. For most regions of the collective variable space, the redshifting and broadening of the peak occur together, but along the β path at $\overline{\beta} \in [0.75, 0.9]$ nm, the redshifting precedes the broadening. This corresponds to the interfacial hydrogen bonds breaking prior to solvation of the Phe^{B24} backbone carbonyl (Figure 2.7). Similarly, for insulin labeled at Glu^{B13} (at the α interface, shown in Figure 2.10C), we see a peak at 1595 cm^{-1}, and it becomes redshifted to 1579 cm^{-1} and broadens along the diagonal as the site containing the label becomes solvated (Figure 2.10D). Again, this behavior was observed along both the α and the β paths (Supplemental Figure 2.12).

By using temperature-jump IR spectroscopy and singular value decomposition as in ref. 58, one can decompose the time series of IR spectra into time-dependent weights of specific spectral components that correlate with molecular features. Our results show that the primary protein component for isotope-labeled spectra would correspond to interfacial solvation of the labeled residue. By measuring the contribution of this component to the overall spectra as a function of time, one can determine the characteristic timescale(s) of solvation for the labeled residue. Thus, by using a construct with the Glu^{B13} isotopic label in a T-jump IR experiment, one should be able to determine a distribution of timescales for α interface solvation by monitoring the relative contribution of the corresponding spectral component. Similarly, one can separately use T-jump IR measurements on a construct with the Phe^{B24} isotope label to measure the distribution of timescales for β interface solvation. By comparing these two distributions, one could determine the order of events during dissociation and in turn if one limiting mechanism or the other predominates; broad, temporally-overlapping

distributions for α helix and β sheet interfacial solvations would be consistent with the diverse ensemble of energetically similar paths we see here. However, slight differences in these distributions could provide insights into preferred pathways, or one particular mechanism could dominate due to kinetic effects not explicitly considered by our simulations.

2.4 Conclusions

A key step in insulin function is its dissociation from dimer to monomer form, and an understanding of this process can aid in design of molecular analogs with desired properties. The practical importance of understanding insulin dimer dissociation has in turn made this system a paradigm for study of complex molecular recognition reactions. Here, we assembled a computational pipeline of methods for the study of complex molecular dynamics and applied it to understanding how the insulin dimer dissociates. This pipeline enabled us to identify collective variables that could promote dissociation with a minimum of monomeric unfolding. The resulting simulations revealed a previously unappreciated diversity of dissociation pathways with comparable free energy barriers. The limiting pathways, in which either the interfacial α helices separate and are solvated first (the α path) or the interfacial β strands separate and are solvated first (the β path), correspond to induced fit and conformational selection mechanisms when considering them from the perspective of association. The similarities in barrier heights for qualitatively different pathways makes clear the importance of achieving chemical precision in the simulations, and an error estimator that we recently introduced allowed us to do so efficiently.

Along the two limiting pathways, the elements at the dimer interface rotate relative to each other as the monomers come apart. Along the energetically preferred α path, the rotation allows the formation of nonnative interactions involving residues on the interfacial α helices; this enables the C-terminal segment of insulin B chain to detach from the nearby interfacial α helix, thereby coupling monomeric unfolding to unbinding. No such unfolding

is observed along the β path. The diversity of paths that we observe encompasses paths previously observed in simulation studies [2, 10]; in this sense, our work reconciles seemingly discordant results in the literature.

Molecular simulations can guide the design and interpretation of experiments. The pathways that we observe provide a microscopic picture of a partially unfolded intermediate with conserved secondary structure previously suggested by T-jump experiments, though differences in conditions between the simulations and experiments make this picture tentative. With a view towards obtaining additional experimental constraints on the mechanism, we use the simulations to test possible sites for isotopic labeling for IR spectroscopy experiments. We predict that two sites in particular, $\mathrm{Phe}^{\mathrm{B24}}$ and $\mathrm{Glu}^{\mathrm{B13}}$, should enable sensitive characterization of the solvation of the interfacial α helices and β strands. Our results thus provide insight into how to pursue the next generation of experiments to achieve residue-level resolution of the dissociation mechanism.

2.5 Supplementary Material

EMUS Asymptotic Error With Replica Exchange. To derive the expression for asymptotic variances of averages given REUS data, we follow the following procedure. The notation corresponds to that in ref. 93.

Assumption 2.5.1. *We assume:*

1. *Sampling over all windows is performed by a single Markov chain Ξ whose state space is L copies of the molecular phase space.*

2. *A central limit theorem holds for the convergence of each sample average $\bar{v}$ estimated in EMUS to its true value v, with asymptotic covariance matrix Σ. The entries in this matrix are given by*

$$\Sigma_{ij} = \frac{1}{2}\int_{-\infty}^{\infty} \mathbb{E}\left[v_i(X_t)v_j(X_0)\right] dt. \tag{2.6}$$

3. The biasing functions ψ_i are chosen so that F is irreducible.

Note that if we sort the averages v by the window in which the average was calculated, then unlike in ref. 93, the matrix Σ is not block diagonal. Rather it has nonzero off-diagonal blocks:

$$\Sigma = \left[\begin{array}{c|c|c} \Sigma_{11} & \Sigma_{12} & \cdots \\ \hline \Sigma_{21} & \Sigma_{22} & \cdots \\ \hline \vdots & \vdots & \ddots \end{array}\right].$$

Under these assumptions, we can still apply Lemma VII.2 in ref. 93 to derive a central limit theorem for EMUS with replica exchange.

Theorem 2.5.2. *As in ref. 93, let B be a quantity of interest and $dB/d\bar{v}$ be the partial derivative of B with respect to each sampled average. Under the assumptions above,*

$$\sqrt{N}\left(B\left(\bar{v}\right) - B\left(v\right)\right) \xrightarrow{\mathrm{d}} N(0, \sigma^2), \tag{2.7}$$

where

$$\sigma^2 = \frac{\partial B}{\partial \bar{v}}^T \Sigma \frac{\partial B}{\partial \bar{v}}. \tag{2.8}$$

To estimate σ^2 using sampled data, we write

$$\begin{aligned} \sigma^2 &= \sum_{i,j} \frac{\partial B}{\partial \bar{v}_i} \Sigma_{ij} \frac{\partial B}{\partial \bar{v}_j} && (2.9) \\ &= \sum_{i,j} \frac{\partial B}{\partial \bar{v}_i} \int_{-\infty}^{\infty} \mathbb{E}\left[v_i(X_t) v_j(X_0)\right] dt \frac{\partial B}{\partial \bar{v}_j} \\ &= \int_{-\infty}^{\infty} \mathbb{E}\left[\left(\sum_i \frac{\partial B}{\partial \bar{v}_i} v_i(X_t)\right)\left(\sum_j \frac{\partial B}{\partial \bar{v}_j} v_j(X_0)\right)\right] dt && (2.10) \end{aligned}$$

This is the the integrated covariance of the trajectory $\sum_i \frac{\partial B}{\partial \bar{v}_i} v_i(X_t)$.

C-terminal detachment is correlated with solvation of the A-chain N-terminal segment and formation of nonnative contacts. The detachment of the C-terminal segment of the B chain (Figure 2.20), discussed extensively in the main text, is also correlated with the increase in SASA of Gly^{A1}-Val^{A3} (Figure 2.21). Gly^{A1}-Val^{A3} are residues that are in contact with Pro^{B28}-Ala^{B30} in the dimer, and these contacts are sacrificed as detachment occurs along the α path. This is consistent with the known binding behavior of the insulin monomer to the insulin receptor, as these Gly^{A1}-Val^{A3} residues are thought to form part of the binding interface [143, 144] and are not in contact with Pro^{B28}-Ala^{B30} when bound to the receptor [139, 145].

Beyond the loss of contacts between Pro^{B28}-Ala^{B30} and Gly^{A1}-Val^{A3} along the α path, one can also characterize the number of contacts that the B-chain C-terminal segment makes with other residues. A subset of these are shown in Figure 2.15. Namely, from the left panels, one sees that along the α path (black), the native contacts of Pro^{B28}-Ala^{B30} with $Gly^{B'20}$-$Gly^{B'23}$ and Gly^{A1}-Val^{A3} are lost. The rotation of the interface associated with both the α and β paths moves the B-chain C-terminal segment away from these both $Gly^{B'20}$-$Gly^{B'23}$ and Gly^{A1}-Val^{A3}.

The right plots, on the other hand, show that some specific nonnative interactions are formed along the α path, but not along the β path. Along the α path, the B-chain C-terminal segment starts to interact with the $Tyr^{B'16}$ on the α helix of the other monomer when $\overline{\alpha} \approx 1.15$ nm. In the same region, Ser^{B9} and $Ser^{B'9}$, serines on opposite interfacial helices, start to interact and form a nonnative contact. Interestingly, comparing this to the core SASA shown in the main text Figure 2.7A, one sees that the core SASA increases sharply after the Ser^{B9} residues lose contact with one another, around $\overline{\alpha} \approx 1.35$ nm. This is also the area where significant detachment starts to occur along the α path. It is possible that the combination of the interfacial rotations and detachment exposes the hydrophobic core of the dimer.

Relation to available insulin therapeutics. To replicate the biphasic activity of insulin *in vivo*, both fast-acting and slow-acting insulin therapeutics have been developed. Two of the most common fast acting therapeutics, lispro ($Pro^{B28}Lys^{B29} \rightarrow Lys^{B28}Pro^{B29}$, Humalog [146]) and aspart ($Pro^{B28} \rightarrow Asp^{B28}$, NovoLog [66]) involve mutation of C-terminal residues of the B chain to destabilize the dimer and favor the monomer. This presumably allows for more rapid and reliable uptake of glucose after injection, as the monomer is the species that binds to the receptor. These mutations have been hypothesized to reduce the interaction of Pro^{B28}-Ala^{B30} of one monomer with the β turn of the other monomer ($Gly^{B'20}$-$Gly^{B'23}$), either through reduced van der Waals attractive forces (lispro and aspart) or the addition of a repulsive electrostatic interaction (aspart) [51]. To the best of our knowledge, no computational work has yet been done to explore these hypotheses as they relate to the mechanism of dissociation. In the previous section, it was described how, along both pathways, the native contacts between Pro^{B28}-Ala^{B30} and both $Gly^{B'20}$-$Gly^{B'23}$ and Gly^{A1}-Val^{A3} are broken. It has been suggested [51, 146] that one or both of the therapeutic mutations could reduce the energetic penalty for breaking these native contacts, specifically the contacts with $Gly^{B'20}$-$Gly^{B'23}$. To explicitly investigate this in our simulations of wildtype insulin, the sum of interaction energies of Pro^{B28}-Ala^{B30} and both $Gly^{B'20}$-$Gly^{B'23}$ and Gly^{A1}-Val^{A3} was computed (Figure 2.16).

β turn angle and disorder of the β turn also varies between dissociation path. We defined the β turn angle, Φ_t, as the angle between the geometrical centers of the backbone atoms in Glu^{B13} , Gly^{B20}, and Phe^{B24} to characterize the β turn. A larger β turn angle corresponds to a wider turn. We also calculated the RMSD of the β turn from its conformation in the dimer by fitting the conserved α helix of the B chain to each monomeric unit, then measuring the RMSD to the solvated crystal structure. The average of both the β turn angle and the RMSD are shown in Figure 2.19.

The β turn angle increases initially and is much wider along the α path than along the β

path. The RMSD also increases as the α helices separate. This increase in RMSD happens before the β sheets are broken for the α path, and after the β sheets are broken for the β path. The different trends in β turn angle and RMSD between paths suggest that these residues on the β turn might be useful candidates for isotopic labeling for 2DIR experiments aimed at distinguishing dissociation pathways.

FTIR to identify labels and solvated species. Here we describe our simulations of FTIR spectra to map the effects of isotopic labels with considerably less computational cost than simulations of 2DIR spectra. As noted in the main text, we computed 50 such spectra along each of the two limiting paths, and each was converted into a difference spectrum by subtracting the unlabeled insulin spectrum at that point along the path from the labeled spectrum. The intensities of these difference spectra were converted into a colormap, with orange being a positive change, and blue being a negative change. This allows one to see the peaks associated specifically with the isotope label. These were then stacked and viewed as a function of path progress along both the α and β paths. Thus, when a red feature shifts to a lower frequency, this corresponds to a redshift in the simulated labeled spectra. These series of spectra, for the $\mathrm{Phe}^{\mathrm{B24}}$ and $\mathrm{Glu}^{\mathrm{B13}}$ isotope-labeled insulins, are shown in Figure 2.12. We simulated all possible constructs with a single interfacial residue isotopically labeled (between $\mathrm{Ser}^{\mathrm{B9}}$ and $\mathrm{Ala}^{\mathrm{B30}}$). The labels on our identified α and β contacts showed the most significant redshifts; of these $\mathrm{Phe}^{\mathrm{B24}}$ and $\mathrm{Glu}^{\mathrm{B13}}$ exemplified these effects.

From these spectra, we identified the ranges of path progress where the redshift (and correlated loss in intensity) first occurs, presumably due to either solvation or change in local secondary structure. These areas are marked areas in Figure 2.12. The left panels show the SASA of the backbone carbonyl group associated with the isotopically labeled residue (either $\mathrm{Phe}^{\mathrm{B24}}$ or $\mathrm{Glu}^{\mathrm{B13}}$). One sees that these marked regions begin once the SASAs for the corresponding backbone carbonyl groups increase from the dimer (with SASAs near 0) to 60% of the monomeric average. One can thus conclude that the redshifting and peak

broadening in these FTIR spectra correlates with the solvation of the backbone.

Table 2.1: Types and descriptions of contacts with high differential SASA between dimer and monomer, with primes indicating intermonomer interactions.

Class	Residue Pairs	Description
β Contacts	Phe^{B24}:$Tyr^{B'26}$, Phe^{B25}:$Phe^{B'25}$, Tyr^{B26}:$Phe^{B'24}$	Phe^{B24} and Tyr^{B26} form hydrogen bonds at the β sheet interface side chains make up part of the hydrophobic core of the dimer. Phe^{B25} is also part of the β sheet, but its side chain is not part of the hydrophobic core.
α Contacts	Ser^{B9}:$Glu^{B'13}$, Ser^{B9}:$Tyr^{B'16}$, Val^{B12}:$Tyr^{B'16}$, Glu^{B13}:$Ser^{B'9}$, Glu^{B13}:$Glu^{B'13}$, Tyr^{B16}:$Ser^{B'9}$, Tyr^{B16}:$Val^{B'12}$	Ser^{B9}, Val^{B12}, Glu^{B13} , Tyr^{B16} are residues from the α-helical portion of the B chain that are at least partially buried in the dimer. Val^{B12} and Tyr^{B16} are part of the hydrophobic core.
Cross Contacts	Val^{B12}:$Phe^{B'24}$, Tyr^{B16}:$Tyr^{B'26}$, Phe^{B24}:$Val^{B'12}$, Tyr^{B26}:$Tyr^{B'16}$	Buried residues across the dimer interface that pair α helices with β strands and vice versa.
Intermittent Contacts	Val^{B12}:$Glu^{B'13}$, Glu^{B13}:$Val^{B'12}$, Glu^{B21}:$Pro^{B'28}$, Gly^{B23}:$Thr^{B'27}$, Thr^{B27}:$Gly^{B'23}$, Pro^{B28}:$Glu^{B'21}$	Thr^{B27} and Pro^{B28} are relatively disordered residues adjacent to the β-sheet that form intermittent contacts with β turn residues Glu^{B21} and Gly^{B23}. Val^{B12} and Glu^{B13} are on the interfacial α helices and also only interact intermittently.
Internal Contacts	Val^{B12}:Tyr^{B26}, Glu^{B13}:Phe^{B24}	Internal contacts within a monomer.

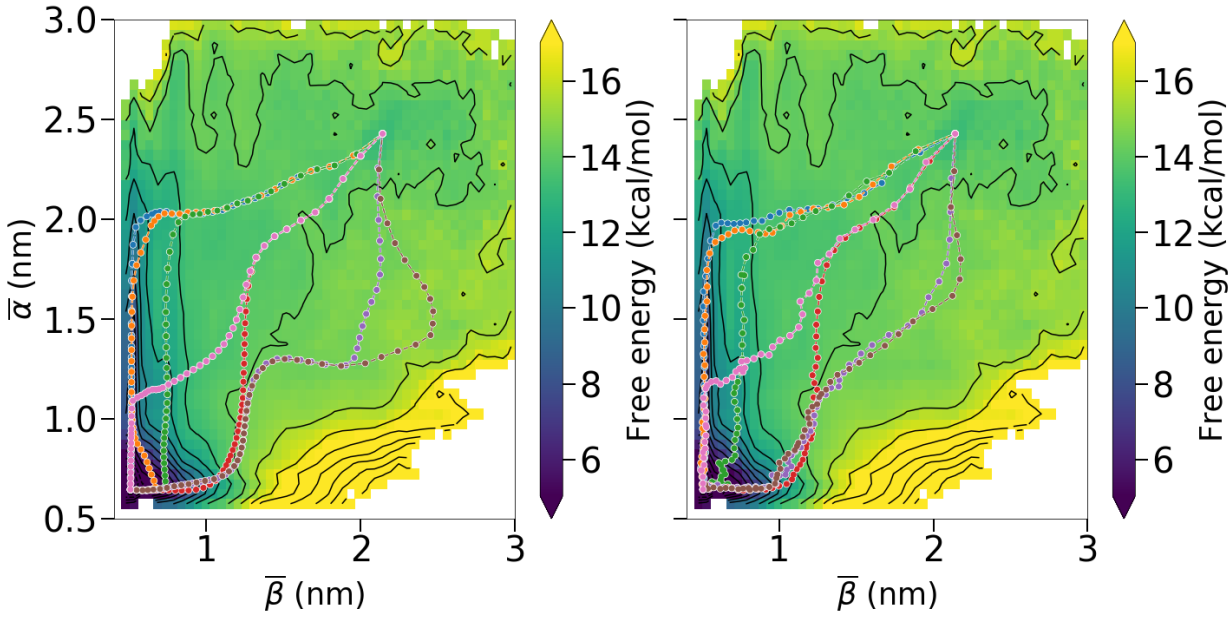


Figure 2.11: Using the string method to confirm stability of dissociation paths. (Left) Minimum free energy paths identified by using the LFEP algorithm from ref. 1. All the paths shown have a maximum free energetic barrier within 4 k_BT of the others. These paths were used as initializations for the string method. (Right) The converged strings after further refinement with the string method. Comparing with the left panel, the orange path has collapsed to the α path, and the purple and brown paths have shifted slightly from their initial positions. Note that while the purple path, which represents the β path, has shifted slightly in CV space, this does not affect the molecular trends discussed in the main text. Overall, the stability of these paths provides evidence that the averaging reducing the 10 interfacial distance dimensions to $\overline{\beta}$ and $\overline{\alpha}$ does not sacrifice mechanistic information.

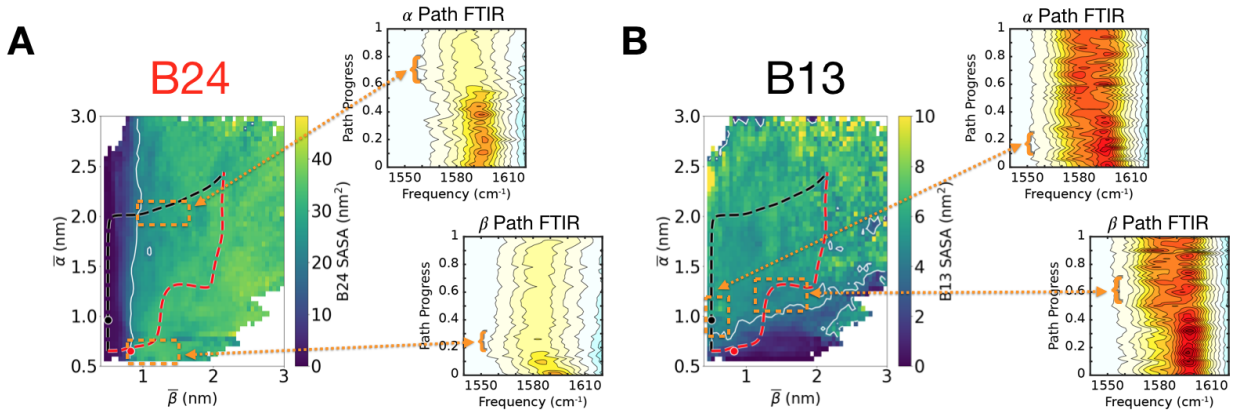


Figure 2.12: Relation between backbone solvation and simulated FTIR spectra. (A) The backbone carbonyl SASA for Phe^{B24}, with the white contour showing where the SASA increases to 60% of the monomeric average (contour at 19.7 nm^2, monomeric average at 32.8 nm^2). The two Phe^{B24}-labeled FTIR plots correspond to simulations along the α (top) and β paths (bottom), with the y-axis being the path progress, or the fractional distance along each specific path. Each value of y corresponds to one FTIR simulation - a FTIR spectra was generated by combining 20 simulations started from a point at that specific value of path progress, and a difference was taken between that isotope-labeled simulated spectrum and the corresponding unlabeled simulated spectrum. 50 such difference spectra were created, and stacked such that the color represents the intensity of the difference. For each path, the spectra that first demonstrated the expected redshift were identified, and then those areas of path space were selected as the solvated species for future study (orange boxes on the SASA graph). (B) Similar graphs for the Glu^{B13} -labeled insulin, showing the backbone carbonyl SASA for Glu^{B13} . The contour is again shown where the SASA increases to 60% of the monomeric average (contour at 4.1 nm^2, monomeric average of 6.8 nm^2).

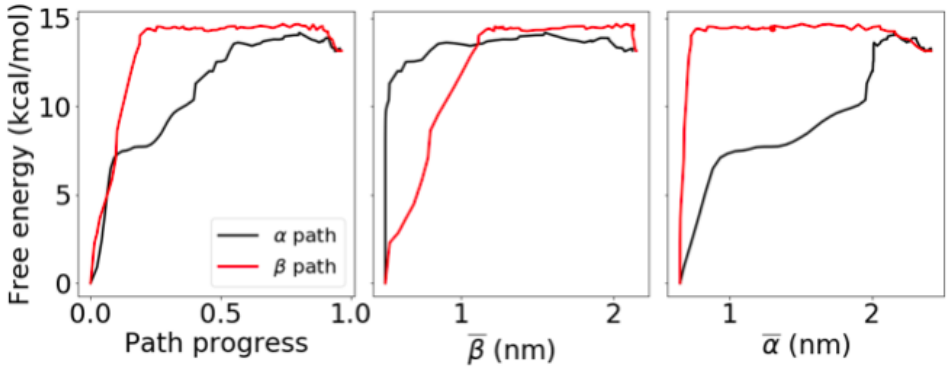


Figure 2.13: 1D cuts of our α and β paths as functions of path progress (left), $\overline{\beta}$ (middle), and $\overline{\alpha}$ (right).

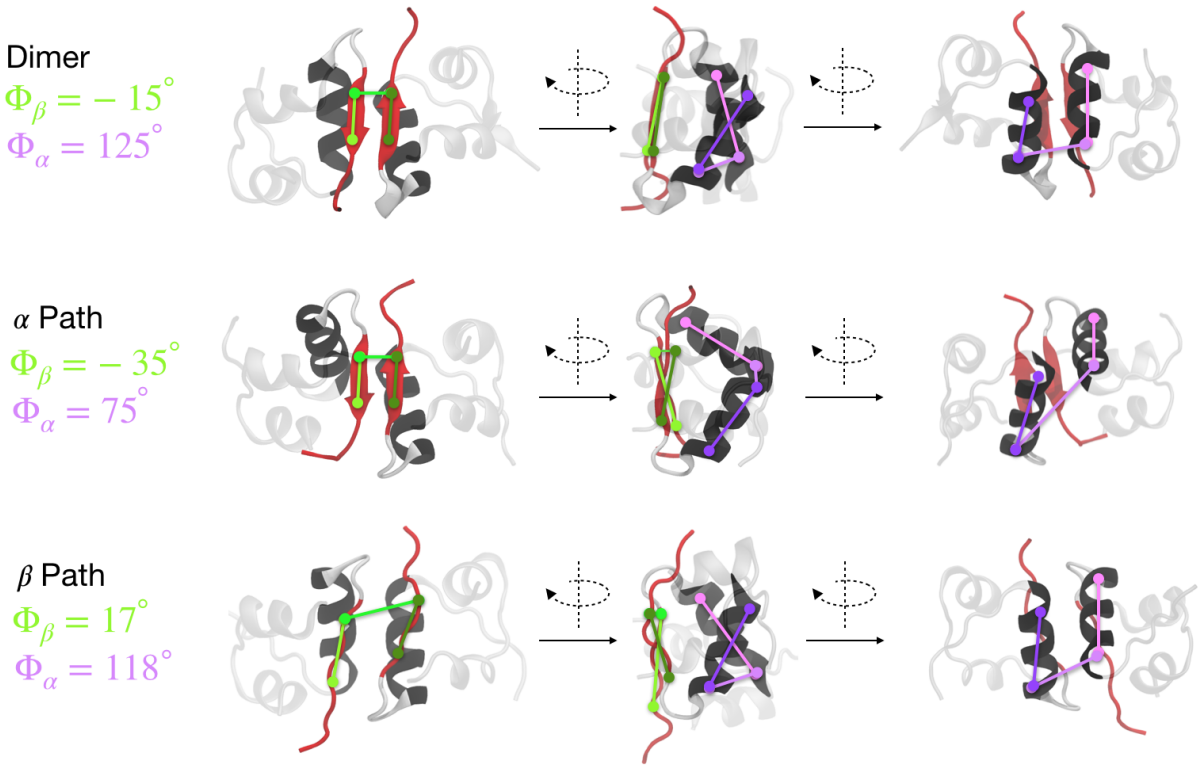


Figure 2.14: Structures representing the dimer (top), initial steps along the α path (middle), and initial steps along the β path (bottom), with lines superimposed to show the interfacial pseudodihedral angles, Φ_β (green) and Φ_α (purple). For these lines, a darker color in the side projection means the residues are in front while a lighter color means they are behind.

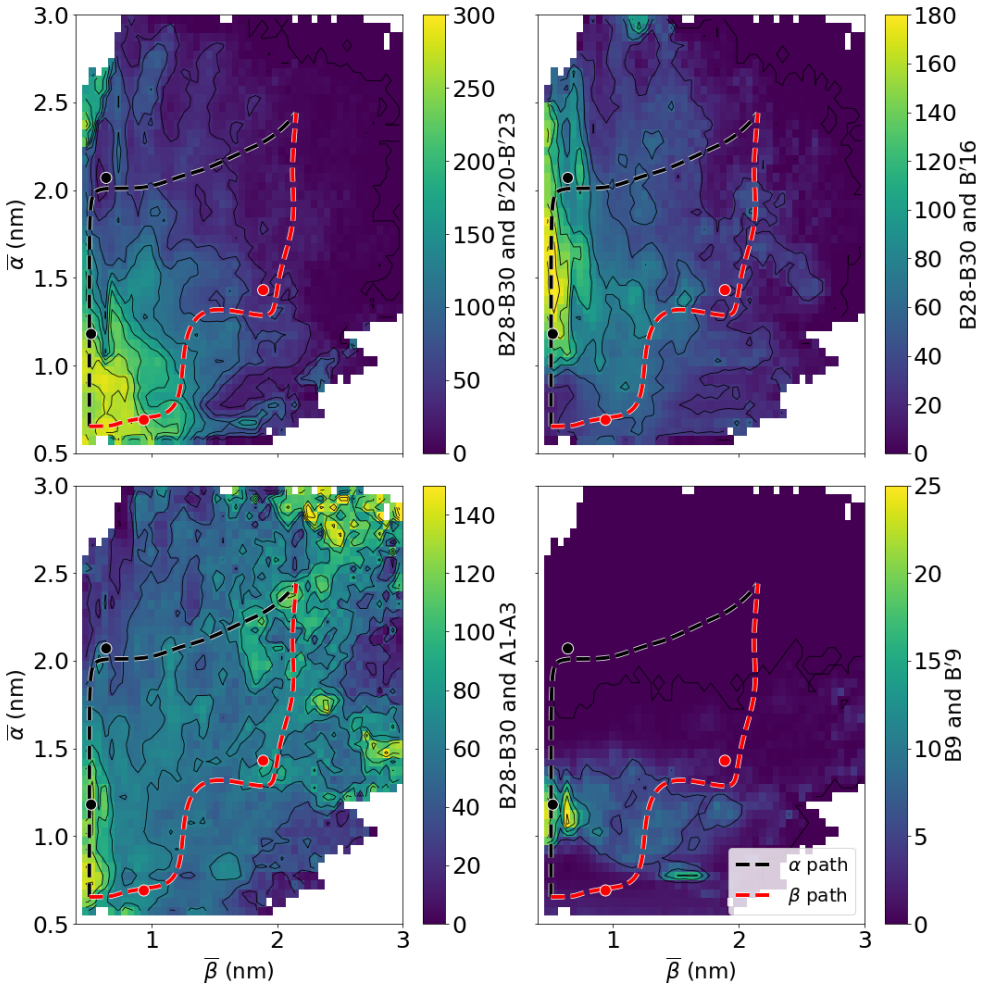


Figure 2.15: Averages of the number of native (left) and nonnative (right) contact pairs, with the specific pair given by the scale bar labels, as a function of $\overline{\alpha}$ and $\overline{\beta}$. The left plots show that along the α path, contacts between $\mathrm{Pro}^{\mathrm{B28}}$-$\mathrm{Ala}^{\mathrm{B30}}$ are lost with both $\mathrm{Gly}^{\mathrm{B'20}}$-$\mathrm{Gly}^{\mathrm{B'23}}$ (top) and $\mathrm{Gly}^{\mathrm{A1}}$-$\mathrm{Val}^{\mathrm{A3}}$ (bottom). The right plots show that along the α path, nonnative contacts start to form between $\mathrm{Pro}^{\mathrm{B28}}$-$\mathrm{Ala}^{\mathrm{B30}}$ and $\mathrm{Tyr}^{\mathrm{B'16}}$ (top), and between $\mathrm{Ser}^{\mathrm{B9}}$ and $\mathrm{Ser}^{\mathrm{B'9}}$.

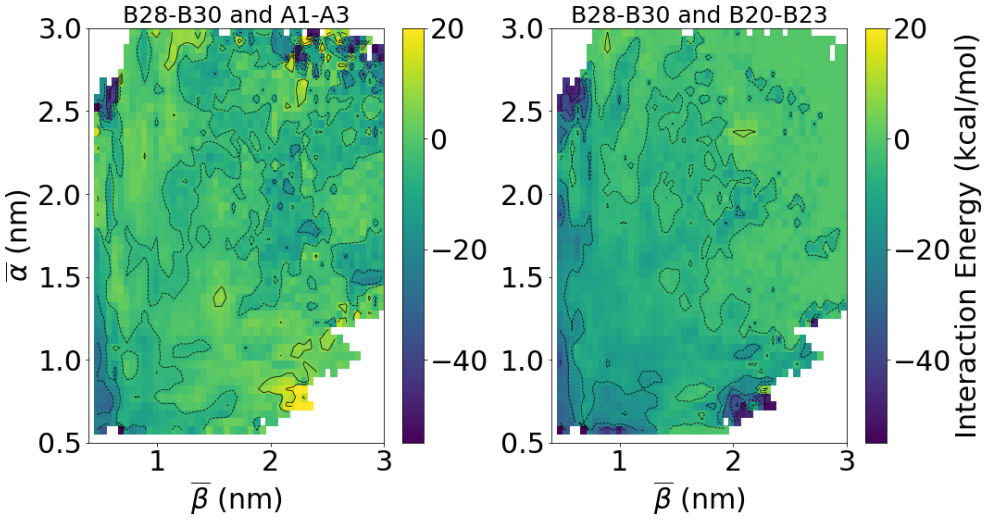


Figure 2.16: Average total interaction energies of $\mathrm{Pro}^{\mathrm{B28}}$-$\mathrm{Ala}^{\mathrm{B30}}$ with $\mathrm{Gly}^{\mathrm{A1}}$-$\mathrm{Val}^{\mathrm{A3}}$ (left) and $\mathrm{Gly}^{\mathrm{B'20}}$-$\mathrm{Gly}^{\mathrm{B'23}}$ (right). Contour lines shown every 10 kcal/mol. Both of these interactions stabilize the dimer state (lower left corner of both panels).

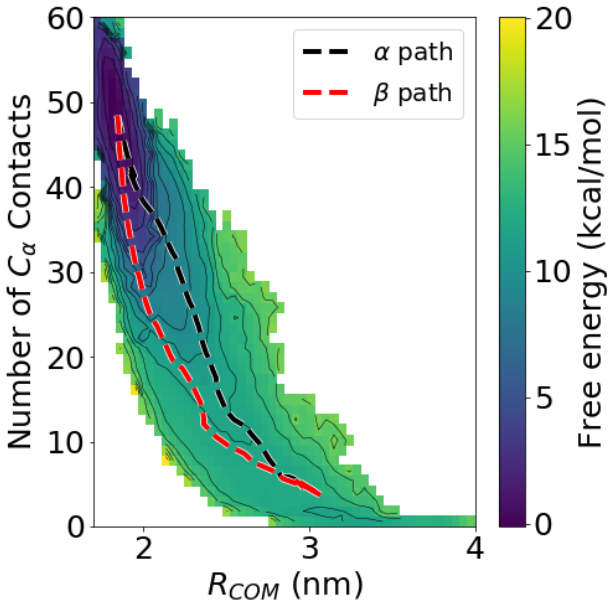


Figure 2.17: PMF as a function of the center of mass distance between the two monomers (R_{COM}) and number of interfacial C_α contacts (cutoff 7 Å,) the coordinates used by Bagchi and coworkers in refs. 2 and 3.

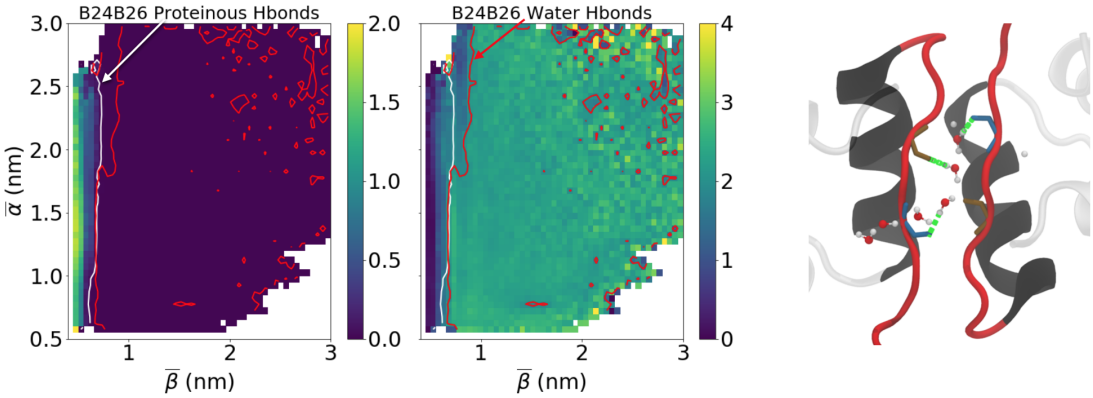


Figure 2.18: Characterizing interfacial hydrogen bonding. (Left) The average number of protein-protein hydrogen bonds on the beta sheet interface averaged on the PMF, compared to (middle) the average number of hydrogen bonds between those same residues and water. The white contour represents when the protein-protein hydrogen bond character drops to 2% of its initial value, while the red contour shows the point where on average 2 hydrogen bonds have been formed with water on the interface. These contours correlate well in CV space. On the right is a representative structure showing this solvation, with hydrogen bonds between the interfacial residues and water shown in green.

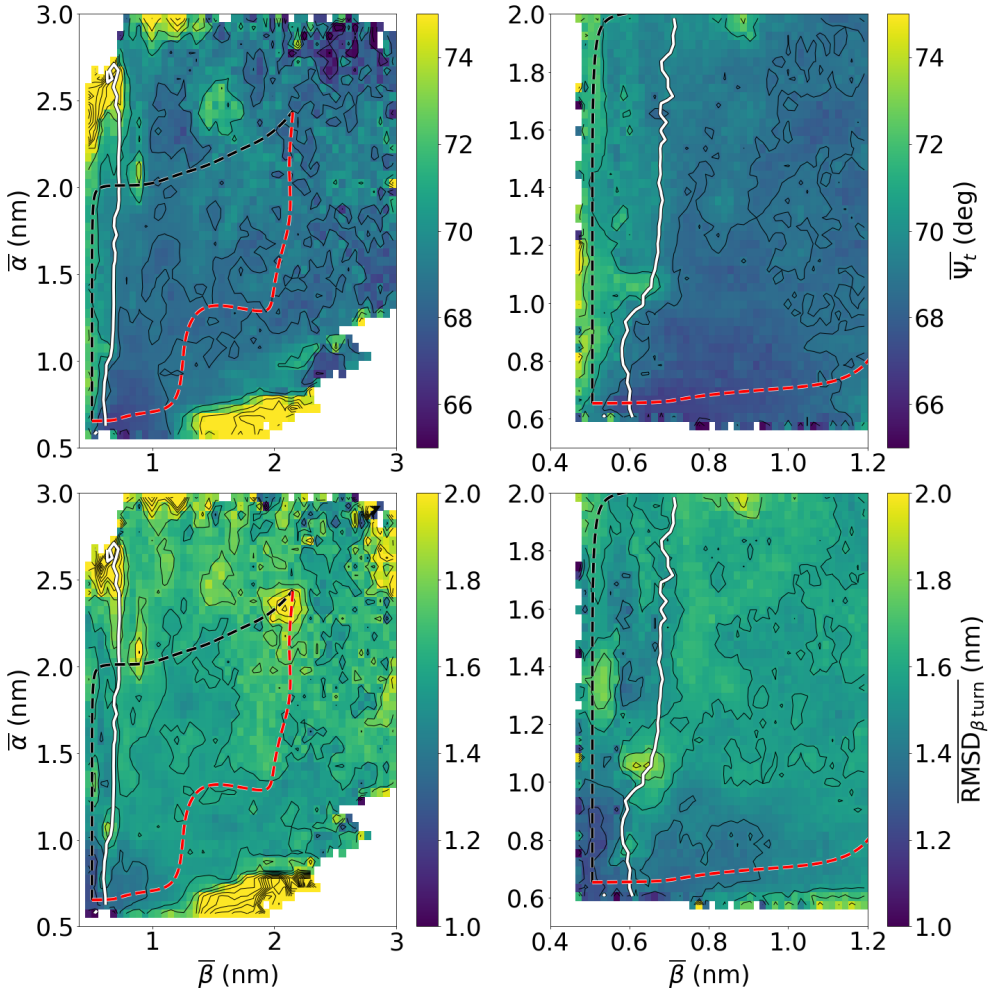


Figure 2.19: Plots showing the behavior of the interfacial β turn during the dissociation. The average β turn angle (top) and β turn RMSD (bottom) as a function of $\overline{\alpha}$ and $\overline{\beta}$, and zoomed in to the near-dimer regime on the right. On all graphs, the α and β paths are superimposed, as well as the white contour that signifies the solvation of the β interface. Here, we see the β turn angle increasing along the α path but not along the β path. Also, the β turn RMSD increases before the β sheets are broken along the α path, while this only occurs after the β sheets are broken for the β path.

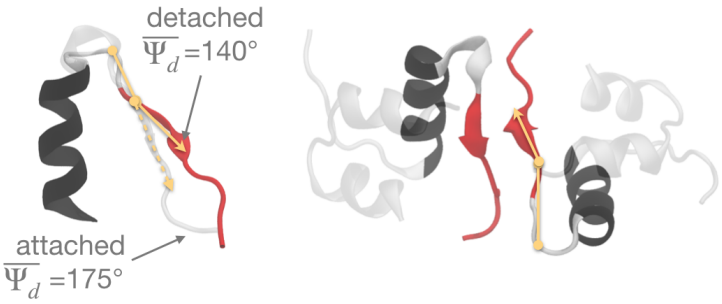


Figure 2.20: Structures showing detachment. (Left) The detached and attached species overlaid, with the detachment angle explicitly overlaid on top of the structure. (Right) This same detached structure, but now showing the entire dimeric species.

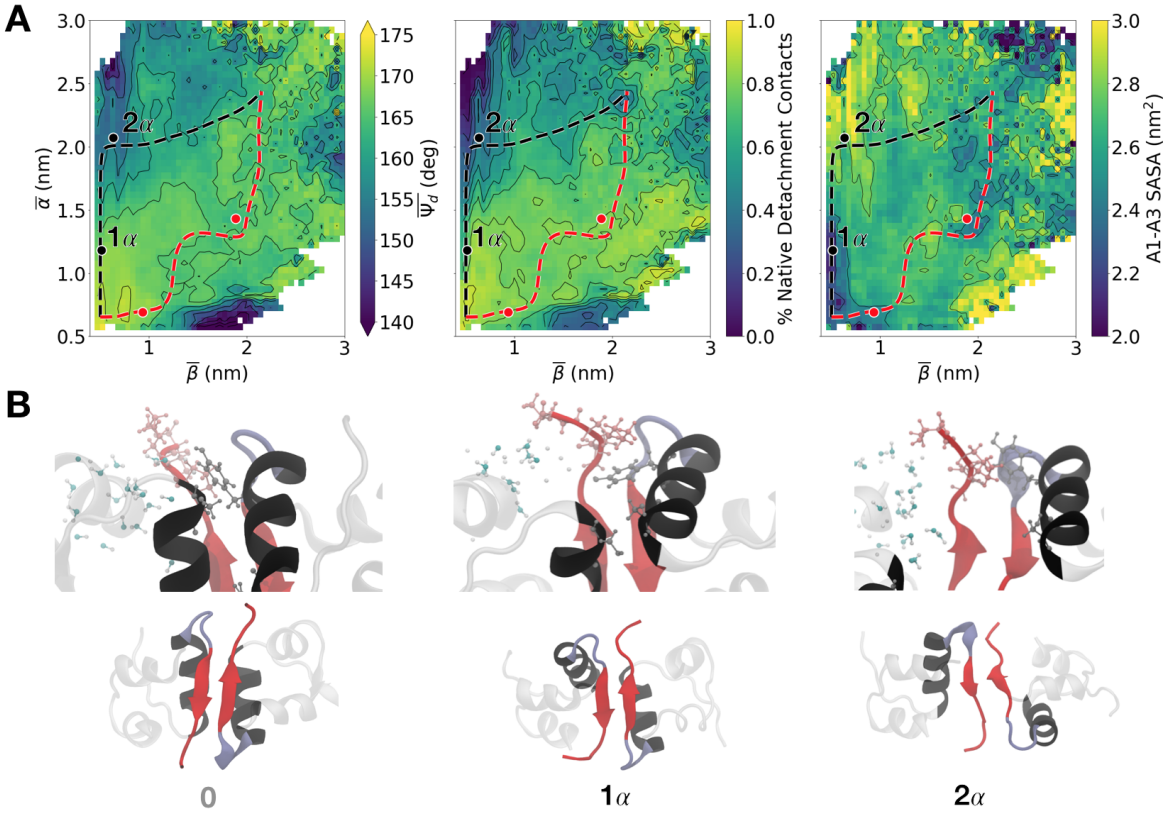


Figure 2.21: Detachment is correlated with solvation of the N-terminal segment of the A chain. (A) Averages of the detachment angle Ψ_d(left), percentage of native contacts between Pro^{B28}-Ala^{B30} and the nearby α helix of the same monomer (middle), and the SASA for Gly^{A1}-Val^{A3} of the same monomer. The similarity between the left and middle plots suggests that the detachment angle is an effective measure of the C-terminal segment moving away from the α helix it is normally tucked against in the native monomeric unit. Furthermore, the similarity to the rightmost plot shows that the solvation of Gly^{A1}-Val^{A3} is correlated to the large detachment of the B chain 's C-terminal segment in the same monomeric unit. (B) Structures showing how the detachment is coupled to the solvation of Gly^{A1}-Val^{A3}.

CHAPTER 3

KINETICS OF PHENOL ESCAPE FROM THE INSULIN R_6 HEXAMER

This chapter was published under this title as Adam Antoszewski, Chatipat Lorpaiboon, John Strahan, and Aaron R. Dinner, *J. Phys. Chem. B*, 125(42), 11637–11649, 2021. [147]

Abstract

Therapeutic preparations of insulin often contain phenolic molecules, which can impact both pharmacokinetics and shelf life. Thus understanding the interactions of insulin and phenolic molecules can aid in designing improved therapeutics. In this study, we use molecular dynamics to investigate phenol release from the insulin hexamer. Leveraging recent advances in methods for analyzing molecular dynamics data, we expand on existing simulation studies to identify and quantitatively characterize six phenol binding/unbinding pathways for wild-type and A10 Ile $\rightarrow$ Val and B13 Glu $\rightarrow$ Gln mutant insulins. A number of these pathways involved large-scale opening of the primary escape channel, suggesting that the hexamer is much more dynamic than previously appreciated. We show that phenol unbinding is a multipathway process, with no single pathway representing more than 50% of the reactive current and all pathways representing at least 10%. We use the mutant simulations to show how the contributions of specific pathways can be rationally manipulated. Predicting the net effects of mutations is more challenging because the kinetics depend on all the pathways, demanding quantitatively accurate simulations and experiments.

3.1 Introduction

The primary therapeutic for diabetes management is the protein hormone insulin. Both the pharmacokinetics and the shelf life of insulin depend on equilibria between conformational

and oligomeric states. Insulin is typically a hexamer in therapeutic formulations, a dimer in the blood, and a monomer when bound to its receptor [51, 52]; therapeutic formulations require cold storage and transport to suppress off-pathway equilibria that lead to fibrillation [61–63]. Much effort is directed toward modulating insulin's equilibria to achieve therapeutics with desired properties [64]. For example, the widely used fast-acting diabetes therapeutics lispro ($Pro^{B28} \rightarrow Lys^{B28}$, and $Lys^{B29} \rightarrow Pro^{B29}$) [65] and aspart ($Pro^{B28} \rightarrow Asp^{B28}$) [66] destabilize the insulin dimer. In contrast, slow-acting basal insulin analogs like glargine [67] and detemir [68, 69] function by either decreasing insulin solubility or adsorption in the body [70]. Longer-acting insulin analogs are also being developed [71, 72], but such formulations are often expensive and complicated. To enable rational design of improved insulin mutants and analogs, better understanding of insulin equilibria at the molecular level is needed.

In the present study, we focus on the insulin hexamer. The hexamer exists in three conformational states that are designated T_6, T_3R_3, and R_6 for the contributing monomer conformational states [148–150]. Each monomer consists of a 21-amino acid A chain joined to a 30-amino acid B chain by two interchain disulfide bonds (Cys^{A7}-Cys^{B7} and Cys^{A20}-Cys^{B19}). Each A chain contains an intrachain disulfide bond (Cys^{A6}-Cys^{A11}) and two α helices: Gly^{A1} - Ser^{A9} and Leu^{A13} - Asn^{A21}. Each B chain has a single α helix, and this secondary structure element differs in length depending on whether the monomer is in the T or R state. It spans Ser^{B9} - Cys^{B19} in the T state and Phe^{B1} - Cys^{B19} in the R state [151]. The hexamer can be viewed as a trimer of dimers, with the dimer interface made up of the B-chain α helix and an anti-parallel β sheet (Phe^{B24}-Tyr^{B26}) from each monomer.

The T_6 and R_6 states of the hexamer have markedly different dissociation kinetics. The T_6 state dominates under physiological conditions [52] and dissociates to dimers in minutes [152]. By contrast, the lifetime of the R_6 state is hours to days [152]. Therapeutic formulations often contain phenol because it prolongs their shelf lives by shifting the equilibrium to R_6 [60], which presumably suppresses dissociation and in turn off-pathway equilibria. The

phenol binds in pockets that are formed in R_6 between the A chains of one insulin dimer and the Gly^{B1} - Gly^{B8} segments from an adjacent dimer [148]. The X-ray crystal structure of the phenol-bound human insulin R_6 hexamer (PDB ID 1ZNJ) [153] suggests that when bound, each phenol forms two hydrogen bonds, one with the backbone carbonyl oxygen of Cys^{A6} and one with the backbone amide NH of Cys^{A11}, and interacts with His^{F5}, among other residues in the binding pocket (Figure 3.1). [148, 154]. These pockets are absent from T_6.

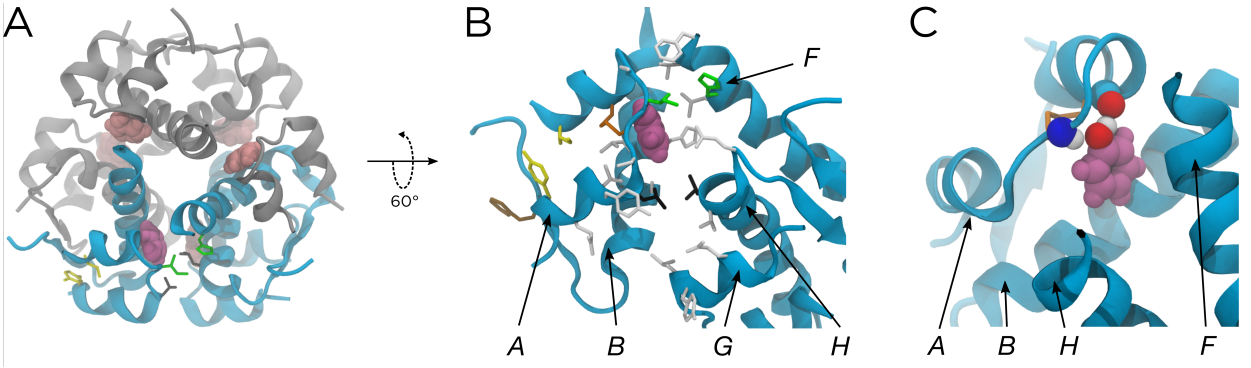


Figure 3.1: The solvated and equilibrated crystal structure for the R_6 insulin hexamer. (A) The full hexamer, with phenols being shown in magenta/pink. The protein is colored to correspond with the other panels. (B) The cyan protein chains from (A), labelled as chains A, B, F, G, and H. For a full description of the protein nomenclature, see the Supplemental Information. Sidechains are shown that define the phenolic binding pocket. Some specific sidechains are highlighted as follows: Ile^{A10} and His^{F5} (green), Leu^{A13} and Leu^{H17} (black), Ile^{A2} and Tyr^{A19} (yellow), Cys^{A6} and Cys^{A11} (orange), and Phe^{B25} (brown). Other residues involved in the binding pocket and escape pathways are shown in white. We omit hydrogens for clarity. (C) The configuration in (B) represented to show the two hydrogen bonds formed by the -OH in the phenol, one with the backbone carbonyl oxygen of Cys^{A6} and one with the backbone amide NH of Cys^{A11}.

Compared to the hexamer dissociation timescales, the phenol unbinding timescale is relatively fast. Existing NMR data suggest that phenol unbinding and rebinding occur on the sub-millisecond timescale [155]. This fast equilibrium is somewhat surprising, as the R_6 binding pockets bury the phenols nearly completely [148, 156]. In this case, the crystal structure is not sufficient to predict the timescales of phenol escape, and we must consider the dynamics of the protein.

NMR data indicate that Ile^{A10} serves as an essential "gatekeeper" residue, whose flexibility is required for the binding/unbinding of phenolic ligands [155]. Two subsequent molecular dynamics studies in which external forces were used to accelerate phenol escape [157, 158] identified three pathways for unbinding. Abrams and Vashisth [158] call these pathways PW1, PW2, and PW3, and we use this nomenclature throughout our paper. PW1 is a gate-pushing mechanism, where the phenol pushes through the gatekeeper residues Ile^{A10} and His^{F5} (green in Figure 3.1). By contrast, PW2 is as a gate-hopping mechanism, with the phenol passing through an existing escape channel between Ile^{A10}/His^{F5} and Leu^{A13}/Leu^{H17} (black in Figure 3.1). Finally, PW3 is an unrelated pathway, where the phenol moves back through the A chain and escapes through the A chain and the β sheet dimeric interface, interacting with Ile^{A2} and Tyr^{A19} (yellow in Figure 3.1).

While the pathways in these simulations suggest possible mechanisms of phenol unbinding, applying external forces can significantly bias the dynamics [159]. Consequently, it is difficult to determine the significance of the pathways and to estimate their kinetics. To address this issue, here, we exploit recent advances in computational methods for estimating kinetic statistics from an ensemble of short, unbiased simulations [25, 26] to compute free energies, escape and rebinding probabilities (commmittors), and reactive currents for phenol unbinding from the R_6 hexamer. By combining these quantities, we are able to characterize six unbinding pathways quantitatively and determine their relative weights. All pathways contribute significantly to unbinding/binding rates, highlighting the importance of methods that naturally account for a diversity of mechanisms [25, 26], in contrast to traditional rate theories [160, 161].

In addition to providing quantitative insights, our simulations reveal qualitatively new dynamics. We observe a large-scale opening of the primary channel for phenol escape and delineate six pathways for phenol unbinding. Based on our observations, we identify and model two mutations that we expect to impact the dynamics. We find that one mutation

(B13 Glu→Gln) encourages phenol unbinding, and the other (A10 Ile→Val) discourages it. The relative weights of our six pathways are dramatically affected for both mutations, with the preferred wildtype (WT) pathway becoming sparsely populated for the B13 Glu→Gln mutation. We thus demonstrate how molecular dynamics simulations can inform the design of therapeutics with improved properties.

3.2 Methods

System setup. All systems were prepared using CHARMM-GUI, version 3.0 for the wildtype (WT) simulations, and version 3.2 for the mutant simulations [99, 100, 162]. The crystal structure for wildtype (WT), phenol-bound human insulin R_6 hexamer was retrieved from the Protein Data Bank (PDB ID 1ZNJ)[153]. All ions (including the two Zn^{2+} ions and two Cl^- ions bound within the hexamer) and 331 crystallographic water molecules were retained. Six of the seven phenols in the structure were included in the simulations; following the procedure in ref. 158, the seventh phenol, located adjacent to one of the bound Cl^- ions, was deleted [158]. Missing Thr residues at the C-termini of the six insulin B chains were added, along with hydrogen atoms. We refer to the A and B chains of the six insulin monomers as chains A through L, with nomenclature specifics given in the Supplemental Information. CHARMM parameter files for phenol were generated from the phenol Structure Data File (SDF) available from the RCSB and the CHARMM General Force Field [163]. Protonation states were chosen by using the PROPKA3 algorithm [164, 165]. The edited crystal structure was solvated in a $(8.2\ \text{nm})^3$ box of TIP3P water [107]. To neutralize the system at a concentration of 150 mM KCl, 53 K^+ and 43 Cl^- additional ions were added [108]. There was a total of 51,064 atoms.

Equilibration. All simulations were performed using GROMACS 2019.4 [98] patched with PLUMED 2.5.3 [118–120], using the CHARMM36m force field [95–97]. Unless otherwise

noted, simulations were carried out in the isochoric isothermal (NVT) ensemble at 303.15 K using a Langevin thermostat [101] with a 2 fs timestep and a friction constant of 10 ps^{-1} applied to all atoms. This friction constant was chosen to be as weak as possible, while still maintaining the correct temperature, so as to minimally perturb the dynamics. The LINCS algorithm [102] was used to constrain all bonds to hydrogen. The particle-mesh Ewald method [103] was used to calculate electrostatic forces, accounting for periodic boundary conditions, with a cutoff distance of 1.2 nm. The Lennard-Jones interactions were smoothly switched off from 1.0 to 1.2 nm through the built-in GROMACS force-switch function. We used VMD [104] for molecular visualization.

To relax each system following its preparation, we used the steepest descent algorithm to minimize the energy until the maximum force felt by the system was below 1000 kJ/mol nm. We then equilibrated the system for 100 ps in the NVT ensemble with a 1 fs timestep, followed by 10 ns in the NPT ensemble at 1 bar using the Parrinello-Rahman barostat [109], with a 2 fs timestep and time constant of 5.0 ps. For the energy minimization and equilibration above, harmonic restraints were used to stabilize the positions of all non-hydrogen protein atoms. The system was equilibrated further for 1 ns in the NPT ensemble without position restraints, and the average box size was determined to be $(7.96\text{ nm})^3$. This box size was used for all further simulations. The system was equilibrated once more without position restraints for 3 ns in the NVT ensemble. The resulting equilibrated structure was used to initialize further simulations as described below.

Comparison of phenols. Although in principle all six phenols are equivalent, in practice their behaviors may differ owing to asymmetries in the initial structure. To determine if this was the case, we used Adiabatic-Bias Molecular Dynamics (ABMD) implemented in PLUMED 2.5.3 [117–120] to drive their dissociation from the WT hexamer. This method uses a half-harmonic potential, moved in a ratchet-and-pawl like mechanism, to trap natural fluctuations toward a specific target in collective variable (CV) space. It thus tends to

bias the dynamics more gently than alternatives. Specifically, we used ABMD because we previously observed ABMD to successfully drive insulin dimer dissociation without melting otherwise stable α helices, in contrast to steered molecular dynamics [76].

Our first guesses for CVs to control phenol dissociation were the six distances d_n, where $1 \leq n \leq 6$ denotes the identity of the phenol being released. Specifically, d_n is the distance between the geometric center of the non-hydrogen atoms of phenol n and the Zn^{2+} ion bound closest to it. We biased the six d_ns independently but simultaneously; since the timing of events in ABMD simulations depends on the specific fluctuations that occur, this led to simulations where varying numbers of the phenols dissociated with no externally imposed order. A total of 276 such simulations were performed (each run for 5 ns with structures saved every 10 ps): 20 simulations for each of nine equally spaced force constants ranging from 20 to 28 kJ/(mol nm), inclusive, and 48 simulations for both 29 and 30 kJ/(mol nm).

For these simulations, phenols were considered dissociated if the number of non-hydrogen atom contacts between them and the protein dropped below five. Phenol 4 was released in all 276 simulations, while phenol 3 was released in only 150/276 simulations. Furthermore, for the 77 simulations where all phenols were released, phenol 4 was released either first or second in 74/77 simulations, while phenol 3 was released either fifth or sixth in 42/77 simulations. This suggested that phenol 4 and phenol 3 are the ligands that are most and least easily released, respectively. To ensure that our results were not specific to any particular feature of initial structure, we investigated the release of each of these two phenols following the procedure described below. However, because we found that the results for them were nearly identical, we present only those for phenol 4.

Generation of starting structures. To generate starting structures for the unbiased trajectories used for later analyses, we again used ABMD, this time to explore various mechanisms of phenol release for a single phenol (phenol 4), as opposed to all six phenols at once. This time, the bias was placed on just d_4. Initially, 120 simulations were performed

(each run for 5 ns with structures saved every 5 ps): 30 simulations for each of the four force constants 12.5, 15.0, 17.5, and 20.0 kJ/(mol nm). Through visual analysis of these trajectories, three CVs were initially chosen to represent the dominant features observed: N_{A10}, N_{A13}, and $\mathrm{RMSD_P}$. N_{A10} and N_{A13} are the number of non-hydrogen atom contacts between phenol 4 and either $\mathrm{Ile^{A10}}$ (green in Figure 3.1) or $\mathrm{Leu^{A13}}$ (black in Figure 3.1), respectively. $\mathrm{RMSD_P}$ is the distance root-mean-squared deviation (RMSD) between the α carbons of residues in the phenolic binding pocket compared to their position in the crystal structure. The distance RMSD is defined by first calculating the pairwise distances between all atoms in a selection for a reference structure, then calculating those same distances for each frame of the simulation, and finally calculating the root mean squared difference of these pairwise distances. The binding pocket residues, pictured in green, orange, black, and white in Figure 3.1, were defined to be the residues most often interacting with phenol both in its crystallographic binding pocket and along the different observed pathways of dissociation. These 22 binding pocket residues, following the naming convention described in the Supplemental Information, are as follows: $\mathrm{Cys^{A6}}$, $\mathrm{Ser^{A9}}$-$\mathrm{Tyr^{A14}}$, $\mathrm{Leu^{A16}}$, $\mathrm{Glu^{A17}}$, $\mathrm{His^{B10}}$, $\mathrm{Leu^{B11}}$, $\mathrm{Ala^{B14}}$, $\mathrm{Phe^{F1}}$, $\mathrm{Val^{F2}}$, $\mathrm{His^{F5}}$, $\mathrm{Leu^{F6}}$, $\mathrm{Leu^{G13}}$, $\mathrm{Tyr^{G14}}$, $\mathrm{Glu^{G17}}$, $\mathrm{Leu^{H17}}$, $\mathrm{Val^{H18}}$, and $\mathrm{Glu^{H21}}$. It should be noted that while N_{A10} and N_{A13} proved sufficient to categorize our initial sampling, we also present analysis based on other contact-based CVs that we subsequently developed to better capture the unbinding pathways.

When building this ABMD data set, we wanted to fully explore unbinding pathways. To do this, we wanted multiple trajectories to sample all populated areas in our CV space of N_{A10}, N_{A13}, and $\mathrm{RMSD_P}$. By doing this, not only did we ensure starting structures that sampled all relevant pathways, but we also decreased the possible correlation between starting structures for our unbiased simulations. From our initial set of ABMD simulations, we observed pathways similar to PW1, PW2, and PW3 as identified by Vashisth and Abrams in ref. 158, although only a handful of trajectories followed PW1, and only one pathway

followed PW3. To supplement the PW1 data, we ran an extra 30 ABMD simulations, each of length 5 ns with force constant $k = 12.5$ kJ/(mol nm), starting from a structure from a previous ABMD run that ended approximately halfway along PW1. We supplement the PW3 data later, in the unbiased data set. Furthermore, we saw a number of trajectories where $\mathrm{RMSD_P}$ spiked from approximately 0.5 Å to approximately 3-4 Å. As discussed in the Results and Discussion, this corresponds to a channel opening mechanism, where two dimers rotate away from one another, exposing more of the phenol directly to solvent. To sample this behavior further, we identified two trajectories that led to channel opening but without phenol release and ran 30 extra ABMD simulations starting from each of these configurations. Each of these trajectories was 5 ns with a force constant of $k = 12.5$ kJ/(mol nm). In total, we thus supplemented our initial data set with 90 additional ABMD simulations, to create a data set of 210 ABMD trajectories.

State definitions. To measure binding and unbinding statistics, one needs to define both the bound and free states. The CVs we used to do this include $\mathrm{RMSD_P}$, as previously defined, and N_{PW1}, N_{PW4}, N_{Prot}, and N_{HP}, which we define here. N_{PW1} and N_{PW4} can be viewed as extensions of the previously defined N_{A10} and N_{A13}, chosen to more fully describe the unbinding behavior in our simulations. N_{PW1} is the number of non-hydrogen atom contacts between phenol and residues $\mathrm{Ile^{A10}}$ and $\mathrm{His^{F5}}$ (green in Figure 3.1). Residue $\mathrm{His^{F5}}$, like $\mathrm{Ile^{A10}}$, helps to define Pathway 1 as in ref. 158. Similarly, N_{PW4} is the number of non-hydrogen atom contacts between phenol and residues $\mathrm{Leu^{A13}}$ and $\mathrm{Leu^{H17}}$ (black in Figure 3.1). Residue $\mathrm{Leu^{H17}}$, along with $\mathrm{Leu^{A13}}$, helps to define Pathway 4 (PW4), a pathway similar to but distinct from ones previously characterized that we discuss in Results and Discussion. Finally, N_{Prot} is the total number of non-hydrogen atom contacts between the phenol and the protein, while N_{HP} is the two-step rolling average of the number of hydrogen bonds between the phenol and both the backbone carbonyl of $\mathrm{Cys^{A6}}$ and the backbone amide nitrogen of $\mathrm{Cys^{A11}}$. For this CV, we defined a hydrogen bond as a distance of less than 4 Å

between the non-hydrogen atoms of the donor (D) and acceptor (A) and a D-H-A angle less than 60° out of line.

The free state is defined so that $N_{\mathrm{PW1}} < 2$, $N_{\mathrm{PW4}} < 2$, and $N_{\mathrm{Prot}} < 5$. The bound state is defined by $N_{\mathrm{Prot}} > 540$, $N_{\mathrm{HP}} > 1$, and $\sum_{\mathrm{CV}}(\mathrm{CV} - \mu_{\mathrm{CV}})^2/(2\sigma_{\mathrm{CV}})^2 \leq 1$ for $\mathrm{CV} = N_{\mathrm{PW1}}$, N_{PW4}, and $\mathrm{RMSD_P}$. Here, μ_{CV} and σ_{CV} represent the average and standard deviation of the corresponding CV as defined from a 10 ns unbiased simulation. For N_{PW1}, N_{PW4}, and $\mathrm{RMSD_P}$, the means were 52.503, 31.924, and 0.0872 Å, and the standard deviations were 5.023, 5.264, and 0.0223 Å, respectively.

DGA data set. Our goal is to estimate equilibrium and dynamical quantities through the Dynamical Galerkin Approximation (DGA) [25, 26]. Using this technique, dynamical statistics like the committor, reactive current, and reaction rate can be calculated from an ensemble of unbiased trajectories. The essential idea is that we cast quantities of interest as solutions to operator equations involving this transition operator, the operator that determines the statistics of the dynamics. In general, the solutions are subject to boundary conditions involving both the bound and free states, defined in the previous section. We approximate the solutions through a basis expansion, choosing our basis so that it captures all of the movements important for phenol binding/unbinding. This enables us to represent the action of the transition operator through its inner products with the basis functions; in turn, these can be estimated from averages over unbiased trajectories. This approach can be thought of as an extension of Markov State Models (MSMs) [20, 22, 23, 166, 167] that directly yields statistics for a specified reaction.

To generate the unbiased trajectories needed for DGA, we first defined a $10 \times 10 \times 10$ grid in cylindrical coordinates that fully covers the area of CV space sampled by our ABMD data set ($N_{\mathrm{A10}} = r\cos\theta$, $N_{\mathrm{A13}} = r\sin\theta$, and $\mathrm{RMSD_P}$). Specifically, r varied between 3 and 50 contacts, θ varied between 0 and 90°, and z varied between 0.08 and 0.38 Å. The structure from the ABMD data set closest to each of these 1000 grid points was found, leading to 326

unique structures (as the ABMD sampling is not uniform in this space, many grid points had duplicate structures; see Supplemental Figure 3.6). From each one of these unique structures, we launched two independent 40 ns unbiased simulations and saved structures every 10 ps. Of these 652 trajectories, only three trajectories partially or fully sampled PW3 as defined by Vashisth and Abrams 158. To supplement the sampling of that pathway, from these three trajectories, we selected a total of 20 structures that captured various features of PW3. From each of these 20 structures, two independent 40 ns simulations were launched. Thus, our total unbiased data set consisted of 692 trajectories of length 40 ns, for an aggregate simulation time of 27.68 μs.

DGA basis choice. As described earlier, DGA solves operator equations for dynamical statistics by expanding them in terms of a set of basis functions. In this work, we chose the modified pairwise distance form described in ref. 26. To construct the appropriate pairwise distances, we first included those between two carbons on opposite sides of the escaping phenol and our 22-atom binding pocket of α carbons described earlier (protein-ligand distances). We also included the pairwise distances between the α carbons of the side chains themselves (protein-protein distances) to account for protein rearrangements. To account for movement along PW3, which involves residues not in the binding pocket, we added the protein-protein and protein-ligand distances between the phenol and the α carbons for residues $\mathrm{Ile^{A2}}$/ $\mathrm{Tyr^{A19}}$ (yellow in Figure 3.1), and $\mathrm{Phe^{B25}}$ (brown in Figure 3.1), as suggested by both ref. 158 and our preliminary results. Additionally, to capture movements of side chains along Pathways 1 and 2, we added the protein-protein and protein-ligand distances between the phenol and C_γ atoms from residues $\mathrm{Ile^{A10}}$, $\mathrm{Leu^{A13}}$, $\mathrm{His^{F5}}$, and $\mathrm{Leu^{H17}}$. Finally, we added one constant basis function, which improved the numerical solutions. Combining these distances led to a 299-dimensional basis set, a summary of which is shown in Supplementary Table 3.3.

For each point in our unbiased data set, we measured the minimum distance in the space

of the 298 pairwise distances to points in both the bound and free states. Using these distances, we followed the procedure in ref. 26 to construct the smoothing function $h(x)$, the guess function(s) $\psi(x)$, and the smoothed basis functions $\phi_i(x)$. These basis functions were then orthogonalized using singular value decomposition. A key choice in applying DGA is the time interval for the transition operator, termed the "lag time." We found most statistics to be insensitive to the lag time over ranges tested (Supplemental Figure 3.7) and report results for a lag time of 500 ps unless otherwise noted because it yielded convergence of estimates with a minimum of apparent statistical error (Supplemental Figure 3.7). The exceptions were rate constants and their ratios. Based on the convergence of our estimates for the relative weights of pathways (Supplemental Figure 3.7), we use lag times between 500 ps and 1.25 ns for rates and their ratios. We discuss further details of these calculations in the Supplemental Information.

Mutants. A similar workflow to the above was followed for both insulin mutants studied, A10 Ile→Val and B13 Glu→Gln. Details on system generation using CHARMM-GUI [99, 100, 162] and workflow for these mutants are given in the Supplemental Information.

3.3 Results and Discussion

The goal of this study was to probe the dynamics of phenol release from the R_6 insulin hexamer to understand how the release mechanism(s) could be altered by mutations. To do this, as described in detail in Methods, we employed a pipeline of multiple simulation techniques. First, we used an array of Adiabatic-Bias Molecular Dynamics (ABMD) simulations to create a large, diverse data set of driven dissociation events. From this, we discovered a small set of physically meaningful collective variables (CVs) that can be best used to visualize the results: $\mathrm{RMSD_P}$ describes the distance RMSD (referenced to the crystal structure) of the 22 residues determined to be in the phenolic binding pocket, while N_{PW1}, N_{PW3}, and

N_{PW4} describe the number of non-hydrogen atom contacts between the released phenol and gatekeeper residues along pathway (PW) 1 ($\mathrm{Ile}^{\mathrm{A10}}$ and $\mathrm{His}^{\mathrm{F5}}$, green in Figure 3.1), PW3 ($\mathrm{Ile}^{\mathrm{A2}}$ and $\mathrm{Tyr}^{\mathrm{A19}}$, yellow in Figure 3.1), and PW4 ($\mathrm{Leu}^{\mathrm{A13}}$ and $\mathrm{Leu}^{\mathrm{H17}}$, black in Figure 3.1), respectively. To escape via each of these pathways, the phenol must pass between the side chains of its two gatekeeper residues. We do not consider an analogous N_{PW2} because there are no specific gatekeeper residues for PW2.

We then ran short (40 ns) unbiased simulations starting from a selection of partially unbound structures from the ABMD data set and used Dynamical Galerkin Approximation (DGA) [25, 26] to estimate long-time statistics from the resulting short trajectories. Using this method, we compute free energies, escape versus rebinding probabilities (committors), and reactive currents. Combining these statistics allows us to to delineate six pathways and their transition states. DGA furthermore enables us to estimate the relative weights for the unbinding pathways and rates, providing insights into kinetics that go beyond the free energies by themselves. Below, we first describe the simulations for WT insulin, followed by those for the A10 Ile→Val and B13 Glu→Gln mutants.

Driven simulations reveal channel opening and six dissociation pathways. When visually analyzing the ABMD simulations with driven phenol release, we were able to distinguish six unbinding pathways. Three of these pathways (PW1, PW2, and PW3) were previously reported by Vashisth and Abrams [158], while the other three (PW1a, PW4, PW4a) are reported here for the first time. An essential feature of two of these pathways (PW1a and PW4a) is a large-scale widening of the primary escape channel (indicated by cyan coloring in Figure 3.2A) that results from two adjacent insulin dimers twisting away from one another. This behavior is discussed in depth in the following section. In addition, while we explicitly discuss the driven ABMD simulations in this section, the same six pathways with the same molecular characteristics are further supported by the unbiased simulations and our quantitative analysis of them.

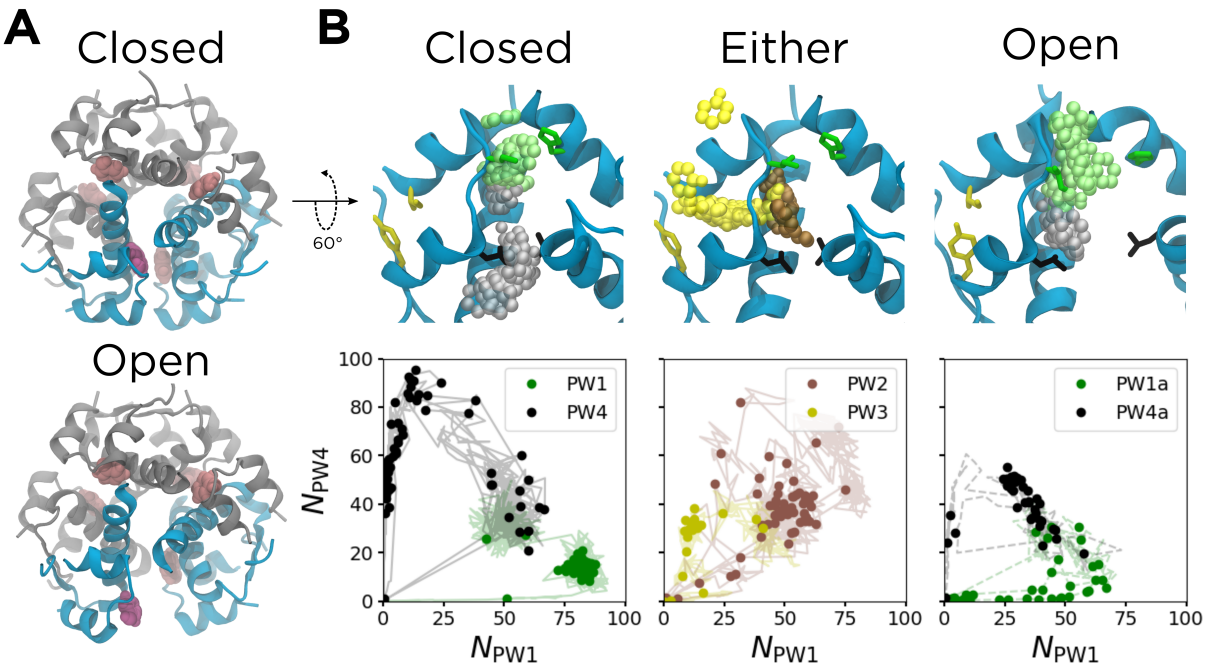


Figure 3.2: Results from the ABMD simulations. (A) The structure of the hexamer, with the phenolic escape channel closed (top) and open (bottom). The chains that form the phenolic binding pocket are shown in cyan; other chains are shown in gray. The released phenol is shown in purple, with other phenols shown in pink. (B) The six unbinding pathways, shown both structurally (top) and as a function of N_{PW1} and N_{PW4} (bottom). The structures shown correspond to the solid data points, and represent the k-medoids cluster centers along each pathway. The solid protein cartoons correspond to the starting structure for each set of driven simulations, and the translucent spheres are the non-hydrogen atoms of the phenols from the cluster centers, all aligned to the A chain backbone of the starting structure. The translucent lines in the bottom panels represent the data used to generate the k-medoids clusters. The cyan chains in the top panels are the same as those in (A). Non-hydrogen atoms of gatekeeper side chains along PW1 (green, $\mathrm{Ile}^{\mathrm{A10}}$ and $\mathrm{His}^{\mathrm{F5}}$), PW3 (yellow, $\mathrm{Ile}^{\mathrm{A2}}$ and $\mathrm{Tyr}^{\mathrm{A19}}$), and PW4 (black, $\mathrm{Leu}^{\mathrm{A13}}$ and $\mathrm{Leu}^{\mathrm{H17}}$) are also shown in the top panels. The configuration of the escape channel is indicated above each panel. For PW2 and PW3, the channel can be either open or closed.

We tested a variety of CVs for their utility in visualizing the six unbinding pathways. As noted earlier, three of the best that we found were N_{PW1}, N_{PW3}, and N_{PW4}, the numbers of non-hydrogen contacts between the phenol and the gatekeeper residues for PW1, PW3, and PW4, respectively. We determined the gatekeeper residues based on the previous literature and our analysis of the ABMD trajectories. The gatekeeper residues for PW1 are $\mathrm{Ile}^{\mathrm{A10}}$ and $\mathrm{His}^{\mathrm{F5}}$; these were identified by both Swegat *et al.*[157] and Vashisth and Abrams [158]

and were confirmed by our own simulations. The gatekeeper residues for PW3 are Ile^{A2} and Tyr^{A19}; these were identified by Vashisth and Abrams [158] and confirmed by our own simulations. The gatekeeper residues for PW4 are Leu^{A13} and Leu^{H17}. Vashisth and Abrams [158] note these residues in conjunction with PW2, but we designate them as gatekeeper residues for PW4 in this study because the phenol passes between them only along PW4 (versus near them along PW2). We present most of our results in terms of N_{PW1} and N_{PW4} because they provide the best separation of all six pathways when projected to two dimensions.

To visualize these pathways, we first identified three ABMD trajectories that corresponded to each pathway other than PW3. We then identified the section of each trajectory that corresponded to phenol unbinding and used k-medoids clustering in the 298-dimensional set of pairwise distances described in Methods to identify 60 cluster centers per pathway. For PW3, we had only one ABMD trajectory (as noted in Methods, we addressed this issue through additional unbiased simulations), so we generated only 20 cluster centers. Regardless, the cluster centers were used to understand the characteristics of each observed pathway. Structural and CV representations of the six unbinding pathways are shown in Figure 3.2B.

As shown in the left panel of Figure 3.2B, PW1 consists of the green phenol moving upwards toward the green gatekeeper residues, Ile^{A10} and His^{F5}, then pushing through them and escaping into the solvent. This corresponds to the green loop in the CV representation. As discussed in the Methods, the bound state exists where $N_{\text{PW1}} \approx 53$ and $N_{\text{PW4}} \approx 32$. Starting from there, PW1 loops down and outward to $N_{\text{PW1}} \approx 85$ and $N_{\text{PW4}} \approx 10$, corresponding to the phenol approaching the green gatekeeper residues. As the phenol pushes through these residues and escapes into the solvent, both N_{PW1} and N_{PW4} decrease to zero.

By contrast, PW4 proceeds in the opposite direction, with the gray phenol passing through the black gatekeeper residues, Leu^{A13} and Leu^{H17}. This leads to the black loop in the CV representation, where the system moves from the bound state to $N_{\text{PW1}} \approx 10$ and

$N_{\mathrm{PW4}} \approx 85$ as the phenol approaches the gatekeeper residues. As the phenol pushes through those residues and escapes into the solvent, both N_{PW1} and N_{PW4} decrease to zero. While PW1 was described in previous simulation studies, [157, 158] PW4 is presented for the first time in this work.

Both PW1 and PW4 occur with a closed escape channel, with each gatekeeper residue remaining in close proximity to its position in the crystal structure. However, very similar pathways can occur with an open escape channel, as the H and F chains twist away from the A chain. This large-scale protein motion, seen in Figure 3.2A, leads to the change in the binding pocket seen in the right panel of Figure 3.2B. Here, the separation between both the green residues and the black residues is increased, meaning they no longer function as gates. For example, along the PW1-like pathway, the phenol still interacts with $\mathrm{Ile}^{\mathrm{A10}}$, but $\mathrm{His}^{\mathrm{F5}}$ is shifted far enough away to allow a layer of water between the side chain and the phenol. We refer to this pathway as PW1a. Similarly, along the PW4-like pathway, the phenol interacts with $\mathrm{Leu}^{\mathrm{A13}}$ but not $\mathrm{Leu}^{\mathrm{H17}}$. We refer to this pathway as PW4a. In the CV representations, both of these pathways are associated with smaller loops in the bottom panels of Figure 3.2B, consistent with the contacts coming from only one gatekeeper residue. Both of these pathways are described for the first time in this work.

Finally, we also observed PW2 and PW3 in our driven simulation data set. These are illustrated in the middle panel of Figure 3.2B. PW2 is a gate-hopping mechanism, where the brown phenol escapes by hopping between the green and black gatekeeper residues. In Figure 3.2B, this is moving out of the plane of the paper, directly toward the reader. In contrast, along PW3, the yellow phenol moves toward the yellow gatekeeper residues, $\mathrm{Ile}^{\mathrm{A2}}$ and $\mathrm{Tyr}^{\mathrm{A19}}$, before pushing through them and escaping into the solvent. Both of these pathways were described by Vashisth and Abrams [158], but we add an important clarification: along both of these pathways, the escape channel can be either closed or open. However, since PW2 involves an escape through a channel that exists in the crystal structure

and PW3 involves escape through an unrelated channel, neither of these two pathways are particularly affected by the opening/closing of the escape channel. For a discussion of the sequence of release of the six phenols and its relation to the experimentally observed cooperativity of phenol binding to the insulin hexamer [168, 169], see Supplemental Figure 3.8 and associated text in Supplemental Information.

Further characterization of the channel opening. The observed channel opening was not an artifact of the ABMD bias, as we also found it in many of our unbiased simulations seeded with a closed channel. Furthermore, by analyzing our unbiased simulations, we can correlate the channel opening with the rotation of the α helices on the dimeric interface, a conformational transition previously found to be accessible to the insulin dimer [76, 170]. In Figure 3.3A, we quantify channel opening by measuring the average of $\mathrm{RMSD_P}$, which is a distance RMSD that quantifies how much the phenolic binding pocket structurally deviates from the binding pocket present in the crystal structure. This quantity was also compared to the average of $\Phi_{\alpha,\mathrm{BD}}$, the pseudodihedral angle of the α helices on the dimeric interface of chains B and D (shown in orange in Figure 3.3B), as defined in ref. 76.

The bound state, marked by the white star in Figure 3.3A and located at $N_{\mathrm{PW1}} \approx 53$ and $N_{\mathrm{PW4}} \approx 32$, is characterized by a relatively low average $\mathrm{RMSD_P}$ of approximately 0.1 Å. The phenolic binding pocket in this state is quite similar to the one in the crystal structure. Across the space defined by N_{PW1} and N_{PW4}, $\mathrm{RMSD_P}$ is highly correlated with $\Phi_{\alpha,\mathrm{BD}}$. The BD dimer, along with the dimer defined by the interface made of chains F and H (shown in blue in Figure 3.3B), makes up the primary escape channel for the phenol studied. The average of $\Phi_{\alpha,\mathrm{FH}}$ is nearly identical to Figure 3.3B, so is omitted for clarity. Both have a value of approximately 132° in the bound state (Figure 3.3B). As discussed previously, this leads to both the green gatekeeper residues ($\mathrm{Ile^{A10}}$ and $\mathrm{His^{F5}}$) and the black gatekeeper residues ($\mathrm{Leu^{A13}}$ and $\mathrm{Leu^{H17}}$) remaining in close proximity, acting as closed gates that define the boundaries of the escape channel.

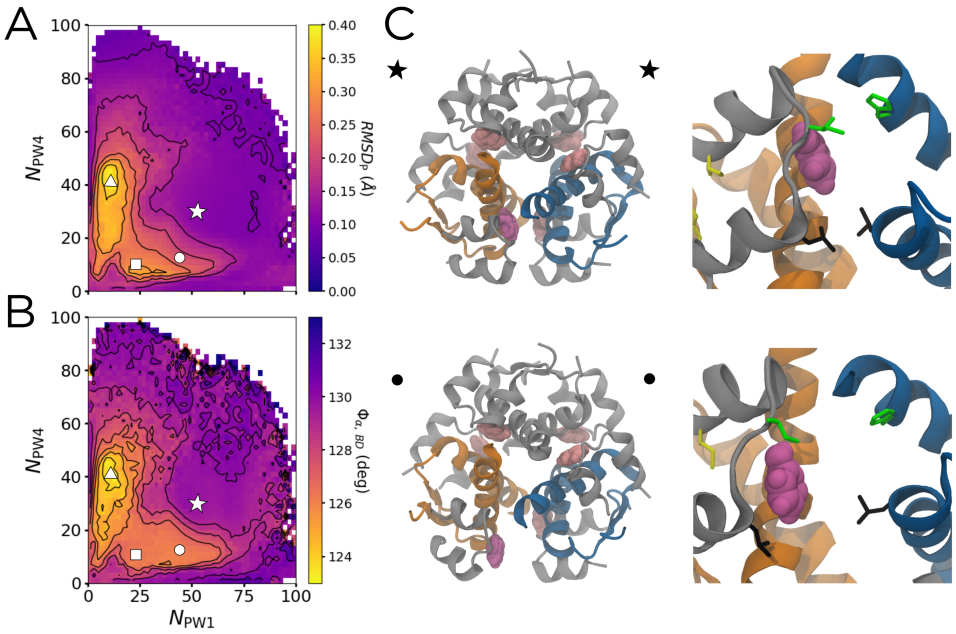


Figure 3.3: Measures of flexibility of the phenolic binding pocket. The bound state is marked by the white star, and the unbound state is marked by the white dashed lines. The circle and square represent the partially-open escape channel with the phenol bound and partially unbound, respectively. The triangle represents a PW3 intermediate in which the A chain is partially melted. (A) The distance RMSD of the 22 binding pocket residues as a function of N_{PW1} and N_{PW4}. Contours spacing is 0.5 Å. (B) The pseudodihedral angle between the α helices at the dimer interface between chains B and D, $\Phi_{\alpha,BD}$, as a function of N_{PW1} and N_{PW4}. Contour spacing is 1°. (C) Hexameric (left) and binding pocket (right) structures showing the closed-to-open transition indicated by the star and circle in (A) and (B), respectively. Chains B and D are shown in orange, while chains F and H are shown in blue.

Moving from the star to circle to square in Figure 3.3A, $RMSD_P$ increases from 0.1 to 0.2 to 0.3Å, corresponding to a three degree rotation of the α helices on the dimer interfaces. Structurally, this dimeric rotation leads to separation of the four gatekeeper residues that define the edges of the phenolic escape channel, as seen in the right panels in Figure 3.3C. Specifically, again moving from star to circle to square, the average separation of the α carbons of Leu^{A13} and Leu^{H17} increases from 0.8 to 1.2 to 1.4 Å, and the average separation of the α carbons of Ile^{A10} and His^{F5} increases from 0.8 to 0.9 to 1.2 Å (Supplemental Figure 3.9A). While the circle and square both correspond to structures with an partially open escape channel, the square represents a slightly more dramatic channel opening that is

paired with the breaking of the hydrogen bonds between the phenol and both the backbone carbonyl of Cys^{A6} and the backbone amide nitrogen of Cys^{A11}; although the phenol is still located in the binding pocket, it is more loosely bound (Supplemental Figure 3.9B). Overall, this channel opening, found in both our biased and unbiased simulations, is also consistent with the displacements observed in the lowest frequency modes of a normal mode analysis of an elastic network representation of the hexamer (Supplemental Figure 3.10), suggesting that this flexibility is an intrinsic feature of the structure. The channel opening allows for increased penetration of solvent closer to the phenol and decreased steric occlusion by the side chains.

Finally, the point marked by the triangle in Figure 3.3 has a $RMSD_P$ between 0.5 - 1.0 Å higher than those for the points marked by the square and the circle. This corresponds to the melting of the A-chain C-terminal α helix along PW3, as the phenol escapes through the A chain (Supplemental Figure 3.9C). As this helix melts, the PW1 and PW4 gatekeeper residues separate even further (Figure 3.3B and Supplemental Figure 3.9A); although this motion affects the binding pocket, it occurs after the phenol is no longer in it.

Combining the potential of mean force, committor, and reactive current enables quantitative characterization of transition states and intermediates. DGA can be used to estimate the relation between the sampled distribution and the stationary distribution and in turn potentials of mean force (PMFs). However, the true power of DGA is in the ability to combine these free energies with estimates for dynamical statistics like the committor, reactive current, and rate. The committor describes the probability of proceeding to a product state before returning to a reactant state. By explicitly defining bound and unbound states (see Methods), we can thus calculate the unbinding committor, q_{unbind}, that describes, at every point in our unbiased data set, the probability of proceeding to the unbound state before returning to the bound state. This is, by construction, the perfect reaction coordinate to track phenol unbinding, as it measures the likelihood of phenol unbinding regardless of

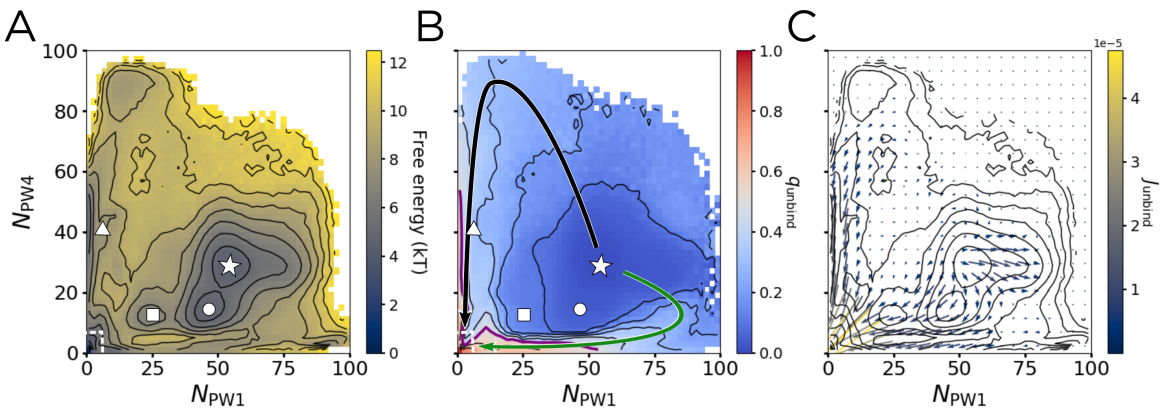


Figure 3.4: The potential of mean force (PMF), unbinding committor (q_{unbind}), and unbinding reactive current (J_{unbind}) projected on N_{PW1} and N_{PW4}. The points marked by the star, circle, square, and triangle are the same as in Figure 3.3, with the unbound state outlined by the dashed white box. (A) The PMF, with contours spaced by $1k_BT$. (B) q_{unbind}, with contours spaced by 0.1 and the $q_{\text{unbind}} = 0.5$ surface marked in purple. Arrows showing PW1 (green) and PW4 (black) are overlaid. (C) J_{unbind} binned into a 22×22 grid spanning from 0 to 100 in both N_{PW1} and N_{PW4}. The results shown are smoothed with a Gaussian filter, using a kernel with standard deviation of 1 bin. Contours are the same as in (A) to aid in comparison.

pathway. By definition, transition states have $q_{\text{unbind}} = 0.5$. By projecting q_{unbind} into various CV spaces, we can distinguish molecular rearrangements (e.g., those giving rise to the metastable states that we identify on a PMF) that increase the probability of unbinding from those that do not.

Another useful quantity is the reactive current, which describes how trajectories that lead to phenol unbinding flow through each point in a CV space. Just like a diagram of fluxes between clusters of states in a Markov State Model [20, 22, 23, 166, 167], a plot of the reactive current can give a sense of the populations of different pathways. Here, as established by recent work [26], we represent the reactive current, J_{unbind}, as a vector field. An advantage to this representation is that it can be compared directly with the PMF and commmittor as functions of the same CVs. By combining these statistics, we can characterize the transition states and intermediates along the six pathways. J_{unbind} can furthermore be used to determine their relative weights; integration of J_{unbind} yields the rate.

We show the free energy, committor, and reactive current projected onto N_{PW1} and N_{PW4} in Figure 3.4. The unbound state (bottom left corner in Figure 3.4A) is set to be the zero of free energy. The DGA-generated PMF is in good agreement with one independently generated through Replica Exchange Umbrella Sampling (Supplemental Figure 3.11), suggesting that the unbiased trajectories cover the space sufficiently to draw quantitative conclusions. The bound state with energy 4.5 k_BT is marked by a white star at $N_{\mathrm{PW1}} \approx 53$ and $N_{\mathrm{PW4}} \approx 32$. As discussed in the Introduction, this state involves the phenol making one hydrogen bond with the backbone carbonyl of $\mathrm{Cys}^{\mathrm{A6}}$ and one with the backbone amide NH of $\mathrm{Cys}^{\mathrm{A11}}$. We find that in this state, $\mathrm{His}^{\mathrm{F5}}$ is rotated outward toward the solvent, with the ring pointing away from the phenol (see Figure 3.2B and 3.3C). The free energetic shoulder directly to the right of the star involves a side chain ring flip of this histidine, so that it is instead facing inward, toward the phenol, which increases the number of contacts between the phenol and the histidine ring. This is consistent with a previous ^{1}H-NMR study that suggested the presence of an unidentified aromatic ring-flip [155], and a computational study which proposed $\mathrm{His}^{\mathrm{H5}}$ as one of the residues that could undergo such a flip [158].

The point marked by a circle in 3.4A corresponds to a free energy basin, confirming that the channel opened state is energetically stable. Furthermore, this transition only involves a free energy barrier of about 1 k_BT, suggesting that these structures can readily interconvert. Corroborating this idea, all of these areas have an average q_{unbind} of less than 0.1 (Figure 3.4B), meaning that both the channel opening and the histidine ring-flip do not markedly increase the probability of escape before returning to the bound state. Interestingly, the channel opening itself does not lead to a marked increase in q_{unbind}. This suggests that while channel opening may play a part in phenol unbinding, such a motion alone is not enough to allow the phenol to escape. The point marked by the square, which corresponds to the phenol breaking its hydrogen bonds to the channel-opened structure, also lies in a free energetic basin, one separated from the channel-opened bound state by a barrier of

about 3 k_BT. This barrier is three times as high as the one for channel opening. and also corresponds to a 0.1 increase in q_{unbind}. The breaking of the hydrogen bonds to Cys^{A6} and Cys^{A11} thus slightly increases the probability of successful unbinding, while the channel opening by itself does not.

As discussed earlier, this projection effectively separates PW1 and PW4, marked by the solid black and green arrows in Figure 3.4B. Comparing these arrows to the PMF in Figure 3.4A, one finds the barriers for PW1 and PW4 to be 4-5 k_BT and 5-6 k_BT, respectively, significantly lower than the 20-30 k_BT barriers found by Vashisth and Abrams [158]. Our free energy barriers change very little when projected into a three-dimensional space that further separates PW1 and PW4 (Supplemental Figure 3.12). Vashisth and Abrams estimate free energies by applying Jarzynski's equality to steered molecular dynamics simulations, which we would expect to overestimate barriers because such simulations will only rarely sample trajectories with protein fluctuations that anticipate the phenol motion. By contrast, our unbiased trajectories, as well as our independent Replica Exchange Umbrella Sampling (Supplemental Figure 3.11), capture these dynamics.

Additionally, comparing Figure 3.4 to Figure 3.3, it is clear that there is very little opening of the escape channel inherent along either PW1 or PW4, as the average RMSD_{P} along these pathways stays below 0.1 Å. Instead, the phenol is directly pushing through the adjacent gatekeeper residues. However, this projection does not allow us to fully distinguish PW2, PW3, PW1a, and PW4a because they partially overlap (see Figure 3.2B). At the same time, the area through which they pass has the highest magnitude of reactive current (Figure 3.4C), suggesting that that the dominant binding pathways lie in this region. To distinguish these pathways, we introduce a third dimension.

Three-dimensional projections show that competing pathways are similarly populated. Since the unbinding committor q_{unbind} tracks phenol unbinding across all pathways, we use it as the third dimension on which to project our data. From our unbiased

data set, we visually identify four trajectories for each of our six unbinding pathways. These trajectories are shown in Figure 3.5A with coloring consistent with Figure 3.2B. We also show the full unbiased data set, colored to show q_{unbind} and with arrows representing the six pathways overlaid in Figure 3.5B. A structural representation of these pathways is shown in Figure 3.5C.

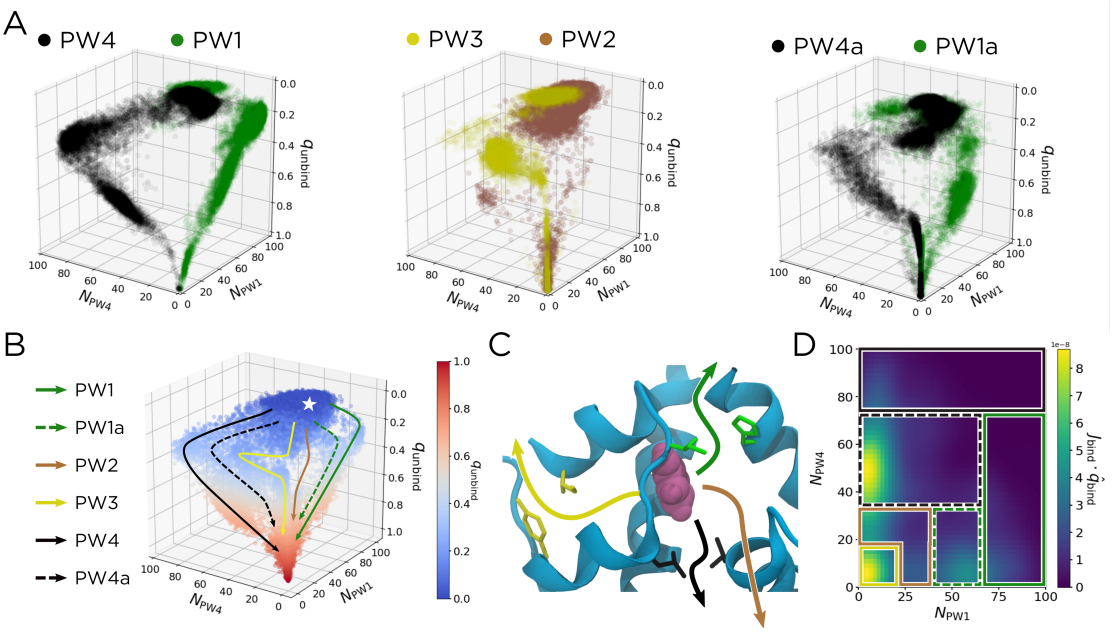


Figure 3.5: Pathways can be more readily distinguished in the space of N_{PW1}, N_{PW4}, and q_{unbind}. (A) Scatter plots of trajectories along each of our six pathways. From the unbiased data set, we identify four trajectories that correspond to each pathway, which are shown for PW1/PW4 (left), PW2/PW3 (middle), and PW1a/PW4a (right), mirroring the conventions used in Figure 3.2B. (B) A scatter plot of q_{unbind} for all the unbiased data. Six pathways are overlaid and labelled. The bound state is represented by the white star. (C) Structural representation of the unbinding pathways in (B). The dashed arrows in (B) correspond to the similarly-colored solid arrows in (C), except with the escape channel being opened as in Figure 3.3. (E) The q_{bind} component of J_{bind}, taken at $q_{\text{bind}} = 0.63$. The patches corresponding to each pathway are overlaid, using the coloring and line styles from (B).

As shown in Figures 3.5A and 3.5B, using q_{unbind} as the third dimension clearly separates the six pathways. The $q_{\text{unbind}} = 0.5$ surface (gray points in Figure 3.5B) is the transition state ensemble, as structures in this region have an equal probability of proceeding to the unbound state as they do of returning to the bound state. Notably, for PW1/PW1a and

PW4/PW4a, this does not occur until after N_{PW1} and N_{PW4} have reached their respective maxima along each pathway. The transition states thus occur when the phenol is just outside the gatekeeper residues, and the numbers of contacts with those residues have started to decrease. In contrast, for PW3, the $q_{\mathrm{unbind}} = 0.5$ surface occurs just as the phenol maximizes its contacts with $\mathrm{Ile}^{\mathrm{A2}}$ and $\mathrm{Tyr}^{\mathrm{A19}}$, the PW3 gatekeeper residues (Supplemental Figure 3.13). Since PW2 encompasses all gate-hopping mechanisms and the phenol can be closer to the gatekeeper residues for PW1 or those for PW4, its transition state is comparatively structurally diverse.

We can also use DGA to calculate the relative weight of each unbinding pathway. To do this, we adapt the method described in ref. 26, using the binding reactive current J_{bind} projected into the space of N_{PW1}, N_{PW4}, and q_{bind}. For this calculation only, we use kinetic statistics for binding as opposed to unbinding, as this slightly improved our ability to distinguish the six pathways (Supplemental Figure 3.14). We choose as a dividing surface the plane corresponding to $q_{\mathrm{bind}} = 0.63$, as this best separates the reaction pathways (Supplemental Figure 3.15). We then partition this surface into patches that encompass the pathways (Figure 3.5D). By binning the reactive current in the direction of q_{bind} across these patches, one can calculate the relative weight of each pathway as the ratio of the reactive current in each patch compared to the total reactive current flowing through the dividing surface. We provide exact definitions of the patches, as well as a discussion of the parameters used to calculate and smooth J_{bind}, in the Supplemental Information. Using this method, we determined the relative weights of the six identified pathways (Table 3.1). These relative weights are robust to changes in the sampling distribution (Supplemental Table 3.4) and choice of dividing surface (Supplemental Figure 3.16).

The pathway with the clear plurality of reactive current (35.0%) is PW4a (dashed black arrow), which corresponds to an unbinding mechanism with an open channel, where the phenol closely passes by $\mathrm{Leu}^{\mathrm{A13}}$ as it leaves the binding pocket. While this is the most likely

Table 3.1: Relative weights of unbinding mechanisms for WT insulin and the A10 Ile→Val and B13 Glu→Gln mutations.

Pathway	WT (%)	A10 (%)	B13 (%)
PW1	11.2	4.3	19.0
PW1a	11.2	7.8	16.0
PW2	16.3	16.7	14.1
PW3	12.6	8.7	7.1
PW4	13.7	18.4	26.7
PW4a	35.0	44.2	17.2

unbinding mechanism, all of the other pathways are significantly populated (ranging from 11% to 17%). Furthermore, the preference for mechanisms with an open escape channel is relatively mild, with the PW1/PW4 percentage being 24.9% compared to the PW1a/PW4a percentage being 46.2%. As a note, as defined in this study, PW2 and PW3 can have either a closed or open escape channel. So while channel opening does play a role in the overall phenol unbinding process, the majority (53.8%) of unbinding events occur through pathways that do not necessarily involve channel opening. As we discuss further below, the manifestly multipathway nature of the unbinding process makes it challenging to predict how point mutations will impact phenol unbinding.

A10 and B13 mutations alter the preferred pathway and the binding/unbinding rates. As discussed in the Introduction, there is much interest in designing insulin mutants and analogs for the management of diabetes, including those which might slow the unbinding of phenol. To this end, we simulated two mutants: A10 Ile→Val and B13 Glu→Gln. The A10 Ile→Val mutation is one of the three sequence differences between human and bovine insulin; since $\mathrm{Ile}^{\mathrm{A10}}$ is one of the gatekeeper residues for PW1/PW1a, we expected its mutation to affect the rate of phenol release. The B13 Glu→Gln mutation removes negative charges from the center of the hexamer and stabilizes the R state, as evidenced by the fact that it leads to the formation of $\mathrm{T_3R_3}$ hexamers even in the absence of zinc [171]. Given the stabilization of the R state, we expected the B13 Glu→Gln mutation to stabilize the $\mathrm{R_6}$ hexamer and

Table 3.2: The inverse unimolecular unbinding rate constant, $k^{-1}_{\text{unbinding}}$, the inverse unimolecular binding rate constant k^{-1}_{binding} and their ratio ($K = k_{\text{unbinding}}/k_{\text{binding}}$). Ranges derive from taking lag times between 500 ps and 1.25 ns.

Statistic	WT	A10	B13
$k^{-1}_{\text{unbinding}}$ (μs)	0.16 - 0.28	0.21 - 0.37	0.17 - 0.27
k^{-1}_{binding} (μs)	0.13 - 0.20	0.12 - 0.18	0.16 - 0.26
K	0.70 - 0.83	0.40 - 0.48	0.97 - 0.99

impact channel opening.

We performed simulations analogous to those above for human insulin with each of these mutations separately and determined their effects on the unimolecular rate constants of unbinding and binding, $k_{\text{unbinding}}$ and k_{binding}, as well as the corresponding ratio, $K = k_{\text{unbinding}}/k_{\text{binding}}$ (Table 3.2 and Supplemental Figure 3.17). We also calculated the relative weights for the six unbinding pathways for each mutant (Table 3.1). Note that k_{binding} is a unimolecular rate constant that does not include the contribution from diffusion, which we expect to be less sensitive to mutations, and consequently K differs from an experimentally measured dissociation constant in that it is a ratio of unimolecular rates. Rate constants that account for diffusion are discussed in the Supplementary Information, with results in Supplemental Figure 3.18 and Supplemental Table 3.5. These estimates provide good quantitative agreement with available experimental values for dissociation constants [172], indicating that DGA yields accurate results. Here, we focus on unimolecular rate constants because we expect them to have fewer sources of error. The binding and unbinding inverse unimolecular rate constants for WT insulin and the two mutants are in the range 0.11-0.31 μs, which is consistent with the expected sub-millisecond phenol unbinding timescale predicted from existing NMR data [155].

We first describe the A10 Ile→Val mutation. We find that the A10 Ile→Val mutation has almost no affect on the phenol binding rate constant, while slightly decreasing the unbinding rate constant (increasing the timescale of unbinding). This, in turn, leads to a 42-43%

decrease in the ratio K compared to WT insulin. To understand how this mutation, which decreases the size of the A10 side chain, inhibits phenol dissociation, we examine the relative weights for the six pathways (Table 3.1). Because A10 is a gatekeeper for PW1/PW1a, we expected the A10 Ile→Val mutation to impact the weights of these pathways most strongly, and indeed this is the case. The decreases in the relative weights of these pathways are corroborated by increases in the free energies and decreases in the unbinding committors and reactive current in associated regions (Supplemental Figure 3.19). Intermediate structures along these pathways are less likely, and when they do occur, they are less likely to lead to the unbound state. The preferred binding pathway remains PW4a, although the relative weights of PW2 and PW4 are all higher than for WT.

Our calculations indicate that the B13 Glu→Gln mutation, while having little effect on phenol unbinding, slows phenol binding. As a result, there is a 19-39% increase in K compared to WT. As seen in Table 3.1, this mutation leads to a dramatic decrease in relative weight for PW4a, the preferred unbinding mechanism for both WT and A10 Ile→Val insulins. This is paired with a corresponding increase in relative weight for both PW1 and PW4, the two mechanisms that explicitly do not include any channel-opening and instead involve the phenol pushing through gatekeeper residues, as well as an increase in relative weight for PW1a, which does involve channel opening. This agrees with our calculations that, after the mutation, the energetic benefit of channel opening is approximately 10 kJ/mol greater along PW1a than along PW4a (Supplemental Figure 3.20). The mutation thus discourages PW4a more than PW1a. The overall shift in relative weight away from PW4a is further corroborated by the PMF, committor, and reactive current (Supplemental Figure 3.19): areas of CV space along PW4a are 2 k_BT higher in free energy compared to the same areas for WT insulin.

Beyond the shifting relative weights between pathways, the overall effect of the mutation is to destabilize the bound state and increase K. Molecularly, this can be explained by the

more favorable interactions the mutated B13 residue can make with the rest of the protein in the free state, particularly with Ser^{B9} and His^{B10} (Supplemental Table 3.6). This prediction agrees with existing experimental data for this mutated species. Dunn and coworkers[172] used UV/Vis spectroscopy and a three-state allosteric model to determine the dissociation constant for WT insulin to be $1.8 \times 10^{-4} \pm 1 \times 10^{-4}$ M, and that for the B13 mutant to be $2.5 \times 10^{-4} \pm 1 \times 10^{-4}$ M, meaning that the mutation caused a 39% increase. Assuming that the main impact of the mutation is on the protein dynamics and not diffusion, this is in agreement with our estimates, which predict a 19-39% increase in K upon B13 Glu→Gln mutation.

For both mutations, there are at least four pathways which each represent at least 10% of the overall reactive current, and no one mechanism ever makes up more than 50%, similar to our findings for WT insulin. As a result, the effects of a mutation on certain pathways can be compensated by those on others. For example, when the A10 Ile→Val mutant discouraged unbinding through PW1 and PW1a, the absolute amount of reactive current flowing through PW4 increased (Supplemental Figure 3.19). Indeed, the multipathway nature of unbinding is the main takeaway from our simulations, and it suggests that future mutation and ligand design studies of the insulin R_6 hexamer need to target multiple pathways at once.

Solvation of phenol and its binding pocket. Given the interplay between the pathways and channel opening, we characterized the solvation of the phenol and select residues by calculating radial distribution functions of water around those species as a function of committor value (Supplemental Figure 3.21). In all cases (WT, A10 Ile→Val, B13 Glu→Gln), the phenol becomes more solvated as it leaves the binding pocket. In general, the radial distribution functions around gatekeeper and pocket residues do not change dramatically as a function of committor. However, we do observe slight changes for three residues: Ile^{A2} (a gatekeeper for PW3), Ile^{A10} (a gatekeeper for PW1/PW1a), and His^{B10} (a pocket residue that also defines the binding site for the Zn^{2+} ions). The increase in solvation of Ile^{A10}

and His^{B10} comes early in the release process, whereas that for Ile^{A2} occurs as the phenol escapes.

We also computed radial distribution functions of water for different $RMSD_P$ values, tracking channel opening (Supplemental Figure 3.22). These data reveal that the gatekeeper residues near the phenol escape channel (Ile^{A10}, His^{F5}, Leu^{A13}, and Leu^{H17}) all become more solvated upon channel opening. The other residues considered are affected less by channel opening. Interestingly, while channel opening does not significantly affect the solvation of Glu^{B13} in WT insulin, it does lead to increased solvation for Glu^{B13} in the B13 Glu→Gln mutant insulin.

Finally, Bagchi and co-workers have argued that water molecules confined in the central cavity stabilize the hexamer [173, 174]. To examine whether the dynamics that we observe can facilitate exchange of water molecules between the cavity and the bulk, we defined the cavity waters as those within 6 Å of the center of mass of the two Zn^{2+} ions and calculated their MSD over 1 ns (right panels in Supplemental Figures 3.21 and 3.22). For comparison, we measure an MSD over 1 ns of 35 nm^2 for bulk waters in our simulations. The MSD of the waters in the cavity shows very little dependence on committor and is generally less than 10 nm^2, meaning that these waters are confined regardless of the dissociation of phenol. By contrast, the MSD of waters in the cavity increases as $RMSD_P$ increases, indicating that channel opening facilitates exchange of solvent between the cavity and the bulk.

3.4 Conclusions

Diabetes management can be improved through both the introduction of insulin analogs with modulated pharmacokinetics, as well as delivery preparations that can facilitate transport and storage. Because phenol stabilizes the R_6 insulin hexamer, understanding the phenol unbinding mechanism can inform the design of improved therapeutics. Here, we use molecular dynamics simulations to investigate this mechanism for WT and two mutant insulins.

We expand on existing simulation studies by Swegat *et al.* [157] and Vashisth and Abrams [158] to identify and quantitatively characterize six phenol binding/unbinding pathways. A number of these pathways involved large-scale opening of the primary escape channel, suggesting that the hexamer is much more dynamic than previously appreciated. Methods that we recently introduced [25, 26] enable us to determine the intermediates, transition states, and relative weights of the pathways. For WT insulin, a pathway in which the channel opens and phenol passes between Leu^{A13} and Leu^{H17} (PW4a) represents 40% of the reactive current, but each of the other pathways represents at least 10% of the reactive current. Phenol unbinding/binding is thus a multipathway process.

Our simulations of mutants show that it is possible to rationally control the prevalence of pathways and the overall unbinding kinetics. The A10 Ile→Val mutant reduced the contributions from pathways for which this residue is a gatekeeper (PW1 and PW1a) and decreased phenol unbinding; the B13 Glu→Gln mutation stabilized the phenol-free state and thus led to increased phenol unbinding. However, because other pathways than those targeted can compensate, the overall effects on rates can be challenging to predict without quantitative simulations like those presented here. By combining computation and experiment, it may be possible to target multiple pathways to achieve larger shifts in kinetics.

3.5 Supplemental Information

Scripts with an implementation of DGA are available at https://github.com/dinner-group/-insulin-hexamer.

Nomenclature. When numbering the six phenols, we followed the conventions in the PDB file. This differs from the numbering convention in ref. 158, which instead assigns phenols 1, 2, and 3 to the "top" trimer, and phenols 4, 5, and 6 to the "bottom" trimer. Furthermore, our nomenclature for protein chains also differs from ref. 158. In particular, for WT insulin,

we choose the nomenclature such that both the A and B chain belong to the insulin monomer closest to phenol 4, which we use to study the phenol unbinding process for WT insulin. From there, the naming trends are identical to the PDB/literature naming convention, with chains ranging from A to L - only the starting point (which chain we designate as A) is shifted. For the mutant systems, we simulate the release of phenol 2, instead of phenol 4. For these systems, then, we followed a similar convention, and denote the monomer closest to phenol 2 as having the A and B chains.

Selecting starting structures for unbiased simulations. To ensure sampling in all areas of our CV space even with relatively short trajectories (40 ns), we initialize the unbiased simulations as follows. We first run a large number of adiabatic-bias molecular dynamics (ABMD) simulations to bias the system toward phenol release, as discussed in the main text. We then define a grid of $10 \times 10 \times 10$ points that covers the sampled regions and find the single frame from our driven database closest to each point (Figure 3.6). Although there are 1000 grid points, the same structure can be the closest to more than one grid point. In the WT case, we obtain 326 unique structures.

DGA details. DGA requires the use of a basis set. We used a basis set of 298 modified pairwise distances and one constant function as described in the main text. A summary of the distances is given in Table 3.3. For categories involving the phenol, we measure distances from both C_1 and C_4; thus there are $2 \times 22 = 44$ distances to the C_α atoms of the 22 residues in the binding pocket and $2 \times 4 = 8$ distances to the C_α atoms of the 4 gatekeeper residues for PW1 and PW4. The PW3 gatekeeper residues in Table 3.3 are A2, A19, and B25.

In this work, we modified the guess functions $\psi_{\text{binding}}(x)$ and $\psi_{\text{unbinding}}(x)$ from their definitions in ref. 26, so $\psi_{\text{binding}}(x) = 1 - \psi_{\text{unbinding}}(x)$:

$$\psi_{\text{binding}}(x) = \frac{d^2_{\text{unbound}}}{d^2_{\text{bound}} + d^2_{\text{unbound}}} \tag{3.1}$$

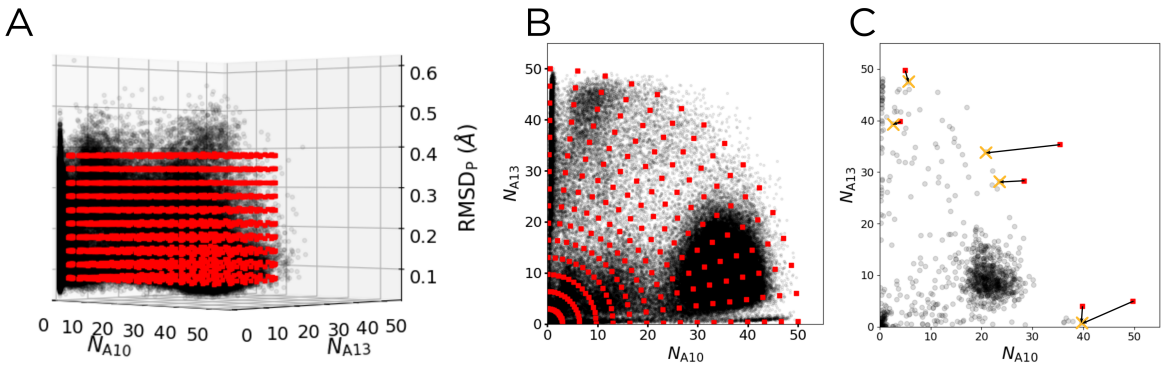


Figure 3.6: A schematic showing how we chose unique starting structures for the unbiased sampling. (A) 3D representation in the space of N_{A10}, N_{A13}, and $\mathrm{RMSP_P}$. Data from our ABMD database is shown by the black dots, and our desired starting points are shown by the red squares. (B) A two-dimensional slice of (A), more clearly showing the r and θ dependence of our desired starting points. (C) A schematic illustrating how we select the closest structures to each desired point. The frame from the ABMD data set closest to each desired starting point is represented by the orange X. For clarity, we only display six desired starting points that lead to five unique starting structures.

$$\psi_{\mathrm{unbinding}}(x) = \frac{d^2_{\mathrm{bound}}}{d^2_{\mathrm{bound}} + d^2_{\mathrm{unbound}}} \tag{3.2}$$

These choices ensure that $q_{\mathrm{unbinding}} = 1 - q_{\mathrm{binding}}$. Above, d_{bound} and d_{unbound} are the smallest Euclidean distances in the 298-dimensional space of pairwise distances (Table 3.3) from any point in the reactive domain (i.e., outside the bound and unbound states) in the data set to any point in the bound and unbound states, respectively.

Table 3.3: Description and number of pairwise distances used as inputs to make our DGA basis functions. Note that to make the eventual 299-dimensional set of basis functions, we also include the constant function.

Type of Distance	Number
Phenol C_1/C_4 atoms to Binding Pocket C_α atoms	44
Binding Pocket C_α atoms to Binding Pocket C_α atoms	231
Phenol C_1/C_4 atoms to PW1/PW4 gatekeeper C_γ atoms	8
PW1/PW4 C_γ atoms to PW1/PW4 gatekeeper C_γ atoms	6
Phenol C_1/C_4 atoms to PW3 gatekeeper C_α atoms	6
PW3 gatekeeper C_α atoms to PW3 gatekeeper C_α atoms	3
Total	298

Since DGA does not enforce the fact that committors are probabilities and thus must be between zero and one, it produces estimates between –0.2 and 1.2. We shift those below zero to zero and those above one to one before using them for further analysis (plotting and reactive current calculations).

In addition, one of the essential parameters for DGA is the lag time [26] (see also refs. 175 and 176). We calculated statistics for lag times ranging from 10 ps to 10 ns. For WT insulin, we found that the $q_{\text{unbind}} = 0.5$ surface was approximately constant as the lag time changed from 500 ps to 5 ns. For the mutant simulations, this was only the case for lag times greater than 1.25 ns. Similar behavior was observed for the relative weights of the six pathways (Figure 3.7).

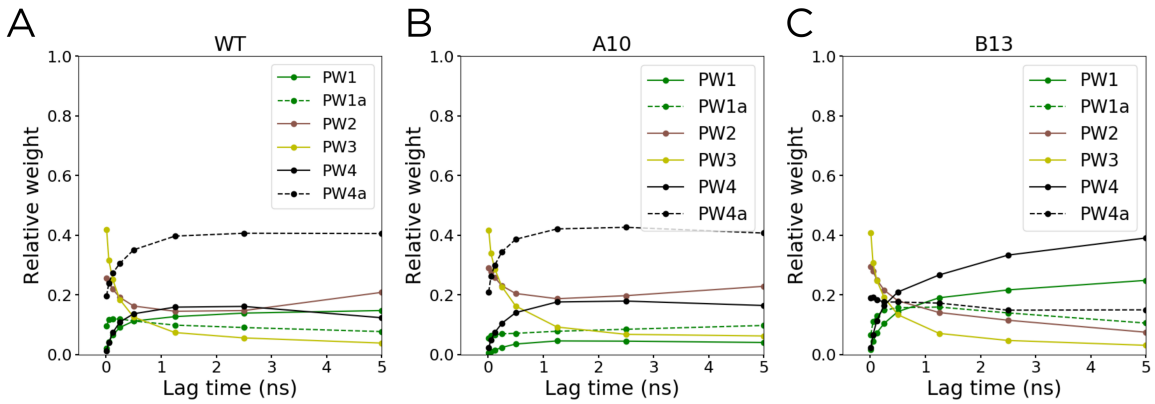


Figure 3.7: The relative weight of the six identified pathways as a function of lag time for (A) WT insulin, (B) A10 Ile → Val insulin, and (C) B13 Glu → Gln insulin.

Cooperativity of phenol release. We examined whether there was evidence in our ABMD simulations of positive intra-trimer cooperativity and negative inter-trimer cooperativity for phenolic binding, as suggested by stopped-flow spectroscopy[169] and isothermal titrating calorimetry[168]. In addition to the 276 simulations described in the main text, which favored increasing the distance d_n $(1 \leq n \leq 6)$ between each phenol and the closest bound zinc ion, we also performed 140 simulations that instead favored decreasing the number of non-hydrogen contacts between the protein and each phenol. 10 simulations (each

of length 5 ns) were run for each of 14 force constants evenly spaced between 3×10^{-9} to 16×10^{-9} kJ/mol. In general, fewer phenols were released, and none of the 140 simulations led to dissociation of all six phenols, compared to 77/276 for the simulations biasing on d_n.

Plots summarizing the sequences of release for the ABMD simulations are shown in Figure 3.8. For clarity, transitions observed in less than 10% of each set of simulations are not drawn. For both sets of simulations, the most common sequence of release is phenol 4, followed by phenol 6 (both in the same trimer, here labeled trimer 1). For the simulations biasing on the number of non-hydrogen contacts between phenol and protein (Figure 3.8B), the next most likely release is that of phenol 2, also in trimer 1. For the simulations biasing of the distance between the phenol and the nearest bound zinc, the most likely third release is either phenol 2 or phenol 1, which is in trimer 2. These data support the existence of negative inter-trimer cooperativity, as the phenols in trimer 1 are preferentially released over the phenols in trimer 2. Because we do not see any large-scale motions that correlate with the sequence of release, we presume that the cooperativity in the simulations reflects subtle differences between the two trimers in the starting structure, but which differences are most important is not apparent.

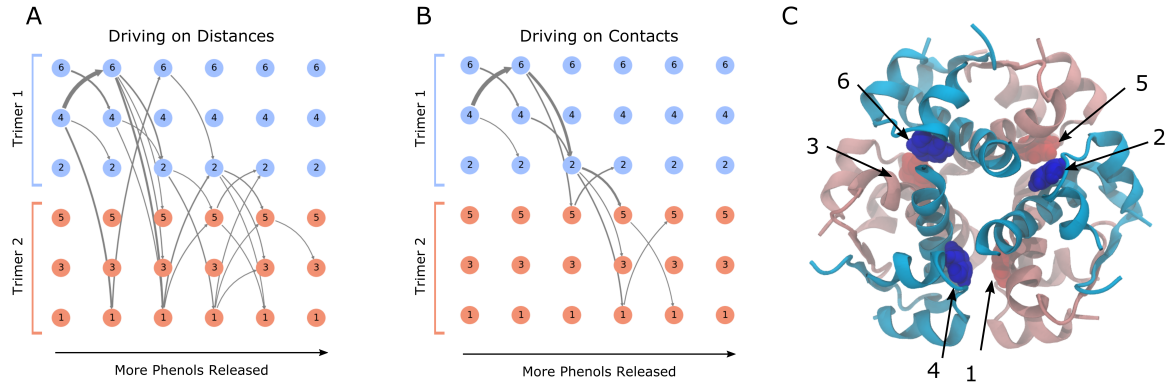


Figure 3.8: The sequence of phenol release for the ABMD simulations biasing on (A) the distance between each phenol and the closest bound zinc, and (B) the number of non-hydrogen contacts between phenol and protein. Phenols bound to trimer 1 and trimer 2 are represented by the blue and red circles, respectively. The size of an arrow represents the relative weight of the indicated transition. (C) The hexamer, colored as in (A) and (B), with phenols labeled.

Channel opening analysis. Using our unbiased data, we can characterize channel opening by describing d_{PW1} and d_{PW4}, the distances between the α carbons of the gatekeeper residues along PW1 and PW4, respectively (Figure 3.9A). We also calculate N_{HP}, the two-step rolling average of the number of hydrogen bonds between the phenol and both the backbone carbonyl of $\mathrm{Cys}^{\mathrm{A6}}$ and the backbone amide NH of $\mathrm{Cys}^{\mathrm{A11}}$ (Figure 3.9B). Finally, we calculate A_{hel}, the helicity of the C-terminal A-chain α helix, A13-A21 (Figure 3.9C). This is the effective number of six-residue segments in the selection in an idealized α helical conformation, based on RMSD [177]; a value of four corresponds to a fully structured C-terminal α helix, while a value of zero corresponds to a fully melted helix.

To further probe the stability of the phenolic escape channel, normal mode analysis was performed on the crystal structure of the WT hexamer. We used WebNMA 3.3 [178], which creates an elastic network model from all of the α carbons in a protein, and solves for the normal modes of this oscillator system. The normalized squared displacements of these α carbons for the four lowest modes (excluding the translational/rotational modes and any identical modes due to the symmetry of the system) are shown in Figure 3.10, where the displacements have been averaged across all 6 monomers to provide a representative set of displacements.

The areas of highest flexibility for the dominant normal modes correlate well with the A-chain α helices and the B-chain β turn, shown in gray on the left and right, respectively. The spikes at B1 and B31 are attributed to the large flexibility of terminal residues. These data suggest an area of flexibility between the A chains and adjacent B-chain β turns in the hexameric structure, the same area that forms the phenolic binding pocket/escape channel.

Comparing the REUS PMF and the DGA PMF. To validate the PMF generated by DGA, we ran Replica Exchange Umbrella Sampling (REUS) simulations and used them to compute an independent PMF. To generate starting structures for the REUS, we first created a 20×20 evenly spaced grid in the cylindrical space of (r, θ), where $N_{\mathrm{PW1}} = r\cos\theta$

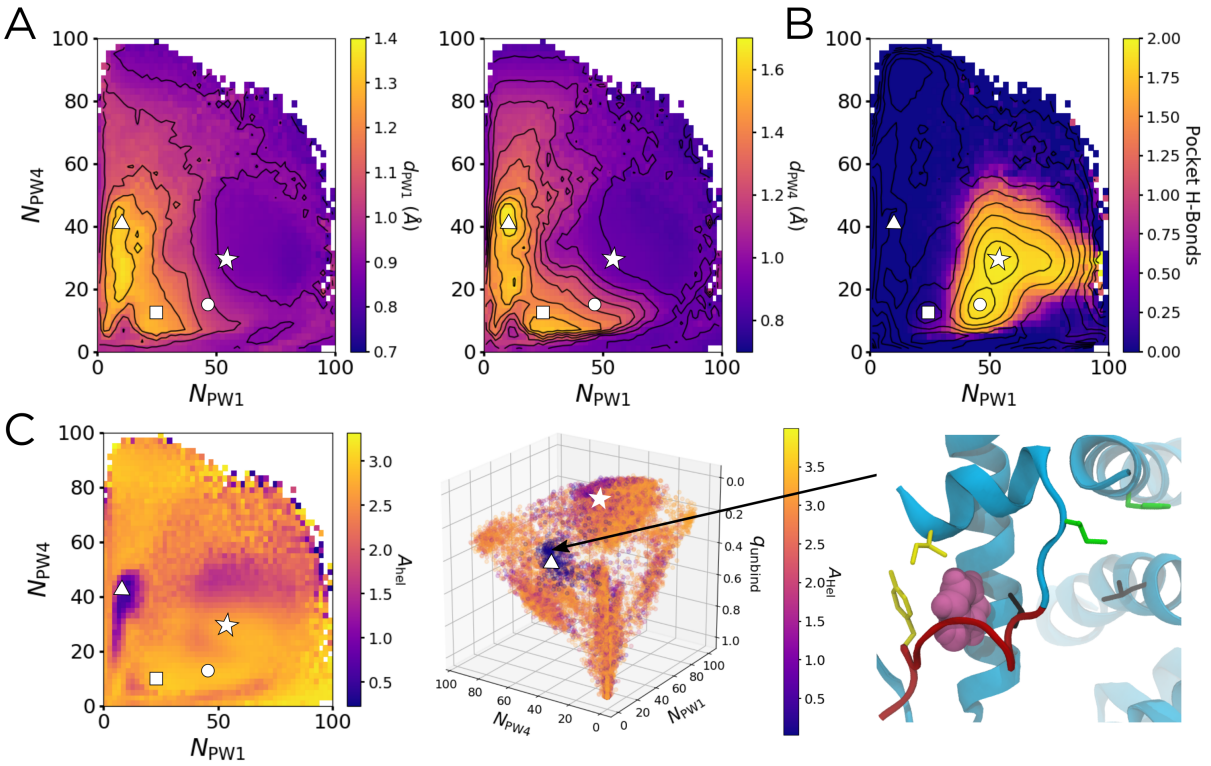


Figure 3.9: Averages of observables, taken from our unbiased data set, associated with different aspects of channel opening, and projected using N_{PW1}, N_{PW4}, and N_{PW3}. The star, circle, square, and triangle mark the same landmarks as in Figures 2 and 3 in the main text. (A) The average of d_{PW1} (left) and d_{PW4} (right) as a function of N_{PW1} and N_{PW4}, with contours shown every 0.1 Å. (B) The average of N_{HP}, with contours from the WT insulin PMF overlaid. (C) The average of A_{hel} as a function of N_{PW1} and N_{PW4} (left) and as a scatter plot in the space of N_{PW1}, N_{PW4}, and $q_{unbinding}$ (middle). A structural representation of the melted C-terminal A-chain α helix (red) along PW3 (triangle) is shown in the right panel. Gatekeeper residues are shown as in the main text.

and $N_{PW4} = r \sin\theta$. Structures closest to each one of these grid points were drawn from our unbiased simulation database and used to initialize the windows. Harmonic biases were then placed on N_{PW1} and N_{PW4}, with window strengths, k, set by applying eq. 4 in ref. 76, with a maximum possible $k = 3$ kJ/mol. The simulation and 2D exchange procedure of ref. 76 was also followed, simulating for a total of 2 ns per window, for an aggregate sampling time of 800 ns. The PMF was constructed from this sampling using the Eigenvector Method for Umbrella Sampling (EMUS) [93] extended to replica exchange data [76]. The resulting PMF and related statistics are shown in Figure 3.11.

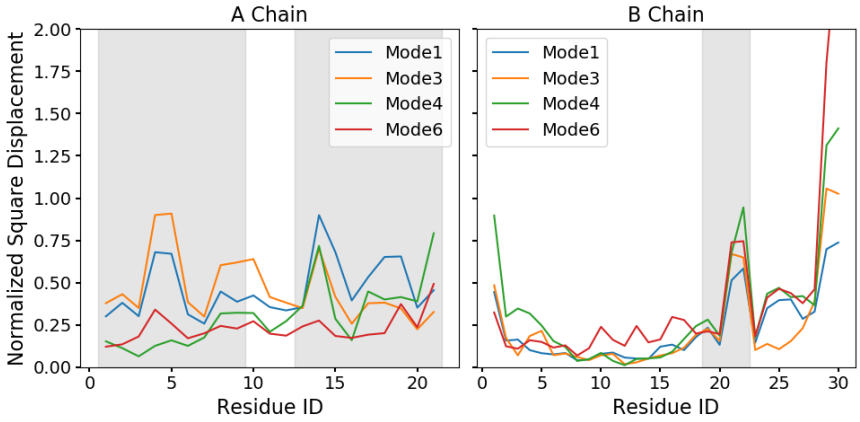


Figure 3.10: Squared displacements from normal mode analysis, where each mode has been normalized so that the sum of the displacements equals 100. The displacements are averaged across the six monomers. The gray areas mark relevant secondary structure elements: the N and C terminal α A-chain helices, A1-A9 and A13-21, respectively, and the B-chain β turn, B18-B22.

The qualitative features of the DGA and REUS PMFs are quite similar, including stable and metastable basin positions. Relative to the REUS PMF, the DGA PMF appears to overestimate the free energy of the bound state by approximately 2 k_BT and to underestimate the free energy in the upper left corner of the PMF (along PW4) by approximately 1-2 k_BT. However, we take the overall agreement to be an indication that our sampling is sufficient.

We also projected the DGA-generated PMF into the space of N_{PW1}, N_{PW4}, and q_{unbind} as discussed in the main text. This PMF is shown for multiple different slices of q_{unbind} in Figure 3.12. In this representation, the free energy barrier for PW4 is the same as for the 2D projection (approximately 5-6 k_BT). In contrast, the free energy barrier for PW1 is somewhat higher (6-8 k_BT in the 3D case compared with 4-5 k_BT in the 2D case). This is not unexpected, as the 3D projection allows us to more easily separate PW1 from PW1a, and from the area of the PMF associated with the (energetically stable) $\mathrm{His}^{\mathrm{F5}}$ ring flip in the bound state. Despite these minor differences, we conclude that the free energy barriers for PW1 and PW4 are comparable.

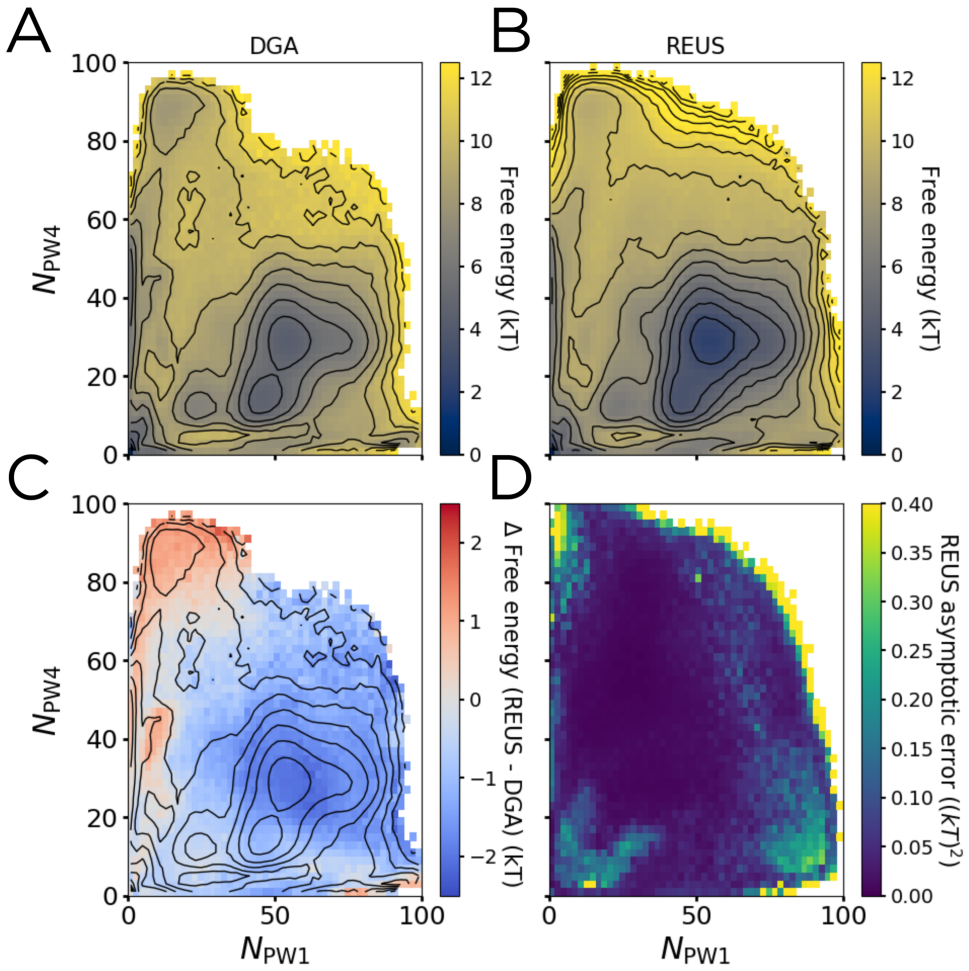


Figure 3.11: Comparison of PMFs generated using REUS and DGA. (A) DGA and (B) REUS PMFs with contours shown every 1 k_BT. (C) The difference of the two PMFs , subtracting (A) from (B), with contours from the DGA PMF superimposed to guide the eye. (D) The asymptotic variance of the REUS PMF.

Describing PW3 in 3D CV space. To help determine the transition state ensemble for PW3, we projected the unbinding committor q_{unbind} onto the space of N_{PW1}, N_{PW4}, and N_{PW3} (Figure 3.13). This shows that the $q_{\text{unbind}} = 0.5$ surface occurs where $N_{\text{PW3}} \approx 80-100$, the maximum value N_{PW3} obtains along PW3. This is in contrast to the transition states along PW1 and PW4, which occur when $N_{\text{PW1}} \approx 60$ or $N_{\text{PW4}} \approx 60$. These states correspond to phenol having already partially escaped from the crystallographic binding pocket, which occurs once the number of contacts with the corresponding gatekeeper residues

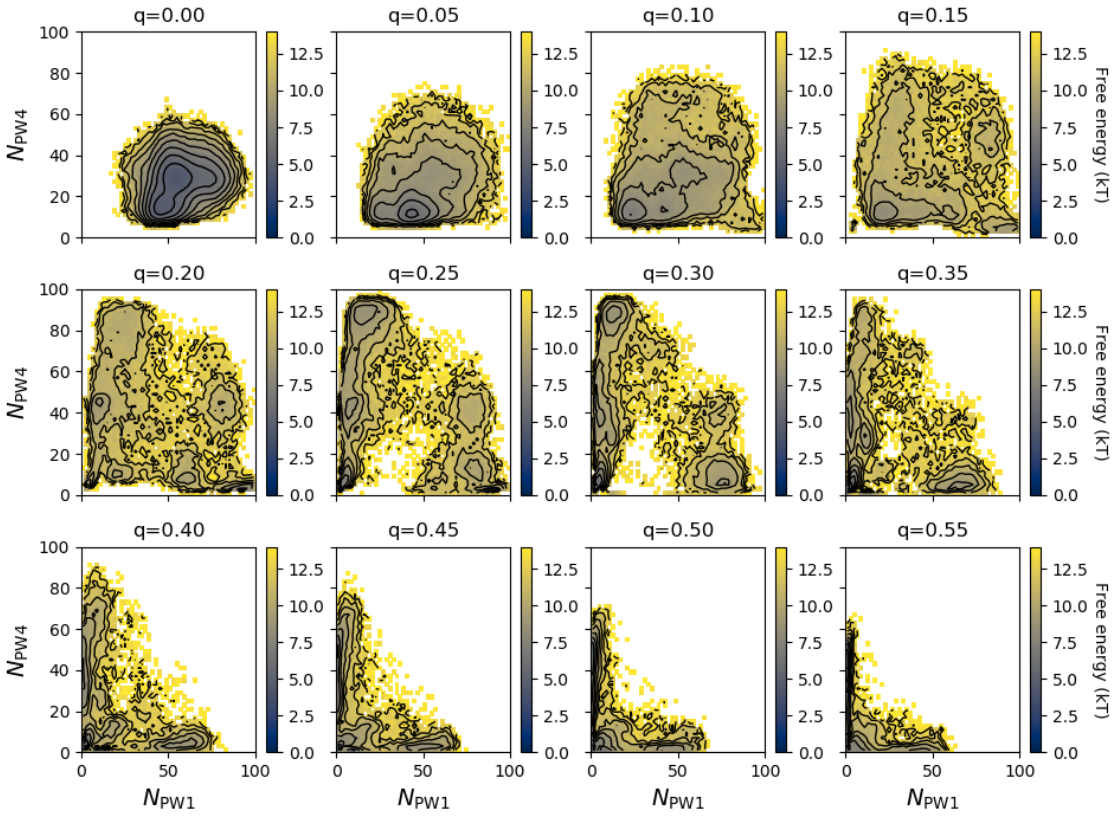


Figure 3.12: The PMF in the space of N_{PW1}, N_{PW4}, and q_{unbind}, shown at indicated slices of q_{unbind}. Contours spacing is 1 k_BT. The minimum free energies for the panels in the first row ($q = 0$ to $q = 0.15$) are 4.0, 5.7, 7.0, and 8.2 k_BT, from left to right.

have begun to decrease. In contrast, along PW3, the phenol has to push through a sterically occluded region of the A chain to reach the gatekeeper residues, meaning that the phenol has already partially escaped the binding pocket by the time it reaches them. Thus, despite the seemingly different locations of the $q_{\mathrm{unbind}} = 0.5$ surface along PW1, PW3, and PW4, all of the transition states exist once the phenol has partially escaped from the binding pocket.

Furthermore, the projection of N_{PW3} into the space of N_{PW1}, N_{PW4}, and $q_{\mathrm{unbinding}}$ (Figure 3.13) shows where this region of large N_{PW3} occurs in the projection used in the main text, further verifying the pathway definitions described there. Similarly, $\mathrm{RMSD_P}$ projected onto the same space shows that the areas of elevated $\mathrm{RMSD_P}$ correspond to PW1a, PW4a, and PW3, with the highest values corresponding to PW3. This is consistent with both the

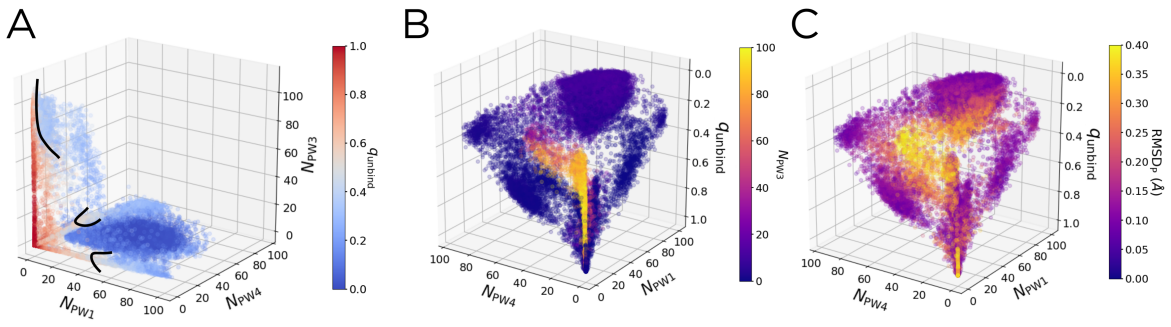


Figure 3.13: The committor and other statistics projected into three dimensions. (A) The unbinding committor q_{unbind} projected into the space of N_{PW1}, N_{PW4}, and N_{PW3}. The $q_{\text{unbind}} = 0.5$ transition state ensemble is highlighted by the black arcs. The large arc near $N_{\text{PW3}} \approx 80 - 100$ corresponds to the transition state along PW3. The two small arcs near $N_{\text{PW1}} \approx 60$ and $N_{\text{PW4}} \approx 60$ correspond to the transition states along PW1 and PW4, respectively. (B) The value of N_{PW3} projected into N_{PW1}, N_{PW4} ,and $q_{\text{unbinding}}$. (C) The value of RMSD_{P} projected into the same space as (B).

channel opening (along PW1a and PW4a) and the melting of the A-chain C-terminal α helix (along PW3) previously noted.

Choice of the dividing surface. For measuring the relative weights of different binding pathways, one needs to introduce a dividing surface between the bound and unbound states, and to partition that surface into patches corresponding to different pathways. The relative weight of each specific pathway is the sum of the current flowing normal through its patch on the surface, normalized by the total amount of current flowing normal to the full surface. The binding current, J_{bind}, was calculated with a lag time of 500 ps in the space of N_{PW1}, N_{PW4}, and q_{bind}. This was binned into a $50\times50\times50$ uniform grid, covering $0 \leq N_{\text{PW1}}, N_{\text{PW4}} \leq 100$, and $0 \leq q_{\text{bind}} \leq 1$. After binning in this CV space, the results were smoothed with a Gaussian filter, using a kernel with standard deviation of 4 bins.

We chose to use the binding statistics (J_{bind} and q_{bind}) to determine the relative weights of the six pathways, instead of using the unbinding statistics (J_{unbind} and q_{unbind}) as in the rest of our analysis. In the limit of infinite sampling, one would expect the binding and unbinding statistics to mirror each other exactly, as the dynamics are reversible. However,

we found that finite sampling led to small but noticeable differences in the reactive currents (J_{bind} and J_{unbind}) that made the determination of the six pathways from J_{unbind} more difficult (Figure 3.14). In particular, the area of high reactive current along the N_{PW4} axis is much more diffuse in Figure 3.14B than Figure 3.14A, as PW4 and PW4a blur together when using the unbinding statistics. While slightly less visible due to the color scale, a similar blurring occurs for PW1 and PW1a when using the unbinding statistics instead of the binding statistics.

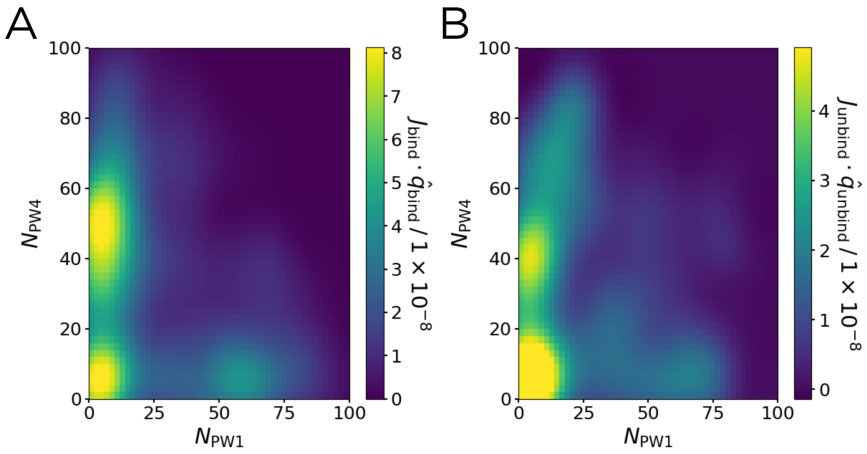


Figure 3.14: Comparison of reactive currents for the unbinding and binding directions. In each case, we show the dividing surface that best separates the six pathways. (A) The q_{bind} component of J_{bind} at $q_{\text{bind}} = 0.67$. (B) The q_{unbind} component of J_{unbind} at $q_{\text{unbind}} = 0.33$.

We use a plane of constant q_{bind} as the dividing surface. Based on visual inspection of Figure 3.15, we determined that $q_{\text{bind}} = 0.63$ provided the best separation of the pathways. We used it to define the patches for both WT and mutant insulins as follows:

- PW4: $N_{\text{PW4}} > 66$
- PW1: $N_{\text{PW1}} > 64$ and not the above
- PW4a: $N_{\text{PW4}} > 32$ and not any of the above
- PW1a: $N_{\text{PW1}} > 40$ and not any of the above
- PW3: $N_{\text{PW1}} < 20$ and $N_{\text{PW4}} < 16$

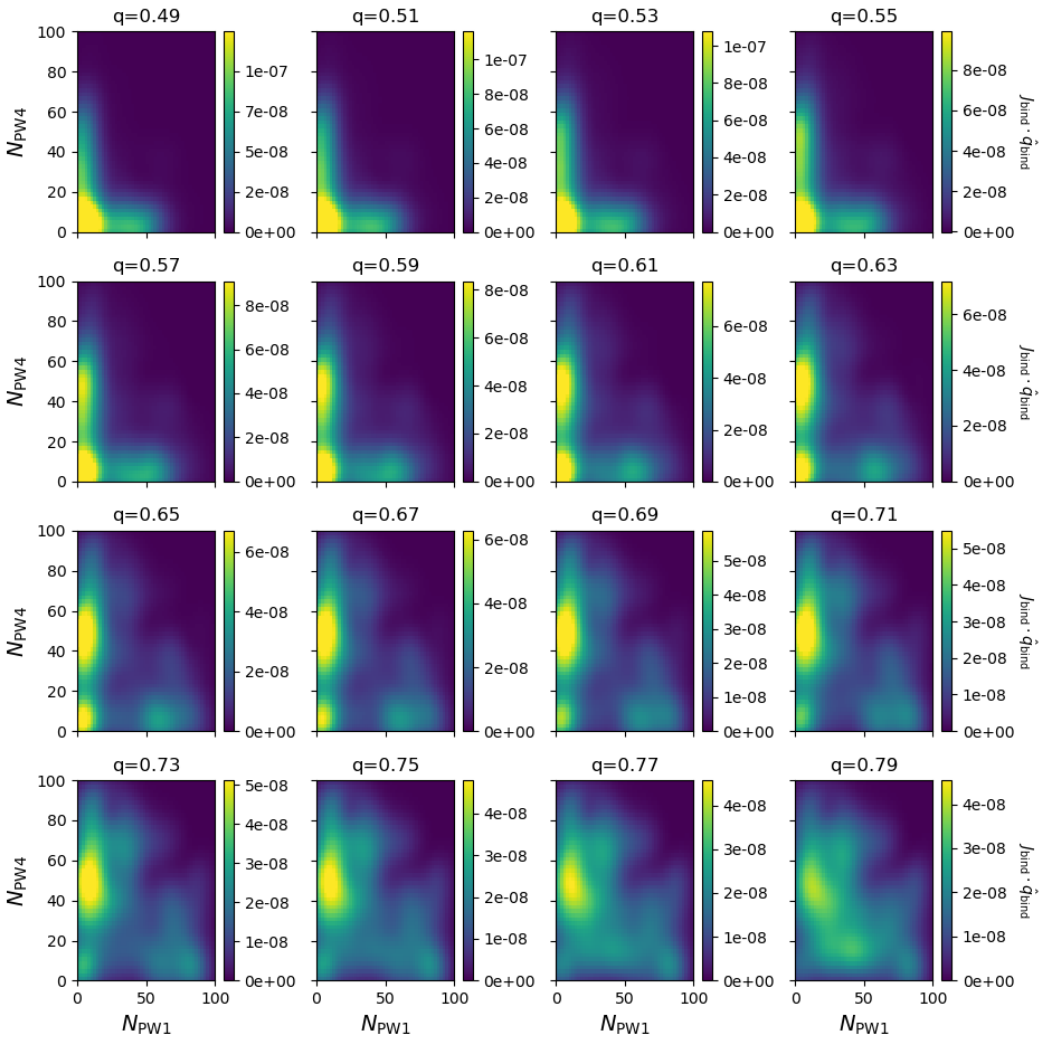


Figure 3.15: The q_{bind} component of J_{bind} at various different slices of q_{bind}.

- PW2: None of the above

These choices are consistent with our ABMD simulations as well. Using these patches, we calculated the relative weights of different pathways as a function of the value of q_{bind}, shown in Figure 3.16. Over the range of q_{bind} values shown, we find the results to be insensitive to the specific choice of dividing surface.

To probe the robustness of these relative weights, we measured them as we varied the sampling distribution. In particular, we identified 50 trajectories from our unbiased data set (692 total trajectories) which started along PW1/PW1a, and we identified a similar set of

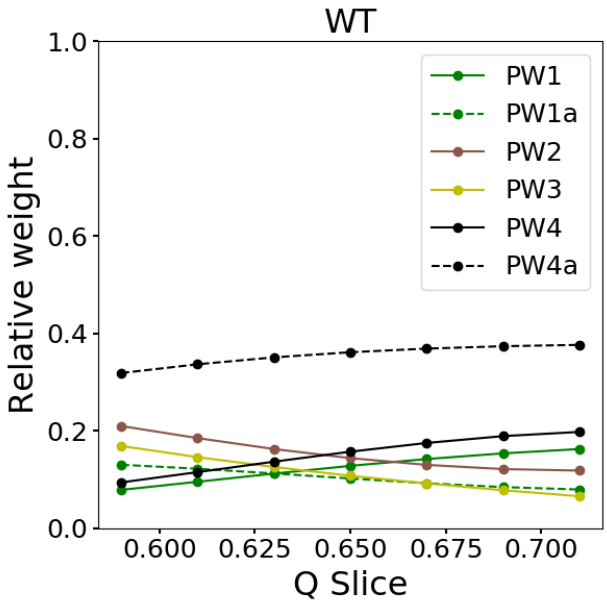


Figure 3.16: The relative weights for the six identified pathways as a function of the value of q_{bind} for the dividing surface.

Table 3.4: The relative weights for phenol unbinding along our six identified pathways, measured after removing trajectories along the described pathways.

Data set	PW1 (%)	PW1a (%)	PW2 (%)	PW3 (%)	PW4 (%)	PW4a (%)
Full 692	11.2	11.2	16.3	12.6	13.7	35.0
-25 PW1	11.3	9.7	16.6	13.2	13.9	35.2
-50 PW1	9.7	9.7	17.0	13.3	13.9	36.2
-25 PW4	10.7	11.6	15.9	13.1	14.1	34.5
-50 PW4	11.3	11.3	16.6	13.1	13.1	34.7

50 trajectories that started along PW4/PW4a. We removed some of these trajectories and re-measured the relative weights for phenol unbinding along each of our six pathways (Supplemental Table 3.4). As we remove trajectories, from either PW1/PW1a or PW4/PW4a, the relative weights of the six observed pathways change very little (with all changes being $\leq 1.5\%$). This provides evidence that our sampling is robust enough to converge the relative weight estimates from DGA.

Mutant simulation details. In addition to that for the wildtype protein, we constructed models for two hexamers with one point mutation each (A10 Ile $\rightarrow$ Val, B13 Glu $\rightarrow$ Gln). In

each case, we used CHARMM-GUI (version 3.2) to modify the 1ZNJ crystal structure at the six sites that differed in sequence and then solvated the system following a similar procedure to that described for WT. Below, we will comment on any differences between the workflow described in the main text and the workflow we followed for each of these mutants.

For the A10 Ile $\rightarrow$ Val mutant, 54 K^+ and 44 Cl^- additional ions were aded to achieve a neutral 150 mM KCl solution, for a total of 51,060 atoms. ABMD simulations were used only to generate initial structures for the unbiased simulations. 140 ABMD simulations of 5 ns each were originally run: 28 simulations for each of five force constants between $k = 6$ and $k = 14$ kJ / (mol nm). For two of these simulations, the binding pocket, which was open after we solvated the system, was observed to close. From each of these closed structures, 16 additional ABMD simulations were run at all five of the previously described force constants, plus $k = 16$ and $k = 18$ kJ / (mol/nm), as stronger force constants were needed to encourage dissociation once the channel had been closed. Finally, supplemental ABMD trajectories (28 at $k = 6$ and 28 at $k = 14$ kJ / (mol nm)) were run from one existing trajectory that sampled PW3 as identified by ref. 158. In total, this driven data set thus consisted of $28 \times 5 + 32 \times 7 + 28 \times 2 = 420$ trajectories, each of length 5 ns.

The grid in the cylindrical space of our CVs (again with $N_{\mathrm{A10}} = r\cos\theta$, $N_{\mathrm{A13}} = r\sin\theta$, and $\mathrm{RMSD_P}$) was changed to a $13\times13\times10$ grid in $(r, \theta, RMSD_{\mathrm{P}})$ space, in order to generate a similar number of unbiased simulations as we did previously; 356 unique structures were selected from the ABMD database. From each of these points, two 40 ns simulations were launched. To further sample PW3, as before, 20 structures from that pathway were identified from the ABMD database, and from each of those structures, two 40 ns simulations were launched. Thus, this unbiased data set consisted of 752 simulations, each of length 40 ns, for an aggregate simulation time of 30.08 μs.

For the B13 Glu $\rightarrow$ Gln mutant, 47 K^+ and 47 Cl^- additional ions were added to achieve a neutral 150 mM KCl solution, for a total of 51,196 atoms. Since the mutated glutamine

residues could form a series of hydrogen bonds with one another, PyMOL 2.3.0 [179] was used to individually rotate each side chain to flip the carbonyl and amine groups, creating $2^6 = 64$ different conformations. We used the steepest descent algorithm to minimize the energy of each of these conformations until the maximum force felt by the system was below 1000 kJ/mol nm. The lowest energy conformation was selected for all further simulations of this mutant.

To generate initial structures for the unbiased simulations, 140 ABMD simulations of 5 ns each were originally run: 28 simulations for each of five force constants between $k = 15$ and $k = 23$ kJ / (mol nm). From these simulations, four structures were chosen that exhibited channel opening. For each of these structures, 40 additional ABMD simulations were run at $k = 23$ kJ / (mol nm), for a total of $28 \times 5 + 40 \times 4 = 300$ trajectories, each of length 5 ns. We used a $15 \times 15 \times 10$ grid in (r, θ, z) space to select 331 unique starting structures. In addition, 24 structures were chosen from the ABMD database that exhibited phenol release along PW3. From each of these 355 structures, two unbiased simulations of length 40 ns were launched, leading to a database of 710 trajectories (aggregate length 28.4 μs).

Unimolecular and bimolecular rate constant estimates. The results for unimolecular rates/equilibrium constants are presented as a function of lag time in Figure 3.17. To calculate the bimolecular association rate, we must account for diffusion. To do this, we adapt the approach of McCammon and coworkers [180]. We define two distances b and c ($b < c$), each measured from the center of mass of the two central Zn^{2+} ions. The distance b should be large enough that interactions between the phenol and hexamer can be considered centrosymmetric; based on our definition of the unbound state in the main text ($N_{\mathrm{PW1}} < 2$, $N_{\mathrm{PW4}} < 2$, and $N_{\mathrm{Prot}} < 5$), we set b to be 3.3 nm. The distance c was set to 5.3 nm, similar to previous simulations [22]. For the statistics in this section, we redefine the unbound state to be the center of mass of the phenol at radii larger than c. The probability of diffusing from a sphere of radius c to a sphere of radius b is $\Omega = b/c$ [181]. Using this, the association

rate constant is $k'_{\text{bind}} = 4\pi Dbp$, where D is the diffusion constant and p is the probability of ultimately binding once the phenol first reaches a distance b. Adapting results from ref. 180, this quantity can be calculated from DGA using the relation:

$$p = \frac{q'_{\text{bind}}}{1 - \Omega(1 - q'_{\text{bind}})}, \tag{3.3}$$

where q'_{bind} is the binding committor with the bound state defined as in the main text and the unbound state defined as above. The factor p can also be used to correct the dissociation rate for re-binding: $k'_{\text{unbind}} = k'_{\text{DGA}}(1 - \Omega p)$, where k'_{DGA} is the unimolecular unbinding rate constant computed by DGA with the bound state as defined in the main text and the unbound state defined as above. Note that q'_{bind} and k'_{unbind} differ from q_{bind} and k_{unbind} in the main text owing to the redefinition of the unbound state.

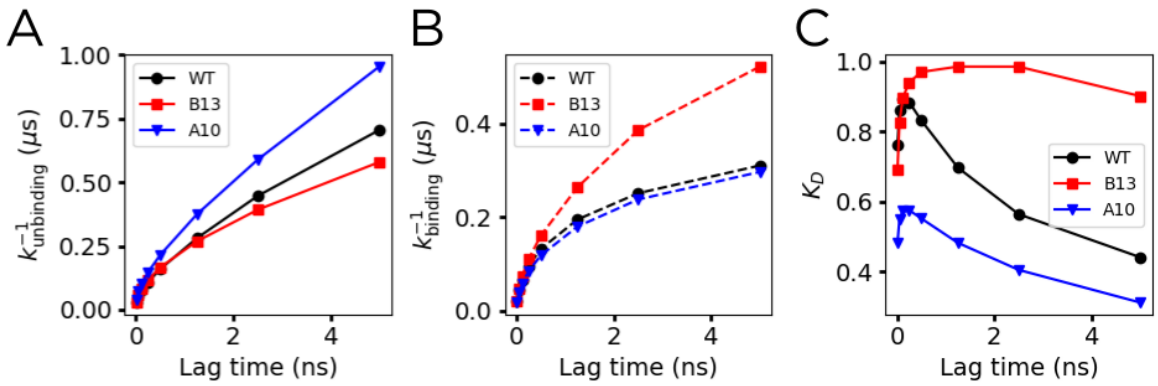


Figure 3.17: Unimolecular rate constants and their ratio at a range of DGA lag times for WT insulin and the two mutants, $\text{Ile}^{\text{A10}} \rightarrow \text{Val}^{\text{A10}}$ (A10 in the legend) and $\text{Glu}^{\text{B13}} \rightarrow \text{Gln}^{\text{B13}}$ (B13 in the legend). We show the inverse unbinding rate constant, $k^{-1}_{\text{unbinding}}$ (A), the inverse binding rate constant k^{-1}_{binding} (B), and their ratio $K = k_{\text{unbinding}}/k_{\text{binding}}$ (C).

The diffusion constant was calculated as a sum of the self-diffusion constants for the phenol and the hexamer determined separately, $D = D_{\text{phenol}} + D_{\text{hexamer}}$. The phenol was solvated and equilibrated using the same procedures as described for the full hexamer in the main text: the total number of atoms was 48,453, including 46 K^+ and 46 Cl^- ions. The box size was $(7.839\ \text{nm})^3$. Both diffusion constants were determined by measuring the slope

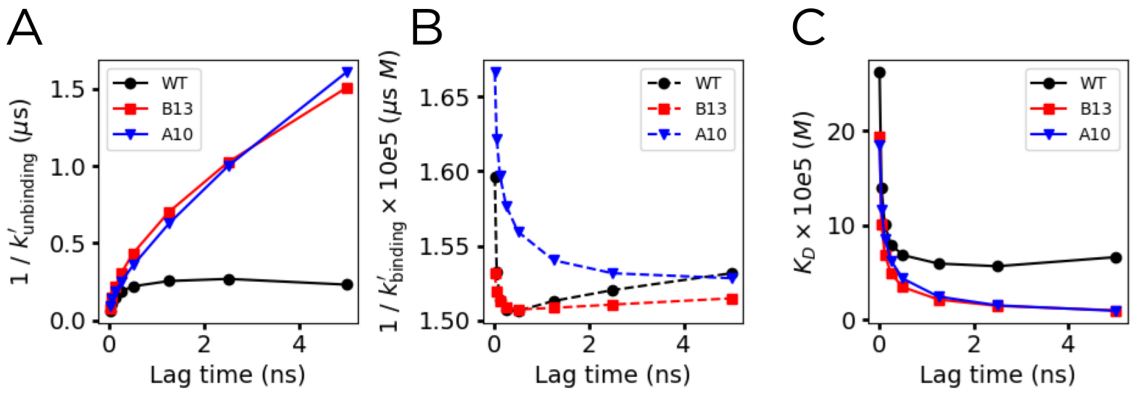


Figure 3.18: Bimolecular rate constant estimates as functions of DGA lag times for WT insulin and the two mutants, $\text{Ile}^{\text{A10}} \rightarrow \text{Val}^{\text{A10}}$ (A10 in the legend) and $\text{Glu}^{\text{B13}} \rightarrow \text{Gln}^{\text{B13}}$ (B13 in the legend). We show (A) the inverse unbinding rate constant, $1/k'_{\text{unbind}}$, (B) the inverse binding rate constant $1/k'_{\text{bind}}$, and (C) the dissociation constant, $K_D = k'_{\text{unbind}}/k'_{\text{bind}}$.

of the mean squared displacement of the center of mass of the relevant species, once it had achieved linearity. For this, the phenol was simulated for 20 ns at 303.15 K, and D_{phenol} was 2.7×10^{-5} cm^2/s. The diffusion constant D_{hexamer} was measured for each insulin species (WT, A10 Ile $\rightarrow$ Val, and B13 Glu $\rightarrow$ Gln) by analyzing a 40 ns trajectory with all six phenols bound, and D_{WT}, D_{A10}, and D_{B13} were 4×10^{-7}, 1×10^{-7}, and 2×10^{-7} cm^2/s, respectively.

The bimolecular results are presented as a function of lag time in Figure 3.18, and for two representative lag times (as in the main text) in Table 3.5. The results are in good agreement with measured values [172] given expected sources of error. First, the choice of b is likely too short for orientational effects to be negligible, which will tend to increase p. Second, diffusion in TIP3P water at 298 K is known to be a factor of 2.45 too high [182]. Third, both k'_{bind} and k'_{unbind} are products, compounding statistical uncertainties. Fourth, the experiments were performed with Co^{2+} and p-aminobenzoate as bound ligands, instead of Zn^{2+} and Cl^- as in these simulations.

Mutant DGA parameter choices. To the greatest extent possible, we used the same simulation parameters for DGA for WT and mutant insulins. Differences were as follows.

Table 3.5: Bimolecular rate constant estimates: The inverse unbinding rate constant, $1/k'_{\text{unbind}}$, the inverse binding rate constant $1/k'_{\text{bind}}$, and the dissociation constant, $K_D = k'_{\text{unbind}}/k'_{\text{bind}}$. Ranges derive from taking lag times between 500 ps and 1.25 ns.

Statistic	WT	A10	B13
$1/k'_{\text{unbinding}}$ (μs)	0.22 - 0.25	0.36 - 0.63	0.44 - 0.70
$1/k'_{\text{binding}}$ (μs M)	1.5 - 1.5×10^{-5}	1.5 - 1.6×10^{-5}	1.5 - 1.5×10^{-5}
K_D (M)	6.0 - 6.8×10^{-5}	1.5 - 2.6×10^{-5}	3.5 - 2.2×10^{-5}
K_D experiment (M) [172]	18×10^{-5}	N/A	25×10^{-5}

For the A10 Ile $\rightarrow$ Val mutant, the bound state was redefined to reflect the means and standard deviations of N_{PW1}, N_{PW4}, and RMSD_{P} in a 10 ns equilibrium simulation of the mutated structure with a closed channel. The means were 52.3, 29.8, and 0.076 Å, and the standard deviations were 5.18, 4.03, and 0.014 Å.

Comparing the PMFs for WT insulin and B13 Glu $\rightarrow$ Gln insulin, the most stable basin moves to where His^{F5} is flipped inward, facing the phenol (Figure 3.19). As a result, the bound state shifted; the corresponding CV means were 77.3, 26.6, and 0.096 Å, and the CV standard deviations were 5.92, 4.40, and 0.016 Å. For the B13 Glu $\rightarrow$ Gln mutant, we also used a longer lag time (1.25 ns compared with 500 ps for WT insulin) because the pathway weights converged more slowly (Figure 3.7). In turn, because the reactive current J_{bind} becomes noisier as the lag time increases, the standard deviation of the Gaussian filter for J_{bind} was increased from 4 to 5 bins.

Statistics for mutant insulins. Comparing the WT and A10 Ile $\rightarrow$ Val PMFs (Figure 3.19A), we see that this mutation stabilizes the basin corresponding to the partially open escape channel (see main text Figure 3 for the location of this region), making it lower in free energy than the normal bound state. At the same time, the mutation causes a shift in the binding reactive current from PW1, PW1a, and PW2 to PW3 and PW4 (Figure 3.19). The current flowing through PW4a generally shifts slightly upward in N_{PW4}. Overall, the

total amount of reactive current flowing through the dividing surface is very similar to that for WT insulin, and as result the rates are similar as well (Figure 3.17).

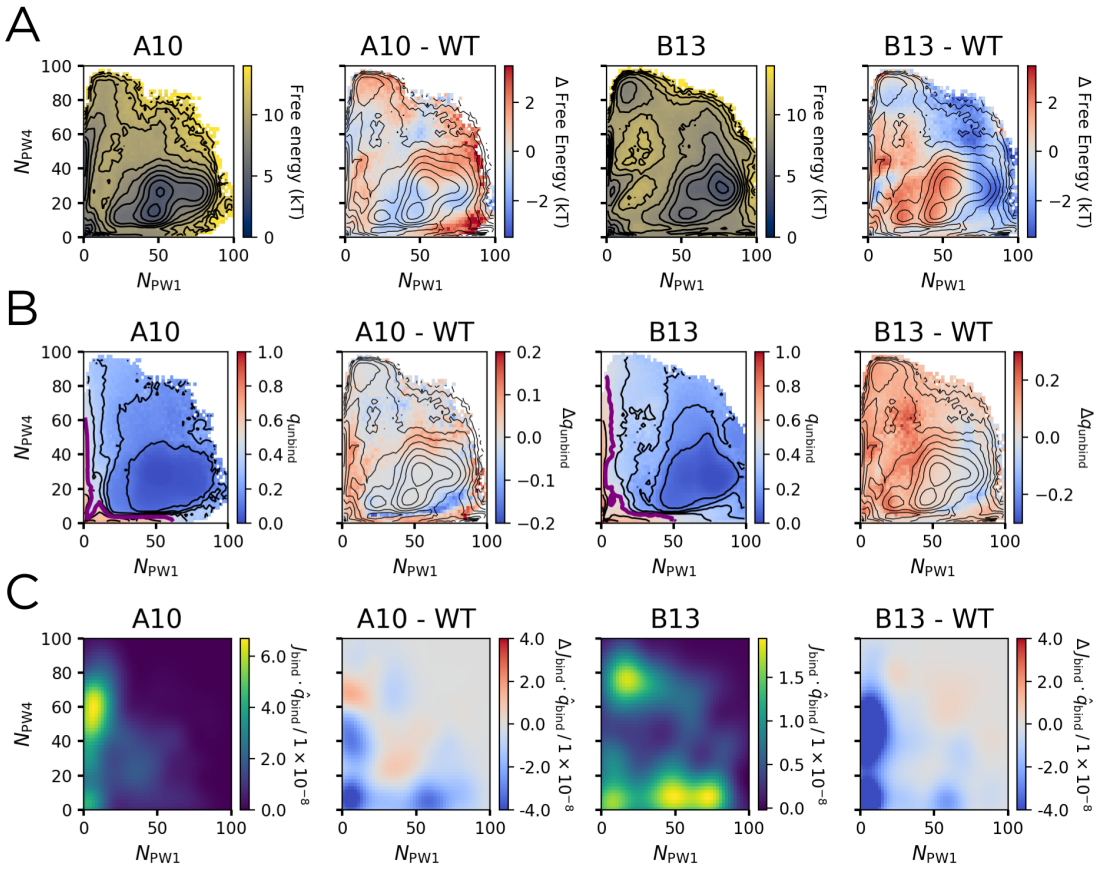


Figure 3.19: Comparing statistics for phenol escape between WT and mutant insulins. In each row, the first and third panels correspond to the $\text{Ile}^{\text{A10}} \rightarrow \text{Val}^{\text{A10}}$ mutant and $\text{Glu}^{\text{B13}} \rightarrow \text{Gln}^{\text{B13}}$ mutant, respectively. The second and fourth panels are the differences between the described mutant and WT insulin. (A) The potential of mean force, with contours shown every $k_B T$. For the differences, the contours from the WT insulin PMF are overlaid. (B) The average unbinding committor q_{unbind}, with contours shown every 0.1, and the $q_{\text{unbind}} = 0.5$ surface shown in purple. For the differences, the contours from the WT insulin PMF are overlaid. (C) The q_{bind} component of the binding reactive current J_{bind}, taken when $q_{\text{bind}} = 0.63$.

By contrast, B13 Glu $\rightarrow$ Gln insulin is dominated by a large-scale loss in J_{bind} through all six pathways. As a result, the binding rate is slower. Looking at the level of individual pathways, the largest area of reactive current loss corresponds to PW4a, and the areas corresponding to PW1 and PW4 lose comparatively little reactive current.

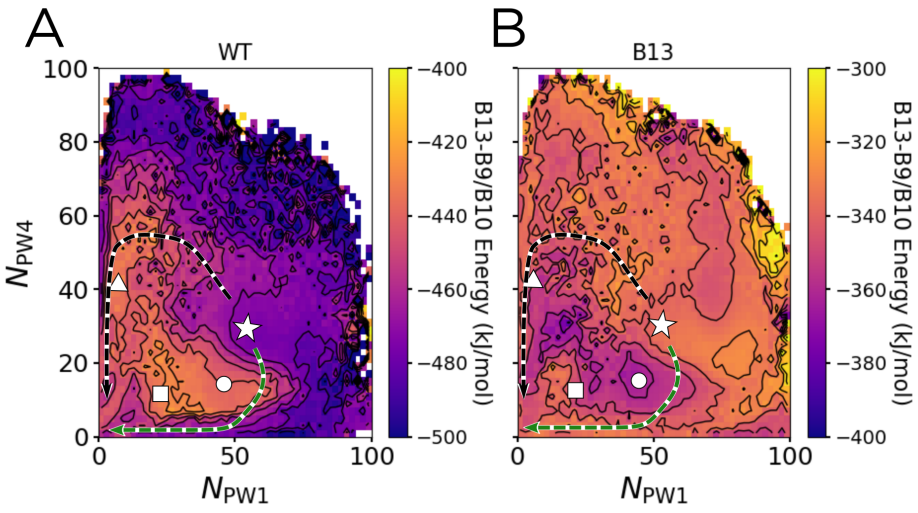


Figure 3.20: Interaction energies between B13 residues and the combination of Ser^{B9}/His^{B10} as a function of N_{PW1} and N_{PW4}, with contours shown every 10 kJ/mol. Arrows representing PW1a and PW4a are overlaid in green and black, respectively. The star, circle, square, and triangle mark the same landmarks as in Figures 2 and 3 in the main text. This is shown for both (A) WT insulin, and (B) the B13 Glu → Gln mutant.

Table 3.6: The change in interaction energies between the free and bound state for WT insulin and the B13 Glu → Gln mutant

Interaction	ΔE_{WT} (kJ/mol)	ΔE_{B13} (kJ/mol)	$\Delta\Delta E$ (kJ/mol)
B13 - Bound Zn^{2+}, Cl^-	–3.9	1.5	5.4
B13 - B13	3.9	5.1	1.3
B13 - B12	–0.9	0.4	1.3
B13 - B10	2.3	–3.5	–5.7
B13 - B9	14.9	–1.0	–15.9

To understand the changes in the unimolecular ratio $K = k_{unbinding}/k_{binding}$ upon mutation, we calculated the interaction energies between the mutated residues and all 50 other protein residues, as well as the bound Zn^{2+}/Cl^- ions, and the solvent (including the phenols). We then averaged these quantities across the free and bound state for both WT insulin and the relevant mutant, and measured $\Delta\Delta E = \Delta E_{mutant} - \Delta E_{WT}$, where $\Delta E = E_{free} - E_{bound}$. For the B13 Glu → Gln mutation, all such interactions where $|\Delta\Delta E| > 1$ kJ/mol are shown in Table 3.6.

The interactions with the largest magnitude of $\Delta\Delta E$ upon the B13 Glu → Gln mutation

Table 3.7: The change in interaction energies between the free and bound state for WT insulin and the A10 Ile → Val mutant

Interaction	ΔE_{WT} (kJ/mol)	ΔE_{A10} (kJ/mol)	$\Delta\Delta E$ (kJ/mol)
A10-B1	–1.6	–0.4	1.2
A10-B5	0.8	1.9	1.1
A10-B2	2.9	1.6	–1.3
A10-Solvent	–0.5	–3.7	–3.2

are B13 with $\mathrm{Ser}^{\mathrm{B9}}$ and $\mathrm{His}^{\mathrm{B10}}$, as discussed in the main text. Phenol unbinding and channel opening are correlated, and in the unbound/open state, the B13 side chain is able to rotate to interact with the backbone residues of $\mathrm{Ser}^{\mathrm{B9}}$ and $\mathrm{His}^{\mathrm{B10}}$. By making the B13 Glu → Gln mutation, we replace a repulsive interaction between the carboxylate side chain and the backbone carbonyl with a energetically-favorable hydrogen bond between the amide side chain and the same backbone carbonyl. These interactions, which stabilize the unbound state upon mutation, partially explain why K increases upon mutation. We calculated a similar set of interactions for the A10 Ile → Val mutant, shown in Table 3.7. For this mutant, no single interaction is comparably dominant.

To understand the effect of the B13 Glu → Gln mutation on the relative weights for the six pathways (and for PW1a and PW4a in particular), we projected the B13 electrostatic and Van der Waals interaction energies with $\mathrm{Ser}^{\mathrm{B9}}$ and $\mathrm{His}^{\mathrm{B10}}$ as functions of N_{PW1} and N_{PW4} for both WT insulin (Figure 3.20A) and the B13 Glu → Gln mutant (Figure 3.20B). The interaction energy decrease along PW1a (green) is approximately 20 kJ/mol, while the interaction energy decrease along PW4a (black) is 10 kJ/mol.

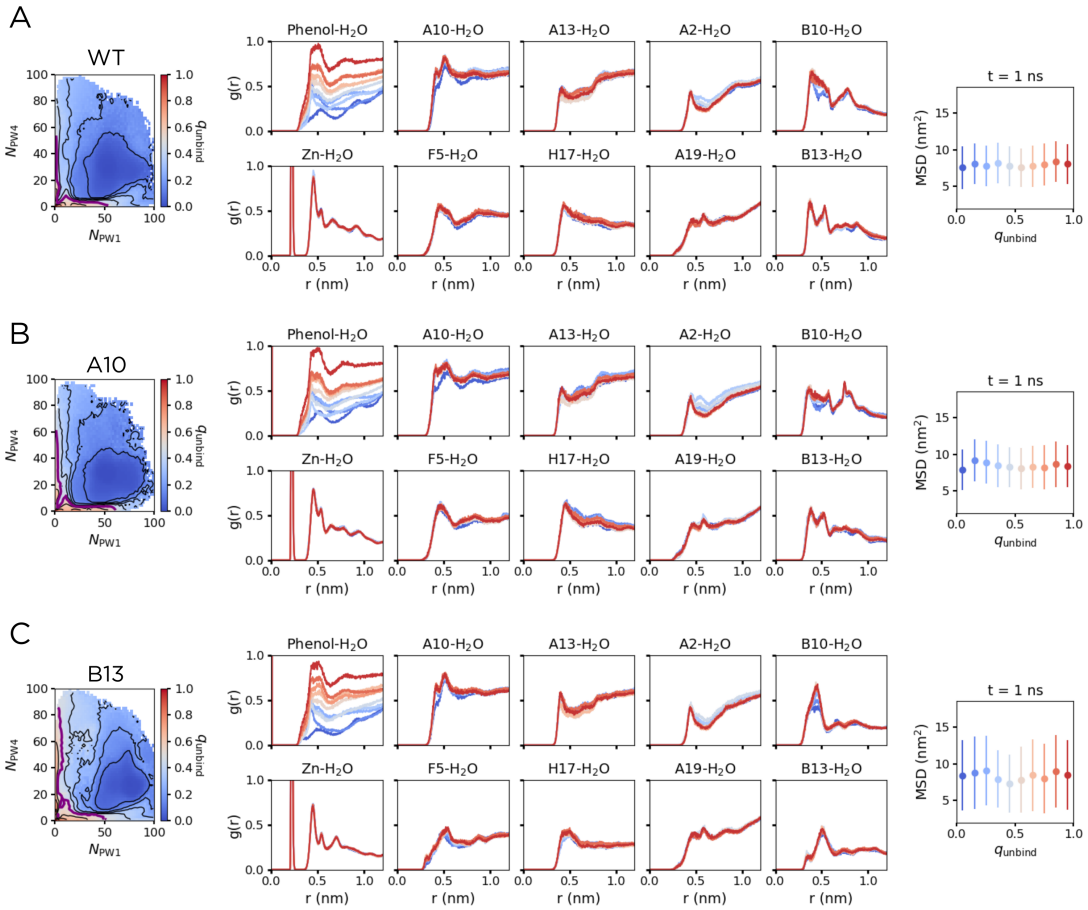


Figure 3.21: Solvation dependence on reaction progress. The (left) average committor, (middle) radial distribution functions for water around the specified species, and (right) mean square displacement (MSD) over 1 ns of waters in the central cavity of the hexamer for (A) WT insulin, (B) A10 Ile→Val insulin, and (C) B13 Glu→Gln insulin. Results are shown for 10 evenly sized bins for committor values between 0 and 1, with color given by the scale in the leftmost panel. We compute the radial distribution function, $g(r)$, from 15,000 structures in each committor value bin; we define r as the distance between the center of mass of the specified species/residue (including main chain atoms) and the center of mass of each water molecule. MSD values are for displacements over 1 ns from 5000 starting structures for each committor value bin.

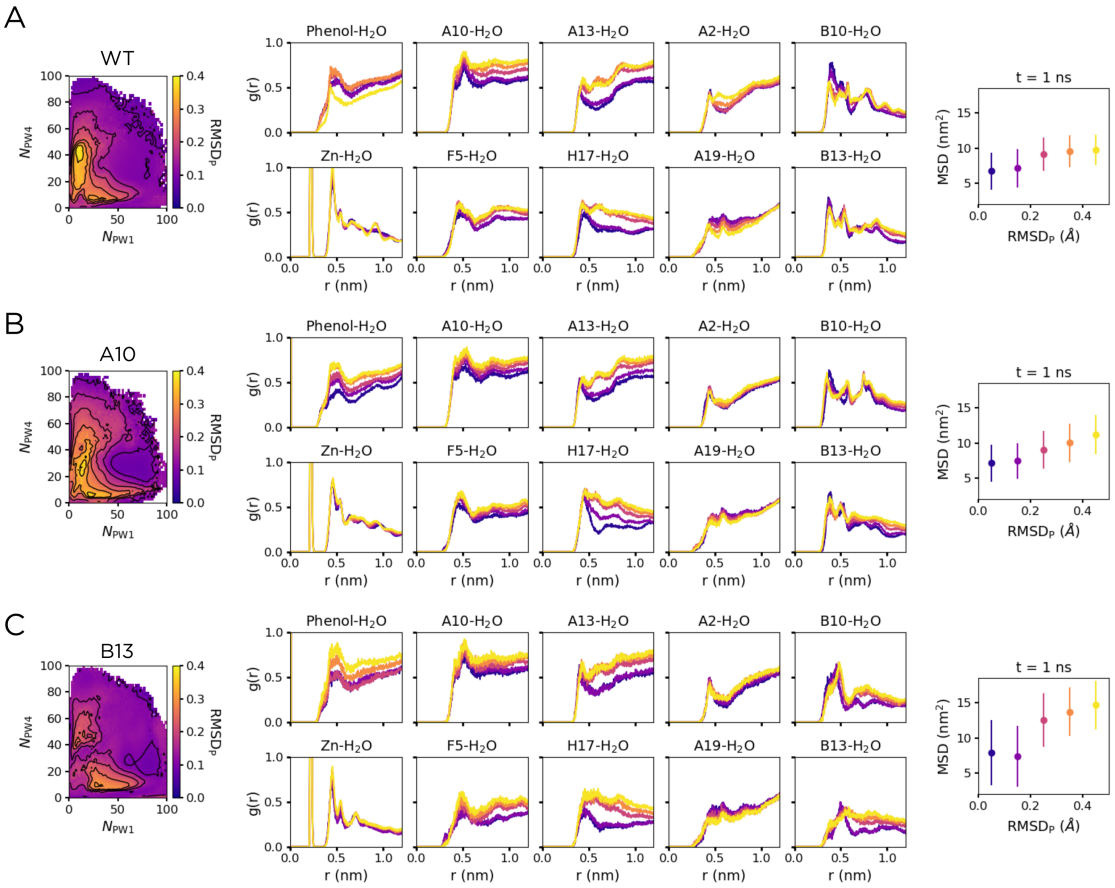


Figure 3.22: Solvation dependence on channel opening. The (left) average $\mathrm{RMSD_P}$, (middle) radial distribution functions for water around the specified species, and (right) mean square displacement (MSD) over 1 ns of waters in the central cavity of the hexamer for (A) WT insulin, (B) A10 Ile→Val insulin, and (C) B13 Glu→Gln insulin. Results are shown for 5 evenly sized bins for $\mathrm{RMSD_P}$ values between 0 and 0.5 Å, with color given by the scale in the leftmost panel. We compute the radial distribution function, $g(r)$, from 5,000 structures in each $\mathrm{RMSD_P}$ value bin; we define r as the distance between the center of mass of the specified species/residue (including main chain atoms) and the center of mass of each water molecule. MSD values are for displacements over 1 ns from 3000 starting structures for each $\mathrm{RMSD_P}$ value bin.

CHAPTER 4

NEAR-ATOMIC MOLECULAR DYNAMICS OF A METAMORPHIC PROTEIN: DYNAMICS OF THE KAIB FOLD SWITCH

Abstract

The cyanobacterial circadian clock protein KaiB is one of the most well-characterized metamorphic proteins. The interconversion time between its two different folds (the ground state and fold switched state) has been suggested to contribute to the period of the associated oscillator. We use unbiased near-atomic level Upside simulations, which employ an energy function obtained from machine learning, to estimate hydrogen-deuterium exchange protection factors. Comparing directly with experimental measurements, we identify the elementary units of fold switching. Using dynamical analysis to extract long-time statistics, we characterize the ensemble of fold switching mechanisms. We find that while the isomerization states of three prolines (P63, P70, and P71) tend to change during fold switching, the primary free energy barrier corresponds to the loss of secondary structure in the fold switched state. The fold switching most often proceeds via subglobal unfolding and refolding events in the C-terminal half of the protein, which can occur in any order. Overall, the secondary structure elements in the C-terminal fold switching domain act as foldons, tending to fold/unfold as units independent of one another. Our work provides the first statistical view of fold switching of a metamorphic protein.

4.1 Introduction

For decades, scientists ranging from structural biologists to computational chemists have used tools like X-ray crystallography, NMR, and Cryo-EM to solve for the structure of proteins.

Although these stable structures can be purified and isolated, it is well-known that both *in vivo* and *in vitro* proteins can exhibit a high degree of flexibility, which is often essential for biological function [29, 77–79]. Furthermore, research into both intrinsically disordered and prion proteins has underscored the important (and sometimes neurologically detrimental) effects of deviations from experimentally solved protein structures [63, 183, 184].

Recent work into a class of proteins deemed "metamorphic proteins" has further expanded the already-dynamic understanding of protein structure and how different structures might interconvert. Instead of irreversibly aggregating like prion proteins, metamorphic proteins can reversibly switch between two or more stable folds, often times accompanied by a change in biological function [31, 32]. For example, the transcriptional antiterminator protein RfaH reversibly switches between an α helix-rich fold and a β sheet-rich fold, which is essential for RfaH to bind to RNA polymerase [185, 186]. For now, this class of proteins remains relatively small, as many fold switching events are dependent on an unknown environmental trigger, like a change in temperature or pH. Because of this, it is possible that a significant fraction of characterized proteins, implicitly assumed to be monomorphic, are instead metamorphic [187]. Beyond their physiological relevance, metamorphic proteins are also promising engineering templates, as one amino acid sequence can encode a switch between two functionally distinct entities, and this switch can be reversibly "flipped" to repeatedly convert between those two entities [32].

A prototypical example of a biologically-relevant metamorphic protein is KaiB. Its reversible interconversion between two stable folds is an essential part of the cyanobacterial circadian clock [75]. Specifically, KaiB is one of three proteins (KaiA, KaiB, and KaiC) that form the core oscillator responsible for circadian rhythms in the cyanobacterium *Synechococcus elongatus*. Much is currently known about the structural biology of the three proteins in this oscillator, and how interactions between them and ATP lead to the observed KaiC phosphorylation cycle. In brief, the binding of KaiA to KaiC promotes the phosphorylation

of KaiC during the daylight hours, while the binding of KaiB leads to the sequestration of KaiA, causing the dephosphorylation of KaiC at night. When bound to KaiC, KaiB adopts a thioredoxin-like fold and exists as a monomer [75]. However, unbound KaiB tends to populate a different fold, one that is biologically inactive and exists largely as a tetramer [74]. In this chapter, we refer to KaiB's biologically inactive fold as the "ground state" (gs), and the biologically-active, thioredoxin-like fold as the "fold switched state" (fs). This work aimed to discover the mechanisms of conversion between gs and fs KaiB, which is an example of how interconversion between stable folds in metamorphic proteins can play a key biological function.

KaiB is a 107-amino acid protein that can be divided into 2 broad domains: the N-terminal domain (amino acids 1-50), which does not change secondary structure upon fold switching, and the C-terminal domain (amino acids 51-107), which is the fold switching domain. In the fold switched state, KaiB adopts the thioredoxin-like secondary structures of $\beta\alpha\beta\alpha\beta\beta\alpha$. In contrast, the last four secondary structure elements of the ground state (all of which are in the C-terminal domain) are inverted, leading to secondary structures of $\beta\alpha\beta\beta\alpha\alpha\beta$ [75]. These two folds are shown in Figure 4.1. KaiB also contains seven prolines. The isomerization of prolines is known to be a slow, often rate-determining factor in protein folding [5, 183, 188]. All seven prolines (P3, P19, P51, P63, P70, P71, and P72) in the WT gs KaiB sequence are in the *trans* state [74]. In contrast, three of these prolines (P63, P70, and P72) are in the *cis* state for a fs-stabilized KaiB mutant [75]. In Figure 4.1, we also show the location of the four prolines in the C-terminal fold switching domain, colored by their isomerization state.

While these structures have been resolved by both NMR and X-ray crystallography, many questions remain as to the mechanism of conversion between the ground state and fold switched state. How does proline isomerization relate to the fold switch? How much unfolding is present during the fold switch? If present, is the unfolding restricted to the fold

switching C-terminal domain, or does it extend into the N-terminal domain? In what order do the secondary structure elements in the fold switching region melt and reform? Is there one dominant mechanism, or an ensemble of available pathways?

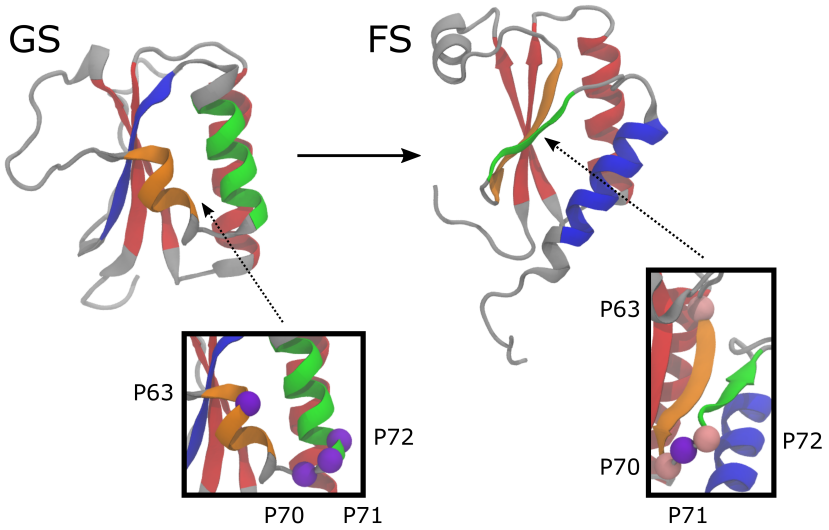


Figure 4.1: The crystal structures for gs KaiB (left, PDB ID 2QKE) and fs KaiB (right, PDB ID 2JYT). The red secondary structures are in the N-terminal domain that does not undergo fold switching. The orange, green, and blue secondary structures are in the fold switching C-terminal domain. The insets show a view of the proline-rich area of the C-terminal domain. C_αs corresponding to *trans* and *cis* prolines are shown as purple and pink spheres, respectively.

Many of these questions have remained heretofore unanswered because, experimentally and computationally, it is difficult to both discover and study fold switching in metamorphic proteins. Computationally, rare events are difficult to model, as the timescale of all-atom molecular dynamics simulations only extends to the order of milliseconds. Recently, we showed how long-time kinetic statistics can be computed from an ensemble of short simulations [25, 26] by Dynamical Galerkin Approximation (DGA). We have also applied this method to study the the dynamics of protein-ligand interactions [147]. In this work, we combine this method with the rapid near-atomic level molecular dynamics package Upside [189, 190] that was recently shown to accurately replicate the PMF and denaturant dependence of small proteins, as experimentally validated by hydrogen-deuterium exchange

experiments [191]. Using these two methods, we can bridge the separation of timescales between theory and experiment, modeling the dynamics of the KaiB fold switch and comparing directly to hydrogen-deuterium exchange NMR experiments.

In this work, we parameterized proline isomerization into Upside and identified three prolines (P63, P70, and P71) that preferentially occupy the *cis* state in fs KaiB and the *trans* state in gs KaiB. We also compared model results to experimental hydrogen-deuterium exchange measurements on a KaiB mutant that favors the monomeric fs KaiB, finding that most residues become solvent-exposed due to local or subglobal unfolding events. Notably, we find almost no residues that exchange via global unfolding events, evidence against fold switching mechanisms that involve a globally melted intermediate. We also found a negligible free energy difference between fs and gs KaiB, in agreement with previous computational predictions [192]. Importantly, we use equilibrium and dynamical statistics to characterize the transition pathways between fs and gs KaiB, finding an ensemble of mechanisms that involve the melting/refolding of C-terminal domain secondary structure elements in various orders. The primary free energy barrier is associated with the breaking of β sheets in the C-terminal domain. Both the isomerization of P63, P70, and P71 and the melting of the C-terminal α helix often precede the β sheet breaking. Furthermore, we observe only small-scale, subglobal unfolding of the N-terminal domain along the pathways of the fold switch. This work clarifies the mechanism of KaiB fold switching, and provides the first statistical view of the fold switching of a metamorphic protein. The combination of near-atomic simulation and dynamical analysis is a powerful tool with great promise to shed light on metamorphic proteins, intrinsically disordered proteins, prion proteins, and beyond.

4.2 Methods

Computational and experimental systems. All computational systems were prepared using CHARMM-GUI version 3.7 [99, 193]. The X-ray crystal structure for tetrameric

wildtype (WT) ground state KaiB was retrieved from the Protein Data Bank (PDB ID 2QKE)[74]. Monomeric gs KaiB was isolated by modeling only chain B, since all residues were resolved. The NMR structure of a fold switched KaiB mutant was also retrieved from the Protein Data Bank (PDB ID 5JYT) [75], and CHARMM-GUI was used to mutate the sequence back to the WT sequence. The same mutant NMR structure (truncated to just amino acids 1-99) was also used to model the mutant used for comparison to the hydrogen-deuterium exchange data. For all small peptide simulations (used for parameterization of the proline double well potential), a 5 amino acid segment was taken from PDB 5JYT, and mutated/capped using CHARMM-GUI to the following structure: Ac-Ala-Xaa-Pro-Ala-Lys-NH2. Here, "Xaa" refers to the identity of the particular amino acid being tested. All simulations were performed and analyzed with a combination of Upside [189, 190] and MDTraj 1.9.6 [194].

Hydrogen-deuterium exchange NMR experiments were performed by Drs. Supratim Dey and Andy LiWang at the University of California, Merced. We used the Y8A-Y94A-G89A-D91R mutant of *Thermosynechococcus elongatus* KaiB, truncated to amino acids 1-99. This mutant stabilizes the monomeric fs state [75]. Experimental ΔG_{HX} values were determined at 12°C and both pH 4.5 and pH 6.5, with standard errors measured across three independent experiments for each pH. m-values were calculated by fitting the slope of ΔG_{HX} (measured at pH 4.5 and 15°C) against urea concentrations of 0.0, 0.5, 1.0, 1.5, and 2.0 M.

Description of Upside. While a full description of Upside can be found elsewhere [189–191], we briefly describe the model and how it treats proline isomerization here. Upside initially treats each amino acid as three beads, representing the N, C, and C_α atoms. The carbonyl oxygen and amide proton are then deterministically placed based on these three beads. One additional oriented bead, which represents the amino acid side chain, is given an initial probability distribution for six pre-set rotamers based on PDB data. The solvent is treated in an implicit fashion. The pairwise probability distributions for all side chains

are optimized to minimize an effective free energy, which combines a machine-learned energy term and an entropic term. The forces are calculated and then applied to the backbone atoms, which then undergo Langevin dynamics for one timestep. The temperature and timescales are both intentionally left as arbitrary. We used Upside force field 1.5 (FF1.5) for the initial simulations investigating proline isomerization. For all later simulations, we used the newly-introduced Upside force field 2.1 (FF2.1). The more modern FF2.1 more accurately captures both the disordered and native state ensembles [191].

DGA Parameters. Dynamical Galerkin Approximation (DGA) has been previously described, in both methods and application [25, 26, 76]. Briefly, dynamical quantities such as the committor and reactive current are cast as solutions to operator equations, and these are then transformed to matrix equations through the introduction of a basis. The matrix elements are estimated from unbiased sampling. In this work, we used a set of indicator basis functions, analogous to the states associated with Markov State Models (MSMs) [20, 22, 23, 166, 167]. Indeed, the DGA can be thought of as a generalization of MSMs that directly yields statistics for a specified reaction.

To generate this basis of indicator functions, we first used PyEMMA 2.5.9 [195] to featurize our existing unbiased data into a set of pairwise distances. We chose the set of every other α carbon from all of the residues defined in Table 4.1 and calculated all of the pairwise distances between them, leading to a 406-dimensional set of input features for each frame in our unbiased data set. We then applied k-means clustering to this featurized data set, processing the data into 150, 200, 250, and 300 clusters. We explicitly set to zero all of the data points that fall into either fs or gs KaiB (defined in *Generation of the unbiased dataset with proline isomerization allowed.*), so as to obey the homogeneous boundary conditions of the DGA operator equations. We also used an indicator for gs KaiB as the DGA guess function, ψ. We then used these basis functions as DGA inputs, generating a set of dynamical quantities for lag times between 1 and 500 time steps for each basis set. These results were

not sensitive to the size of the basis set; in the rest of this work, we present results from the basis set of 300 indicator functions. Estimates of the committor and reactive current also showed little dependence on the lag time, other than slight numerical instability at high lag times. For the rest of this work, we chose a lag time of 2500 Upside time steps, which was the largest value tested with no observed numerical instabilities. To help control numerical errors, averages of the committor were generated after first propagating the committor value forward in time either by the lag time or by the stopping time, whichever one is shorter. The 2D estimates of the reactive current were binned into a 21×21 grid in CV space, and these binned values were smoothed with a Gaussian filter, using a kernel with standard deviation of 2 bins. For the 3D estimates of the reactive current, the binning was $50 \times 50 \times 50$ in CV space, and the kernel's standard deviation was 4 bins.

4.3 Results and Discussion

Through this work, we aimed to explore the mechanism of the KaiB fold switch. To do this, we employed a pipeline of experimental and computational techniques, which we divide into two sections. First, in *Upside Parameterization and Comparison to Hydrogen-Deuterium Exchange Data*, we sought to calibrate Upside and our sampling techniques to ensure the model was experimentally reasonable and produced results that were physically interpretable. Second, in *Elucidation of KaiB Fold Switch Mechanism By Application of the Dynamical Galerkin Approximation*, we applied dynamical analysis to structurally and quantitatively describe the mechanisms of the KaiB fold switch.

4.3.1 Upside Parameterization and Comparison to Hydrogen-Deuterium Exchange Data

Before investigating the mechanism of the KaiB fold switch, we first explored the KaiB system itself and validated Upside as a model capable of simulating the fold switch. First, we used Upside to generate a suite of unbiased simulations that locked prolines in either the *cis* or *trans* isomer in order to generate initial sampling and discover physically-interpretable collective variables (CVs). Second, we implemented a double well potential to allow prolines to isomerize during the course of a simulation. Third, we simulated the free energies of hydrogen-deuterium exchange (HX) for an experimentally-available KaiB mutant and compared them to experimental values.

Isomerizing prolines generated robust KaiB sampling. To discover the collective variables that best capture the KaiB fold switch, we first sought to control the system and generate sampling that spanned state space between gs and fs KaiB. Hypothesizing that proline isomerization might determine the kinetics of the fold switch, we manipulated P63, P70, and P72 to try to generate sampling of the fold switch. We chose these three prolines because they switch from *trans* in gs KaiB to *cis* in fs KaiB (Figure 4.1). Early versions of Upside used force field 1.5 (FF1.5) with a single-well potential for backbone isomerization, meaning that over the course of the simulation, prolines were effectively locked into the isomerization state in which they were initialized. We thus used this package to isomerize and then simulate KaiB, hypothesizing that isomerizing fs KaiB prolines from *cis* to *trans* might be sufficient to drive the simulations from fs to gs KaiB. This would give us the sampling we needed to determine appropriate collective variables.

First, we needed to set a simulation temperature, as the Upside temperature scale is arbitrary. To find a suitable simulation temperature that encouraged fold switching, we ran a set of 28 temperature replica-exchange (T-REMD) simulations of fs KaiB with P63, P70,

and P72 locked in the *cis* conformation. Each simulation was 1.5 million Upside time units long, saving every 100 time units, at temperatures ranging between $T = 0.8$ and $T = 1.0$ (in arbitrary units), attempting exchanges every 10 time units evaluated by the Metropolis criterion. After reweighting with MBAR [196], the melting temperature of fs KaiB was determined to be $T = 0.94$, the point where the number of KaiB hydrogen bonds as a function of simulation temperature dropped dramatically. This temperature was chosen for our preliminary simulations to increase the likelihood of seeing fold switching events.

765 unbiased simulations were then launched at $T = 0.94$, each one starting from fs KaiB, and running for 1.5 million Upside time units. This time, P63, P70, and P72 were locked into *trans* to encourage transition from fs to gs. From this, we observed 55 fold switching transitions, where the ground state structure was successfully recovered. We also launched 242 simulations investigating the reverse transition, starting from gs KaiB but locking P63, P70, and P72 in the *cis* isomer. Here, too, we observed multiple (four) fold switching events. This is in contrast to all of our T-REMD simulations of KaiB, where even at high temperature, we observed no fold switching events as long as P63, P70, and P72 were maintained in their native isomerization state. In these simulations, KaiB remained near its native structure, before it eventually unfolded completely at higher temperatures. We thus concluded that switching the isomerization state of these three prolines (P63, P70, and P72) was sufficient to drive fold switching in either direction.

This procedure, using T-REMD followed by unbiased sampling based on isomerized starting structures, was repeated when Upside force field 2.1 (FF2.1) was released. We aimed to ensure that we achieved the same control of the fold switch with this new force field. Following the T-REMD, we determined the melting temperature to be $T = 0.915$. This time, we performed the unbiased sampling at three temperatures: $T = 0.85$, $T = 0.89$, and $T = 0.91$. Again, for each of these trajectories, we started in the fs structure and isomerized P63, P70, and P72 from *cis* to *trans*. We observed no fold switching events for $T = 0.85$. For $T = 0.89$

and $T = 0.91$, however, we observed eight and sixteen fold switching events, respectively. For $T = 0.89$, only one of these eight events passed through a globally unfolded intermediate (defined as when the RMSD of the regions that make up the conserved N-terminal secondary structure elements exceeded 2 nm). For $T = 0.91$, 8/16 fold switching events passed through a globally unfolded intermediate. This sampling, although providing multiple fold switching events, starts from a perturbed state of the isomerized fs KaiB structure. This likely perturbs the dynamics of the fold switch. Additionally, by locking the prolines as either *cis* or *trans* for the duration of the simulation, we lose the ability to discern the role of proline isomerization during the actual fold switch. We discuss the development of a potential that allows for proline isomerization during a simulation in a later section.

Three secondary structure variables and three RMSDs describe the fold switch. By analyzing the unbiased trajectories generated in the previous section, we developed a set of collective variables (CVs) that best capture the change from fs to gs KaiB while remaining physically interpretable. Three of these variables, the Blue Transition (BT), the Green Transition (GT), and the Orange Transition (OT), measure the change in secondary structure in the fold switching C-terminal region from fs to gs KaiB. The colors in these variables refer to the colors chosen for secondary structures in Figure 4.1 and throughout this work. We defined these residues by inspecting the crystal structures and preliminary equilibrium simulations of the fs KaiB and gs KaiB ensembles; the identities of the amino acids (AAs) involved in colored secondary structure elements are shown in Table 4.1. We also defined a set of three RMSDs: RMSD_{fs}, RMSD_{gs}, and RMSD_{red}. RMSD_{fs} measures the RMSD relative to the fs KaiB structure for all the backbone heavy atoms (C, N, O, CA) in the fs row of Table 4.1. RMSD_{gs} is a similar variable for gs KaiB. RMSD_{red} measures the heavy atom RMSD relative to gs KaiB of the secondary structures in the N-terminal region, the red residues in Table 4.1. This is a proxy for global unfolding, since the N-terminal region likely remains well ordered unless the protein is globally melted. Since this region is

Table 4.1: Residue definitions, in terms of amino acid (AA) numbers, for the colored secondary structure elements seen in Figure 4.1 and referenced throughout this work.

KaiB Fold	Red AAs	Orange AAs	Green AAs	Blue AAs
Gs	9-14, 21-35, 41-46	62-67	72-81	87-93
Fs	9-14, 21-35, 41-46	64-69	73-77	84-96

conserved between fs and gs KaiB, the choice of reference crystal structure did not noticeably change $\mathrm{RMSD}_{\mathrm{red}}$.

The mathematical form for the three secondary structure CVs (with $x = B$, G, and O) is as follows:

$$xT = Q^{\gamma}_{x,\mathrm{gs}} - Q^{\gamma}_{x,\mathrm{fs}} \tag{4.1}$$

This equation represents a difference between two measures of contacts, Q^{γ}, for fold switching secondary structures of type $\gamma = \alpha$ or $\gamma = \beta$. Each term in the above equation is based on a different set of contacts, but both are normalized to vary between 0 and approximately 1. For α helices, $Q^{\gamma} = Q^{\alpha}$ is simply the fraction of heavy atom native contacts that the particular helix makes with all other heavy atoms in KaiB, as defined by Best *et al.* [197]. The native structures used to compute $Q^{\alpha}_{x,\mathrm{gs}}$ and $Q^{\alpha}_{x,\mathrm{fs}}$ were the crystal structure for gs KaiB (PDB ID 2QKE) and fs KaiB (PDB ID 5JYT), respectively. We set the distance cutoff to determine initial native contacts as 0.45 nm, and the β and λ parameters to be 50 nm^{-1} and 1.5, respectively. β determines the "softness" of the switching function, determining how quickly the contact function decays from 1 to 0 after the reference distance is passed. λ is the tolerance for the reference distance, providing slack to allow atoms at near-native distance to still be considered contacts. Heavy atoms in close proximity were only considered native contacts if they were more than 3 residues apart. Q^{α} thus ranges from 0 to 1, with $Q^{\alpha} = 1$ indicating a fully-folded α helix in the correct location, and $Q^{\alpha} = 0$ indicating a fully melted α helix making no native contacts.

While this variable accurately captured the melting/repositioning of KaiB α helices, we found this definition too restrictive when measuring β sheets. In particular, this definition was too sensitive to β sheet register shifts, measuring a 1- or 2-residue register shift in KaiB's antiparallel β sheets (often seen in our preliminary simulations) as a near-complete melting transition. Thus, for β sheets, $Q^\gamma = Q^\beta$ is a modified version of what is defined in ref. 197, and instead measures the number of total contacts made between nearby β sheets, normalized to the total number of contacts between those β sheets in the crystal structures. This definition now measures a β sheet register shift as only a slight decrease in Q, as new non-native β sheet contacts compensate for the lost native contacts. Q^β parameters (both β and λ) were the same as Q^α. For $Q^\beta_{O,\text{fs}}$, we measured contacts between the fs orange residues in Table 4.1 and the red β sheet residues (AAs 8-13). For $Q^\beta_{G,\text{fs}}$, we measured contacts between the fs green and fs orange residues in Table 4.1. For $Q^\beta_{B,\text{gs}}$, we measured contacts between the gs blue residues in Table 4.1 and the red β sheet residues (AAs 8-13). The specific definitions for GT, BT, and OT are given below:

$$GT = Q^\alpha_{G,\text{gs}} - Q^\beta_{G,\text{fs}} \tag{4.2}$$

$$BT = Q^\beta_{B,\text{gs}} - Q^\alpha_{B,\text{fs}} \tag{4.3}$$

$$OT = Q^\alpha_{O,\text{gs}} - Q^\beta_{O,\text{fs}} \tag{4.4}$$

Because it is possible for non-native conformations to have more overall contacts between specified β sheets than the crystal structure does, Q^β ranges from 0 to approximately 1.3. Since Q_α is explicitly bounded between 0 and 1, this means that the typical ranges for GT and OT were between –1.3 and +1, while the typical range for BT was between –1 and +1.3.

Double well potential allows for proline isomerization during the course of a simulation. As described previously, Upside FF1.5 and FF2.1 by default handle the proline

Table 4.2: Proline isomerization based on Ac-Ala-Xaa-Pro-Ala-Lys-NH2. Experimental data (%cis (exp)) taken from refs. 4 and 5, measured at 23° at 20 mM sodium phosphate and pH 6.0, except for His, which was measured at pH 8.0. Theoretical results from Upside (%cis (ups)) were generated with $T = 0.89$ using FF2.1. Xaa candidates found in KaiB are marked with an asterisk. For Xaa = Pro, the proline at amino acid position 2 was also allowed to isomerize and was parametrized with the data from Xaa = Ala.

Xaa	%cis (exp)	%cis (ups)	$\Delta E_{\text{cis/trans}}$ (k_BT)
Pro*	6.0	6.2	2.1
Lys	6.8	7.9	2.4
Arg	7.2	7.6	1.5
Asp	7.3	6.7	0.6
Ala*	7.7	7.6	2.0
Cys	8.7	8.5	1.1
Glu	9.0	8.8	1.1
Thr*	9.4	9.2	1.1
His	9.5	9.6	1.1
Met	10.0	9.5	1.2
Val	10.4	10.3	1.2
Gln	11.5	11.1	0.9
Asn	11.6	11.4	0.5
Leu*	12.0	11.8	1.1
Ile	12.0	11.6	1.1
Gly	13.7	14.2	1.6
Phe	23.0	23.9	0
Tyr	24.0	23.3	0
Trp	37.7	37.6	–0.6

dihedral angle with a harmonic potential centered at either 180° or 0°, depending on whether the proline was initialized in the *trans* or *cis* state, respectively. It is possible, however, that some or all of the prolines in KaiB isomerize during the course of the fold switch. To allow for the the explicit isomerization of the prolines during individual simulations, we implemented a two-well proline isomerization potential, with minima at 180° and 0°, separated by a 3 k_BT barrier. We optimized the difference in energy between the *cis* and *trans* wells to match literature values that are dependent on the preceding amino acid [4, 5].

To set this parameter, we run a suite of simulations for five-amino acid fragments, each with at least one proline that is allowed to isomerize. In KaiB, the seven prolines are preceded

by one of four unique amino acids (Ala, Leu, Thr, Pro). For each one of these amino acids, a five-amino acid peptide was built as described in *Methods*. For each peptide, 8 independent unbiased simulations of length 20 million Upside time units were run from each initialization state (*cis* or *trans*), where $\Delta E_{\text{cis/trans}} = E_{\text{cis}} - E_{\text{trans}}$ ranged from –1.0 to 2.7 $k_B T$. The proline dihedral angle was calculated, and the proline was considered *cis* if $\psi \in [-\pi/2, \pi/2]$. We observed no difference in the probability distributions based on *cis*/*trans* initialization, so for all other amino acids, we only initialized with *trans* prolines. Table 4.2 describes the $\Delta E_{\text{cis/trans}}$ value that best reproduced experimental values. In this work, we used these parameters for all simulations where we implemented the double well proline potential.

Experimental hydrogen-deuterium exchange (HX) measurements reveal 16 EX2 residues. To help validate Upside as a model and determine the stability of fs KaiB secondary structures, ΔG_{HX} values were experimentally measured and compared to the values estimated from Upside simulations. The classical model of hydrogen-deuterium exchange assumes that exchange occurs after an amide hydrogen (H_N) undergoes a transition from a protected (closed) to a solvent-exposed (open) state [198, 199]:

$$\text{H}_\text{N}^{\text{closed}} \underset{k_{\text{cl}}}{\overset{k_{\text{op}}}{\rightleftharpoons}} \text{H}_\text{N}^{\text{open}} \xrightarrow{k_{\text{int}}} \text{D}_\text{N}^{\text{open}} \tag{4.5}$$

where k_{op} and k_{cl} are the opening and closing rates, respectively, and k_{int} is the intrinsic rate of exchange for a given residue. Since k_{int} is proportional to hydroxide concentration at pH 4 and above [200], this intrinsic rate is pH-dependent. Using either NMR or mass spectrometry, experiments measure the apparent rate at which protons exchange (k_{obs}), which can be expressed as

$$k_{\text{obs}} = \frac{k_{\text{op}} k_{\text{int}}}{k_{\text{op}} + k_{\text{cl}} + k_{int}} \simeq \frac{k_{\text{op}} k_{\text{int}}}{k_{\text{op}} + k_{\text{cl}}} = \frac{k_{\text{int}}}{\text{PF}} \tag{4.6}$$

where the so-called Protection Factor (PF) is defined as PF = $1+k_{\mathrm{cl}}/k_{\mathrm{op}}$. The approximation made in Eq. 4.6 is that $k_{\mathrm{cl}} \gg k_{int}$, which is known as the EX2 regime [201]. The opposite extreme, where $k_{\mathrm{int}} \gg k_{\mathrm{cl}}$, is known as the EX1 regime. As described in ref. 191, we compute ΔG_{HX} values by using

$$\Delta G_{\mathrm{HX}} = RT \ln(K_{\mathrm{eq}}) = RT \ln\left(\frac{k_{\mathrm{cl}}}{k_{\mathrm{op}}}\right) \simeq RT \ln(PF - 1) \tag{4.7}$$

Because this assumes the exchanging $\mathrm{H_N}$s are in the EX2 regime, we only compare residues between simulation and experiment that were EX2. Experimentally, in the EX2 regime $k_{\mathrm{obs}} \propto k_{\mathrm{int}}$ is pH-dependent. In contrast, in the EX1 regime $k_{\mathrm{int}} \gg k_{\mathrm{cl}}$, which means $k_{\mathrm{obs}} \simeq k_{\mathrm{op}}$ is not pH-dependent. Comparing experimental measures of the observed exchange rate across pH 4.5 and pH 6.5, we determined the following sixteen residues to be in the EX2 regime: V9, L25 - L32, Y40, A41, L65, R74, I76, V86, R91-L93. Comparing to the fs row of Table 4.1, we note that V9, L25-L32, and A41 map to secondary structures in the N-terminal domain, while L65, R74, I76, V86, and R91-L93 map to secondary structures in the fold switching C-terminal domain. These residues open and close quickly compared to the timescale of hydrogen-deuterium exchange, confirming that these secondary structures are thermodynamically stable.

ΔG_{HX} and its denaturant dependence. To generate computational estimates of ΔG_{HX}, we used Upside FF2.1 to run 48 T-REMD simulations of fs KaiB with P63, P70, and P72 locked in the *cis* conformation. We did this, instead of allowing the prolines to isomerize, because the rate of hydrogen exchange is much faster than the rate of proline isomerization [4]. Thus, experimental HX measurements correspond to a single isomerization state, which here was measured to be *cis* via X-ray crystallography. Each simulation was 2 million Upside time units long, saving every 100 time units, at temperatures ranging between $T = 0.7$ and $T = 1.1$ (in arbitrary units), attempting exchanges every 10 time units evaluated by

the Metropolis criterion. The protection factor and $\Delta G_{\mathrm{HX},i}$ for each non-proline residue (labelled with i) was calculated as in ref. 191:

$$\Delta G_{\mathrm{HX},i} = RT \ln \left(\frac{\overline{w \cdot PS_i}}{1 - \overline{w \cdot PS_i}} \right). \tag{4.8}$$

The overlines indicate an average across all N simulation frames, and w is the MBAR reweighting factor used to combine the T-REMD data [196]. PS_i indicates the protection state for the ith residue, determined by adapting Persson and Halle's criteria for exchange competent amide nitrogens (H-bond is broken and the NH is coordinated to at least 2 nearby waters [202]). Instead of hydrogen bonds, we used Upside to calculate an H-bond score, which is a geometric measure of nearby backbone carbonyl and amide oxygens. This score has been transformed to vary between 0 (no H-hond) and 1 (H-bond). The distribution of the H-bond score is effectively bimodal, with peaks near 0 and 1. Also, since Upside does not contain solvent molecules, we instead calculated the burial level (BL) of each residue. This burial level is a geometric measure of the number of nearby residues, with contributions from backbone beads and side chain beads. If $BL > 5$, the residue is considered "buried", meaning that it would not be accessible to solvent. Combining these two conditions, $PS_i = 1$ if $BL > 5$ or if the amide is H-bonded. Otherwise, $PS_i = 0$. A full description is found in ref. 191. The Upside temperature (which, again, is on an arbitrary scale) was chosen to minimize the mean squared error of the computational predictions for EX2 residues. Doing that, we found that $T = 0.87$, which is slightly higher than the estimated room temperature for Upside FF2.1 ($T = 0.85$).

The denaturant dependence of ΔG_{HX} for EX2 residues also provides information as to the origin of the unfolding events that lead to amide exposure [203]. These unfolding events are in three broad categories: global unfolding, subglobal unfolding, and local opening. The degree of denaturant sensitivity is measured as the m-value, the slope of ΔG_{HX} against denaturant concentration. The m-value of EX2 residues is proportional to the amplitude

of the amide-exposing unfolding event [204]. Residues that exchange via global, subglobal, and local unfolding events have a high, intermediate, and low/zero m-value, respectively. In practice, different regimes can dominate at different concentrations of denaturant. For example, an opening event that transitions from subglobal to global unfolding as the denaturant concentration increases corresponds to a biphasic ΔG_{HX} vs. denaturant concentration graph, with an intermediate slope at low denaturant concentrations transitioning to a large slope at high denaturant concentrations.

To measure the degree of local, subglobal, and global unfolding, m-values were calculated by fitting the initial (low denaturant) slope of $\Delta G_{\mathrm{HX},i}$ against the concentration of urea. The denaturant dependence of the computational prediction of $\Delta G_{\mathrm{HX},i}$ was calculated as in ref. 191. For this, the reweighting factor w in equation 4.8 was redefined to be

$$w = w_{\mathrm{MBAR}} \exp\left(\frac{-s \cdot N^{\mathrm{HN}}_{\mathrm{closed}} \cdot [\mathrm{den}]}{RT}\right). \tag{4.9}$$

This reweighting accounts for simulated denaturant (with concentration [den]) by assuming it destabilizes conformations proportional to the number of solvent-protected backbone amides ($N^{\mathrm{HN}}_{\mathrm{closed}}$); w_{MBAR} is the original MBAR reweighting term. The parameter that measures the sensitivity of KaiB to this denaturant (s) is left as a free parameter that is set to best match experiment. We vary this parameter between 0.02 and 0.52. We also re-optimize T around our previous value of $T = 0.87$, as there might be coupling between these two parameters. Together, the combination of (s, T) that minimized the mean squared error of the computational predictions for EX2 residues was $T = 0.87$ and $s = 0.25$.

Comparing computational and experimental HX measurements validates Upside and reveals primarily subglobal unfolding. We sought to compare experimental and simulated ΔG_{HX} and m-value measurements to validate Upside as a model, and to determine how fs KaiB tends to unfold. We show the simulated ΔG_{HX} as a function of

simulated denaturant (using equations 4.8 and 4.9) in Figure 4.2A, and explicitly compare to experimental data in Figure 4.2B.

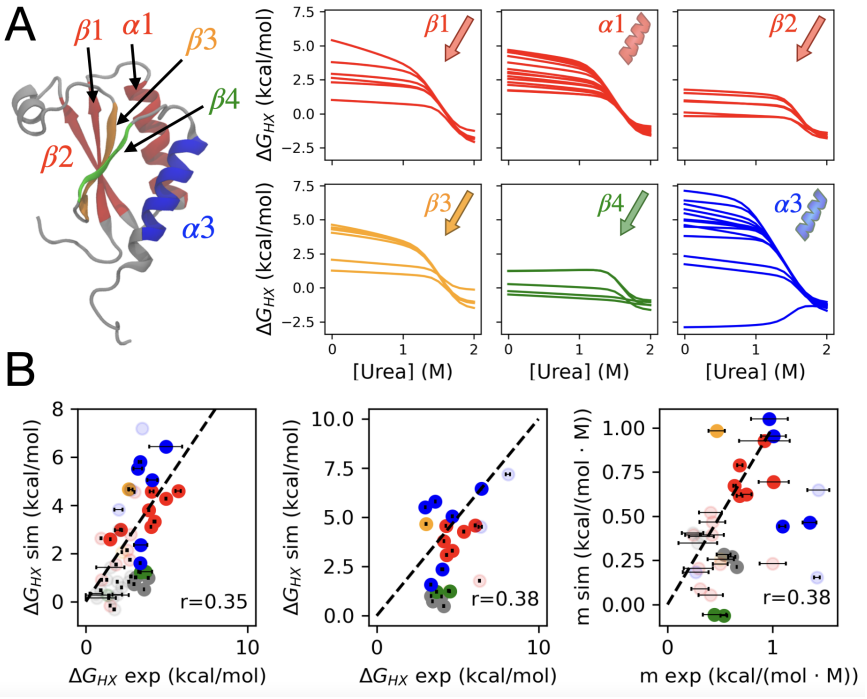


Figure 4.2: Comparing results for simulated and experimental HX measurements. (A) Simulated values of ΔG_{HX} as a function of denaturant for residues in the labelled secondary structures. We display the structure of the tested KaiB fs mutant for reference, with secondary structure elements labelled. (B) The comparison of simulated to experimental results of the fs KaiB mutant. We show comparisons of ΔG_{HX} at experimental pH=4.5 (left) and pH=6.5 (middle). We also show comparison of m-values (right), which is the initial slope of ΔG_{HX} against denaturant concentration. In all cases, data is colored by secondary structure as in panel A, and error bars are given for experimental data as the standard error across three runs. Opaque data points were experimentally determined to be in the EX2 regime, while translucent data points were either EX1 or undetermined. Also displayed is the line where $x = y$ (black dotted line). For the simulated results, we show results for $T = 0.87$ and $s = 0.25$, parameters that set the internal Upside temperature and sensitivity to simulated denaturant, respectively. We also show the Pearson correlation coefficient for EX2 residues.

The ΔG_{HX} values are well-clustered within each secondary structure (Figure 4.2A), providing evidence that Upside is consistently modeling HX for similarly protected residues. The residues with lower ΔG_{HX} compared to the rest of the secondary structure in $\beta 1$ (V9), $\beta 3$ (V68 and L69), and $\alpha 3$ (E84, E95, and E96) are all at the edges of their respective

secondary structure element, making it plausible that they are more solvent-exposed than other residues within that element. Furthermore, most of these traces exhibit a biphasic denaturant dependence, with relatively small slope at low simulated urea concentrations and a larger slope at higher concentrations, with the conversion point being at [Urea] $\approx$ 1.5 M. Beyond being consistent with experimental observations, this type of behavior is a hallmark of solvent exposure that proceeds through subglobal unfolding events, as discussed earlier. Here, at low/intermediate concentrations of urea, amide hydrogens can become solvent exposed through partial unfolding of secondary structures, as opposed to global unfolding of the entire protein (which is only found at high denaturant concentrations).

Turning to Figure 4.2B, we see that Upside predicts a larger range of ΔG_{HX} for EX2 residues compared to experimental estimates. It is possible that Upside is more sensitive to amino acid position within a secondary structure, as opposed to experimental estimates which more tightly cluster these ΔG_{HX} values. In particular, Upside seems to underestimate both ΔG_{HX} and m for residues that are either in $\beta 4$ (green) or not within a defined secondary structure element (gray). More work is needed to fully untangle the sources of error, whether they arise from force field inaccuracies or the lack of explicit solvent. Despite these quantitative discrepancies, Upside still captures qualitative experimental features like the biphasic ΔG_{HX} behavior discussed earlier and the lack of any residues that only become solvent-exposed via global unfolding events. If some residues exchanged only via global unfolding events, some ΔG_{HX} vs. [Urea] traces would be monotonic with a large slope, which we do not find in either experiment or simulation. This suggests that the fs KaiB mutant under experimental conditions (namely, 12 °C) always has a subglobal unfolding event that is more dynamically accessible than a global unfolding event at low denaturant concentrations. Additionally, comparison to experiment validated Upside as a model that is capable of measuring an experimentally-reasonable ensemble of structures, providing confidence in our statistical estimates moving forward.

4.3.2 Elucidation of KaiB Fold Switch Mechanism By Application of the Dynamical Galerkin Approximation

After CV discovery and Upside calibration, we moved on to characterizing the fold switching mechanism. We first generated a large set of sampling that explored the fold switch and explicitly allowed prolines to isomerize. We then used the Dynamical Galerkin Approximation (DGA) [25, 26] to estimate equilibrium averages and long-time dynamical statistics from this data set of unbiased simulations. From the DGA, we were able to compute a potential of mean force, a set of committors, and a set of reactive currents. The committor measures the probability of, for example, reaching gs KaiB before returning to fs KaiB, and the committor-1/2 surface defines the transition state ensemble for the fs-to-gs transition. The reactive current measures how trajectories that move from fs-to-gs (and vice versa) move across CVs, providing information as to specific mechanisms and their relative weights. Doing this, we statistically compared fs and gs KaiB and described how the two folds interconvert. We also computed the likelihood of individual mechanisms and identified structural features that corresponded to the transition state ensemble.

Generation of the unbiased dataset with proline isomerization allowed. To perform dynamical analysis, we wanted sampling that thoroughly covered the transition state regions between the fs and gs states. We also wanted to allow the prolines to switch their isomerization state during the simulation, as previously described. Thus, we needed to initialize a large set of simulations that explored CV space between fs and gs KaiB, and we needed to choose a simulation temperature. We do not use the temperature determined from HX analysis, as this corresponded to the experimental temperature ($\approx 12^{\circ}$C) for a fs KaiB mutant. Instead, we wanted a temperature for WT KaiB that was high enough to encourage the fs-to-gs transition. Furthermore, from the HX measurements, we knew that fs KaiB tends to unfold primarily through subglobal unfolding routes. Thus, we also wanted

this temperature to be low enough to not cause global KaiB melting. To this end, we chose $T = 0.89$, as during previous sampling we obtained eight fold switching events, and only one of them proceeded through a globally unfolded intermediate (see *Isomerizing prolines generated robust KaiB sampling*).

These eight fold switching simulations, even though they locked prolines into the *trans* state, provide a database of possible starting structures for new sampling that allowed the prolines to isomerize. Thus, from each of these 8 transitions, we selected 32 starting structures that were evenly spaced in simulation time, providing good coverage in the space of the CVs described earlier. From each of these starting structures, we launched 2 independent unbiased simulations of length 1.5 million time units, one initialized with P63, P70, and P72 in the *cis* state, and one initialized with the same prolines in the *trans* state. Regardless of the initialization, the simulations were parametrized with the double-well potential that allows for proline isomerization during the simulation itself. For all of these simulations, we used Upside FF2.1 at $T = 0.89$, saving configurations every 50 time units, with the previously described proline double-well parameters. This gave us a total of $8 \times 32 \times 2 = 512$ simulations, for an aggregate length of 768 million Upside time units. We observed the timescale of proline isomerization to be much faster than that of secondary structure melting/formation, such that the simulations can be meaningfully considered to sample a single ensemble. By comparing averages of CVs in sampling density maxima associated with fs and gs KaiB to the available crystal structures, we defined fs and gs KaiB in terms of our previously described CVs. In particular, we defined fs KaiB to be where $GT < -0.78$, $BT < -0.83$, $OT < -0.75$, and $\mathrm{RMSD}_{\mathrm{fs}} < 0.35$ nm. Additionally, we defined gs KaiB to be where $GT > 0.57$, $BT > 0.98$, $OT > 0.75$, and $\mathrm{RMSD}_{\mathrm{gs}} < 0.45$ nm.

P63, P70, and P71 (and not P72) are primarily *cis* in fs KaiB. We measured the backbone proline dihedral ω for all structures in our unbiased DGA dataset, considering a proline in the *cis* conformation if $|\omega| < 90°$. We then calculate the percentage of prolines in

Table 4.3: Percentage of prolines in *cis* conformation for the seven prolines in the WT KaiB sequence, histogrammed across both fs and gs structures from our unbiased DGA data set.

	P3	P19	P51	P63	P70	P71	P72
$\%\text{cis}_{\text{gs}}$	9.4	1.0	4.9	0.0	0.0	0.1	1.1
$\%\text{cis}_{\text{fs}}$	9.7	1.2	10.1	35.1	15.7	68.3	5.3

the *cis* conformation for each proline in the WT sequence, data summarized in Table 4.3.

For P63, P70, and P71, the *cis* isomer is preferentially associated fs KaiB, while the *trans* isomer is preferentially associated with gs KaiB. This is in comparison to data from available crystal structures, which resolve P63, P70, and P72 as the three prolines that switch isomerization state from gs to fs KaiB. While we see such behavior for P63 and P70, our data instead assigns P71 as the third proline to favor the *cis* isomer in fs KaiB, instead of P72. Indeed, P71 is 68.2% more likely to be *cis* in fs KaiB than it is in gs KaiB, according to our simulations. Of all WT KaiB prolines, P71 has the largest such preference, almost double that of P63. In contrast, P72 is dominantly *trans* for both gs and fs KaiB. This discrepancy between our Upside data and available crystallographic data could be due either to the sequence difference between WT KaiB and the fs-stabilized mutant, or an inadequacy of the model. Further experimental and simulation results, perhaps from all-atom molecular dynamics, are needed to further investigate proline behavior in KaiB.

Regardless, it is clear that the isomerization state of the prolines in the fold switching C-terminal domain is strongly correlated to the fold KaiB adopts. This is consistent with the observation that proline isomerization is an important element of protein folding. Instead of being driven to one stable fold by the correct isomerization of prolines, KaiB instead can adopt one of two different stable folds, each one associated with a different set of proline isomerization states. We discuss how proline isomerization is related to the mechanism of the fold switch itself in a later section.

2D averages correlate free energy basins to secondary structure, and reveal little free energy difference between fs and gs KaiB. To first investigate the mechanism of the KaiB fold switch, we built a 2D potential of mean force (PMF) in the space of *GT* and *BT*, CVs that track two secondary structure transformations in the C-terminal fold switching domain. This PMF, along with structures corresponding to various basins in this 2D space, are shown in Figure 4.3A.

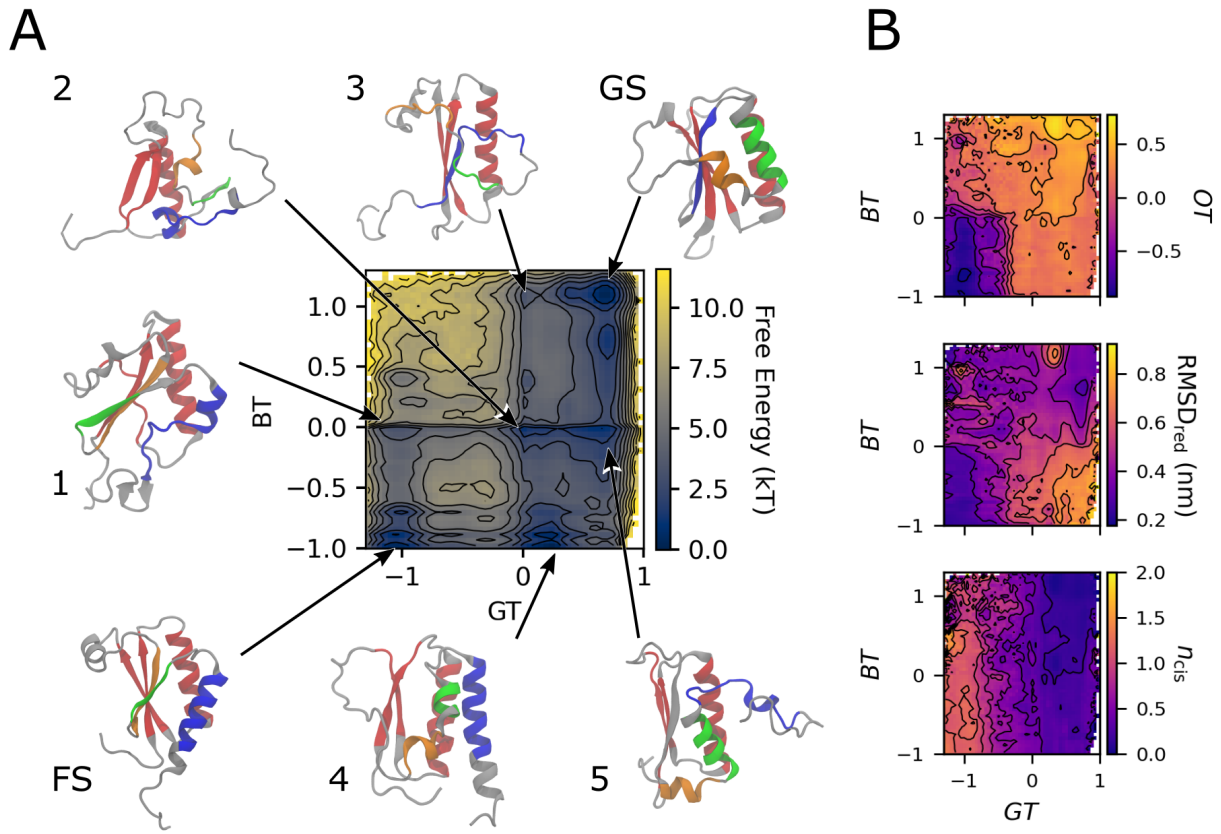


Figure 4.3: The potential of mean force and other equilibrium averages of collective variables. (A) The PMF as a function of *GT* and *BT*, with contours shown every k_BT. Fs KaiB is in the bottom left corner, while gs KaiB is in the top right corner. Other intermediate structures are labelled and displayed around the PMF. (B) Equilibrium averages of *OT* (top, contours every 0.2), RMSD_{red} (middle, contours every 0.1 nm), and the number of *cis* prolines (n_{cis}, contours every 0.2) in the C-terminal domain (P63, P70, P71, and P72).

First, we measured the average free energies of the basins corresponding to fs KaiB (bottom left corner of Figure 4.3A) and gs KaiB (top right corner), and found them to be

0.5 and 0.4 k_BT. This near-identical free energy of the two stable folds agrees with the results from Rivera and coworkers, who used confine-convert-release molecular dynamics [205, 206] and a thermodynamic cycle to measure the free energy difference between gs and fs KaiB as -1 ± 3 kcal/mol [192]. The two folds have near-identical stability, which is consistent with KaiB's metamorphic character. Interestingly, the global free energy minimum of our 2D PMF (set to 0 k_BT) is near Structure 4 of Figure 4.3A, where the green fs β sheet has been broken, and has slightly refolded to form a small, three-residue α helix, smaller than the green α helix in gs KaiB. While all the structures shown correspond to free energy basins, Structure 4 represents a particularly stable intermediate.

The PMF is largely dominated by vertical and horizontal free energy troughs, corresponding to the melting and refolding of individual elements of secondary structure. The lack of troughs along diagonals provide evidence that it is most energetically favorable for the secondary structure elements to fold and refold as units independently of one another, rather than in a concerted fashion. This is consistent with the idea of protein folding proceeding via foldons, small cooperative units that fold together in various orders [47, 48, 50, 207]. We first focus on describing motions along individual axes, and then describe global features of the PMF and its associated averages. As one moves from –1 to 0 to +1 along the GT axis of Figure 4.3A , the green β sheet in fs KaiB first melts to form a near-disordered structure, and then folds to the extended α helix. This corresponds to the transition in the green regions moving from FS to Structure 4 to Structure 5. Structure 5 is a metastable intermediate that corresponds to the green fs β sheet having completely refolded to a near-gs KaiB α helix, but with the blue fs α helix having melted into a disordered strand, here sticking out away from the rest of the protein.

Along the other axis, as one moves from –1 to 0 to +1 along BT, the blue α helix in fs KaiB first melts to form a near-disordered strand, before it refolds to a β sheet like the one seen in gs KaiB. This corresponds to the transition in the blue regions moving from

FS to Structure 1 to Structure 3. In terms of the Green Transition, Structure 1 has a fs-like green β sheet, while Structure 3 has a disordered green strand that has yet to adopt a secondary structure. The final structure displayed in Figure 4.3A is Structure 2, which has an C-terminal domain that is nearly completely disordered, as the green and blue secondary structures of both folds have melted.

The largest free energy barrier separating all of these structures is where $GT \approx -0.3$, corresponding to where the green fs β sheet melts. The two most stable energetic troughs that traverse this barrier are (1) between fs KaiB and Structure 4, and (2) between Structures 1 and 2. The maximum free energy barriers for these crossings are 3.8 and 4.3 k_BT, respectively. Thus, in terms of free energies, it is slightly more favorable to first melt the green fs β and then the blue fs α helix, compared to the reverse ordering. The similar maximum free energy barriers of these two pathways, however, suggest that both are likely populated at room temperature. In later sections, we use dynamical statistics to more completely describe the pathways of melting the fs KaiB secondary structures.

Comparing the PMF to the averages in Figure 4.3B, we note that the maximum free energy barrier is well correlated to a sharp increase in OT. Structurally, this increase means the orange fs β sheet has melted and is now disordered, or even beginning to fold into an α helix, similar to the one found in gs KaiB. Structures 2, 4, and 5 show such an orange α helix, while Structure 3 instead shows a disordered orange region. This provides evidence that the melting of green and orange β sheets in fs KaiB are correlated. Structurally, this makes sense, as these β sheets are interacting with one another, so once one melts, the other is more likely to melt. Furthermore, this free energy barrier is also correlated to an increase in average RMSD_{red}, from approximately 0.2 nm to approximately 0.5 nm. This level of unfolding is primarily manifested in the partial or complete melting of one red β sheet or the red α helix. The complete unfolding of the red region corresponds to an RMSD_{red} of 0.75 - 3 nm. While we chose the Upside temperature $T = 0.89$ to minimize this type of global

unfolding, we do see some such behavior in the bottom right quadrant of the PMF. In a later section, we describe the contribution of this region to the overall progress of the fold switch.

Finally, where $GT \approx -0.5$, we also find that the average number of *cis* prolines in the C-terminal domain (n_{cis}) decreases from approximately 1 to approximately 0.2. P63, P70, and P71 are all located either at the beginning or end of the green/orange fs KaiB β sheets. The isomerization of these residues could thus disrupt green/orange β sheet hydrogen bonds, thereby encouraging the sheet to break. This is consistent with the isomerization occurring where $GT \approx -0.5$, since GT tracks contacts between green/orange residues. However, proline isomerization in the C-terminal domain (dominated in our case by P71) largely precedes the primary free energy barrier. The melting of the green and orange β sheets has a higher free energy cost than the isomerization of the prolines does.

Once the free energy barrier has been crossed and the green and orange β sheets have been broken (where $GT > 0$), the free energy surface is relatively smooth. The observed free energy basins are relatively shallow and are connected by free energetic troughs with maximum barriers of 2-3 k_BT. In particular, starting from gs KaiB, the barrier to melting either the blue β sheet or the green α helix is approximately 2 k_BT, lower than the 3-4 k_BT barrier required to melt the fs KaiB elements of secondary structure.

2D dynamical statistics reveal transition state ensemble that corresponds to the breaking/formation of fs KaiB secondary structures. Dynamical statistics like the committor, q, and the reactive current, J, provide useful mechanistic information about rare events. Since the committor from fs to gs KaiB (q_{fs2gs}) tracks the probability of next entering gs KaiB instead of fs KaiB, it is by construction the ideal reaction coordinate to track the fold switch. In addition, the committor-1/2 surface defines the transition state ensemble. Finally, the reactive current from fs to gs KaiB (J_{fs2gs}) provides a view into how reactive trajectories move across our set of CVs, and in particular across GT and BT, discussed at length in the previous section. Estimates of the committor and reactive current are shown

in Figure 4.4A and B, respectively. Furthermore, since q_{fs2gs} tracks the progress of the fold switch itself, by histogramming our secondary structure CVs as a function of committor, one can gain information as to the population of KaiB secondary structures at different stages of reaction. These histograms, particularly highlighting the transition state ensemble ($q_{\text{fs2gs}} = 0.5$), are shown in Figure 4.4C.

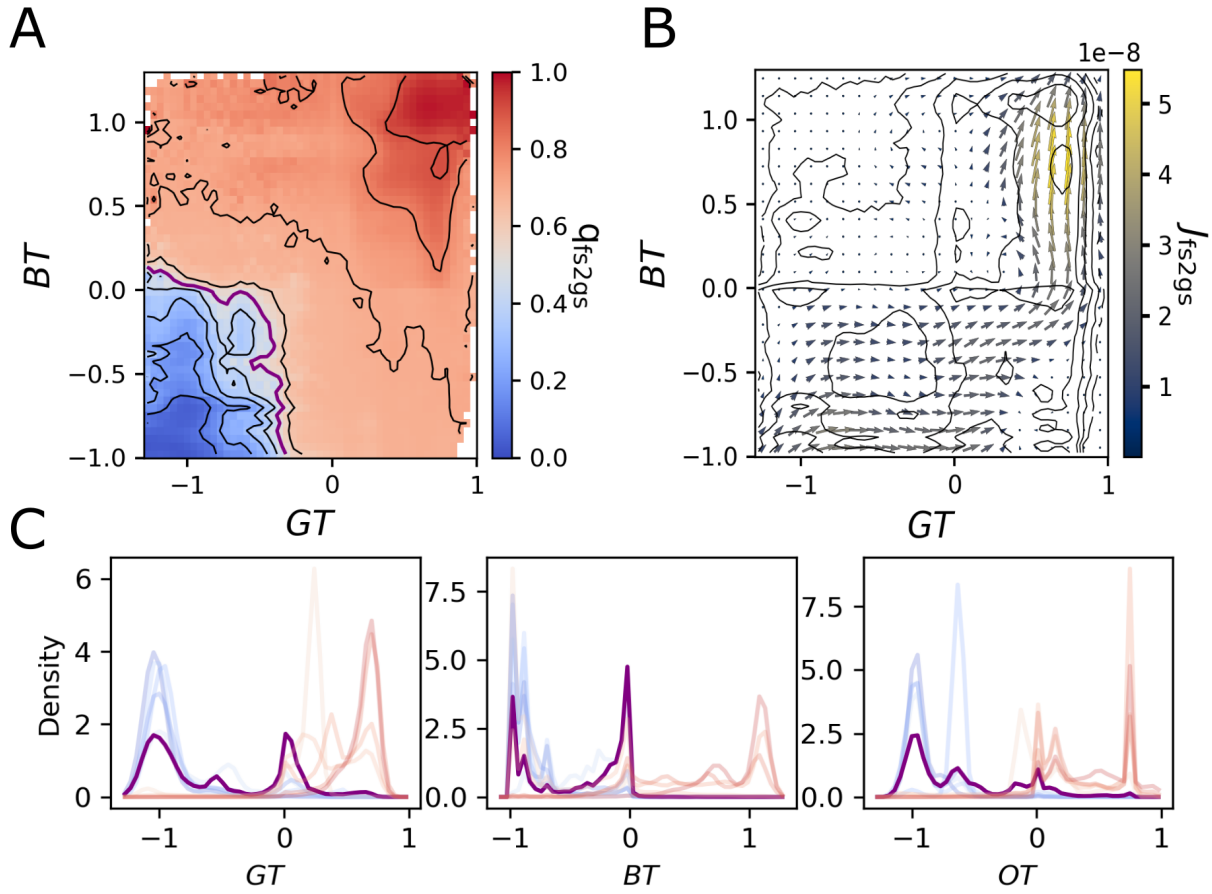


Figure 4.4: 2D dynamical statistics of the KaiB fold switch. (A) The average committor from fs to gs KaiB as a function of GT and BT. Contours are shown every 0.1, with the committor-1/2 surface represented by the thicker, purple contour. (B) The reactive current (binned into a 21×21 grid in GT and BT) from fs to gs KaiB, represented by arrows whose color and length represent the magnitude of the current. PMF contours, spaced every 2 k_BT, are also displayed. (C) Normalized histograms for GT (left), BT (middle), and OT (right) for eleven different bins of the committor q_{fs2gs} between 0 and 1. The colors of each translucent line represent the committor value in that bin, and are the same as in the left panel of A. The committor-1/2 line is drawn in opaque purple.

The location of the committor-1/2 surface for q_{fs2gs} correlates well with the location of the free energy barrier described in the last section. Thus, the transition state ensemble correlates with the CV behavior previously described; importantly, this surface is crossed regardless of which element of fs KaiB secondary structure melts. Melting either the β sheet or the α helix in the N-terminal domain is sufficient to reach the transition state, after which it becomes more probable to reach gs KaiB than to return to fs KaiB. Furthermore, comparing to Figure 4.3B, the initial isomerization of P63, P70, and P71 likely occurs before the transition state is reached. Thus, while the transition state ensemble is populated by mostly *trans* prolines, the isomerization of prolines in the C-terminal domain does not determine the kinetics of the fold switch.

The reactive current in Figure 4.4B shows that reactive trajectories can cross the free energy barrier with the blue fs KaiB α helix either melted or folded, with comparable currents along both pathways. This current also reveals that either reactive trajectories avoid the area where RMSD_{red} is high (the lower right corner), or trajectories that enter that area must first leave before proceeding on to gs KaiB. This confirms that in general, reactive trajectories tend to avoid fully unfolded intermediates altogether, or must refold to a near-native like N-terminal domain before continuing in the fold switch. Additionally, most of the current that has crossed the transition state flows up the right edge of Figure 4.4B once $GT \approx 1$, moving from Structure 5 to gs KaiB in terms of the structures shown in Figure 4.3A. This means that the likely last step of the fs-to-gs fold switch is the recruitment of the unstructured blue strand, which has already melted from the fs α helix, into the β sheet like what is seen in gs KaiB. The green and orange α helices have most likely already been formed before this step.

We can also look at the sequence of secondary structure peaks at the committor-1/2 surface to help characterize the transition state ensemble. As shown by the purple traces in Figure 4.4C (which represent histograms at the transition state, where $q_{\text{fs2gs}} = 0.5$), both

BT, GT and OT all exhibit local peaks at $xT = 0$. Defining $xT = -0.5$ as the point where a secondary structure element transitions from folded to melted, 60% of transition state structures contain a melted blue fs KaiB helix, compared to the 40% and 30% that contain melted green or orange fs KaiB β sheets, respectively. The transition state thus likely contains a mix of folded and unfolded secondary structure elements in the C-terminal fold switching domain. However, the multimodality of the transition state histograms in Figure 4.4C, combined with the wide area of significant reactive current in Figure 4.4B and similar free energy barriers for secondary structure melting in Figure 4.3A, tells us that the fold switch must be multipathway.

3D dynamical statistics reveal an ensemble of pathways, with the melting of C-terminal fs KaiB helix playing a central role. To visualize and quantify the ensemble of pathways available for the fs-to-gs fold switch, we computed our committor and reactive currents in various three-dimensional spaces. First, in addition to GT and BT which we have heretofore used as our 2D axes, we consider OT, to fully characterize the secondary structure behavior of the system. This projection is seen in Figure 4.5A. We see that once either GT, BT, or $OT > 0$, we have crossed over the committor-1/2 surface. In other words, loss of any one fs KaiB secondary structure element is sufficient for commitment to the ground state. Because of the multitude of ways to cross the transition state, using a reactive current in this 3D space to determine relative weights of pathways becomes quite difficult to interpret. Without some measure of progress of the fold switch, currents at very different stages of the reaction are averaged together because they exist in similar areas of this 3D space. Thus, we also use as a third dimension the committor itself, q_{fs2gs}, since it by construction tracks the fs-to-gs transition. The committor on this projection is seen in Figure 4.5B.

Using the same CV space as Figure 4.5B, we calculate 3D reactive currents that we bin into a $50 \times 50 \times 50$ grid and smooth with a Gaussian filter, as described in *Methods*. This gives us information as to how reactive trajectories flow in this CV space. This is

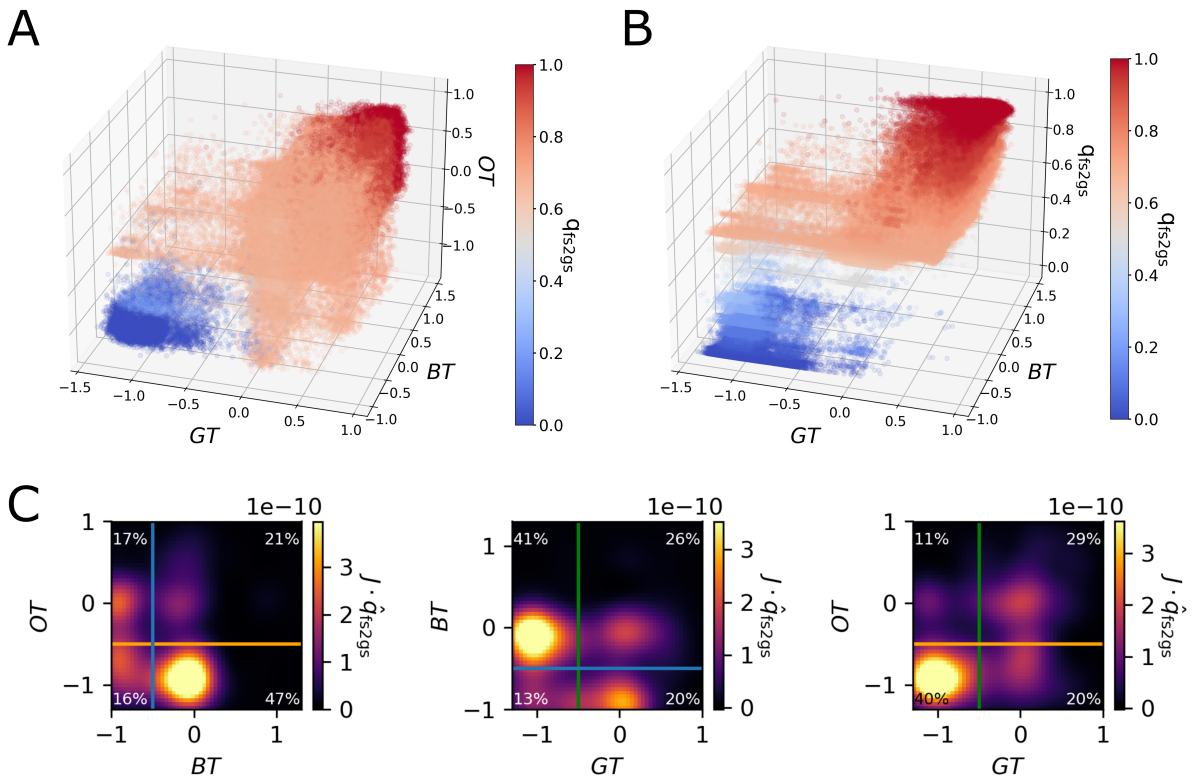


Figure 4.5: Three-dimensional dynamical statistics of the KaiB fold switch. Scatter plot of the committor from fs to gs KaiB as a function of (A) GT, BT, and OT and (B) GT, BT, and q_{fs2gs}. (C) 2D slices of a 3D reactive current in the direction of the committor at $q_{\text{fs2fs}} = 0.49$. The three panels correspond to three different current calculations. All three have q_{fs2gs} as the third dimension. The other two dimensions are pairwise combinations of secondary structure variables: OT and BT (left), BT and GT (middle), and OT and GT (right). Overlaid are lines where GT, BT, or $OT = 0.5$, colored appropriately, representing when these secondary structures transition from folded to melted. These lines divide each slice into quadrants that represent pathways, and the weights of the quadrants relative to each individual slice are shown on the graphs.

useful information because, by using the committor as the third dimension, we have clearly separated out various pathways based on their probability of proceeding to gs KaiB. We can then use this 3D reactive current to calculate reactive weights of pathways that exhibit various combinations of GT and BT. We do this by selecting a dividing surface that separates fs and gs KaiB, and calculate the reactive flux through patches of the surface relative to the entire surface. In this case, we choose $q_{\text{fs2gs}} = 0.49$ as a dividing surface, to characterize

the formation of the transition state. We choose $BT = 0.5$ and $GT = 0.5$ as the boundaries to define the patches. We chose these boundaries to be an intermediate value between -1 (folded) and 0 (unfolded). We repeat this procedure for the other pairwise combinations of BT, GT, and OT to fully characterize the secondary structure behavior at the formation of the transition state. The resulting set of three 2D slices, which each cleanly separate 4 reactive tubes, are shown in Figure 4.5C.

A caveat to this analysis is that the three slices come from three separate current calculations, meaning that in all cases, the third secondary structure variable is averaged over, provided the committor value is the same. This means that, using the above methods, is impossible to rigorously calculate relative weights for the pathways that melt these three secondary structure elements in arbitrary order. We note that performing this procedure in four dimensions (GT, BT, OT, and q_{fs2gs}) can solve this problem, but such a projection can often be numerically unstable and difficult to visualize. This is left to future work. However, using Figure 4.5C, we can still glean some information as to the likelihood of various pathways. For example, in the left and middle panels of Figure 4.5C, 47% and 41% of current flows through pathways where the blue α helix has melted, while the other element of secondary structure (either the orange or green β sheet) is maintained in a near-native fold. It is likely that this behavior also corresponds to the bottom left quadrant of the rightmost panel. Here, the current is flowing through the transition state with both GT and OT having near-fs KaiB values, meaning that the blue α helix has likely melted. This further confirms the analysis from the previous section, that much of the time (40-50% of the time), the blue fs KaiB α helix has melted to generate the transition state, shortly after which the green and orange β sheets break. However, we also now clearly see the diversity of pathways available to generate this transition state. All other displayed combinations of secondary structure contain between 10-30% of the reactive current as it flows through the transition state ensemble. Thus, the secondary structures can melt in various orders during the initial

steps of the fs-to-gs transition.

4.4 Conclusions

Organisms in all kingdoms of the tree of life have circadian rhythms, and the Kai system is the simplest known circadian oscillator. Notably, KaiB is a metamorphic protein, an important and growing class of proteins that can reversibly switch between two stable folds. In the case of KaiB, these are the ground state (gs) and fold switched state (fs). Because of the rarity of metamorphic transitions, however, such fold switching proteins are often experimentally and computationally intractable to investigate. Here, we use a combination of coarse-grained molecular dynamics, dynamical analysis, and hydrogen-deuterium exchange (HX) NMR, to investigate the mechanism for the KaiB fold switch. The simulated and experimental HX measurements, in qualitative agreement, revealed a lack of global unfolding in a fs-stabilized KaiB mutant. Using this information to set simulation parameters, we used a dynamical analysis pipeline that we recently introduced [25, 26] to extract long-time statistics to determine the mechanism of the KaiB fold switch. We also explicitly allow for proline isomerization by building and parameterizing a new term to the coarse-grained model's potential.

We identify three prolines (P63, P70, and P71) that preferentially populate the *cis* isomer when in fs KaiB and the *trans* isomer when in gs KaiB. In particular, our P71 assignment disagrees with available crystal structures, which instead identify P72 as the third proline that switches isomerization state between gs KaiB and a fs-stabilized KaiB mutant [74, 75]. We find the free energy difference between fs and gs KaiB to be negligible, in agreement with previous calculations. The three elements of secondary structure in the C-terminal fold switching domain can fold and refold in various orders. Regardless of the order, these secondary structures act as foldons, folding and unfolding near-independently of one another. The melting of the C-terminal α helix of fs KaiB occurs first in between 40-50%

or transitions, but the melting of any element of fs KaiB secondary structure is enough to cross the primary free energy barrier for fold switching. Beyond revealing the previously-uncharacterized mechanisms of KaiB fold switching, this work is the first statistical view of transitions in a metamorphic protein.

CHAPTER 5

CONCLUSIONS AND OUTLOOK

At its heart, protein folding is the process where disordered amino acids search for and eventually adopt a stable structure. This general idea cuts across all of the work presented in this thesis. We have presented three case studies where specific proteins undergo transitions between order and disorder, either in isolation or when paired with other species. In all cases, we found an ensemble of pathways where the proteins (and assemblies thereof) melted, rearranged, or reorganized. There is no single mechanism for any of these processes. Indeed, trying to simplify complex biological equilibria to a single mechanism is misguided. By ignoring the diversity of pathways available to proteins, we sacrifice a true understanding of the underlying process in favor of perceived conceptual elegance. As we have shown through studies in this work, understanding the effects of mutations, for example, requires a thorough knowledge of all available pathways. Development in the fields of structure-based and motion-based drug design depend on exactly this type of knowledge.

In Chapter 2, we described the ensemble of pathways for the insulin dimer dissociation. These range from induced fit to conformational selection, and include multiple intermediate pathways that blend elements of both extremes. As the insulin monomers start to separate, they sometimes undergo an order-to-disorder transition primarily described by the detachment of insulin's B-chain C-terminus. This is not surprising, as the protein-protein interactions in the broader assembly (the dimer) are slowly being broken. As normally-adjacent and mutually-stabilizing elements of secondary structure separate, the monomer has to now search for a new stable structure. In that sense, coupled folding and binding reduces to a protein folding problem. Instead of a dichotomy between induced fit and conformational selection pathways, we should expect to see a diverse ensemble of pathways like we do for protein folding. Each one of these pathways involve elements of secondary structure breaking and reforming in various orders, accompanied by varying degrees of disorder. This

is exactly what we describe in Chapter 2 when we characterize the α and β paths, along with their various intermediate pathways. This is also quite similar to the fold switching pathways we observe in Chapter 4.

There is currently ongoing work that aims to calculate dynamical statistics for the dimer dissociation, much like what we present in Chapters 3 and 4 of this thesis. A statistical understanding of reactive pathways could provide key insights into the potential design of new diabetes therapeutics. While current fast-acting insulin analogs are designed to destabilize the dimer interface [65, 66], it is currently unknown how these mutations affect the mechanisms of the dimer dissociation. Comparing the pathway ensemble for WT and mutant insulins could generate new ideas for targeted insulin mutations that affect the dynamics of the dissociation, as opposed to just dimeric structure. Furthermore, some single-chain insulin analogs that tether the C-terminus of insulin's B chain to the N-terminus of insulin's A chain have been observed to have near-WT insulin levels of biological activity [140, 208, 209]. It is currently unknown how these analogs might associate and dissociate, and if the introduction of the tether prevents the detachment of the B-chain C-terminus. Simulations investigating these concepts could prove enlightening. Finally, there is much opportunity to pair molecular dynamics with computational infrared (IR) spectroscopy, extending the work done in Chapter 2 and elsewhere [125, 170, 210–212]. The ability to use DGA to characterize the dynamics of rare events complements the ability of ultrafast 2D IR spectroscopy to track these same dynamics on the picosecond scale. Reweighting molecular dynamics trajectories to generate a kinetic model consistent with experiment could provide a powerful avenue for interpretation and validation of both simulation and experiment [213, 214].

In Chapter 3, we quantified and characterized six pathways for the phenol release from the insulin hexamer. Two of these pathways involve the opening of phenol's primary escape channel, as the gatekeeper residues that surround the bound phenol separate from one another. We also explicitly described how two targeted mutations affected the relative weights

of these six pathways. For all systems, the phenol can either escape by pushing through relatively dense areas of protein side chains, or it can escape through a widened channel that connects the binding pocket to the bulk solvent. This channel opening corresponds to α helices on the trimer interface unpacking and separating. Here, again, we find elements of protein secondary structure reorganizing as part of an unbinding process. The unbinding here is of a small molecule ligand instead of another protein, as in Chapter 2. Regardless, as the stabilizing phenol leaves the hexamer, the protein searches for a stable overall structure.

This work showed the potential of combining multiple protein mutations that simultaneously affect multiple pathways of phenol unbinding. Moving forward, it would be of interest to explore mutants/analogs that include unnatural amino acids and/or covalent linkers, which could further discourage phenol release and extend hexamer lifetimes. Furthermore, it is possible to design different phenolic ligands that bind in a similar pocket as phenol [172, 215]. Using DGA to model the release of these ligands could provide dynamical insight as to why substitution with different ligand scaffolds and/or functional groups leads to markedly different hexameric lifetimes. This type of motion-based ligand design would be another avenue for the generation of diabetes therapeutics that take advantage of the hexamer-ligand equilibrium to achieve desired pharmacokinetics. Finally, this type of potential study demonstrates the need for an analysis method that generates dynamical statistics for a family of ligands/mutants without the need for extensive additional sampling. Drug discovery relies on rapid, high-throughput methods for virtual screening of initial drug candidates [7]. Building a large DGA dataset for each such candidate would be computationally intractable. Instead, we would want to generate estimates of dynamical statistics for multiple ligands from a single data set. There has been previous work using path reweighting to do exactly that for Markov State Models [216–218]. Developing a similar method for DGA is an intriguing area of future research.

In Chapter 4, we used near-atomic simulation and DGA to explore the diverse mech-

anisms of the KaiB fold switch. We compared simulated and experimental hydrogen-deuterium exchange measurements, finding that fs KaiB primarily unfolds through sub-global unfolding pathways. Furthermore, the mechanism of the fold switch involves the unfolding/refolding of cooperative, near-independent elements of secondary structure. These foldons can melt and form in various orders. This view of fold switching is in agreement with existing models of protein folding, much like the other chapters of this work. An immediate next step would be to quantitatively compare the relative weights of all transition pathways by calculating reactive currents in four dimensions. Furthermore, we can seed all-atom molecular dynamics simulations from our near-atomic starting structures (after reconstructing side chains) to further validate the model. Another interesting future direction is to calibrate the free energy barrier of proline switching using experimental techniques capable of measuring the rate of isomerization, like single-molecule spectroscopy [219, 220]. We can also apply this computational pipeline to study the fold-switching of other metamorphic proteins, like the RfaH transcription factor [185, 186].

Throughout this thesis, we demonstrated that computational pipelines involving umbrella sampling and DGA are specifically well suited for investigations of multipathway processes. Determining good collective variables is many times prohibitively difficult for such processes. As seen throughout this thesis, the existence of multiple pathways implies a wide variety of molecular motions important to the overall process. In turn, this means that the set of collective variables needed to describe multipathway reactions is quite large, a challenge for most enhanced sampling techniques. The methods we displayed in this thesis overcome this challenge in multiple ways. In Chapter 2, we presented a computational pipeline that combined Replica Exchange Umbrella Sampling (REUS) and the Eigenvector Method for Umbrella Sampling (EMUS [93]) to characterize a multipathway reaction. It is not obvious that umbrella sampling would be amenable to such a study, since it is highly dependent on the CVs chosen for biasing. However, our pipeline presents three chief advantages. First,

the addition of replica exchange helps control for slowly-relaxing degrees of freedom, which are common in multipathway systems. Second, EMUS enables the calculation of asymptotic variances in the PMF, allowing us to monitor the convergence in free energy as we add sampling. Furthermore, we can decompose the per-window contributions to high error regions, allowing for computationally efficient adaptive sampling. In Chapter 2, these features allowed us to both identify and sample pathways that otherwise would have remained hidden. Third, EMUS allows for the re-projection of the PMF into any CV space without the need for additional sampling. This feature allowed us to examine the free energy surface in a wide variety of physically-motivated CVs, independent of the two CVs chosen for biasing. This was an invaluable tool for identifying and characterizing many intermediate pathways.

In Chapters 3 and 4, we instead characterized multipathway processes using DGA, a powerful framework that allowed us to identify and quantitatively compare the relative weights of multiple pathways. This method relies on unbiased sampling, meaning that compared to methods like umbrella sampling that bias on a few chosen CVs, there is much less dependence on CVs in DGA. Indeed, as DGA inputs, we used hundreds of pairwise backbone distances to account for a broad set of molecular motions that we did not explicitly know *a priori*. Again, the ability to project the DGA-generated PMF and associated averages into any CV space was a very useful tool. Even more, the DGA allows for the calculation of the committor at every point in the data set, meaning the committor itself can be used as a CV. The committor, by construction, is the perfect reaction coordinate. It explicitly tracks the reaction progress between any two arbitrary-defined sets, regardless of the underlying rate of said transition. Thus, the committor is effectively a data-driven CV that can be defined to perfectly track any transition of interest. Because it both considers dynamics and does not only measure slow modes, it is a more direct measure of a transition than the results from other data-driven dimensionality reduction techniques like Principal Component Analysis (PCA [221, 222]) or the Variational Approach to Conformational Dynamics (VAC

[223–225]). We note that VAC encompasses the popular method of Time-lagged Independent Component Analysis (TICA [226–228]). In both Chapter 3 and Chapter 4, we showed how calculating the reactive current in a CV space that includes the committor cleanly separates multiple tubes of reactive current, each corresponding to an individual mechanism. These tubes were difficult to discern (and even more difficult to quantitatively compare) using purely physical CVs. We believe this approach of separating pathways using the reactive current as a function of committor will prove to be a general method of separating and comparing mechanisms in multipathway reactions.

Finally, a particular promising area of future research is on the mechanism of insulin fibrillation. Insulin forms amyloid fibrils under conditions that disfavor the dimer, such as elevated temperature, low pH, and increased ionic strength [229–231]. Fibril formation complicates production of insulin and has been a limiting factor in its long-term storage. Furthermore, amyloid fibrils have been observed to form in patients at sites of frequent injections [232]. A model of the insulin fibril was built from a crystal structure of the smallest fragment of the insulin B-chain that fibrillates. This model implies that amino acids normally in or near α helices when in monomeric insulin (B11-B17 and A12-19) instead form β sheets to form the central backbone of the fibril [63]. Essentially, monomeric insulin undergoes a fold switch and then associates into a polymeric assembly rich in β sheets along a central, fibril-like backbone. In this thesis, we have shown that our computational pipelines are particularly effective at characterizing both fold switching and coupled folding and binding. Thus, we are well positioned to computationally investigate the mechanisms of insulin fibril formation. Given that insulin fibrils exhibit the cross-β diffraction pattern common to amyloid fibrils [233], these potential studies promise to have broad implications beyond insulin, to neurodegenerative disorders like Parkinson's and Alzheimer's diseases.

REFERENCES

[1] Mahmoud Moradi. Codes and Scripts, Theoretical and Computational Biophysics Group. https://www.ks.uiuc.edu/mahmoud/codes.html, (accessed July 15, 2019).

[2] Puja Banerjee, Sayantan Mondal, and Biman Bagchi. Insulin Dimer Dissociation in Aqueous Solution: A Computational Study of Free Energy Landscape and Evolving Microscopic Structure Along the Reaction Pathway. *J. Chem. Phys.*, 149(11):114902, 2018.

[3] Puja Banerjee, Sayantan Mondal, and Biman Bagchi. Effect of Ethanol on Insulin Dimer Dissociation. *J. Chem. Phys.*, 150(8):084902, 2019.

[4] Beatrice M P Huyghues-Despointes, J Martin Scholtz, and C Nick Pace. Protein conformational stabilities can be determined from hydrogen exchange rates. *Nature Struct. Biol.*, 6:910–912, 1999.

[5] Ulf Reimer, Gerd Scherer, Mario Drewello, Susanne Kruber, Mike Schutkowski, and Gunter Fischer. Side-chain effects on peptidyl-prolyl cis/trans isomerisation. *J. Mol. Biol.*, 279:449–460, 6 1998.

[6] J. Towns, T. Cockerill, M. Dahan, I. Foster, K. Gaither, A. Grimshaw, V. Hazlewood, S. Lathrop, D. Lifka, G. D. Peterson, R. Roskies, J. R. Scott, and N. Wilkins-Diehr. XSEDE: Accelerating Scientific Discovery. *Comput. Sci. Eng.*, 16(5):62–74, 2014.

[7] Jacob D Durrant and J Andrew McCammon. Molecular dynamics simulations and drug discovery. *BMC Biol.*, 9:71, 12 2011.

[8] David E. Shaw, Peter J. Adams, Asaph Azaria, Joseph A. Bank, Brannon Batson, Alistair Bell, Michael Bergdorf, Jhanvi Bhatt, J. Adam Butts, Timothy Correia, Robert M. Dirks, Ron O. Dror, Michael P. Eastwood, Bruce Edwards, Amos Even, Peter Feldmann, Michael Fenn, Christopher H. Fenton, Anthony Forte, Joseph Gagliardo, Gennette Gill, Maria Gorlatova, Brian Greskamp, J.P. Grossman, Justin Gullingsrud, Anissa Harper, William Hasenplaugh, Mark Heily, Benjamin Colin Heshmat, Jeremy Hunt, Douglas J. Ierardi, Lev Iserovich, Bryan L. Jackson, Nick P. Johnson, Mollie M. Kirk, John L. Klepeis, Jeffrey S. Kuskin, Kenneth M. Mackenzie, Roy J. Mader, Richard McGowen, Adam McLaughlin, Mark A. Moraes, Mohamed H. Nasr, Lawrence J. Nociolo, Lief O'Donnell, Andrew Parker, Jon L. Peticolas, Goran Pocina, Cristian Predescu, Terry Quan, John K. Salmon, Carl Schwink, Keun Sup Shim, Naseer Siddique, Jochen Spengler, Tamas Szalay, Raymond Tabladillo, Reinhard Tartler, Andrew G. Taube, Michael Theobald, Brian Towles, William Vick, Stanley C. Wang, Michael Wazlowski, Madeleine J. Weingarten, John M. Williams, and Kevin A. Yuh. Anton 3: Twenty Microseconds of Molecular Dynamics Simulation Before Lunch. *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–11, 11 2021.

[9] Vladislav V Kiselyov, Soetkin Versteyhe, Lisbeth Gauguin, and Pierre De Meyts. Harmonic oscillator model of the insulin and IGF1 receptors' allosteric binding and activation. *Mol. Syst. Biol.*, 5:243, 1 2009.

[10] Albert C. Pan, Daniel Jacobson, Konstantin Yatsenko, Duluxan Sritharan, Thomas M. Weinreich, and David E. Shaw. Atomic-Level Characterization of Protein–Protein Association. *Proc. Natl. Acad. Sci.*, 116(10):4244–4249, 2019.

[11] Glenn M. Torrie and John P. Valleau. Monte Carlo free energy estimates using non-Boltzmann sampling: Application to the sub-critical Lennard-Jones fluid. *Chem. Phys. Let.*, 28(4):578–581, 1974.

[12] G. M. Torrie and J. P. Valleau. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.*, 23(2):187–199, 1977.

[13] Alessandro Laio and Michele Parrinello. Escaping free-energy minima. *P. Natl. Acad. Sci.*, 99(20):12562–12566, 2002.

[14] G.A. Huber and S. Kim. Weighted-ensemble Brownian dynamics simulations for protein association reactions. *Biophys. J.*, 70(1):97–110, 1996.

[15] Rosalind J. Allen, Patrick B. Warren, and Pieter Rein ten Wolde. Sampling Rare Switching Events in Biochemical Networks. *Phys. Rev. Lett.*, 94:018104, Jan 2005.

[16] Sergei V. Krivov and Martin Karplus. Hidden complexity of free energy surfaces for peptide (protein) folding. *P. Natl. Acad. Sci.*, 101:14766–14770, 10 2004.

[17] Nina Singhal, Christopher D. Snow, and Vijay S. Pande. Using path sampling to build better Markovian state models: Predicting the folding rate and mechanism of a tryptophan zipper beta hairpin. *J. Chem. Phys.*, 121:415, 2004.

[18] Jan-Hendrik Prinz, Hao Wu, Marco Sarich, Bettina Keller, Martin Senne, Martin Held, John D. Chodera, Christof Schütte, and Frank Noé. Markov models of molecular kinetics: Generation and validation. *J. Chem. Phys.*, 134:174105, 5 2011.

[19] Vijay S. Pande, Kyle Beauchamp, and Gregory R. Bowman. Everything you wanted to know about Markov State Models but were afraid to ask. *Methods*, 52(1):99–105, SEP 2010.

[20] Brooke E Husic and Vijay S Pande. Markov State Models : From an Art to a Science. *J. Am. Chem. Soc.*, 140:2386–2396, 2018.

[21] Wei Wang, Siqin Cao, Lizhe Zhu, and Xuhui Huang. Constructing Markov State Models to elucidate the functional conformational changes of complex biomolecules. *WIREs Comput. Mol. Sci.*, 8, 1 2018.

[22] Nuria Plattner, Stefan Doerr, Gianni De Fabritiis, and Frank Noé. Complete Protein-Protein Association Kinetics in Atomic Detail Revealed by Molecular Dynamics Simulations and Markov Modelling. *Nat. Chem.*, 9(10):1005–1011, 2017.

[23] Fabian Paul, Frank Noe, and Thomas R Weikl. Identifying Conformational-Selection and Induced-Fit Aspects in the Binding-Induced Folding of PMI from Markov State Modeling of Atomistic Simulations. *J. Phys. Chem. B*, 122:5649–5656, 2018.

[24] Maxwell I. Zimmerman, Justin R. Porter, Michael D. Ward, Sukrit Singh, Neha Vithani, Artur Meller, Upasana L. Mallimadugula, Catherine E. Kuhn, Jonathan H. Borowsky, Rafal P. Wiewiora, Matthew F. D. Hurley, Aoife M. Harbison, Carl A. Fogarty, Joseph E. Coffland, Elisa Fadda, Vincent A. Voelz, John D. Chodera, and Gregory R. Bowman. SARS-CoV-2 simulations go exascale to predict dramatic spike opening and cryptic pockets across the proteome. *Nature Chem.*, 13:651–659, 7 2021.

[25] Erik H. Thiede, Dimitrios Giannakis, Aaron R. Dinner, and Jonathan Weare. Galerkin Approximation of Dynamical Quantities Using Trajectory Data. *J. Chem. Phys.*, 150(24):244111, 2019.

[26] John Strahan, Adam Antoszewski, Chatipat Lorpaiboon, Bodhi P. Vani, Jonathan Weare, and Aaron R. Dinner. Long-Time-Scale Predictions from Short-Trajectory Data: A Benchmark Analysis of the Trp-Cage Miniprotein. *J. Chem. Theory Comput.*, 17(5):2948–2963, 2021.

[27] Daniel E. Koshland. The Key–Lock Theory and the Induced Fit Theory. *Angew. Chem., Int. Ed. Engl.*, 33(23-24):2375–2378, 1995.

[28] Yaakov Levy, Samuel S. Cho, José N. Onuchic, and Peter G. Wolynes. A Survey of Flexible Protein Binding Mechanisms and their Transition States Using Native Topology Based Energy Landscapes. *J. Mol. Biol.*, 346(4):1121–1145, 2005.

[29] G. Schreiber, G. Haran, and H.-X. Zhou. Fundamental Aspects of Protein-Protein Association Kinetics. *Chem. Rev.*, 109(3):839–860, 2009.

[30] Peter Csermely, Robin Palotai, and Ruth Nussinov. Induced Fit, Conformational Selection and Independent Dynamic Segments: an Extended View of Binding Events. *Trends Biochem. Sci.*, 35(10):539–546, 2010.

[31] Alexey G. Murzin. Metamorphic Proteins. *Science*, 320:1725–1726, 6 2008.

[32] Andy LiWang, Lauren L. Porter, and Lee-Ping Wang. Fold-switching proteins. *Biopolymers*, 112, 10 2021.

[33] José Nelson Onuchic, Zaida Luthey-Schulten, and Peter G. Wolynes. Theory Of Protein Folding: The Energy Landscape Perspective. *Annu. Rev. Phys. Chem.*, 48(1):545–600, 1997. PMID: 9348663.

[34] A Sali, E Shaknovich, and M Karplus. How Does a Protein Fold. *Nature*, 369(6477):248–251, MAY 19 1994.

[35] CM Dobson, A Sali, and M Karplus. Protein folding: A perspective from theory and experiment. *Angew. Chem. Int. Edit.*, 37(7):868–893, APR 20 1998.

[36] Aaron R Dinner, Andrej Šali, Lorna J Smith, Christopher M Dobson, and Martin Karplus. Understanding protein folding via free-energy surfaces from theory and experiment. *Trends Biochem. Sci.*, 25:331–339, 7 2000.

[37] Kresten Lindorff-Larsen, Stefano Piana, Ron O. Dror, and David E. Shaw. How fast-folding proteins fold. *Science*, 334(6055):517–520, OCT 28 2011.

[38] Ken A. Dill and Justin L. MacCallum. The Protein-Folding Problem, 50 Years On. *Science*, 338(6110):1042–1046, NOV 23 2012.

[39] José Nelson Onuchic and Peter G Wolynes. Theory of protein folding. *Curr. Opin. Struc. Biol.*, 14(1):70–75, 2004.

[40] Peter G. Wolynes, William A. Eaton, and Alan R. Fersht. Chemical physics of protein folding. *P. Natl. Acad. Sci.*, 109:17770–17771, 10 2012.

[41] Janghyun Yoo, John M. Louis, and Hoi Sung Chung. Diverse Folding Pathways of HIV-1 Protease Monomer on a Rugged Energy Landscape. *Biophys. J.*, 117(8):1456–1466, OCT 15 2019.

[42] Martin Karplus and David L. Weaver. Diffusion-collision model for protein folding. *Biopolymers*, 18:1421–1437, 6 1979.

[43] Martin Karplus and David L. Weaver. Protein folding dynamics: The diffusion-collision model and experimental data. *Protein Sci.*, 3:650–668, 4 1994.

[44] Donald B. Wetlaufer. Nucleation, Rapid Folding, and Globular Intrachain Regions in Proteins. *P. Natl. Acad. Sci. USA*, 70(3):697–701, 1973.

[45] RR Matheson Jr and HA Scheraga. A method for predicting nucleation sites for protein folding based on hydrophobic contacts. *Macromolecules*, 11(4):819–829, 1978.

[46] Zhuyan Guo and D Thirumalai. The nucleation-collapse mechanism in protein folding: evidence for the non-uniqueness of the folding nucleus. *Fold. Des.*, 2:377–391, 12 1997.

[47] Wenbing Hu, Zhong-Yuan Kan, Leland Mayne, and S Walter Englander. Cytochrome c folds through foldon-dependent native-like intermediates in an ordered pathway. *P. Natl. Acad. Sci.*, 113(14):3809–3814, 2016.

[48] S. Walter Englander and Leland Mayne. The case for defined protein folding pathways. *P. Natl. Acad. Sci.*, 114:8253–8258, 8 2017.

[49] Geoffrey C. Rollins and Ken A. Dill. General Mechanism of Two-State Protein Folding Kinetics. *J. Am. Chem. Soc.*, 136:11420–11427, 8 2014.

[50] Eric Johnson. A Maximum Caliber analysis of the Foldon Hypothesis. *Proteins: Struct. Funct. Genet.*, 90:1170–1178, 5 2022.

[51] M R DeFelippis, R E Chance, and B H Frank. Insulin Self-Association and the Relationship to Pharmacokinetics and Pharmacodynamics. *Crit. Rev. Ther. Drug Carrier Syst.*, 18(2):201–264, 2001.

[52] Michael A. Weiss. In Gerald Litwack, editor, *Insulin and IGFs*, volume 80 of *Vitamins and Hormones*, chapter 2, The Structure and Function of Insulin: Decoding the TR Transition, pages 33–49. Academic Press, London, 2009.

[53] John G. Menting, Jonathan Whittaker, Mai B. Margetts, Linda J. Whittaker, Geoffrey K.W. Kong, Brian J. Smith, Christopher J. Watson, Lenka Žáková, Emília Kletvíková, Jiří Jiráček, Shu Jin Chan, Donald F. Steiner, Guy G. Dodson, Andrzej M. Brzozowski, Michael A. Weiss, Colin W. Ward, and Michael C. Lawrence. How Insulin Engages its Primary Binding Site on the Insulin Receptor. *Nature*, 493(7431):241–245, 2013.

[54] Vincent Zoete, Markus Meuwly, and Martin Karplus. A Comparison of the Dynamic Behavior of Monomeric and Dimeric Insulin Shows Structural Rearrangements in the Active Monomer. *J. Mol. Biol.*, 342(3):913–929, 2004.

[55] Anne Marie M. Jørgensen, Søren M. Kristensen, Jens J. Led, and Per Balschmidt. Three-Dimensional Solution Structure of an Insulin Dimer: A Study of the B9(Asp) Mutant of Human Insulin using Nuclear Magnetic Resonance, Distance Geometry and Restrained Molecular Dynamics. *J. Mol. Biol.*, 227(4):1146–1163, 1992.

[56] Danielle Keller, Rikke Clausen, Knud Josefsen, and Jens J. Led. Flexibility and Bioactivity of Insulin: an NMR Investigation of the Solution Structure and Folding of an Unusually Flexible Human Insulin Mutant with Increased Biological Activity. *Biochemistry*, 40(35):10732–10740, 2001.

[57] Ziad Ganim, Kevin C. Jones, and Andrei Tokmakoff. Insulin Dimer Dissociation and Unfolding Revealed by Amide I Two-Dimensional Infrared Spectroscopy. *Phys. Chem. Chem. Phys.*, 12(14):3579–3588, 2010.

[58] Xin Xing Zhang, Kevin C. Jones, Ann Fitzpatrick, Chunte Sam Peng, Chi Jui Feng, Carlos R. Baiz, and Andrei Tokmakoff. Studying Protein-Protein Binding through T-Jump Induced Dissociation: Transient 2D IR Spectroscopy of Insulin Dimer. *J. Phys. Chem. B*, 120(23):5134–5145, 2016.

[59] Luis Busto-Moner, Chi jui Feng, Adam Antoszewski, Andrei Tokmakoff, and Aaron R Dinner. Structural Ensemble of the Insulin Monomer. *Biochemistry*, 60:3125–3136, 10 2021.

[60] Harald Berchtold and Rolf Hilgenfeld. Binding of Phenol to R6 Insulin Hexamers. *Biopolymers*, 51(2):165–172, 1999.

[61] Jens Brange, Lennart Andersen, Erik D. Laursen, Giorgio Meyn, and Eigil Rasmussen. Toward Understanding Insulin Fibrillation. *J. Pharm. Sci.*, 86(5):517–525, 1997.

[62] Qing Xin Hua and Michael A. Weiss. Mechanism of Insulin Fibrillation: The Structure of Insulin Under Amyloidogenic Conditions Resembles a Protein-Folding Intermediate. *J. Biol. Chem.*, 279(20):21449–21460, 2004.

[63] M. I. Ivanova, S. A. Sievers, M. R. Sawaya, J. S. Wall, and D. Eisenberg. Molecular Basis for Insulin Fibril Assembly. *Proc. Natl. Acad. Sci.*, 106(45):18990–18995, 2009.

[64] Alexander N. Zaykov, John P. Mayer, and Richard D. DiMarchi. Pursuit of a Perfect Insulin. *Nat. Rev. Drug Discov.*, 15(6):425–439, 2016.

[65] R.D. DiMarchi, R.E. Chance, H.B. Long, J.E. Shields, and L.J. Slieker. Preparation of an Insulin with Improved Pharmacokinetics Relative to Human Insulin through Consideration of Structural Homology with Insulin-Like Growth Factor I. *Horm. Res.*, 41(2):93–96, 1994.

[66] Stephen M. Setter, Cynthia F. Corbett, R. Keith Campbell, and John R. White. Insulin Aspart: A New Rapid-Acting Insulin Analog. *Ann. Pharmacother.*, 34(12):1423–1431, 2000.

[67] J. Rosenstock, S. L. Schwartz, C. M. Clark, G. D. Park, D. W. Donley, and M. B. Edwards. Basal Insulin Therapy in Type 2 Diabetes: 28-Week Comparison of Insulin Glargine (HOE 901) and NPH Insulin. *Diabetes Care*, 24(4):631–636, 2001.

[68] Svend Havelund, Anne Plum, Ulla Ribel, Ib Jonassen, Aage Vølund, Jan Markussen, and Peter Kurtzhals. The Mechanism of Protraction of Insulin Detemir, a Long-Acting, Acylated Analog of Human Insulin. *Pharm. Res.*, 21(8):1498–1504, 2004.

[69] Kjeld Hermansen, Melanie Davies, Taudeusz Derezinski, G. Martinez Ravn, Per Clauson, and Philip Home. A 26-Week, Randomized, Parallel, Treat-to-Target Trial Comparing Insulin Detemir With NPH Insulin as Add-On Therapy to Oral Glucose-Lowering Drugs in Insulin-Naive People With Type 2 Diabetes. *Diabetes Care*, 29(6):1269–1274, 2006.

[70] Kitty Poon and Allen B King. Glargine and Detemir: Safety and Efficacy Profiles of the Long-Acting Basal Insulin Analogs. *Drug Healthc. Patient Saf.*, 2(1):213–223, 2010.

[71] Jeremy Pettus, Tricia Santos Cavaiola, William V. Tamborlane, and Steven Edelman. The Past, Present, and Future of Basal Insulins. *Diabetes Metab. Res.*, 32(6):478–496, 2016.

[72] Alice Cheng, Timothy S. Bailey, Didac Mauricio, and Ronan Roussel. Insulin Glargine 300 U/mL and Insulin Degludec: A Review of the Current Evidence Comparing These Two Second-Generation Basal Insulin Analogues. *Diabetes Metab. Res.*, pages 1–10, 2020.

[73] Harish Vashisth and Cameron F Abrams. Ligand Escape Pathways and (Un)Binding Free Energy Calculations for the Hexameric Insulin-Phenol Complex. *Biophys. J.*, 95(9):4193–4204, 2008.

[74] Rekha Pattanayek, Dewight R Williams, Sabuj Pattanayek, Tetsuya Mori, Carl H Johnson, Phoebe L Stewart, and Martin Egli. Structural model of the circadian clock KaiB–KaiC complex and mechanism for modulation of KaiC phosphorylation. *EMBO J.*, 27:1767–1778, 6 2008.

[75] Roger Tseng, Nicolette F. Goularte, Archana Chavan, Jansen Luu, Susan E. Cohen, Yong-Gang Chang, Joel Heisler, Sheng Li, Alicia K. Michael, Sarvind Tripathi, Susan S. Golden, Andy LiWang, and Carrie L. Partch. Structural basis of the day-night transition in a bacterial circadian clock. *Science*, 355:1174–1180, 3 2017.

[76] Adam Antoszewski, Chi-Jui Feng, Bodhi P Vani, Erik H Thiede, Lu Hong, Jonathan Weare, Andrei Tokmakoff, and Aaron R Dinner. Insulin Dissociates by Diverse Mechanisms of Coupled Unfolding and Unbinding. *J. Phys. Chem. B*, 124:5571–5587, 2020.

[77] Annmarie Surprenant and R. Alan North. Signaling at Purinergic P2X Receptors. *Annu. Rev. Physiol.*, 71(1):333–359, 2009.

[78] Robert C. Baxter. How IGF-1 Activates its Receptor. *J. Cell Commun. Signaling*, 9(1):87, 2015.

[79] Pierre De Meyts. Insulin/Receptor Binding: The Last Piece of the Puzzle? What Recent Progress on the Structure of the Insulin/Receptor Complex Tells Us (Or Not) About Negative Cooperativity and Activation. *BioEssays*, 37(4):389–397, 2015.

[80] Timothy R. Dafforn and Corinne J I Smith. Natively Unfolded Domains in Endocytosis: Hooks, Lines and Linkers. *EMBO Rep.*, 5(11):1046–1052, 2004.

[81] Harish Vashisth and Cameron F Abrams. Docking of Insulin to a Structurally Equilibrated Insulin Receptor Ectodomain. *Proteins: Struct., Funct., Bioinf.*, 78:1531–1543, 2010.

[82] Harish Vashisth and Cameron F Abrams. All-atom Structural Models of Insulin Binding to the Insulin Receptor in the Presence of a Tandem Hormone-Binding Element. *Proteins: Struct., Funct., Bioinf.*, 81(6):1017–1030, 2013.

[83] Michael A. Weiss and Michael C. Lawrence. A Thing of Beauty: Structure and Function of Insulin's "Aromatic Triplet". *Diabetes, Obes. Metab.*, 20(April):51–63, 2018.

[84] H. Chen, M. Shi, Z. Y. Guo, Y. H. Tang, Z. S. Qiao, Z. H. Liang, and Y. M. Feng. Four New Monomeric Insulins Obtained by Alanine Scanning the Dimer-Forming Surface of the Insulin Molecule. *Protein Eng.*, 13(11):779–782, 2000.

[85] Vincent Zoete, Markus Meuwly, and Martin Karplus. Study of the Insulin Dimerization: Binding Free Energy Calculations and per-Residue Free Energy Decomposition. *Proteins: Struct., Funct., Bioinf.*, 61(1):79–93, 2005.

[86] Qiankun Gong, Haomiao Zhang, Haozhe Zhang, and Changjun Chen. Calculating the Absolute Binding Free Energy of the Insulin Dimer in an Explicit Solvent. *RSC Adv.*, 10(2):790–800, 2020.

[87] Wojciech Bocian, Jerzy Sitkowski, Elżbieta Bednarek, Anna Tarnowska, Robert Kawęcki, and Lech Kozerski. Structure of Human Insulin Monomer in Water/Acetonitrile Solution. *J. Biomol. NMR*, 40(1):55–64, 2008.

[88] F.S. Legge, A. Budi, H. Treutlein, and I. Yarovsky. Protein Flexibility: Multiple Molecular Dynamics Simulations of Insulin Chain B. *Biophys. Chem.*, 119(2):146–157, 2006.

[89] Richa Singh, Rohit Bansal, Anurag Singh Rathore, and Gaurav Goel. Equilibrium Ensembles for Insulin Folding from Bias-Exchange Metadynamics. *Biophys. J.*, 112(8):1571–1585, 2017.

[90] J. G. Menting, Y. Yang, S. J. Chan, N. B. Phillips, B. J. Smith, J. Whittaker, N. P. Wickramasinghe, L. J. Whittaker, V. Pandyarajan, Z.-l. Wan, S. P. Yadav, J. M. Carroll, N. Strokes, C. T. Roberts, F. Ismail-Beigi, W. Milewski, D. F. Steiner, V. S. Chauhan, C. W. Ward, M. A. Weiss, and M. C. Lawrence. Protective Hinge in Insulin Opens to Enable its Receptor Engagement. *Proc. Natl. Acad. Sci.*, 111(33):E3395–E3404, 2014.

[91] Dolev Rimmerman, Denis Leshchev, Darren J. Hsu, Jiyun Hong, Irina Kosheleva, and Lin X. Chen. Direct Observation of Insulin Association Dynamics with Time-Resolved X-ray Scattering. *J. Phys. Chem. Lett.*, 8(18):4413–4418, 2017.

[92] Puja Banerjee and Biman Bagchi. Dynamical Control by Water at a Molecular Level in Protein Dimer Association and Dissociation. *Proc. Natl. Acad. Sci.*, 117(5):2302–2308, 2020.

[93] Erik H. Thiede, Brian Van Koten, Jonathan Weare, and Aaron R. Dinner. Eigenvector Method for Umbrella Sampling Enables Error Analysis. *J. Chem. Phys.*, 145(8):084115, 2016.

[94] Aaron R Dinner, Erik H Thiede, Brian Van Koten, and Jonathan Weare. Stratification as a General Variance Reduction Method for Markov Chain Monte Carlo. *SIAM-ASA J. Uncertain.*, 8(3):1139–1188, 2020.

[95] A. D. MacKerell, 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. K. 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. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B*, 102(18):3586–3616, 1998.

[96] Robert B. Best, Xiao Zhu, Jihyun Shim, Pedro E. M. Lopes, Jeetain Mittal, Michael Feig, and Alexander D. MacKerell. Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone ϕ, ψ and Side-Chain χ_1 and χ_2 Dihedral Angles. *J. Chem. Theory Comput.*, 8(9):3257–3273, 2012.

[97] Jing Huang, Sarah Rauscher, Grzegorz Nawrocki, Ting Ran, Michael Feig, Bert L. De Groot, Helmut Grubmüller, and Alexander D. MacKerell. CHARMM36m: An Improved Force Field for Folded and Intrinsically Disordered Proteins. *Nat. Methods*, 14(1):71–73, 2016.

[98] Mark James Abraham, Teemu Murtola, Roland Schulz, Szilárd Páll, Jeremy C. Smith, Berk Hess, and Erik Lindah. GROMACS: High Performance Molecular Simulations Through Multi-Level Parallelism from Laptops to Supercomputers. *SoftwareX*, 1-2:19–25, 2015.

[99] Sunhwan Jo, Taehoon Kim, Vidyashankara G. Iyer, and Wonpil Im. CHARMM-GUI: A Web-Based Graphical User Interface for CHARMM. *J. Comput. Chem.*, 29(11):1859–1865, 2008.

[100] Jumin Lee, Xi Cheng, Jason M. Swails, Min Sun Yeom, Peter K. Eastman, Justin A. Lemkul, Shuai Wei, Joshua Buckner, Jong Cheol Jeong, Yifei Qi, Sunhwan Jo, Vijay S. Pande, David A. Case, Charles L. Brooks, Alexander D. MacKerell, Jeffery B. Klauda, and Wonpil Im. 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, 2016.

[101] N. Goga, A. J. Rzepiela, A. H. De Vries, S. J. Marrink, and H. J.C. Berendsen. Efficient Algorithms for Langevin and DPD Dynamics. *J. Chem. Theory Comput.*, 8(10):3637–3649, 2012.

[102] Berk Hess, Henk Bekker, Herman J.C. Berendsen, and Johannes G.E.M. Fraaije. LINCS: A Linear Constraint Solver for Molecular Simulations. *J. Comput. Chem.*, 18(12):1463–1472, 1997.

[103] Tom Darden, Darrin York, and Lee Pedersen. Particle Mesh Ewald: An N·log(N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.*, 98(12):10089–10092, 1993.

[104] William Humphrey, Andrew Dalke, and Klaus Schulten. VMD: Visual Molecular Dynamics. *J. Mol. Graphics*, 7855(October 1995):33–38, 1996.

[105] James C. Phillips, Rosemary Braun, Wei Wang, James Gumbart, Emad Tajkhorshid, Elizabeth Villa, Christophe Chipot, Robert D. Skeel, Laxmikant Kalé, and Klaus Schulten. Scalable Molecular Dynamics with NAMD. *J. Comput. Chem.*, 26(16):1781–1802, 2005.

[106] N. Sakabe, K. Sakabe, K. Sasaki, and M. Murayoshi. 0.92A structure of 2Zn human insulin at 100K. www.rcsb.org/structure/3w7y, 2013.

[107] William L Jorgensen, Jayaraman Chandrasekhar, Jeffry D Madura, Roger W Impey, and Michael L Klein. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.*, 79(2):926–935, 1983.

[108] Dmitrii Beglov and Benoît Roux. Finite Representation of an Infinite Bulk System: Solvent Boundary Potential for Computer Simulations. *J. Chem. Phys.*, 100(12):9050–9063, 1994.

[109] M. Parrinello and A. Rahman. Polymorphic Transitions in Single Crystals: A New Molecular Dynamics Method. *J. Appl. Phys.*, 52(12):7182–7190, 1981.

[110] Barry Isralewitz, Mu Gao, and Klaus Schulten. Steered Molecular Dynamics and Mechanical Functions of Proteins. *Curr. Opin. Struct. Biol.*, 11(2):224–230, 2001.

[111] Luca Maragliano, Alexander Fischer, Eric Vanden-Eijnden, and Giovanni Ciccotti. String Method in Collective Variables: Minimum Free Energy Paths and Isocommittor Surfaces. *J. Chem. Phys.*, 125(2):024106, 2006.

[112] Eric Vanden-Eijnden and Maddalena Venturoli. Revisiting the Finite Temperature String Method for the Calculation of Reaction Tubes and Free Energies. *J. Chem. Phys.*, 130(19):194103, 2009.

[113] MichaelA Weiss. Design of Ultra-Stable Insulin Analogues for the Developing World. *J. Health Spec.*, 1(2):59–70, 2013.

[114] Brian Van Koten and Mitchell Luskin. Stability and Convergence of the String Method for Computing Minimum Energy Paths. *Multiscale Model. Simul.*, 17(2):873–898, 2019.

[115] Ronald R. Coifman and Stéphane Lafon. Diffusion Maps. *Appl. Comput. Harmonic Anal.*, 21(1):5–30, 2006.

[116] Tyrus Berry and John Harlim. Variable Bandwidth Diffusion Kernels. *Appl. Comput. Harmonic Anal.*, 40(1):68–96, 2016.

[117] Massimo Marchi and Pietro Ballone. Adiabatic Bias Molecular Dynamics: A Method to Navigate the Conformational Space of Complex Molecular Systems. *J. Chem. Phys.*, 110(8):3697–3702, 1999.

[118] Massimiliano Bonomi, Davide Branduardi, Giovanni Bussi, Carlo Camilloni, Davide Provasi, Paolo Raiteri, Davide Donadio, Fabrizio Marinelli, Fabio Pietrucci, Ricardo A. Broglia, and Michele Parrinello. PLUMED: A Portable Plugin for Free-Energy Calculations with Molecular Dynamics. *Comput. Phys. Commun.*, 180(10):1961–1972, 2009.

[119] Gareth A. Tribello, Massimiliano Bonomi, Davide Branduardi, Carlo Camilloni, and Giovanni Bussi. PLUMED 2: New Feathers for an Old Bird. *Comput. Phys. Commun.*, 185(2):604–613, 2014.

[120] Massimiliano Bonomi, Giovanni Bussi, Carlo Camilloni, Gareth A. Tribello, Pavel Banáš, Alessandro Barducci, Mattia Bernetti, Peter G. Bolhuis, Sandro Bottaro, Davide Branduardi, Riccardo Capelli, and Paolo Carloni. Promoting Transparency and Reproducibility in Enhanced Molecular Simulations. *Nat. Methods*, 16(8):670–673, 2019.

[121] Soohyung Park, Taehoon Kim, and Wonpil Im. Transmembrane Helix Assembly by Window Exchange Umbrella Sampling. *Phys. Rev. Lett.*, 108:108102, 2012.

[122] Soohyung Park and Wonpil Im. Two Dimensional Window Exchange Umbrella Sampling for Transmembrane Helix Assembly. *J. Chem. Theory Comput.*, 9(1):13–17, 2013.

[123] Yuji Sugita, Akio Kitao, and Yuko Okamoto. Multidimensional Replica-Exchange Method for Free-Energy Calculations. *J. Chem. Phys.*, 113(15):6042–6051, 2000.

[124] Mike Reppert and Andrei Tokmakoff. Computational Amide I 2D IR Spectroscopy as a Probe of Protein Structure and Dynamics. *Annu. Rev. Phys. Chem.*, 67(1):359–386, 2016.

[125] C. J. Feng, B. Dhayalan, and A. Tokmakoff. Refinement of peptide conformational ensembles by 2d ir spectroscopy: Application to ala–ala–ala. *Biophys. J.*, 114(12):2820–2832, 2018.

[126] Balamurugan Dhayalan, Ann Fitzpatrick, Kalyaneswar Mandal, Jonathan Whittaker, Michael A. Weiss, Andrei Tokmakoff, and Stephen B. H. Kent. Efficient Total Chemical Synthesis of 13 C= 18 O Isotopomers of Human Insulin for Isotope-Edited FTIR. *ChemBioChem*, 17(5):415–420, 2016.

[127] M. Reppert and A. Tokmakoff. Communication: Quantitative Multi-Site Frequency Maps for Amide I Vibrational Spectroscopy. *J. Chem. Phys.*, 143(6):061102, 2015.

[128] T. la Cour Jansen, A. G. Dijkstra, T. M. Watson, J. D. Hirst, and J. Knoester. Modeling the Amide I Bands of Small Peptides. *J. Chem. Phys.*, 125(4):44312, 2006.

[129] T. la Cour Jansen and J. Knoester. A Transferable Electrostatic Map for Solvation Effects on Amide I Vibrations and its Application to Linear and Two-Dimensional Spectroscopy. *J. Chem. Phys.*, 124(4):044502, 2006.

[130] Peter Hamm, Manho Lim, and Robin M. Hochstrasser. Structure of the Amide I Band of Peptides Measured by Femtosecond Nonlinear-Infrared Spectroscopy. *J. Phys. Chem. B*, 102(31):6123–6138, 1998.

[131] H. Torii. Effects of Intermolecular Vibrational Coupling and Liquid Dynamics on the Polarized Raman and Two-Dimensional Infrared Spectral Profiles of Liquid N,N-Dimethylformamide Analyzed with a Time-Domain Computational Method. *J. Phys. Chem. A*, 110(14):4822–4832, 2006.

[132] C. Liang and T. L. Jansen. An Efficient N(3)-Scaling Propagation Scheme for Simulating Two-Dimensional Infrared and Visible Spectra. *J. Chem. Theory. Comput.*, 8(5):1706–1713, 2012.

[133] C. J. Feng and A. Tokmakoff. The Dynamics of Peptide-Water Interactions in Dialanine: An Ultrafast Amide I 2D IR and Computational Spectroscopy Study. *J. Chem. Phys.*, 147(8):085101, 2017.

[134] Shampa Raghunathan, Krystel El Hage, Jasmine L. Desmond, Lixian Zhang, and Markus Meuwly. The Role of Water in the Stability of Wild-type and Mutant Insulin Dimers. *J. Phys. Chem. B*, 122(28):7038–7048, 2018.

[135] FHC Crick. The Packing of α-Helices: Simple Coiled-Coils. *Acta Crystallogr.*, 6(8-9):689–697, 1953.

[136] Timothy J Richmond and Frederic M Richards. Packing of α-Helices: Geometrical Constraints and Contact Areas. *J. Mol. Biol.*, 119(4):537–555, 1978.

[137] Cyrus Chothia, Michael Levitt, and Douglas Richardson. Helix to Helix Packing in Proteins. *J. Mol. Biol.*, 145(1):215–250, 1981.

[138] Dmitrij Frishman and Patrick Argos. Knowledge-Based Protein Secondary Structure Assignment. *Proteins: Struct., Funct., Genet.*, 23(4):566–579, 1995.

[139] Giovanna Scapin, Venkata P. Dandey, Zhening Zhang, Winifred Prosise, Alan Hruza, Theresa Kelly, Todd Mayhood, Corey Strickland, Clinton S. Potter, and Bridget Carragher. Structure of the Insulin Receptor-Insulin Complex by Single-Particle Cryo-EM Analysis. *Nature*, 556(7699):122–125, 2018.

[140] Michael D. Glidden, Yanwu Yang, Nicholas A. Smith, Nelson B. Phillips, Kelley Carr, Nalinda P. Wickramasinghe, Faramarz Ismail-Beigi, Michael C. Lawrence, Brian J. Smith, and Michael A. Weiss. Solution Structure of an Ultra-Stable Single-Chain Insulin Analog Connects Protein Dynamics to a Novel Mechanism of Receptor Binding. *J. Biol. Chem.*, 293(1):69–88, 2018.

[141] Carlos R. Baiz, Yu-Shan Lin, Chunte Sam Peng, Kyle A. Beauchamp, Vincent A. Voelz, Vijay S. Pande, and Andrei Tokmakoff. A Molecular Interpretation of 2D IR Protein Folding Experiments with Markov State Models. *Biophys. J.*, 106(6):1359–1370, 2014.

[142] Jasmine L. Desmond, Debasish Koner, and Markus Meuwly. Probing the Differential Dynamics of the Monomeric and Dimeric Insulin from Amide-I IR Spectroscopy. *J. Phys. Chem. B*, 123(30):6588–6598, 2019.

[143] Claus Kristensen, Thomas Kjeldsen, Finn C. Wiberg, Lauge Schäffer, Morten Hach, Svend Havelund, Joseph Bass, Donald F. Steiner, and Asser S. Andersen. Alanine Scanning Mutagenesis of Insulin. *J. Biol. Chem.*, 272(20):12978–12983, 1997.

[144] Karina Sinding Thorsøe, Morten Schlein, Dorte Bjerre Steensgaard, Jakob Brandt, Gerd Schluckebier, and Helle Naver. Kinetic Evidence for the Sequential Association of Insulin Binding Sites 1 and 2 to the Insulin Receptor and the Influence of Receptor Isoform,. *Biochemistry*, 49(29):6234–6246, 2010.

[145] Felix Weis, John G. Menting, Mai B. Margetts, Shu Jin Chan, Yibin Xu, Norbert Tennagels, Paulus Wohlfart, Thomas Langer, Christoph W. Müller, Matthias K. Dreyer, and Michael C. Lawrence. The Signalling Conformation of the Insulin Receptor Ectodomain. *Nat. Commun.*, 9(1):4420, 2018.

[146] Ewa Ciszak, John M. Beals, Bruce H. Frank, Jeffrey C. Baker, Nancy D. Carter, and G.David Smith. Role of C-terminal B-chain Residues in Insulin Assembly: The Structure of Hexameric LysB28ProB29-Human Insulin. *Structure*, 3(6):615–622, 1995.

[147] Adam Antoszewski, Chatipat Lorpaiboon, John Strahan, and Aaron R. Dinner. Kinetics of Phenol Escape from the Insulin R6 Hexamer. *J. Phys. Chem. B*, 125(42):11637–11649, 2021.

[148] U. Derewenda, Z. Derewenda, E. J. Dodson, G. G. Dodson, C. D. Reynolds, G. D. Smith, C. Sparks, and D. Swenson. Phenol Stabilizes More Helix in a New Symmetrical Zinc Insulin Hexamer. *Nature*, 338(6216):594–596, 1989.

[149] Wonjae E. Choi, Mark L. Brader, Valentin Aguilar, Niels C. Kaarsholm, and Michael F. Dunn. Allosteric Transition of the Insulin Hexamer is Modulated by Homotropic and Heterotropic Interactions. *Biochemistry*, 32(43):11638–11645, 1993.

[150] Edgar Jacoby, Peter Kruger, Yaşar Karatas, and Axel Wollmer. Distinction of Strucutral Reorganization and Ligand Binding in the T-R Transition of Insulin on the Basis of Allosteric Models. *Biol. Chem. H-S.*, 374(September):877–885, 1993.

[151] Edward N Baker, Thomas Leon Blundell, John F Cutfield, Eleanor Joy Dodson, George Guy Dodson, Dorothy Mary Crowfoot Hodgkin, Roderick E Hubbard, Neil W Isaacs, Colin D Reynolds, Kiwako Sakabe, Norioshi Sakabe, and Numminate M Vijayan. The Structure of 2Zn Pig Insulin Crystals at 1.5 Å Resolution. *Philos. T. Roy. Soc. B*, 319(1195):369–456, 1988.

[152] Ulrich Hassiepen, Matthias Federwisch, Thomas Mülders, and Axel Wollmer. The Lifetime of Insulin Hexamers. *Biophys. J.*, 77(3):1638–1654, 1999.

[153] J.P. Turkenburg, J.L. Whittingham, U. Derewenda, Z.S. Derewenda, E.J. Dodson, G.G. Dodson, G.D. Smith, and B. Xiao. Structure Determination and Refinement of Two Crystal Forms of Native Insulins. www.rcsb.org/structure/1znj, 1998. (accessed 4/20/2020).

[154] M. Roy, M. L. Brader, R. W.K. Lee, N. C. Kaarsholm, J. F. Hansen, and M. F. Dunn. Spectroscopic Signatures of the T to R Conformational Transition in the Insulin Hexamer. *J. Biol. Chem.*, 264(32):19081–19085, 1989.

[155] Edgar Jacoby, Qing Xin Hua, Alan S. Stern, Bruce H. Frank, and Michael A. Weiss. Structure and Dynamics of a Protein Assembly. ^{1}H-NMR Studies of the 36 kDa R_6 Insulin Hexamer. *J. Mol. Biol.*, 258(1):136–157, 1996.

[156] G.David Smith. The Phenolic Binding Site in T3Rf3 Insulin. *J. Mol. Struct.*, 470(1-2):71–80, 1998.

[157] Wolfgang Swegat, Jürgen Schlitter, Peter Krüger, and Axel Wollmer. MD Simulation of Protein-Ligand interaction: Formation and Dissociation of an Insulin-Phenol Complex. *Biophys. J.*, 84(3):1493–1506, 2003.

[158] Harish Vashisth and Cameron F Abrams. Ligand Escape Pathways and (Un)Binding Free Energy Calculations for the Hexameric Insulin-Phenol Complex. *Biophys. J.*, 95(9):4193–4204, 2008.

[159] Jie Hu, Ao Ma, and Aaron R Dinner. Bias Annealing: A Method for Obtaining Transition Paths de Novo. *J. Chem. Phys.*, 125(11):114101, 2006.

[160] Ao Ma, Ambarish Nag, and Aaron R Dinner. Dynamic Coupling Between Coordinates in a Model for Biomolecular Isomerization. *J. Chem. Phys.*, 124(14):144911, 2006.

[161] Jie Hu, Ao Ma, and Aaron R Dinner. A Two-Step Nucleotide-Flipping Mechanism Enables Kinetic Discrimination of DNA Lesions by AGT. *Proc. Natl. Acad. Sci.*, 105(12):4615–4620, 2008.

[162] B. R. Brooks, C. L. Brooks, A. D. Mackerell, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and M. Karplus. CHARMM: The Biomolecular Simulation Program. *J. Comput. Chem.*, 30(10):1545–1614, 2009.

[163] K. Vanommeslaeghe, E. Hatcher, C. Acharya, S. Kundu, S. Zhong, J. Shim, E. Darian, O. Guvench, P. Lopes, I. Vorobyov, and A. D. Mackerell. CHARMM General Force Field: A Force Field for Drug-Like Molecules Compatible With the CHARMM All-Atom Additive Biological Force Fields. *J. Comput. Chem.*, 32:671–690, 2009.

[164] Mats H. M. Olsson, Chresten R. Søndergaard, Michal Rostkowski, and Jan H. Jensen. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical p K a Predictions. *J. Chem. Theory Comput.*, 7(2):525–537, 2011.

[165] Chresten R. Søndergaard, Mats H. M. Olsson, Michał Rostkowski, and Jan H. Jensen. Improved Treatment of Ligands and Coupling Effects in Empirical Calculation and Rationalization of p K a Values. *J. Chem. Theory Comput.*, 7(7):2284–2295, 2011.

[166] Frank Noé, Christof Schütte, Eric Vanden-Eijnden, Lothar Reich, and Thomas R Weikl. Constructing the Equilibrium Ensemble of Folding Pathways from Short Off-Equilibrium Simulations. *Proc. Natl. Acad. Sci.*, 106(45):19011–19016, 2009.

[167] Fabian Paul, Christoph Wehmeyer, Esam T Abualrous, Hao Wu, Michael D Crabtree, Johannes Schöneberg, Jane Clarke, Christian Freund, Thomas R Weikl, and Frank Noé. Protein-Peptide Association Kinetics Beyond the Seconds Timescale from Atomistic Simulations. *Nat. Commun.*, 8(1095):1–9, 2017.

[168] Duane T. Birnbaum, Steven W. Dodd, Bo E. H. Saxberg, Alexander D. Varshavsky, and John M. Beals. Hierarchical Modeling of Phenolic Ligand Binding to 2Zn-Insulin Hexamers. *Biochemistry*, 35(17):5366–5378, 1996.

[169] Duane T. Birnbaum, M. A. Kilcomons, M. R. DeFelippis, and John M. Beals. Assembly and Dissociation of Human Insulin and LysB28ProB29-Insulin Hexamers: A Comparison Study. *Pharm. Res.*, 14:25–36, 1997.

[170] Chi-Jui Feng, Anton Sinitskiy, Vijay Pande, and Andrei Tokmakoff. Computational IR Spectroscopy of Insulin Dimer Structure and Conformational Heterogeneity. *J. Phys. Chem. B*, 125(18):4620–4633, 2021.

[171] G.A. Bentley, J. Brange, Z. Derewenda, E.J. Dodson, G.G. Dodson, J Markussen, A.J. Wilkinson, A Wollmer, and B. Xiao. Role of B13 Glu in insulin assembly. *J. Mol. Biol.*, 228(4):1163–1176, 1992.

[172] Curtis R Bloom, Wonjae E Choi, Peter S Brzovic, Julie J. Ha Sheng-Tung Huang, Niels C Kaarsholm, and Michael F Dunn. Ligand Binding to Wild-type and E-B13Q Mutant Insulins: A Three-state Allosteric Model System Showing Half-site Reactivity. *J. Mol. Biol.*, 245(4):324–330, 1995.

[173] Saumyak Mukherjee, Sayantan Mondal, Ashish Anilrao Deshmukh, Balasubramanian Gopal, and Biman Bagchi. What Gives an Insulin Hexamer Its Unique Shape and Stability? Role of Ten Confined Water Molecules. *J. Phys. Chem. B*, 122(5):1631–1637, 2018.

[174] Saumyak Mukherjee, Ashish A. Deshmukh, Sayantan Mondal, Balasubramanian Gopal, and Biman Bagchi. Destabilization of Insulin Hexamer in Water–Ethanol Binary Mixture. *J. Phys. Chem. B*, 123(49):10365–10375, 2019.

[175] Robert J Webber, Erik H Thiede, Douglas Dow, Aaron R Dinner, and Jonathan Weare. Error Bounds for Dynamical Spectral Estimation. *SIAM J. Math. of Data Sci.*, 3(1):225–252, 2021.

[176] Chatipat Lorpaiboon, Erik Henning Thiede, Robert J Webber, Jonathan Weare, and Aaron R Dinner. Integrated Variational Approach to Conformational Dynamics: A Robust Strategy for Identifying Eigenfunctions of Dynamical Operators. *J. Phys. Chem. B*, 124(42):9354–9364, 2020.

[177] Fabio Pietrucci and Alessandro Laio. A Collective Variable for the Efficient Exploration of Protein Beta-Sheet Structures: Application to SH3 and GB1. *J. Chem. Theory Comput.*, 5(9):2197–2201, 2009.

[178] Sandhya P. Tiwari, Edvin Fuglebakk, Siv M. Hollup, Lars Skjærven, Tristan Cragnolini, Svenn H. Grindhaug, Kidane M. Tekle, and Nathalie Reuter. WEBnm@ v2.0: Web Server and Services for Comparing Protein Flexibility. *BMC Bioinformatics*, 15(1):427, 2014.

[179] Schrödinger, LLC. The PyMOL Molecular Graphics System, Version 2.3.0, November 2015.

[180] Scott H Northrup, Stuart A Allison, and J. Andrew McCammon. Brownian Dynamics Simulation of Diffusion-Influenced Bimolecular Reactions. *J. Chem. Phys.*, 80(4):1517–1524, 1984.

[181] Adithya Vijaykumar, Peter G. Bolhuis, and Pieter Rein ten Wolde. The Intrinsic Rate Constants in Diffusion-Influenced Reactions. *Faraday Discuss.*, 195:421–441, 2016.

[182] Pekka Mark and Lennart Nilsson. Structure and Dynamics of the TIP3P, SPC, and SPC/E Water Models at 298 K. *J. Phys. Chem. A*, 105(43):9954–9960, 2001.

[183] Christopher M Dobson. Protein folding and misfolding. *Nature*, 426(December), 2003.

[184] Matthew G. Iadanza, Matthew P. Jackson, Eric W. Hewitt, Neil A. Ranson, and Sheena E. Radford. A new era for understanding amyloid structures and disease. *Nat. Rev. Mol. Cell Biol.*, 2018.

[185] Jeevan B. GC, Yuba R. Bhandari, Bernard S. Gerstman, and Prem P. Chapagain. Molecular Dynamics Investigations of the α-Helix to β-Barrel Conformational Transformation in the RfaH Transcription Factor. *J. Phys. Chem. B*, 118:5101–5108, 5 2014.

[186] Bahman Seifi and Stefan Wallin. The C-terminal domain of transcription factor RfaH: Folding, fold switching and energy landscape. *Biopolymers*, 112, 10 2021.

[187] Madhurima Das, Nanhao Chen, Andy LiWang, and Lee-Ping Wang. Identification and characterization of metamorphic proteins: Current and future perspectives. *Biopolymers*, 112, 10 2021.

[188] William J. Wedemeyer, Ervin Welker, and Harold A. Scheraga. Proline Cis-Trans Isomerization and Protein Folding. *Biochemistry*, 41:14637–14644, 12 2002.

[189] John M. Jumper, Nabil F. Faruk, Karl F. Freed, and Tobin R. Sosnick. Trajectory-based training enables protein simulations with accurate folding and Boltzmann ensembles in cpu-hours. *PLOS Comput. Biol.*, 14:e1006578, 12 2018.

[190] John M. Jumper, Nabil F. Faruk, Karl F. Freed, and Tobin R. Sosnick. Accurate calculation of side chain packing and free energy with applications to protein molecular dynamics. *PLOS Comput. Biol.*, 14:e1006342, 12 2018.

[191] Xiangda Peng, Michael Baxa, Nabil Faruk, Joseph R. Sachleben, Sebastian Pintscher, Isabelle A. Gagnon, Scott Houliston, Cheryl H. Arrowsmith, Karl F. Freed, Gabriel J. Rocklin, and Tobin R. Sosnick. Prediction and Validation of a Protein's Free Energy Surface Using Hydrogen Exchange and (Importantly) Its Denaturant Dependence. *J. Chem. Theory Comput.*, 18:550–561, 1 2022.

[192] Maira Rivera, Pablo Galaz-Davison, Ignacio Retamal-Farfán, Elizabeth A. Komives, and César A. Ramírez-Sarmiento. Dimer dissociation is a key energetic event in the fold-switch pathway of KaiB. *Biophys. J.*, 121:943–955, 3 2022.

[193] Sunhwan Jo, Xi Cheng, Shahidul M. Islam, Lei Huang, Huan Rui, Allen Zhu, Hui Sun Lee, Yifei Qi, Wei Han, Kenno Vanommeslaeghe, Alexander D. MacKerell, Benoît Roux, and Wonpil Im. CHARMM-GUI PDB Manipulator for Advanced Modeling and Simulations of Proteins Containing Nonstandard Residues. *Adv. Protein Chem. Str.*, 96:235–265, 2014.

[194] Robert T. McGibbon, Kyle A. Beauchamp, Matthew P. Harrigan, Christoph Klein, Jason M. Swails, Carlos X. Hernández, Christian R. Schwantes, Lee-Ping Wang, Thomas J. Lane, and Vijay S. Pande. MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys. J.*, 109(8):1528 – 1532, 2015.

[195] Martin K. Scherer, Benjamin Trendelkamp-Schroer, Fabian Paul, Guillermo Pérez-Hernández, Moritz Hoffmann, Nuria Plattner, Christoph Wehmeyer, Jan-Hendrik Prinz, and Frank Noé. PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models. *J. Chem. Theory Comput.*, 11:5525–5542, October 2015.

[196] Michael R. Shirts and John D. Chodera. Statistically optimal analysis of samples from multiple equilibrium states. *J. Chem. Phys.*, 129:124105, 9 2008.

[197] Robert Best, Gerhard Humer, and William Eaton. Native contacts determine protein folding mechanisms in atomistic simulations. *Proc. Nat. Acad. Sci.*, 110(44):17874–17879, 2013.

[198] Aase Hvidt and K. Linderstrøm-Lang. Exchange of hydrogen atoms in insulin with deuterium atoms in aqueous solutions. *Biochim. Biophys. Acta*, 14:574–575, 1 1954.

[199] Aase Hvidt and Sigurd O. Nielsen. Hydrogen Exchange in Proteins. volume 21 of *Advances in Protein Chemistry*, pages 287–386. Academic Press, 1966.

[200] Manfred Eigen. Proton transfer, acid-base catalysis, and enzymatic hydrolysis. Part I: elementary processes. *Angew. Chem. Int. Edit.*, 3(1):1–19, 1964.

[201] S. W. Englander, L. Mayne, Y. Bai, and T. R. Sosnick. Hydrogen exchange: The modern legacy of Linderstrøm-Lang. *Protein Sci.*, 6:1101–1109, 5 1997.

[202] Filip Persson and Bertil Halle. How amide hydrogens exchange in native proteins. *P. Natl. Acad. Sci.*, 112:10383–10388, 8 2015.

[203] Yawen Bai, John S Milne, Leland Mayne, and S Walter Englander. Protein stability parameters measured by hydrogen exchange. *Proteins: Struct. Funct. Genet.*, 20(1):4–14, 1994.

[204] Jeffrey K. Myers, C. Nick Pace, and J. Martin Scholtz. Denaturant m values and heat capacity changes: Relation to changes in accessible surface areas of protein unfolding. *Protein Sci.*, 4:2138–2148, 10 1995.

[205] M. Cecchini, S. V. Krivov, M. Spichty, and M. Karplus. Calculation of Free-Energy Differences by Confinement Simulations. Application to Peptide Conformers. *J. Phys. Chem. B*, 113:9728–9740, 7 2009.

[206] Victor Ovchinnikov, Marco Cecchini, and Martin Karplus. A Simplified Confinement Method for Calculating Absolute Free Energies and Free Energy and Entropy Differences. *J. Phys. Chem. B*, 117:750–762, 1 2013.

[207] Yawen Bai, Tobin R. Sosnick, Leland Mayne, and S. Walter Englander. Protein Folding Intermediates: Native-State Hydrogen Exchange. *Science*, 269:192–197, 7 1995.

[208] Qing Xin Hua, Satoe H. Nakagawa, Wenhua Jia, Kun Huang, Nelson B. Phillips, Shi Quan Hu, and Michael A. Weiss. Design of an active ultrastable single-chain insulin analog: Synthesis, structure, and therapeutic implications. *J. Biol. Chem.*, 283(21):14703–14716, 2008.

[209] Michael D. Glidden, Khadijah Aldabbagh, Nelson B. Phillips, Kelley Carr, Yen Shan Chen, Jonathan Whittaker, Manijeh Phillips, Nalinda P. Wickramasinghe, Nischay Rege, Mamuni Swain, Yi Peng, Yanwu Yang, Michael C. Lawrence, Vivien C. Yee, Faramarz Ismail-Beigi, and Michael A. Weiss. An ultra-stable single-chain insulin analog resists thermal inactivation and exhibits biological signaling duration equivalent to the native protein. *J. Biol. Chem.*, 293(1):47–68, 2018.

[210] Mike Reppert, Anish R Roy, Jeremy O B Tempkin, Aaron R Dinner, and Andrei Tokmako. Refining Disordered Peptide Ensembles with Computational Amide I Spectroscopy : Application to Elastin-Like Peptides. *J. Phys. Chem. B*, 120(44):11395–11404, 2016.

[211] Seyedeh Maryam Salehi, Debasish Koner, and Markus Meuwly. Dynamics and infrared spectrocopy of monomeric and dimeric wild type and mutant insulin. *J. Phys. Chem. B*, 124(52):11882–11894, 2020.

[212] Seyedeh Maryam Salehi and Markus Meuwly. Site-selective dynamics of azidolysozyme. *J. Chem. Phys.*, 154(16):165101, 2021.

[213] Adam K Nijhawan, Arnold M Chan, Darren J Hsu, Lin X Chen, and Kevin L Kohlstedt. Resolving Dynamics in the Ensemble: Finding Paths through Intermediate States and Disordered Protein Structures. *J. Phys. Chem. B*, 125(45):12401–12412, 2021.

[214] Markus Meuwly. Atomistic Simulations for Reactions and Vibrational Spectroscopy in the Era of Machine Learning- Quo Vadis? *J. Phys. Chem. B*, 126(11):2155–2167, 2022.

[215] Kasper Huus, Svend Havelund, Helle B Olsen, Marco van de Weert, and Sven Frokjaer. Chemical and Thermal Stability of Insulin: Effects of Zinc and Ligand Binding to the Insulin Zinc-Hexamer. *Pharm. Res.*, 23(11):2611–2620, 2006.

[216] Christof Schütte, Adam Nielsen, and Marcus Weber. Markov state models and molecular alchemy. *Mol. Phys.*, 113:69–78, 1 2015.

[217] L. Donati, C. Hartmann, and B. G. Keller. Girsanov reweighting for path ensembles and Markov state models. *J. Chem. Phys.*, 146, 2017.

[218] Stefanie Kieninger, Luca Donati, and Bettina G. Keller. Dynamical reweighting methods for Markov models. *Curr. Opin. Struc. Biol.*, 61:124–131, 4 2020.

[219] Lorenz Rognoni, Tobias Möst, Gabriel Žoldák, and Matthias Rief. Force-dependent isomerization kinetics of a highly conserved proline switch modulates the mechanosensing region of filamin. *P. Natl. Acad. Sci.*, 111(15):5568–5573, 2014.

[220] Franziska Zosel, Davide Mercadante, Daniel Nettels, and Benjamin Schuler. A proline switch explains kinetic heterogeneity in a coupled folding and binding reaction. *Nature Commun.*, 9:3332, 12 2018.

[221] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.*, 24(6):417, 1933.

[222] Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572, 1901.

[223] Guillermo Pérez-Hernández, Fabian Paul, Toni Giorgino, Gianni De Fabritiis, and Frank Noé. Identification of slow molecular order parameters for markov model construction. *J. Chem. Phys.*, 139(1):07B604_1, 2013.

[224] Feliks Nuske, Bettina G Keller, Guillermo Pérez-Hernández, Antonia SJS Mey, and Frank Noé. Variational approach to molecular kinetics. *J. Chem. Theory Comput.*, 10(4):1739–1752, 2014.

[225] Frank Noé and Feliks Nuske. A variational approach to modeling slow processes in stochastic dynamical systems. *Multiscale Model. Sim.*, 11(2):635–655, 2013.

[226] Lutz Molgedey and Heinz Georg Schuster. Separation of a mixture of independent signals using time delayed correlations. *Phys. Rev. Let.*, 72(23):3634, 1994.

[227] Christian R Schwantes and Vijay S Pande. Improvements in markov state model construction reveal many non-native interactions in the folding of ntl9. *J. Chem. Theory Comput.*, 9(4):2000–2009, 2013.

[228] Stefan Klus, Feliks Nüske, Péter Koltai, Hao Wu, Ioannis Kevrekidis, Christof Schütte, and Frank Noé. Data-driven model reduction and transfer operator approximation. *J. Nonlinear Sci.*, 28(3):985–1010, 2018.

[229] Mario Bouchard, Jesus Zurdo, Ewan J Nettleton, Christopher M Dobson, and Carol V Robinson. Formation of insulin amyloid fibrils followed by ftir simultaneously with cd and electron microscopy. *Protein Sci.*, 9(10):1960–1967, 2000.

[230] Atta Ahmad, Ian S Millett, Sebastian Doniach, Vladimir N Uversky, and Anthony L Fink. Partially folded intermediates in insulin fibrillation. *Biochemistry*, 42(39):11404–11416, 2003.

[231] Qing-xin Hua and Michael A Weiss. Mechanism of insulin fibrillation the structure of insulin under amyloidogenic conditions resembles a protein-folding intermediate. *J. Biol. Chem.*, 279(20):21449–21460, 2004.

[232] F. E. Dische, C. Wernstedt, G. T. Westermark, P. Westermark, M. B. Pepys, J. A. Rennie, S. G. Gilbey, and P. J. Watkins. Insulin as an amyloid-fibril protein at sites of repeated insulin injections in a diabetic patient. *Diabetologia*, 31(3):158–161, 1988.

[233] Matthew G Iadanza, Matthew P Jackson, Eric W Hewitt, Neil A Ranson, and Sheena E Radford. A new era for understanding amyloid structures and disease. *Nat. Rev. Mol. Cell Bio.*, page 1, 2018.