MOLECULAR SIMULATIONS OF PATHWAYS AND KINETICS FOR PROTEIN-PROTEIN BINDING PROCESSES

by

Ali Sinan Saglam

B.A. Chemistry, Marmara University, 2011

Submitted to the Graduate Faculty of

the Dietrich School of Arts and Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2018

UNIVERSITY OF PITTSBURGH

DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

Ali Sinan Saglam

It was defended on

April 3 2018

and approved by

Lillian Chong, Professor, Chemistry

Kenneth Jordan, Professor, Chemistry

Seth Horne, Professor, Chemistry

Thomas Kiefhaber, Professor, Faculty of Natural Sciences I, Martin-Luther-Universitat Halle-Wittenberg

Dissertation Director: Lillian Chong, Professor, Chemistry

MOLECULAR SIMULATIONS OF PATHWAYS AND KINETICS FOR PROTEIN-PROTEIN BINDING PROCESSES

Ali Sinan Saglam, PhD

University of Pittsburgh, 2018

Protein-protein binding processes are crucial for biological functions and characterizing these processes fully has been a challenge in biophysics. In this work I use weighted ensemble path sampling method coupled with molecular simulations of varying levels of detail to answer long standing questions regarding protein-protein binding. In Chapter 3, I investigate the effects of preorganization on association between an intrinsically disordered peptide fragment of tumor suppressor p53 and the MDM2 protein using flexible residue level models. I simulated the binding process between p53 and MDM2 with varying degrees of preorganization in p53 and determined that the association rate constant of p53 peptide does not depend on the extent to which the peptide is preorganized for binding MDM2. In Chapter 4, I apply simulations with flexible molecular models to directly compute the "basal" k_{on} for the association of the two proteins barnase and barstar, in the absence of electrostatics. I simulated the binding process between exact hydrophobic analogues barnase and barstar and determined the extent with which the electrostatics enhance the basal k_{on}. Finally, in Chapter 5, I have generated binding pathways of barnase and barstar using all-atom simulations with explicit solvent. This study not only enabled a more detailed characterization of the binding mechanism but also provided an opportunity to determine the role of solvent in the binding process. Water molecules are proposed to play a crucial role in binding of barnase and barstar since water molecules can be found at the binding interface in the crystal structure and they increase the interfacial complementarity. Overall, the work presented here demonstrates the power of the weighted ensemble strategy in making it practical to

characterize binding processes that are otherwise unfeasible for standard simulations.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

Throughout my graduate career I have learned one thing above all else: I fail, a lot, and when I do I can rely on people who care about me to keep on moving forward.

I thank my advisor Prof. Lillian T. Chong for her support through more than a few rough patches I had during my graduate career. At a time when I didn't realize I needed help she did and helped me more than I thought was possible. I thank my committee for their guidance. I thank my friends and Chong group members, in particular A. J. Pratt, Alex DeGrave, Corinn Durham, Karl Debiec and Matt Zwier all of whom helped me tremendously. I thank my parents for their consistent patience and support, despite knowing how much they miss me. I thank my friend Vedat Sinan Ural for without him I wouldn't have taken on this endeavor. I thank my friend Tom Brinzer who has been consistently supportive every time I needed it. I thank my girlfriend Kelly Lenhart who patiently supported me through this time and without whom I wouldn't have had the strength to finish.

1.0 PROTEIN-PROTEIN ASSOCIATION

1.1 INTRODUCTION

Many biological processes involve the formation of complexes involving two or more proteins. Formation of these complexes controls the assembly of cellular structures, signal transduction and inhibition, immune response and more. Furthermore, protein-protein interactions have been a focus of the field of drug design due to the attractiveness of protein-protein interfaces as drug targets.

Characterizing the mechanisms of protein-protein binding processes has been a challenge in biophysics. Typical biophysical experiments can provide ensemble-averaged observables for the binding processes as well as high-resolution structures of stable states. As an ideal complement to such experiments, molecular dynamics simulations can function as a computational "microscope" to provide atomically detailed views of complete pathways for the binding processes, including states that are too transient to be captured by experiment. However, due to the long-timescales of protein binding processes, it has not been practical to access these timescales using standard simulations. The overarching goal of this work is to couple the enhanced sampling of the weighted ensemble strategy with atomistic molecular dynamics simulations to characterize the pathways and kinetics of protein-protein binding processes.

1.2 RESEARCH QUESTIONS

In the primary body of my work I have investigated different aspects of protein-protein binding, including the effect of conformational changes on the binding process. Many proteins are either partially or completely unfolded when not bound to their partners and are thus known as intrinsically disordered peptides (IDPs). An unanswered question regarding IDPs is the effect of preorganization of a disordered binding partner have on the binding process. Previous efforts to answer this question involved analogues of the IDP with varying degrees of preorganization; however, experimentally, it is difficult to vary the secondary structure of an IDP without chemical alterations which can affect the binding process. In contrast, molecular simulations enable changes to a single aspect of the binding process (e.g. degree of preorganization) without perturbing other aspects (e.g. chemical sequence). In Chapter 3, I discuss the binding simulations of p53 and MDM2 where I have tuned the protein model of p53 to obtain completely preorganized and completely disordered variants of p53 without chemical alterations.

A crucial unanswered question for protein-protein association was the effect of electrostatics on the basal k_{on}, the rate constant of association in the abscence of electrostatic interactions. To answer this question, I have investigated the mechanism of one of the most rapid protein-protein binding processes involving the extracellular ribonuclease barnase and its intracellular inhibitor barstar. I have simulated, in molecular detail, the wild-type barnase-barstar and the exact barnase-barstar hydrophobic isosteres, in which the partial charges are set to zero but the shapes are identical. In Chapter 4, I discuss the association simulations of hydrophobic isosteres of barnase and barstar and the effect of electrostatics on binding.

In Chapter 5, I have simulated barnase and barstar binding using atomically detailed simulations with explict solvent, characterized the binding process and investigated the role of solvent in the protein-protein binding process. In particular, while it is well known that desolvation of the binding interfaces occur during binding, it is not known when during binding it occurs.

Finally, while this thesis is focused on my work in protein-protein binding simulations, I have also worked together with another member of the lab on characterizing alternate

folded states of the fast-folding villin headpiece subdomain. Although the folding process of this subdomain has been long characterized as a two-state process, recent experimental studies by our collaborator, Thomas Kiefhaber (Martin Luther University Halle-Wittenberg) have demonstrated that the subdomain adopts two distinct folded states. The goal of our simulation study is to provide atomically detailed structures of these two alternate states.

1.3 ACCESSING LONGER TIMESCALES

While there are many ways to use molecular simulations, in this work I focus on methods that provide complete pathways leading to binding so that I can directly look at the binding process rather than indirect models. While the difficulty of generating events is completely dependent on the binding process and the size of the proteins, generating complete protein-protein binding pathways is generally computationally expensive. This difficulty can be somewhat mitigated by the use of simpler models that provide less detail, but, depending on the process even simpler models can be challenging. Furthermore, the model detail has to be selected carefully depending on the scientific question that is being asked.

Standard simulations, which are carried out for sufficiently long times to capture a large number of the events of interest, can only routinely access process as long as a microsecond. To access timescales beyond microseconds, a variety of strategies have been developed that enhance the sampling of long-timescale processes while maintaining rigorous kinetics. My thesis work has focused on the development of simulation protocols involving the weighted ensemble path strategy to enable the generation of complete pathways for protein binding process without introducing any bias in the dynamics. The strengths and limitations of the WE strategy are covered in the next chapter.

2.0 REVIEW OF PATH-SAMPLING STRATEGIES

The text in this chapter has been adapted from L. T. Chong, A. S. Saglam and D. M. Zuckerman, *Curr. Opin. Struc. Biol.*, **2017**, *43*, 88-94.

2.1 CHAPTER SUMMARY

Despite more than three decades of effort with molecular dynamics simulations, long-timescale (ms and beyond) biologically relevant phenomena remain out of reach in most systems of interest. This is largely because important transitions, such as conformational changes and (un)binding events, tend to be rare for conventional simulations (<10 µs). That is, conventional simulations will predominantly dwell in metastable states instead of making large transitions in complex biomolecular energy landscapes. In contrast, path sampling approaches focus computing effort specifically on transitions of interest. Such approaches have been in use for nearly 20 years in biomolecular systems and enabled the generation of pathways and calculation of rate constants for ms processes, including large protein conformational changes, protein folding, and protein (un)binding.

2.2 INTRODUCTION

Advances in computing hardware and software[1–3] along with record-setting molecular dynamics (MD) simulations, in terms of both length[4] and system size[5] bode well for the future of simulation. Nevertheless, the capacity of MD for investigating long timescales of biological

interest remains inadequate, particularly as investigators set their sights on ever larger and more complex systems.[6,7]

Path sampling approaches can substantially increase the 'reach' of MD in simulating rare events such as protein conformational changes, (un)folding, and (un)binding, by focusing computational effort on the functional transitions rather than the stable states (Figure 1) — without introducing bias in the results. In particular, such approaches exploit the fact that for rare events, the duration of the transition event itself (t_b) is much shorter than the dwell time (t_{dwell}) in the preceding metastable region ($t_b \ll t_{dwell}$). Even when there is not a clear separation of timescales between t_b and t_{dwell}, path sampling may offer a considerable advantage over straight-ahead MD, as described in the next section ('Path sampling methods and recent advances').

In addition to providing rigorous estimates of rate constants, a key strength of path sampling approaches is the generation of an ensemble of transition trajectories. The trajectories themselves yield the full sequence of intermediate configurations of a transition, which are essential for characterizing the mechanism of a complex biological process and too fleeting to be captured by laboratory experiments. Further, the probabilistic description intrinsic to an ensemble quantifies pathway heterogeneity, the importance of which remains to be understood in biomolecular processes of different types.

Path-sampling methods have been advanced significantly in recent years and appear to have reached a state of maturity where theoretical underpinnings have been clarified, and where essential commonalities can be discerned. However, the reader is cautioned that all of the approaches have intrinsic limitations, sketched below, and that path-sampling data must be critically analyzed for undersampling to prevent unfounded interpretation.

We take this opportunity to survey key ideas and recent progress in the field. We cover only approaches that are well-founded in non-equilibrium statistical mechanics and hence capable of yielding, for example, unbiased estimates of rate constants and a true sample of the transition path ensemble. We note that the related Markov state modeling approach will be addressed separately in this issue.

2.3 PATH SAMPLING METHODS AND RECENT ADVANCES

2.3.1 Conceptual framework

Path sampling approaches exploit the separation of timescales that typically occurs in biomolecular systems. Consider the extreme example of attempting to observe transient unfolding of a stable protein under native conditions: unfolding events will be few and far between. Path sampling approaches can explicitly focus computational effort on the unfolding event, bypassing the lengthy dwells in the folded state.

Path sampling can be useful for rare events even when the separation of timescales is ambiguous. Consider another extreme case where a single uncharged receptor and ligand occupy a large volume, so that the probability of complexation is very small on MD timescales. The time for binding by diffusion arguably is the same as the 'transition time' (t_b) in such a system and there is no clear timescale separation. Yet path sampling approaches can focus simulation effort on successful events, and even account for the rareness of binding without bias[8]. Likewise the conformational sampling of stable states separated by low barriers can be efficiently accomplished using path sampling[9,10].

Though path sampling approaches can yield equilibrium state populations and potentials of mean force, their primary strength is a capacity to estimate non-equilibrium observables such as rate constants. In the latter context, the ability to account for directionality and history is critical — particularly tracing back any given trajectory to the most recently occupied state (A or B, 'initial' or 'target' state), which enables unbiased rate calculation[11–13]; see also[14,15]. This insight from path theory has important practical implications for analyzing ordinary MD simulations and avoiding the Markov assumption[16].

Current path sampling approaches can be divided into the following three categories for conceptual clarity.

2.3.2 Methods using complete paths

Two approaches work directly with complete A-to-B transition paths (Figure 2a). *Transition path sampling* (TPS) is based on Pratt's suggestion to run Monte Carlo (MC) simulations on

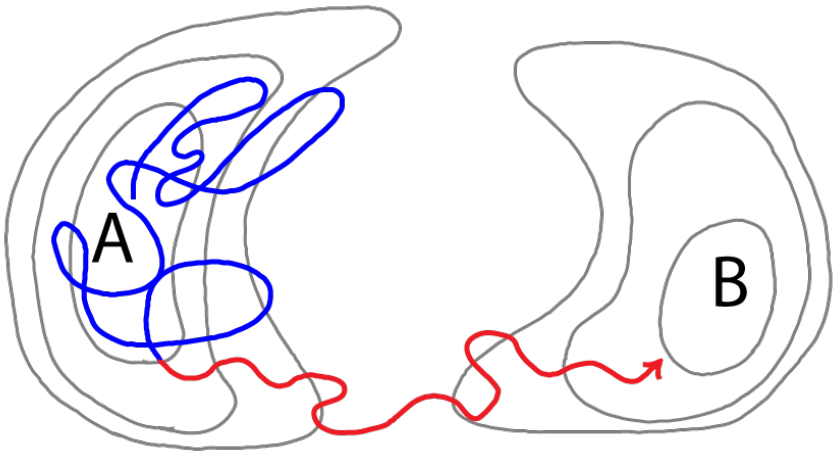

Figure 1: A schematized very long MD trajectory which successfully transitions to basin B after starting in A is superimposed over energy contours (gray lines). By definition, every unbiased transition trajectory consists of (i) a dwell period (blue) of duration t_{dwell} prior to the last exit from the initial state and (ii) the transition event itself (red) of duration t_b. If $t_b \ll t_{dwell}$, then path sampling strategies may be useful in focusing computational effort on the transition process.

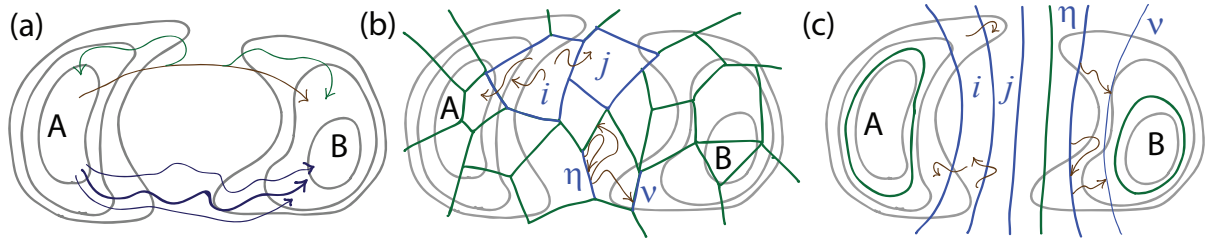

Figure 2: An energy landscape (gray contours) is shown for which the transition from basin A to B is rare on the timescale of typical MD simulations. **(a)** Some methods use full-length transition trajectories. In transition path sampling, an initial unphysical trajectory (brown) is perturbed via random trials (green) using a Metropolis Monte Carlo procedure in trajectory space, whereas in dynamic importance sampling, a set of biased trajectories (dark blue) are reweighted to conform with unbiased behavior. **(b)** Many methods use fully unbiased trajectory segments (brown) connecting bins (i and j), such as the weighted ensemble, or connecting interfaces (η and ν), such as milestoning and non-equilibrium umbrella sampling. **(c)** Other approaches, such as transition interface sampling and forward flux sampling, use strictly nested interfaces interpolating from A to B. Generally speaking, shorter transitions among bins or interfaces are much more probable than full A-to-B transitions, and trajectory segments can be connected using rigorous statistical mechanics to infer longer-time behavior.

entire trajectories[17] rather than on the more familiar MC for configurations. Advanced by Chandler and coworkers[18–20], TPS uses trial perturbations to an existing A-to-B trajectory and a Metropolis acceptance criterion. *Dynamic importance sampling* (DIMS), proposed by Woolf[21] based on earlier work[22,23], also uses complete paths. In DIMS, however, independent transition trajectories are generated using biased dynamics, and are then reweighted using the ratio of sampled to true probability[24].

2.3.3 Methods using trajectory segments: region-to-region

Most current path-sampling approaches work procedurally with trajectory segments, even if fully or nearly continuous A-to-B transitions ultimately are produced. As shown in Figures 2b,c, segment-based methods can be categorized accordingly to whether partial transitions are sampled between regions ('bins') or between interfaces. Bin-to-bin transitions typically are sampled via trajectory segments of fixed duration, whereas interfacial transitions require 'catching' trajectories in the act of crossing.

Huber and Kim proposed the *weighted ensemble* (WE) approach in 1996[25], which was essentially a rediscovery of the 'splitting' strategy described by Kahn in 1951[26]. The basic idea is to classify configuration space into bins among which transitions are affordably likely. A set of unbiased trajectories is run in parallel, with replication of segments that reach new bins, encouraging progress toward B. Statistical weighting ensures unbiased results[27], and the approach has been extended for steady state and rate-constant calculations[28,29]. The related *adaptive multilevel splitting* (AMS) approach uses trajectory splitting within a different statistical formulation without bins[30]. See also[31,32].

Underscoring the methodological convergence occurring in the field, some interfacial approaches have now been adapted for bin-to-bin sampling[33,34]. Markov state models also operate in a bin-to-bin framework (see review by Noe in this issue). The *discrete path sampling approach* uses energy basins instead of bins[35–37]; see also[38,39].

2.3.4 Approaches using trajectory segments: interface–interface

Most current methods sample trajectory segments of heterogeneous lengths that start and end on interfaces. Some approaches require fully nested interfaces that interpolate from initial to target state and others can use nearly arbitrary interfaces — surfaces of arbitrary bins tiling configuration space (Figure 2b,c).

With *transition interface sampling* (TIS), van Erp, Moroni, and Bolhuis[11] introduced an extension of TPS which attempted to improve the rate-constant calculation by using a series of partial-flux calculations for a set of nested interfaces separating states A and B — see Figure 2c. Intermediate TPS calculations are used to generate the necessary TIS path ensembles. There have been a number of TIS extensions[40,41]. *Forward flux sampling* (FFS) uses a similar formalism but instead runs standard (not TPS) simulations between interfaces[42], and FFS has been generalized[33]. See also[43].

Interfaces which may not be nested (e.g., boundaries of Voronoi cells — see Figure 2b) are used in some approaches. *Non-equilibrium umbrella sampling* (NEUS), introduced by Dinner and coworkers, first showed how to use interfaces for arbitrary cells which tile configuration space in steady-state calculations[44] and was further developed[13,45,46]. *Milestoning*, although originally introduced by Faradjian and Elber for nested interfaces[47], was later generalized for use with arbitrary interfaces[48,49].

2.3.5 Limitations

All the approaches discussed here share the goal of generating an ensemble of transition trajectories, and hence they also share certain limitations. The focusing of sampling on transition regions instead of stable states in an unbiased manner typically requires that the transition trajectories are correlated with one another (e.g.,[19,27]). Such correlations imply a reduction in information content: perhaps one in 100 transitions is truly independent. Therefore, trajectories should be analyzed carefully for correlations and sampling quality[8,11,29,49]. For methods where the path-sampled trajectories are not correlated, there generally is another type of statistical inefficiency[24].

Another practical concern regards software. Several pathsampling packages are publicly

available[50–53], and most require some parameter tuning. Algorithms which examine trajectories at fixed time intervals, such as WE, lend themselves to facile interoperability with a variety of MD engines. Interface-based methods require 'catching' trajectories in the act of crossing boundaries, which already has been hard-wired in some packages[53,54], but could represent a significant barrier for users desiring alternative dynamics.

2.4 SUCCESSES

In recent years, path sampling approaches have enabled the simulation of several types of long-timescale biological processes that would not have been practical using conventional simulation: large protein conformational transitions, protein folding, and protein–ligand (un)binding.

2.4.1 Protein conformational transitions and folding processes

Notable successes involving large protein conformational transitions include simulations of substrate-induced conformational changes in enzymes and large conformational transitions in membrane transport proteins. In studies involving enzymes, milestoning has generated ms conformational transitions between the open and closed states of the HIV reverse transcriptase[55,56], yielding rate constants that are consistent with experiment (Figure 3a). In studies involving membrane transport proteins, the WE approach has generated pathways for outward-to-inward-facing transitions in the sodium symporter Mhp1 using coarse-grained simulations[57] and the DIMS approach has generated transitions between the cytoplasmic open conformation and perisplamic open conformation of the lactose permease transporter using atomistic simulations in implicit solvent[58]. For the related problem of ion permeation, the WE approach 90 has enabled the calculation of current–voltage relationships for a simple model ion channel[59].

Applications of path sampling approaches to protein folding — the most extreme protein conformational transition — have been focused on mini-proteins that fold on the ms

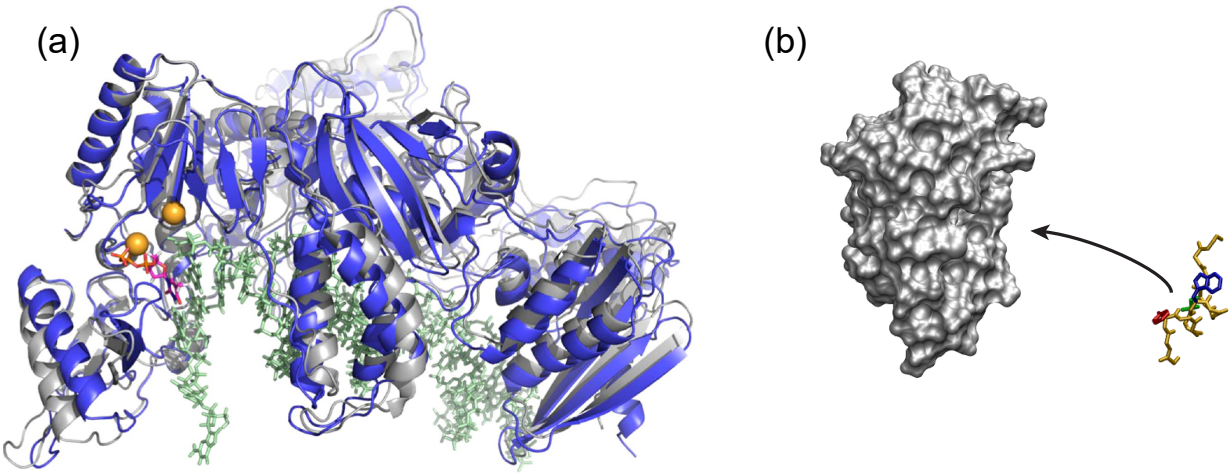

Figure 3: **(a)** Milestoning has generated pathways and calculated rate constants for substrate-induced transitions between the open (gray) and closed (blue) conformations of HIV reverse transcriptase in complex with Mg^{2+} ions (yellow) and duplex DNA (green); for clarity, only the p66 subunit is shown, although both p66 and p51 subunits were included in the simulations [4]. **(b)** The WE approach has generated pathways and calculated rate constants for the protein–peptide binding process involving the MDM2 protein (gray) and an intrinsically disordered p53 peptide (yellow)[10].

timescale. For example, the single-replica multistate TIS method has enabled efficient simulation of both folding and unfolding processes for Trp-cage[60] while the FFS method has been used to simulate a loop unfolding transition in Trp-cage[61] that was revealed by a previous TIS study to be rate-limiting for the unfolding process[62]. In addition, the single-replica multistate TIS method has been applied to the ms-folding process of the villin headpiece as well as its much slower sub-ms unfolding process (mean first passage time of 0.8 ms), demonstrating that path sampling approaches can be effective in estimating rate constants for protein unfolding processes as well as folding processes[63]. Of future interest are the application of these approaches to the (un)folding processes of entire proteins (e.g., NTL9 and ubiquitin) at experimental temperatures; due to their long-timescales (ms or beyond), such folding processes have typically been characterized at the (considerably higher) melting temperatures by straightforward simulations[64,65].

2.4.2 Protein (un)binding processes

The characterization of protein (un)binding mechanisms is not only fundamental to biology, but of great interest to the field of drug design. The simulation of protein binding processes with rigorous kinetics is particularly challenging due to the presence of metastable intermediates (e.g., the encounter complex).

Path sampling has yielded initial successes with models at different levels of resolution. For example, the WE approach has enabled the first atomistic simulations (to our knowledge) of protein–peptide binding pathways with rigorous rate constants; these simulations involved the MDM2 protein and an intrinsically disordered p53 peptide, which adopts an α-helical conformation upon binding MDM2[10]. In addition, two studies have demonstrated the power of path sampling strategies in generating atomistic pathways for protein–ligand unbinding processes and the corresponding k_{off} values, which are of great interest for drug design efforts. These studies involve, firstly, the application of the WE approach to the FK506 binding protein and several low-affinity, small molecule inhibitors, which unbind on timescales up to tens of ns, resulting in the first analysis of ligand-exit distributions[66], and second, the application of the AMS approach to trypsin and the benzamidine inhibitor, which unbinds

on the ms timescale[67]. In addition, it has been demonstrated that experimental k_{on} values can be efficiently reproduced for various protein–ligand systems using milestoning as part of an atomistic MD/Brownian Dynamics approach[68].

Even coarse-grained models may not be amenable to complete sampling via straight-ahead simulation. For example, the WE strategy has been of great benefit to even Brownian Dynamics simulations involving coarsegrained, albeit flexible protein models that have been parameterized to reproduce the molecular shapes, electrostatic potentials, and diffusion properties of all-atom models. The resulting WE simulations enabled not only the efficient reproduction of experimental k_{on} values for wild-type and mutant complexes of barnase and barstar, but a statistically robust estimate of the much slower 'basal' k_{on} involving the hydrophobic isosteres of the two proteins — a quantity of fundamental interest to the field of molecular recognition[8] (Figure 3b).

2.5 CHALLENGES

As path sampling approaches are used to target more complex systems and slower processes, which seems inevitable, a number of challenges remain. The most basic difficulty hinges on intrinsic timescales of the systems themselves: for example, if the transition event duration (see *Introduction* section) for a certain process exceeds 1 µs, then sampling an ensemble of uncorrelated transition events would be almost impossible given a total budget of 10 µs. Of course, the intrinsic timescales would not be known ahead of time, suggesting caution is necessary for complex systems.

Coordinates and correlations present the primary methodological challenge. The problem of generating correlated transition trajectories was discussed above in 'Path Sampling Methods and Recent Advances', but it is closely connected to the difficulty of constructing suitable coordinates (or bins or interfaces) for methods requiring them. Consider a system which is not readily described by a one dimensional reaction coordinate (i.e., which has slow orthogonal coordinates). If one-dimensional bins or interfaces are used, it can be expected that fully sampling the orthogonal space will be slow and may render the results unreliable

— the sampled trajectory segments may be overly correlated. Fortunately, investigators are already beginning to make progress in adaptively developing bins and interfaces[27,69,70].

It will be important to develop software resources further. As noted in '*Path sampling methods and recent advances*' section, several highly scalable packages are currently available, including WESTPA, AWE-WQ and FRESHS, which have demonstrated inter-operability with a variety of dynamics engines[50–52]. A competitive software ecosystem with additional robust packages should be a boon to the field. Nevertheless, we caution that path sampling tools are likely to continue to require considerable user expertise in yielding reliable results.

On a final note, another frontier that has already been addressed by initial studies is the application of path sampling approaches to problems at other scales. Several approaches have already been applied to signaling networks, gene regulation, and spatially resolved cell models[42,71–77].

2.6 ACKNOWLEDGEMENTS

This work was supported by NIH grant 1RO1GM115805 to L.T.C. and D.M.Z.; NIH Grant P41GM103712 and NSF Grant MCB-1119091 to D.M.Z.

3.0 FLEXIBILITY VS PREORGANIZATION: DIRECT COMPARISON OF BINDING KINETICS FOR A DISORDERED PEPTIDE AND ITS EXACT PREORGANIZED ANALOGUES

The text in this chapter has been adapted from A. S. Saglam, D. W. Wang, M. C. Zwier, and L. T. Chong., *J. Phys. Chem. B*, **2017**, *121* (43), pp 10046–10054.

3.1 CHAPTER SUMMARY

Many intrinsically disordered proteins, which are prevalent in nature, fold only upon binding their structured partner proteins. Such proteins have been hypothesized to have a kinetic advantage over their folded, preorganized analogues in binding their partner proteins. Here we determined the effects of ligand preorganization on the k_{on} for a biomedically important system: an intrinsically disordered p53 peptide ligand and the MDM2 protein receptor. Based on direct simulations of binding pathways, computed k_{on} values for fully disordered and preorganized p53 peptide analogues were within error of each other, indicating little if any kinetic advantage to being disordered or preorganized for binding the MDM2 protein. We also examined the effects of increasing the concentration of MDM2 on the extent to which its mechanism of binding to the p53 peptide is induced fit vs conformational selection. Results predict that the mechanism is solely induced fit if the unfolded state of the peptide is more stable than its folded state; otherwise, the mechanism shifts from being dominated by conformational selection at low MDM2 concentration to induced fit at high MDM2 concentration. Taken together, our results are relevant to any protein binding process that involves a disordered peptide of a similar length that forms a single α-helix upon binding a

partner protein. Such disorder-to-helix transitions are common among protein interactions of disordered proteins and are therefore of fundamental biological interest.

3.2 INTRODUCTION

Many proteins that are either partially or completely unfolded in their unbound states[78,79] fold only upon binding their structured partner proteins. Such "intrinsically disordered" proteins (IDPs) have been proposed to have a kinetic advantage over their preorganized, folded analogues for binding their partners,[80,81] which challenges the long-standing assumption that the preorganization of a ligand to its receptor-bound conformation results in a faster association rate constant (k_{on}). Potential mechanisms by which this kinetic advantage might be achieved are (i) the "fly-casting" mechanism, in which the IDP collides more rapidly with the partner receptor due to a larger "capture" radius,[80] and (ii) the "dock-and-coalesce" mechanism for IDPs with two or more segments in which the initial docking of one segment results in a more rapid, pseudointramolecular docking of the remaining segments.[81] Throughout this work, the term "ligand" refers to a molecule (e.g., small molecule, peptide, or protein) that binds to a larger molecule that serves as the target receptor.

While experimental studies have provided informative insights about the effects of preorganization on the binding kinetics of IDP ligands,[82–87] these studies have not been able to provide definitive proof of a kinetic advantage (or lack thereof) to being disordered vs preorganized. Existing experimental studies indicate differing results on the effect of ligand preorganization on binding kinetics. For example, preorganization has resulted in faster binding for certain IDPs (ACTR and Y507A mutant of the E3 rRNase domain),[82,83] and no significant effect on the binding kinetics for other IDPs (PUMA and cMyb).[84,85] In addition, an unfolded variant of the Fyn SH3 domain that was engineered via truncation of only four residues has achieved the same k_{on} as the full-length, folded domain for a high-affinity peptide,[86] and the preorganization of the disordered monomers of an engineered GCN4-p1 leucine zipper variant has resulted in slower dimerization.[87] Ideally, the effect of ligand preorganization on binding kinetics would be assessed by engineering peptide analogues that

differ only in their degree of preorganization without altering the chemical structures, which is not possible in experiments.

Molecular simulations provide the only practical means to compute k_{on} values for both IDPs and their exact preorganized analogues — which have been engineered *in silico*—by directly generating the corresponding binding pathways. Furthermore, while experiments can typically measure only the k_{on}, simulations can be used to directly compute the rate constants of individual steps. However, due to the relatively long time scales of protein binding processes, only one simulation study has reported atomistic binding pathways along with the k_{on} for an IDP ligand (p53 peptide) and its protein receptor (MDM2), and these simulations did not sample fully disordered analogues.[10] Both atomistic and residue-level models have been used to characterize solely the late stages of binding, i.e., after the IDPs have collided with their partner proteins.[88–90] Residue-level simulation studies of binding pathways for IDPs have been reported,[91,92] including the only study that has determined the effects of preorganization on the binding kinetics of an IDP, focusing on the intrinsically disordered, phosphorylated KID (pKID) domain and its folding into a pair of linked-together α-helices upon binding the KIX protein.[92]

Here, we focused on an IDP ligand that adopts a single α-helix upon binding its folded protein receptor: the intrinsically disordered, N-terminal peptide fragment of tumor suppressor p53 and MDM2 protein. We determined the effects of ligand preorganization on the k_{on} by directly simulating binding pathways of the disordered p53 peptide and several of its exact analogues with various extents of preorganization. In addition, we used the computed k_{on} values to predict the effect of increasing the concentration of MDM2 on the extent to which the binding mechanism proceeds through induced fit and conformational selection. Based on atomistic simulations, the binding mechanism of the MDM2 receptor and p53 peptide ligand is predicted to shift from being dominated by conformational selection at low receptor concentration to induced fit at high receptor concentration.[93] Likewise, based on experimental rate constants, this shift in mechanism is expected to occur upon increasing the ligand concentration for systems involving disordered protein receptors and their small organic ligands.[94,95] Given the prevalence of single α-helix binding motifs among protein-ligand interactions,[96] the mechanism of MDM2-p53 binding is not only of biomedical

importance[97] but fundamental to biology.

3.3 METHODS

Key features of our simulation strategy are the following. First, we employed minimal residue-level models (C_α models) along with a Gō-type potential energy function,[98,99] which enables tuning of the extent of preorganization of the IDP (in our case, the p53 peptide) from fully disordered to fully preorganized. Second, dynamics were propagated using a Brownian dynamics algorithm with the inclusion of appropriately parametrized hydrodynamic interactions (HIs) between protein residues to yield realistic diffusion properties.[100] Third, we applied the weighted ensemble (WE) path sampling strategy,[25,27,101] which has been demonstrated to be orders of magnitude more efficient than standard Brownian dynamics simulations in generating pathways and rate constants for protein binding processes.[8] Full details of the protein model, simulations, and analysis are below.

3.3.1 The Protein Model

Residue-level protein models were used in which each residue was represented by a single pseudoatom at its C_α position, yielding 85 pseudoatoms for the MDM2 protein (residues 25-109) and 13 pseudoatoms for the p53 peptide (residues 17-29). Coordinates for the unbound and bound conformations of MDM2 and p53 peptide were taken from the crystal structure of the MDM2-p53 peptide complex (PDB code: 1YCR).[102]

A Gō-type potential energy function[98,99] was used to govern the conformational dynamics of the protein model. In this energy function, bonded interactions between pseudoatoms are modeled by standard molecular mechanics terms for bonds, angles, and dihedrals; and nonbonded interactions between pseudoresidues separated by four or more pseudobonds were treated as either native or non-native contacts. A native contact was defined as a residue-residue contact in which the heavy atoms of the two residues are within 5.5 Å of each other in the crystal structure of the native complex. In addition to 57 intermolecular native contacts

between p53 and MDM2, the p53 peptide and MDM2 consisted of 10 and 266 intramolecular contacts, respectively.

The protein model was parametrized by focusing separately on the following three contributions to the total energy function:

$$E_{\text{total}} = E_{\text{p53}} + E_{\text{MDM2}} + E_{\text{MDM2/p53}} \tag{3.1}$$

where E_{p53} and E_{MDM2} correspond to intramolecular contributions from p53 and MDM2, respectively, and $E_{\text{MDM2/p53}}$ corresponds to the intermolecular MDM2/p53 contributions.

As others have done,[103] we tuned the degree of structure and backbone flexibility of the IDP (in our case, the p53 peptide) by applying a single scaling factor α to the pseudoangle, pseudodihedral, and intramolecular nonbonded terms of the energy function involving solely the IDP:

$$\begin{aligned} E_{\text{p53}} = & \sum_{\text{bonds}} k_{\text{bond}}(r - r_{\text{eq}})^2 \\ & + \alpha \Bigg\{ \sum_{\text{angles}} k_{\text{angles}}(\theta - \theta_{\text{eq}})^2 \\ & + \sum_{\text{dihedrals}} V_1 \left[1 + \cos(\varphi - \varphi_1)\right] + V_3 \left[1 + \cos(3\varphi - \varphi_3)\right] \\ & + \sum_{i<j-4,\text{non-native}}^{\text{p53}} \varepsilon^{\text{non-native}} \left(\frac{\sigma_{ij}^{\text{non-native}}}{r_{ij}} \right)^{12} \\ & + \sum_{i<j-4,\text{native}}^{\text{p53}} \varepsilon^{\text{native}} \left[5 \left(\frac{\sigma_{ij}^{\text{native}}}{r_{ij}} \right)^{12} - 6 \left(\frac{\sigma_{ij}^{\text{native}}}{r_{ij}} \right)^{6} \right] \Bigg\} \end{aligned} \tag{3.2}$$

where r, θ, φ are pseudo bond lengths, pseudoangles, and pseudodihedrals, respectively; V_1 and V_3 are potential barriers for the dihedral terms; $\varepsilon^{\text{native}}$ is the energy well depth for native contacts, r_{ij} is interatomic distance between pseudoatoms i and j during simulation, and $\sigma_{ij}^{\text{native}}$ is the corresponding distance in the crystal structure; $\sigma_{ij}^{\text{non-native}}$ and $\varepsilon^{\text{non-native}}$ for non-native contacts were set to 4.0 Å and 1 kcal/mol, respectively. Equilibrium bond lengths (r_{eq}), angles (θ_{eq}), and dihedral phase angles (φ_1 and φ_3) were taken from the crystal structure. The force constants, k_{bond} and k_{angle}, were set to 100 kcal/mol/Å and 20 kcal/mol/rad,

respectively, and V_1 and V_3 were set to 1 and 0.5 kcal/mol, respectively. The scaling factor α was set to 0.1, 0.5, 1.0, and 2.0 to model analogues of the p53 peptide that exhibit, on average, a fraction of native contacts (Q_{p53}) of 0.25, 0.5, 0.85, and 0.99, respectively, based on 10 µs standard simulations of the isolated peptide (Figures S7 and S8). Thus, α values of 0.1 and 2.0 represent the fully disordered and fully preorganized versions of the p53 peptide, respectively.

The same potential function was used for MDM2 (E_{MDM2}) and nonbonded MDM2-p53 interactions ($E_{\mathrm{MDM2/p53}}$), except for the omission of the scaling factor α. An $\varepsilon^{\mathrm{native}}$ of 1.0 kcal/mol was used for intramolecular native contacts of MDM2, yielding a fraction of native contacts $Q_{\mathrm{MDM2}} > 0.8$ based on five 10 µs simulations. To ensure that the fully disordered p53 peptide folds upon binding MDM2, the $\varepsilon^{\mathrm{native}}$ for native MDM2-p53 interactions was set to the minimum value (2.0 kcal/mol) required to ensure that the peptide folds upon binding MDM2 ($Q_{p53} > 0.7$ throughout a 10 µs standard simulation (no WE sampling); Figure S9). Following others,[92] the same $\varepsilon^{\mathrm{native}}$ value for intermolecular contacts (in our case, MDM2-p53 contacts) was used for all analogues of the IDP (the p53 peptide). The same $\varepsilon^{\mathrm{native}}$ was also used for native contacts within the fully preorganized p53 peptide.

3.3.2 Weighted Ensemble Simulations

To generate MDM2-p53 peptide binding pathways, we applied the weighted ensemble (WE) path sampling strategy,[25] as implemented in the WESTPA software package (https://westpa.github.io/westpa),[50] to orchestrate a large set of Brownian dynamics trajectories that were carried out using the framework of the Northrup-Allison-McCammon (NAM) method.[104] In this hybrid WE/ NAM approach, two concentric spherical surfaces are first defined with radii b and q that correspond to separation distances between MDM2 and the p53 peptide. The inner sphere, or b surface, represents the initial unbound state, and the outer sphere, or q surface, represents a much larger separation distance ($q \gg b$) at which trajectories are terminated to avoid wasting computing time sampling any indefinite drifting apart of the binding partners. The next step of the WE/ NAM approach is to define a progress coordinate between the unbound and bound states and to divide this coordinate into bins

with the goal of populating each bin with N trajectories, each of which is assigned a statistical weight. Starting from N trajectories in the initial unbound state, the dynamics of each trajectory are simultaneously propagated in parallel and occasionally coupled by replication and combination events at fixed time intervals τ based on their progress toward the target state (e.g., the bound state), splitting and combining the statistical weights, respectively, such that no bias is introduced into the dynamics.[25] To maintain steady-state conditions, any trajectory that reaches the q surface is "recycled" by terminating the trajectory and starting a new trajectory from an initial, unbound state with the same statistical weight.

In our WE simulations, the radii b and q were set to 35 and 50 Å, respectively; as required for the WE/NAM approach, b is sufficiently large such that the intermolecular forces between the binding partners can be assumed to be isotropic (as mentioned above, only short-range residue-residue interactions were modeled in our simulations). Initial unbound states were generated by randomly reorienting the binding partners with respect to each other at a separation of 35 Å using their corresponding conformations from the crystal structure of MDM2-p53 complex.[102] For the progress coordinate, we used the C_α RMSD of the p53 peptide after alignment of MDM2 ranging from 0 to 100 Å. This progress coordinate was evenly divided into 29 bins with a target number of 6 trajectories/bin, yielding a maximum total of 390 trajectories at any point in the WE simulation. The fixed time interval τ for each WE iteration was set to 100 ps, which allowed for at least one trajectory to advance to the next bin after each WE iteration.

For each p53 peptide analogue (each α value), 10 independent WE simulations of the MDM2-p53 binding process were carried out under pseudoequilibrium conditions in which trajectories were recycled at the q surface, but not the bound state, to allow for refinement of the bound-state definition after completion of the simulations. Once this was refined, we effectively recycled trajectories that reached the refined definition of the bound state by removing the trajectories from subsequent analysis with proper renormalization of the remaining probabilities. This renormalization was straightforward given that no trajectories in the reverse, unbinding direction were generated in our Gō-type simulations. Each WE simulation was carried out for a maximum trajectory length of 200 ns (2000 WE iterations), which was sufficiently long for obtaining converged estimates of the k_{on} (Figure S11).

Conformations were sampled every 1 ps for analysis.

3.3.3 Propagation of Dynamics

The dynamics of our WE simulations were propagated using a standard Brownian dynamics algorithm[105] with the inclusion of hydrodynamic interactions (HI),[100] as implemented in the UIOWA_BD software.[100,106] Hydrodynamic radii were set to 5.3 Å, which has been found to reproduce the translational and rotational diffusion coefficients of all-atom models of folded proteins when using the residue-level models of this study.[100] The solvent viscosity was set to 0.89 cP to represent water at 25 °C. To enable the use of a 50 fs time step, all pseudobonds between residues were constrained to their native bond lengths by applying the LINCS algorithm.[107]

3.3.4 Calculation of Bimolecular Rate Constants

All bimolecular rate constants k were calculated using the Northrup-Allison-McCammon (NAM) equation:[104]

$$k = \frac{k_D(b)\beta}{1 - (1-\beta)k_D(b)/k_D(q)} \tag{3.3}$$

where k_D(r) is the diffusion rate constant for the two binding partners achieving a separation distance r, and β is the probability that a simulation starting from the unbound state with a separation distance of b (35 Å) reaches the target state before drifting apart to a separation distance of q (50 Å). To calculate the rate constant k_1, the target state is the encounter complex; likewise, to calculate k_{on}, the target state is the native, bound state (see definitions in Results).

Assuming that the motions of the two binding partners are isotropic, the diffusion rate constants were calculated using the Smoluchowski equation: $k_D = 4\pi \text{Dr}$, where D is the relative translational diffusion coefficient of the two partners (i.e., the sum of their corresponding diffusion coefficients). Therefore, eq 3 reduces to

$$k = \frac{4\pi D b \beta}{(1 - (1-\beta)b/q)} \tag{3.4}$$

The translational diffusion coefficient of MDM2 was calculated using five 10 µs standard simulations of isolated MDM2, and the translational diffusion coefficient for each analogue of the p53 peptide was calculated using conformations sampled every 100 ps from a single 10 µs standard simulation of the corresponding isolated p53 peptide. The β value was estimated using the following equation:[108]

$$\beta = \frac{f_{\mathrm{SS}}^{\mathrm{target}}}{f_{\mathrm{SS}}^{\mathrm{target}} + f_{\mathrm{SS}}^{\mathrm{qsurf}}} \tag{3.5}$$

where $f_{\mathrm{SS}}^{\mathrm{target}}$ is the steady-state flux into the target state (encounter complex or bound state) and $f_{\mathrm{SS}}^{\mathrm{qsurf}}$ is the steady-state flux across the q surface in the WE simulation. All rate constants were calculated from each of 10 independent WE simulations, and then averaged. Uncertainties in the averaged rate constants represent two standard errors of the mean (SEM).

3.3.5 Calculation of the Percentage of Productive Collisions

The percentage of productive collisions (i.e., encounter complexes that succeed in rearranging to the bound state) was calculated according to the following equation:

$$\%\ \text{productive collisions} = \frac{f_{\mathrm{SS}}^{\mathrm{native}}}{f_{\mathrm{SS}}^{\mathrm{encounter}}} \tag{3.6}$$

where $f_{\mathrm{SS}}^{\mathrm{native}}$ is the steady-state flux into the native, bound state and $f_{\mathrm{SS}}^{\mathrm{encounter}}$ is the steady-state flux into the encounter complex; both fluxes were evaluated only after an approximate steady state was achieved (Figure S11). Reported percentages of productive collisions are averages over 10 independent WE simulations with uncertainties representing two SEM.

3.4 RESULTS

The goals of this study were to determine (i) the effects of preorganizing the p53 peptide ligand on its k_{on} for binding the MDM2 protein receptor and (ii) the effect of increasing the concentration of the MDM2 receptor on the binding mechanism. As shown in Figure 4A,

the extent of preorganization in the p53 peptide was tuned by applying a scaling factor α to the components of the energy function that involve solely the p53 peptide (see Methods) and setting the α values to 0.1 (fully disordered), 0.5, 1.0, and 2.0 (fully preorganized). To enable the calculation of statistically robust rate constants, we applied the WE path sampling strategy[25,27] in conjunction with molecular simulations to enhance the sampling of binding events while maintaining rigorous kinetics. For each p53 peptide analogue (i.e., each α value), a set of 10 independent WE simulations were carried out, yielding > 3000 binding events per simulation to achieve highly precise rate constants with relative errors of $\leq$16%, which amounts to a $\leq$ 0.1 kcal/mol difference in the corresponding free energy barrier at 25 °C as estimated by $-RT\ln(1/1.16)$. The simulations required one month to complete using 128 CPU cores of 2.3 GHz AMD Interlagos processors.

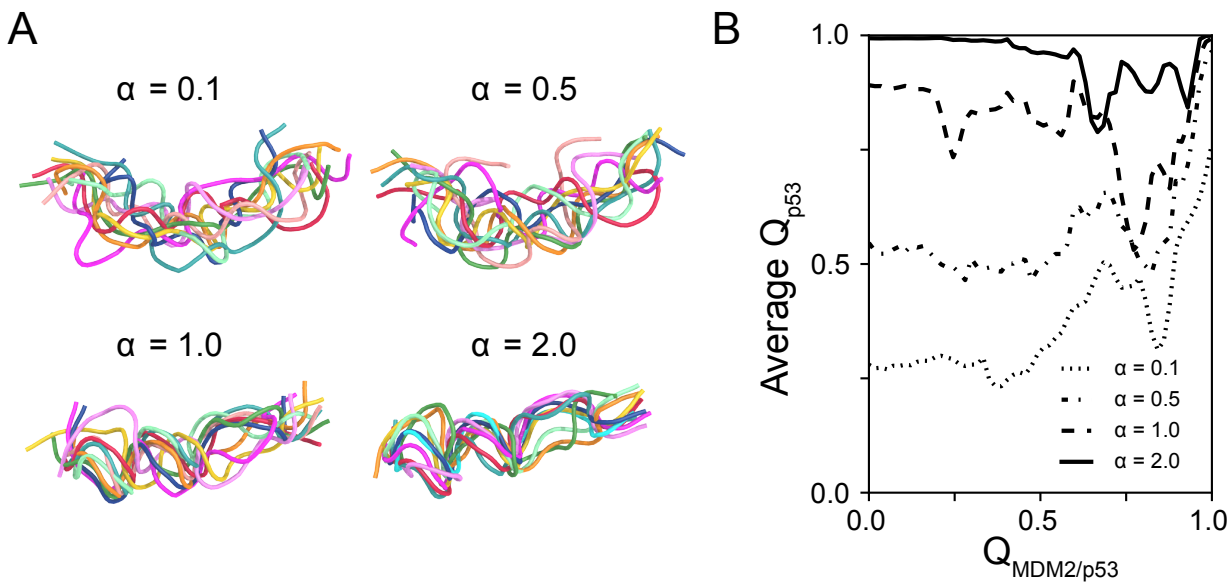

Figure 4: Tuning of the protein model to yield p53 peptide analogues with varying extents of preorganization. (A) Representative conformations of p53 peptide analogues that range from fully disordered ($\alpha = 0.1$) to fully preorganized ($\alpha = 2.0$). Conformations were sampled every 1 µs from 10 µs BF simulations of the corresponding unbound p53 peptide. (B) The fully disordered p53 analogue folds only upon binding the MDM2 protein as revealed by monitoring the average fraction of native contacts in the p53 peptide (Q_{p53}) as a function of the fraction of native contacts between MDM2 and p53 peptide ($Q_{\text{MDM2/p53}}$) for all of the p53 peptide analogues. Data shown for each p53 peptide analogue is based on 10 independent WE simulations.

3.4.1 Is There a Kinetic Advantage to Being Disordered vs Preorganized?

To directly compare the binding kinetics of the fully disordered p53 peptide relative to the other more preorganized analogues, it was essential to ensure that the fully disordered peptide was able to fold into an α-helical conformation upon binding MDM2. As shown by Figure 4B, all of the p53 peptide analogues are folded when bound to the MDM2 protein.

By construction, our model of the fully disordered peptide ($\alpha = 0.1$) results in an induced fit (folding-after-binding) mechanism[109] in which the peptide folds only upon binding MDM2 in our simulations; likewise, the fully preorganized peptide ($\alpha = 2.0$) results in a conformational selection (binding-after-folding) mechanism in which the peptide is fully folded before binding MDM2 in our simulations (Figure 4B).

For all of the p53 peptide analogues, ranging from fully disordered to fully preorganized, our simulations reveal that the mechanism of binding to the MDM2 receptor involves a two step process in which diffusive collisions of the binding partners first form a metastable "encounter" complex followed by rearrangement of the encounter complex to the native, bound state (Figure 4; Figure S9):

$$\mathrm{p_{53}\,peptide + MDM_2} \underset{\mathrm{k_{-1}}}{\overset{\mathrm{k_1}}{\rightleftharpoons}} \mathrm{encounter\,complex} \underset{\mathrm{k_{-2}}}{\overset{\mathrm{k_2}}{\rightleftharpoons}} \mathrm{bound\,state} \tag{3.7}$$

where k_1 is the rate constant for formation of the encounter complex, k_{-1} is the rate constant for the dissociation of the encounter complex to the unbound state, k_2 is the rate constant for rearrangement of the encounter complex to the bound state, and k_{-2} is the rate constant for rearrangement of the bound state to the encounter complex.

For our calculations of rate constants, we used the most stringent definitions of the encounter complex and bound state that encompassed the corresponding basins in the probability distributions of both the fully disordered and preorganized p53 peptides in Figure 5. The encounter complex was defined as those conformations satisfying the following criteria: (i) the binding partners are within van der Waals contact (< 6 Å), (ii) the C_α RMSD for the p53 peptide after alignment of MDM2 is > 2 Å, and (iii) at least one MDM2-p53 native contact is formed. The bound state was defined as having the binding partners within van der Waals contact and a C_α RMSD ≤ 2 Åof the p53 peptide after alignment of MDM2.

To assess whether there is a kinetic advantage to the peptide ligand being disordered or preorganized, we computed the k_{on} values of the exact ordered and disordered analogues using the NAM framework in conjunction with WE simulations (see Methods). As shown in Table 1, the ratio of the k_{on} for the fully disordered peptide relative to that of the fully

preorganized peptide is 0.9 ± 0.2 (uncertainties represent two SEM), with a percent uncertainty that amounts to only a 0.1 kcal/mol difference in the corresponding free energy barrier as estimated by $-RT\ln(k_{\mathrm{on}}^{\alpha=2.0}/k_{\mathrm{on}}^{\alpha=0.1})$. Thus, given the high precision of these computed values, any kinetic advantage to being disordered (or preorganized) is very small.

We next examined the extent to which ligand preorganization influences the individual steps of the binding process. The computed bimolecular rate constant for formation of the encounter complex, k_1, of the fully disordered p53 peptide is within error of that of its fully preorganized analogue with a ratio of 1.0 ± 0.1, indicating that being disordered (or preorganized) did not enable more rapid initial collisions.

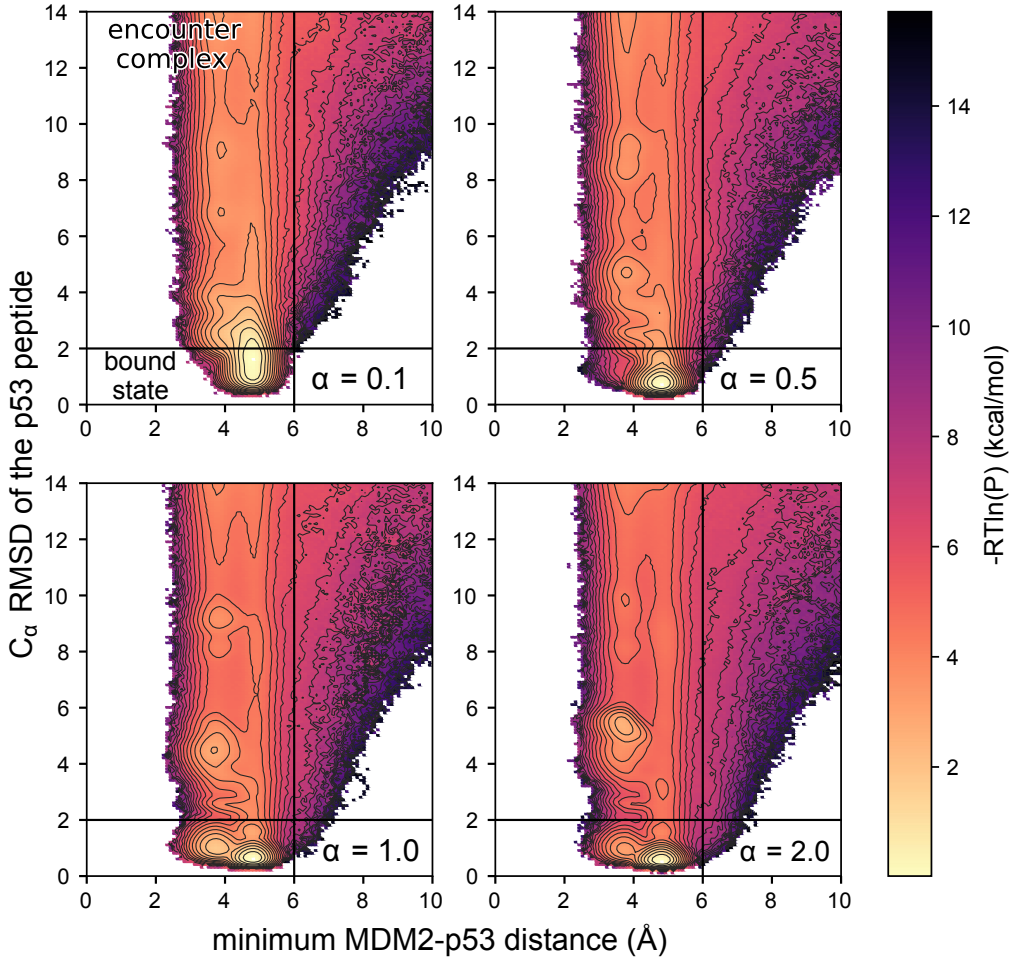

Figure 5: Zoomed-in views of probability distributions over the WE progress coordinate for various extents of structure in the p53 peptide, ranging from fully disordered ($\alpha = 0.1$) to fully preorganized ($\alpha = 2.0$) (for a representative full view of the probability distribution, see Figure S10). The progress coordinate consisted of the C_α RMSD of the p53 peptide after alignment of MDM2 from the crystal structure of the MDM2-p53 peptide complex29 and minimum MDM2-p53 distance. Definitions of the encounter complex and bound state are delineated by the solid black lines (for a representative full view of the probability distribution, see Figure S10). The color scale represents $-RT \ln P$ where P is the pseudoequilibrium probability density based on trajectory weights from each of 10 independent WE simulations that were carried out for the corresponding MDM2-p53 system (see Methods). Contour lines represent intervals of 0.5 kcal/mol.

					relative to $\alpha = 0.1$		
	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 1.0$	$\alpha = 2.0$	$\alpha = 0.5$	$\alpha = 1.0$	$\alpha = 2.0$
k_{on} ($10^7 M^{-1}s^{-1}$)	5.7 ± 0.6	5.7 ± 0.3	5.6 ± 0.6	5.1 ± 0.8	1.0 ± 0.1	1.0 ± 0.2	0.9 ± 0.2
k_1 ($10^7 M^{-1}s^{-1}$)	6.1 ± 0.5	6.1 ± 0.3	6.2 ± 0.6	5.9 ± 0.6	1.0 ± 0.1	1.0 ± 0.1	1.0 ± 0.1
lifetime of the encounter complex (ps)	80 ± 20	90 ± 30	130 ± 50	80 ± 20	1.1 ± 0.5	1.6 ± 0.8	1.0 ± 0.4
% productive collisions	65 ± 3	64 ± 6	68 ± 3	66 ± 2	1.0 ± 0.1	1.1 ± 0.1	1.0 ± 0.1
$D(10^{-6} cm^2/s)$	4.2 ± 0.7	3.9 ± 0.4	3.9 ± 0.5	4.0 ± 0.4	0.9 ± 0.2	0.9 ± 0.2	1.0 ± 0.2

Table 1: Computed k_{on}, k_1 for Formation of the Encounter Complex, Lifetime of the Encounter Complex, % Productive Collisions, and Relative Translational Diffusion Coefficients D for the MDM2-p53 Binding Process and p53 Peptide Analogues Ranging from Fully Disordered ($\alpha = 0.1$) to Fully Preorganized ($\alpha = 2.0$) in the Presence of Hydrodynamic Interactions (HIs). Data shown are averages from 10 independent WE simulations; uncertainties represent two SEM.

Given that native contacts are rewarded and non-native contacts are penalized in our simulation model (a Gō-type model), $k_{-2} \ll k_2$ such that the expression for the overall association rate constant is $k_{on} = (k_1 k_2/(k_{-1} + k_2))$. Since k_{on} and k_1 are within error of each other for all of the peptide analogues [e.g., for the fully disordered peptide, the k_{on} and k_1 are $(5.7 \pm 0.6) \times 10^7\ M^{-1}s^{-1}$ and $(6.1 \pm 0.5) \times 10^7\ M^{-1}s^{-1}$, respectively], the kinetics of the binding processes must be close to the limiting case where $k_{-1} \ll k_2$, such that $k_{on} = (k_1 k_2/(k_{-1} + k_2)) \cong k_1$.[109] The formation of the encounter complex is therefore rate-limiting for all of the p53 peptide analogues (k_2 was not computed since the hybrid WE/NAM approach permits calculation of bimolecular rate constants, but not unimolecular rate constants). Interestingly, the most preorganized peptide analogues ($\alpha = 1.0$ and $\alpha = 2.0$) undergo partial loss of structure upon forming the encounter complex (Figure 4B). This result suggests that the MDM2 receptor might aid the process of binding by disrupting preformed interactions within the p53 peptide that hinder rearrangement of the encounter complex to the bound state.

To gain further insight into the similarity in the k_{on} values among all of the p53 peptide

analogues, we calculated the percentage of productive collisions (i.e., those collisions that eventually reach the bound state) and the lifetime of the encounter complex. As shown in Table 1, the percentage of productive collisions for the fully disordered and fully preorganized p53 peptides are within error of each other (a ratio of 1.0 ± 0.1 for the percentage of productive collisions of the fully disordered peptide relative to that of the fully preorganized peptide) as are the lifetimes of the encounter complex (ratio of 1.0 ± 0.4). The high percentages of productive collisions (65 ± 3% and 66 ± 2% for the fully disordered and fully preorganized peptides, respectively) are consistent with our conclusion above that $k_{-1} \ll k_2$. Given that our simulations were carried out under steady-state conditions, generating pathways in only the binding direction, it was possible to obtain statistically robust estimates of nonequilibrium observables (e.g., rate constants and percentage of productive collisions), but not equilibrium observables (e.g., populations and lifetimes of the encounter complex), which would require sampling of unbinding as well as binding pathways. Nonetheless, since both the percentage productive collisions and lifetimes of the encounter complex are similar for the fully disordered and fully preorganized peptides, k_{-1} as well as k_2 must be similar for the peptides. Thus, the folding of the fully disordered p53 peptide upon binding MDM2 does not appear to affect k_2 relative to that of the fully preorganized peptide. It is worth noting that the k_2 step may be slower in all-atom simulations due to attractive non-native interactions that are missing in our Gō-type simulations and that such nonnative interactions would likely result in additional benefits to the p53 peptide being preorganized relative to being disordered.

Our computed k_{on} values are within error of the computed k_{on} from atomistic simulations [$(7 \pm 4) \times 10^7 M^{-1}s^{-1}$] 11 and 6× faster than the experimental value ($9.2 \times 10^6 M^{-1}s^{-1}$).[110] Thus, while the use of the Gō-type potential energy function[98,99] would be expected to artificially accelerate the dynamics,[111,112] the inclusion of appropriately parametrized HIs yields realistic rate constants.[100] In particular, the computed relative translational diffusion coefficients for MDM2 and the p53 peptide for all of the peptide analogues are in excellent agreement with that predicted for the corresponding all-atom models by the hydrodynamics program HYDROPRO,[113] $3.7 \times 10^{-6} cm^2/s$. As others have shown,[100] the translational diffusion coefficients of proteins are underestimated in molecular simulations that neglect

HIs in our case, by 10× (Table 1; Table 2) underscoring the importance of including HIs in simulations that lack explicit solvent.[100] Interestingly, the extent of structure in the p53 peptide has no significant effect on the relative translational diffusion coefficient for the p53 peptide and MDM2 protein.

3.4.2 Effect of Including Hydrodynamic Interactions (HIs)

The inclusion of HIs in our simulations increased the k_{on} by 30× (Table 1; Table 2). This result may appear at odds with previous simulation studies of protein-protein associations in which the inclusion of HIs was found to slow down the approach of the proteins.[8,114] However, our results are in fact consistent with these studies since the effect of including HIs on the k_{on} depends on the extent to which the intramolecular and intermolecular HIs have opposing effects on the diffusion of the binding partners. Whereas intramolecular HIs speed up the diffusion of binding partners that have no interactions with each other, yielding larger translational diffusion coefficients, intermolecular HIs slow down the diffusion of the binding partners when they are close to one another and have the tendency to move together. Our results involving the MDM2–p53 system reveal that the net effect of including both intramolecular and intermolecular HIs is a faster k_1 as well as slower dissociation of the encounter complex (k_{-1}), the latter being evident from longer lifetimes of the encounter complex and a greater percentage of productive collisions.

3.4.3 Effect of Increasing Receptor Concentration

As demonstrated by previous experimental studies, the mechanism by which a small organic ligand binds a disordered protein receptor shifts from conformational selection to induced fit with increasing ligand concentration.[94,95] Here, we examined the effects of increasing the concentration of a protein receptor (MDM2) on its mechanism of binding to a disordered peptide (p53 peptide), i.e., the relative fluxes through conformational selection and induced-fit mechanisms (Figure 63A).

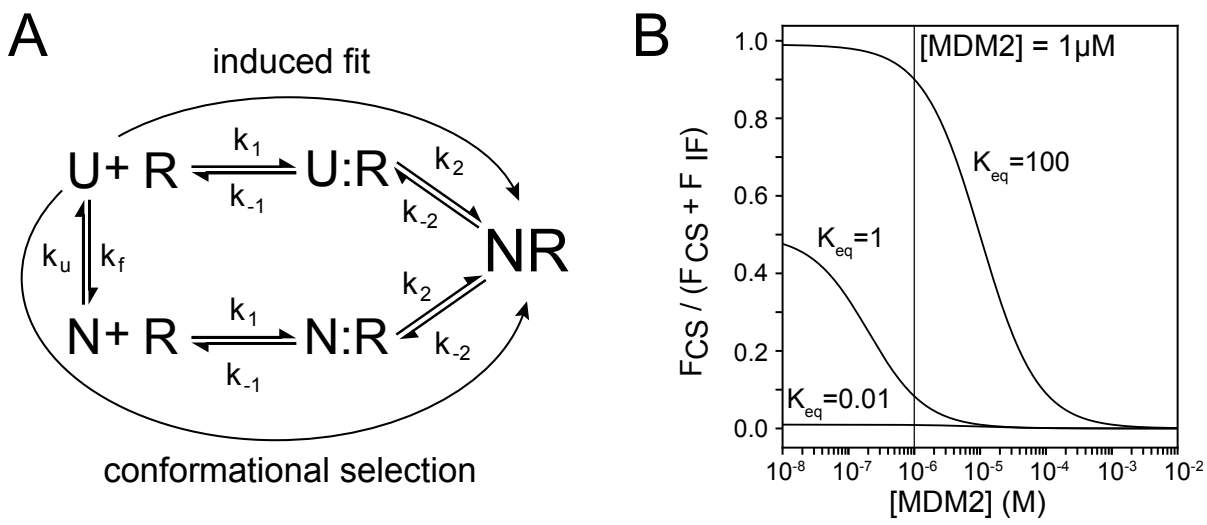

Figure 6: Conformational selection and induced-fit mechanisms of binding, and the effects of increasing receptor concentration. (A) Conformational selection and induced fit mechanism of binding for an IDP ligand and its folded receptor. N is the folded (fully preorganized) state of the IDP, U is the unfolded (fully disordered) state of the IDP, and R is the receptor, U:R and N:R are the encounter complexes resulting from diffusional collisions of the unfolded and folded states, respectively, with the receptor, and NR is the native, bound conformation. (B) Fractional flux through conformational selection (CS) for the binding process as a function of receptor (MDM2) concentration. Given that the equilibrium constant K_{eq} of the IDP ($K_{\mathrm{eq}} = k_{\mathrm{f}}/k_{\mathrm{u}}$) is not known, the fractional flux is estimated for three K_{eq} values (0.01, 1, and 100). The black line represents the [MDM2] used in an experimental study of the MDM2-p53 peptide binding mechanism.[110]

Given that the computed k_{on} k_1 for all of the p53 peptide analogues in this study, the binding mechanism for the MDM2/p53 peptide system can be approximated as a twostep mechanism with a very fast second step (the k_2 step; Figure S12) such that the fractional flux can be calculated using the following equation:

$$\frac{\mathrm{F_{CS}}}{\mathrm{F_{CS}} + \mathrm{F_{IF}}} = \frac{k_{\mathrm{f}}}{(k_{\mathrm{u}} + k_{\mathrm{f}}) + k_{\mathrm{on}}[\mathrm{R}]} \tag{3.8}$$

where k_{on} is set to an order-of-magnitude estimate ($10^7\ M^{-1}s^{-1}$) since the computed k_{on} values are essentially the same for the fully disordered and fully preorganized p53 peptides; FCS and FIF are the fluxes through the conformational selection and induced-fit mechanisms, respectively; [R] is concentration of the folded receptor (MDM2); as shown in Figure 6A, k_f is the rate constant for folding of the ligand (p53 peptide) from the fully disordered, unfolded (U) state, and k_U is the rate constant for unfolding of the ligand from the fully preorganized, native folded (N) state. Thus, in this scenario, the fractional flux through conformational selection depends only on the concentration of the receptor and is therefore independent of ligand concentration. A detailed derivation of eq 7 can be found in the Supporting Information.

Since the equilibrium constant K_{eq} (ratio of $k_f\ /k_U$) for the folding of the isolated p53 peptide is not known, we tested three different scenarios: (i) $K_{eq} = 1$ for equally stable unfolded and folded states, (ii) $K_{eq} = 100$ for an unfolded state that is much less stable than the folded state, and (iii) $K_{eq} = 0.01$ for an unfolded state that is much more stable than the folded state (Figure 6B). When the folded state is much less stable than the unfolded state ($K_{eq} = 0.01$), the mechanism of binding would be solely induced fit, regardless of MDM2 concentration. Substantial flux through conformational selection would be expected only when the folded state is equal or greater in stability to the unfolded state ($K_{eq} \geq 1$). For example, if $K_{eq} = 1$, 10% flux through conformational selection would be expected at the MDM2 concentration (1 µM) in binding kinetics experiments.[110] In the regime where $K_{eq} \geq 1$, the mechanism of binding is predicted to shift from being dominated by conformational selection to induced fit with increasing MDM2 concentration (Figure 6B). These results are consistent with those from atomistic simulations in which a Markov state model[115,116] was constructed to estimate rate constants for the MDM2-p53 peptide binding process and relative fluxes through conformational selection and induced fit were estimated (i) using a mechanism consisting of four instead of the three states used here and (ii) for various extent of helical content of the p53 peptide, which is analogous to varying K_{eq} values for the unfolding/folding equilibrium of the peptide.[93] In particular, the dominant binding mechanism becomes induced fit as the concentration of MDM2 increases and the extent of helical content decreases (or K_{eq} decreases).

3.5 DISCUSSION

To our knowledge, the only other study that has directly compared the binding kinetics of an IDP relative to its exact preorganized analogue is a simulation study that focused on the binding of the disordered pKID domain to its partner protein, KIX.[92] In this study, the disordered pKID domain was found to have a modest kinetic advantage ($>> 2.5x$) for binding relative to the preorganized analogue due to a more rapid k_2 step, which corresponds to the rearrangement of the encounter complex to the native, bound state. In contrast, our study yielded similar computed k_{on} values for the disordered and preorganized analogues of the p53 peptide in binding the MDM2 protein, revealing that the folding of the disordered p53 peptide upon binding MDM2 is very fast such that the k_2 step is just as rapid as that of the preorganized analogue.

As noted above, the pKID domain is significantly larger than the p53 peptide: upon binding its partner protein, the pKID domain adopts two α-helices while the p53 peptide adopts only a single α-helix. Given its larger size, the folding of the fully disordered pKID domain is slower and may therefore have a more significant influence on k_2. In particular, since the fully disordered pKID consists of two segments, the folding of the domain can take advantage of a dock-and-coalesce mechanism[81] in which the docking of one segment facilitates the folding process in the k_2 step.

The fact that our computed k_1 values for the formation of the encounter complex are the same for the disordered and preorganized p53 peptides indicates that the MDM2-p53 binding process does not involve the "fly-casting" mechanism in which the disordered peptide would be predicted to collide more rapidly with its partner protein due to a greater capture radius.[80] The lack of a fly-casting effect in our molecular simulations is underscored by our use of a Gō-type potential, which creates the optimal scenario for capturing the effect, i.e., the fully disordered p53 peptide folds only upon binding (forming $\geq$70% of intramolecular p53 native contacts only upon forming $\geq$98% of intermolecular MDM2-p53 native contacts; Figure 4B). Furthermore, we observed no differences in the capture radius of the fully disordered p53 peptide relative to its fully preorganized analogue as quantified by the radius of gyration R_g (most probable values of 7.7 and 7.3 Å, respectively) as well as a more sensitive

metric, the maximum principal axis radius RM (6.6 and 6.9 Å, respectively; see Figure S11), despite the fact that the disordered conformations were generated with no rewarding of native contacts. The lack of differences in the capture radius and therefore the hydrodynamic radius is consistent with the fact that the computed translational diffusion coefficients of the fully disordered and fully preorganized p53 peptides are indistinguishable from each other (Table 1). Regardless, based on the Stokes-Einstein equation in which the translational diffusion coefficient is inversely proportional to the hydrodynamics radius, any kinetic advantage that could result from a larger capture radius (and therefore hydrodynamics radius) of the disordered peptide relative to its preorganized analogue might be canceled out by the effects of a slower translational diffusion coefficient.

3.6 CONCLUSIONS

We have determined the effects of preorganization of the intrinsically disordered, N-terminal p53 peptide on the kinetics of binding its partner protein, MDM2, using molecular simulations. In particular, our application of the WE strategy enabled the generation of > 3000 of binding events, yielding statistically robust k_{on} values for the fully disordered p53 peptide and exact analogues of the peptide that have been preorganized to various extents.

The resulting computed k_{on} values are in reasonable agreement with experiment. Notably, the k_{on} for the fully disordered p53 peptide is within error of that for its fully preorganized analogue, indicating no kinetic advantage to being disordered or preorganized for binding MDM2. Given that the rate constant k_1 for formation of the encounter complex is essentially the same for the fully disordered and fully preorganized peptides, fly-casting is not a significant effect in our simulations of the MDM2-p53 peptide system, even though the ideal scenario for this effect was modeled, i.e., using a Gō-type potential that ensured folding of the fully disordered peptide only upon binding MDM2. Furthermore, since the percentages of productive collisions and lifetimes of the encounter complex are similar for the fully disordered and preorganized p53 peptides, the rate constant k_2 for rearrangement of the encounter complex to the bound state must also be similar. Thus, folding of the fully disordered p53 peptide upon binding MDM2 during the k_2 step must be very rapid. In contrast, the slower folding of larger IDPs may have a more significant effect on k_2 relative to that for their fully preorganized analogues, as predicted for the pKID domain[92] and by the dock-and-coalesce mechanism.[81] Interestingly, the two most preorganized p53 peptide analogues undergo partial loss of structure upon forming the encounter complex, implying that the MDM2 receptor might "erase" preformed interactions within the p53 peptide that hamper the k_2 step.

Finally, based on our k_{on} values, we determined the effect of increasing the concentration of MDM2 on its mechanism of binding to the disordered p53 peptide ligand. When the unfolded state is much less stable than the folded state of the isolated p53 peptide, the mechanism for the binding of the MDM2 receptor to the disordered p53 peptide is predicted to switch from being dominated by conformational selection to induced-fit with increasing

concentration of MDM2. On the other hand, when the unfolded state is either equal to or much greater in stability than the folded state, the mechanism of binding is solely induced fit, regardless of the MDM2 concentration. These results are consistent with those from recent atomistic simulations of the binding process involving the MDM2 receptor and p53 peptide ligand.[93] Given the general features of our residue-level simulation models, results from our molecular simulations are relevant to any protein binding process involving a disordered peptide of a similar length to the p53 peptide that folds into a single α-helix upon binding its partner protein. Such disorder-to-helix transitions are common among molecular recognition events, including protein interactions of IDPs that play crucial cellular roles.[79,96,117] Our results provide a valuable set of simulation data for testing future hypotheses that might be proposed for the binding mechanisms of IDPs and their preorganized analogues.

3.7 ACKNOWLEDGEMENTS

This work was supported by NSF CAREER Award MCB0845216 and NIH 1R01GM115805-01 to L.T.C.; a DAAD graduate research grant to A.S.S.; University of Pittsburgh James V. Harrison Fund and Honors College Brackenridge Research Fellowships to D.W.W.; and University of Pittsburgh Arts & Sciences and Mellon Fellowships to M.C.Z. Computational resources were provided by NSF CNS-1229064 and the University of Pittsburgh's Center for Research Computing. We thank Karl Debiec for making available his myplotspec Python library for generating plots, and Adam Pratt, Alex DeGrave, Adrian Elcock (University of Iowa), and Thomas Kiefhaber (Martin Luther University Halle-Wittenberg) for helpful discussions.

3.8 SUPPORTING INFORMATION

3.8.1 SI Figures

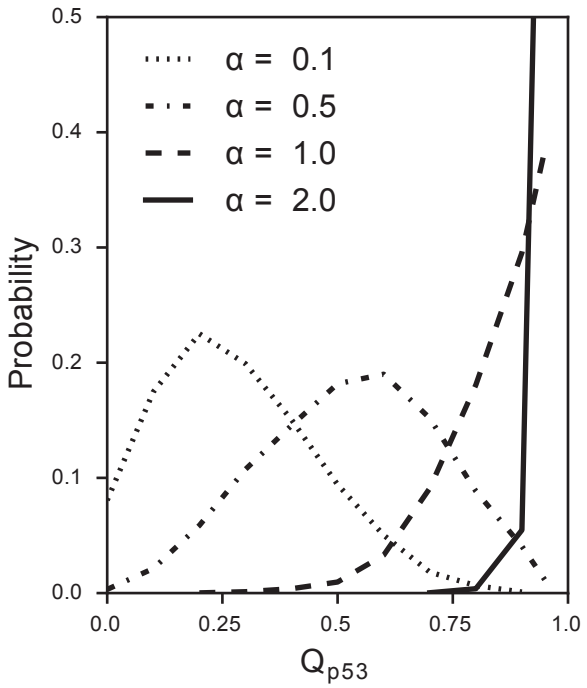

Figure 7: Probability distributions of the fraction of native contacts of the p53 peptide (Q_{p53}) in the absence of MDM2, ranging from fully disordered ($\alpha = 0.1$) to fully preorganized ($\alpha = 2.0$). Distributions for each value of the scaling factor α were generated using conformations sampled every 100 ps from a single 10 µs standard simulation starting from the MDM2-bound conformation.

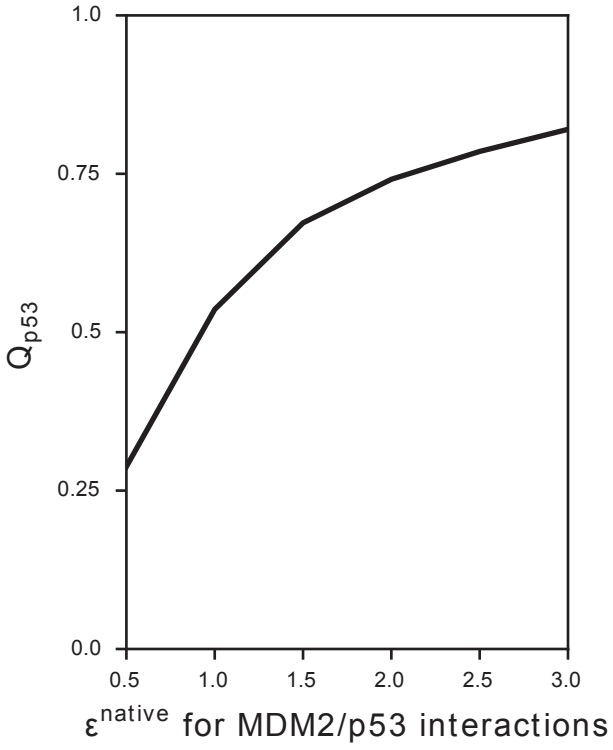

Figure 8: Average fraction of native contacts in p53 (Q_{p53}) as a function of the $\varepsilon^{\mathrm{native}}$ for MDM2-p53 native contacts. For each $\varepsilon_{\mathrm{native}}$ value, the average Q_{p53} was calculated using conformations sampled every 100 ps from a single 10 µs standard simulation of the MDM2-p53 peptide complex with the fully disordered peptide ($\alpha = 0.1$) starting from the MDM2-bound conformation.

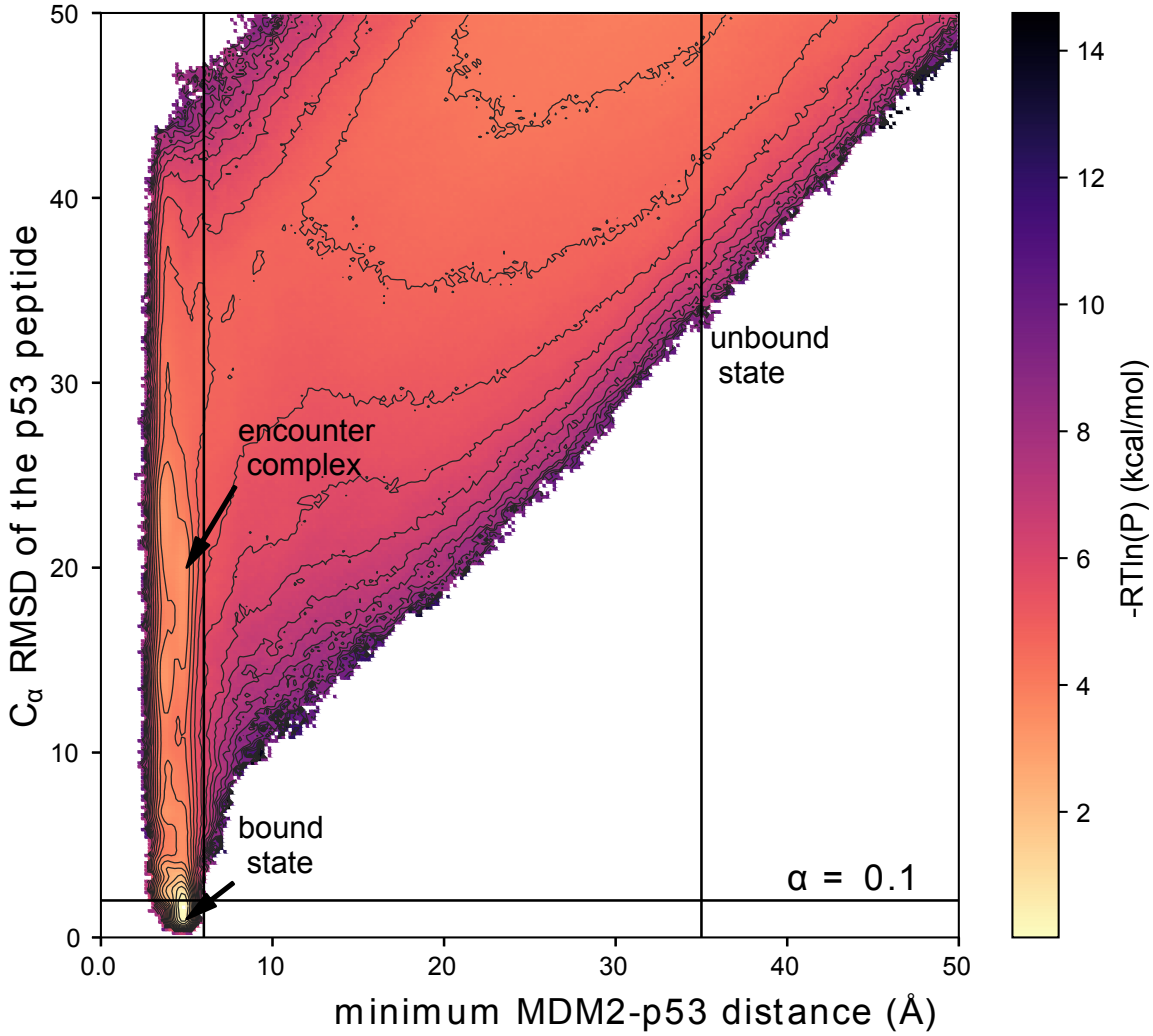

Figure 9: Full view of the free energy landscape of the MDM2-p53 binding process as a functionof the C_α RMSD of the p53 peptide after alignment of MDM2 from the crystal structure of the MDM2-p53 peptide complex 1 and the minimum MDM2-p53 distance for the fully disordered p53 peptide ($\alpha = 0.1$). Data shown is based on conformations sampled every 1 ps from 10 independent WE simulations under steady-state conditions. Contour lines represent intervals of 0.5 kcal/mol.

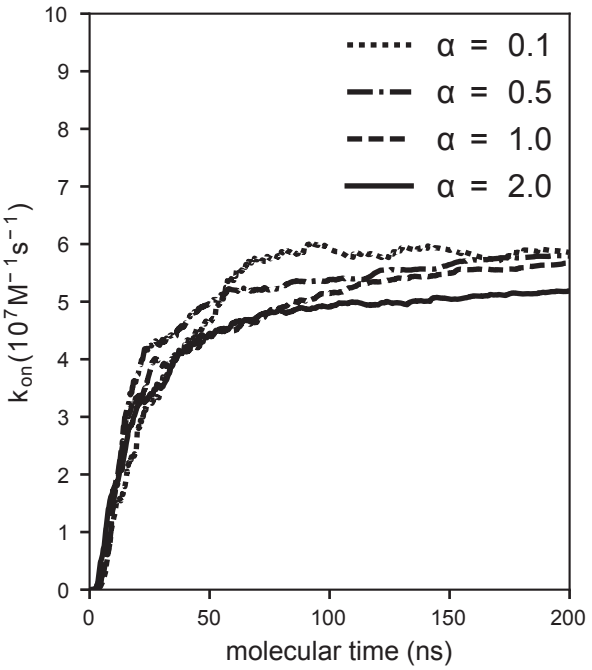

Figure 10: Computed k_{on} as a function of WE iteration for p53 peptide analogues with various extents of structure, ranging from fully disordered ($\alpha = 0.1$) to fully preorganized ($\alpha = 2.0$). The molecular time is defined as Nτ where N is the number of WE iterations and τ is the fixed time interval of each iteration.

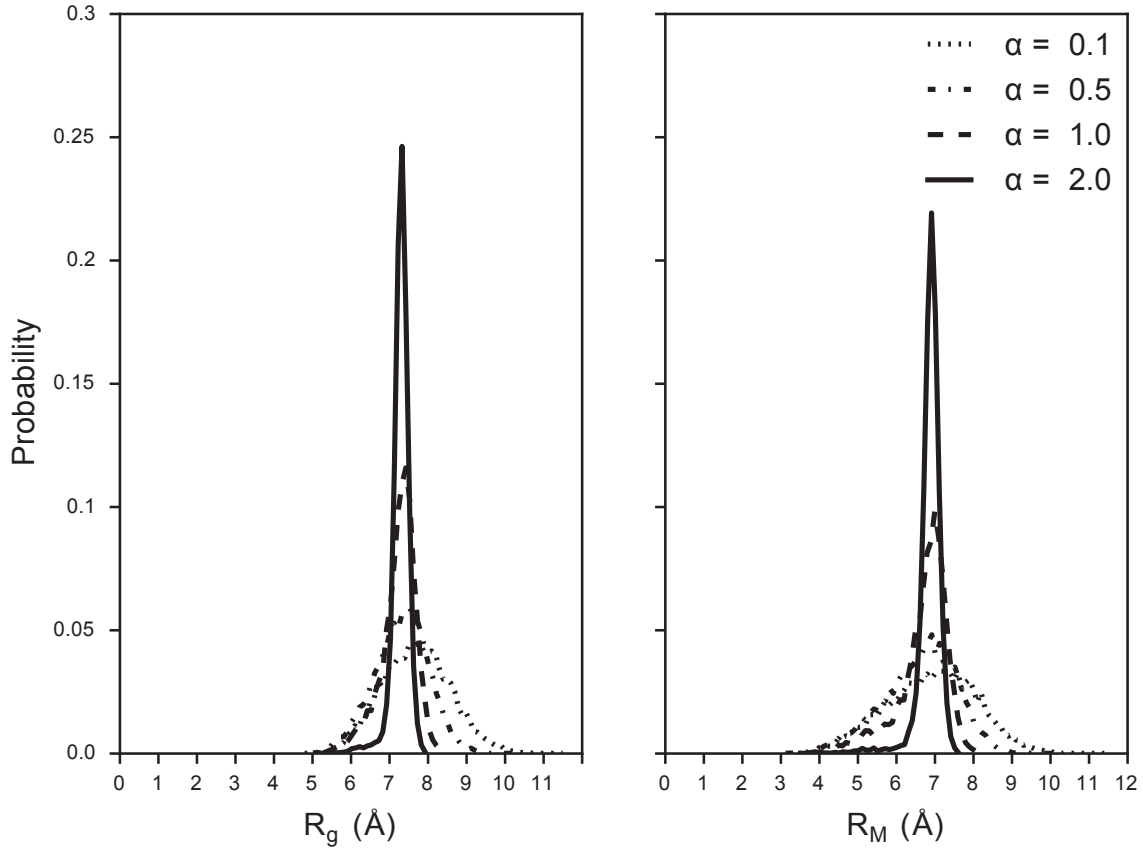

Figure 11: Probability distributions of the "capture" radius for the p53 peptide with various extents of structure, ranging from fully disordered ($\alpha = 0.1$) to fully preorganized ($\alpha = 2.0$), as monitored by the radius of gyration R_{g}, and maximum principal axis radius R_{M}. For each a value, the probability distribution was calculated from a 10 µs BF simulation of the isolated peptide. Based on the R_{g} metric, it may appear that the fully disordered p53 peptide achieves a significantly larger maximum value than that of the fully preorganized peptide (11.6 Å vs. 8.0 Å). However, the more sensitive R_{M} metric reveals that the fully disordered p53 peptide not only assumes more expanded conformations (maximum value of 10.2 Å), but also more contracted conformations (minimum value of 3.3 Å).

α (without HI)	0.1	0.5	1.0	2.0
k_{on} ($10^6 M^{-1}s^{-1}$)	1.5 ± 0.4	2.0 ± 0.3	1.4 ± 0.4	1.3 ± 0.3
$k_1(10^6 M^{-1}s^{-1})$	1.8 ± 0.3	2.1 ± 0.3	1.7 ± 0.4	1.3 ± 0.2
lifetime of the encounter complex (ps)	50 ± 10	90 ± 20	60 ± 20	80 ± 20
% productive collisions	45 ± 6	49 ± 4	43 ± 5	49 ± 6
$D(10^{-6} cm^2/s)$	0.4 ± 0.1	0.4 ± 0.1	0.4 ± 0.1	0.4 ± 0.1

Table 2: Computed k_{on}, k_1 for formation of the encounter complex, lifetime of the encounter complex, % productive collisions, and relative translational diffusion coefficients for the MDM2-p53 binding process and various analogues of the p53 peptide, ranging from fully disordered ($\alpha = 0.1$) to fully preorganized ($\alpha = 2.0$) in the absence of hydrodynamic interactions (HIs). Data shown are averages from 10 independent WE simulations; uncertainties represent 95% confidence intervals.

3.8.2 SI Methods

3.8.2.1 Calculation of the "capture" radius. To quantify the extent that the p53 peptide can reach out to contact its partner protein —termed the "capture radius" — we computed the radius of the longest principal axis of an approximate ellipsoid surrounding the peptide. A radius of gyration tensor was first constructed as follows:

$$R = \begin{bmatrix} \sum x_n^2 & \sum x_n y_n & \sum x_n z_n \\ \sum y_n x_n & \sum y_n^2 & \sum y_n z_n \\ \sum z_n x_n & \sum z_n y_n & \sum z_n^2 \end{bmatrix} \tag{3.9}$$

where $\mathbf{R}$ is the gyration tensor, and x_n,y_n,z_n are the coordinates of the nth pseudoatom assuming the center of geometry is located at the origin. The eigenvalues of $\mathbf{R}$, $\{\lambda_1,\ \lambda_2,\ \lambda_3\}$, give the principal moments of the gyration tensor along the principal axes of the peptide. Assuming $\lambda_3 > \lambda_2 > \lambda 1$, then the radius of the longest principal axis, R_M, is given by:

$$R_M = 2\sqrt{\lambda_3} \tag{3.10}$$

Probability distributions of R_M as well as the radius of gyration R_g for each p53 peptide analogue were calculated using conformations sampled every ns from a 10 µs BF simulation of the peptide in its unbound, isolated state.

3.8.2.2 Derivation of equation for fractional flux through conformational selection. Given that the computed $k_{on} \cong k_1$ for both the fully disordered (unfolded) and fully preorganized (folded) p53 peptide analogues, we can approximate the corresponding binding mechanisms with those shown in Fig. S12.

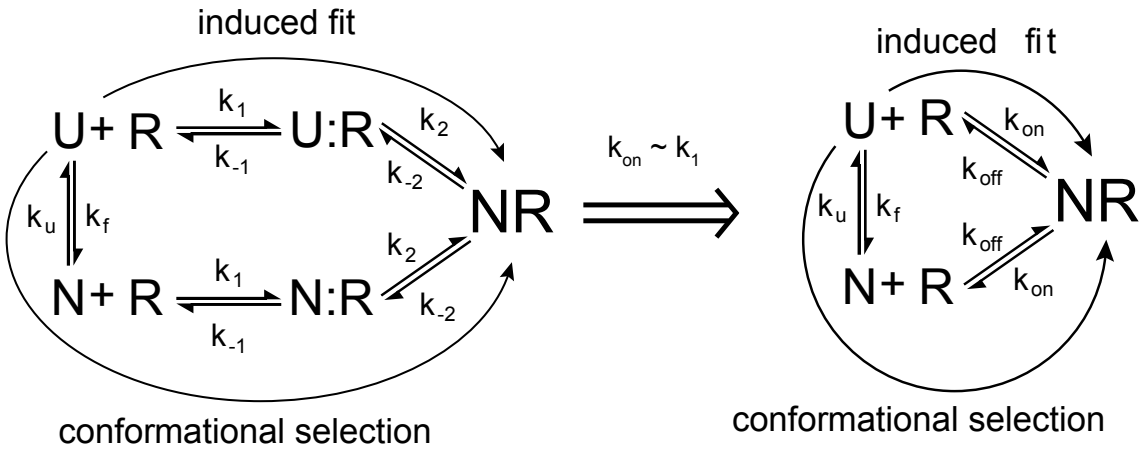

Figure 12: Approximation of the binding mechanism of the disordered p53 peptide ligand to the folded MDM2 protein receptor when $k_{on} \cong k_1$. States are defined as in Fig. 6A.

As done by others[95], fluxes through the conformational selection and induced-fit mechanisms can be calculated using the following equations for parallel and serial paths:
Parallel reaction paths : $F_{total} = \sum F_i$
Serial reaction paths : $F_{total} = \sum \frac{1}{1/F_i}$
where F_i is the flux of the ith reaction step. The fluxes through conformational selection and induced fit (F_{CS} and F_{IF}, respectively) are therefore predicted by the following:

$$F_{CS} = (\frac{1}{k_f[U]} + \frac{1}{k_{on}[N][R]})^{-1} = (\frac{k_{on}[N][R] + k_U[U]}{k_f k_{on}[N][R][U]})^{-1} = \frac{k_f k_{on}[N][R][U]}{k_{on}[N][R] + k_f[U]}$$

$$F_{IF} = k_{on}[U][R]$$

where U is the unfolded state of the IDP ligand, N is the native, folded state of the IDP ligand, and R is the folded receptor; k_f is the rate constant for folding of the ligand from the unfolded (fully disordered) state; and the k_{on} is the association rate constant, which was found to be essentially the same value for both the unfolded (U) and folded (N) analogues of the p53 peptide. We can then derive the expression for predicting the fractional flux $[F_{CS}/(F_{CS} + F_{IF})]$ through conformational selection:

$$\begin{aligned} F_{CS} + F_{IF} &= \frac{k_f k_{on}[N][R][U]}{k_{on}[N][R] + k_f[U]} + k_{on}[U][R] \\ &= \frac{k_f k_{on}[N][R][U] + k_{on}^2[U][N][R]^2 + k_f k_{on}[R][U]^2}{k_{on}[N][R] + k_f[U]} \end{aligned}$$

$$\begin{aligned} \frac{F_{CS}}{F_{CS} + F_{IF}} &= \frac{\frac{k_f k_{on}[N][R][U]}{k_{on}[N][R]+k_f[U]}}{\frac{k_f k_{on}[N][R][U]+k_{on}^2[U][N][R]^2+k_f k_{on}[R][U]^2}{k_{on}[N][R]+k_f[U]}} \\ &= \frac{k_f k_{on}[N][R][U]}{k_f k_{on}[N][R][U] + k_{on}^2[U][N][R]^2 + k_f k_{on}[R][U]^2} \\ &= \frac{k_f[N]}{k_f[N] + k_{on}[N][R] + k_f[U]} \end{aligned}$$

Since U and N are in equilibrium (i.e. $k_f[U] = k_U[N]$), we can replace $k_f[U]$ in the denominator:

$$\frac{F_{CS}}{F_{CS} + F_{IF}} = \frac{k_f[N]}{k_f[N] + k_{on}[N][R] + k_U[N]} = \frac{k_U}{(k_f + k_U) + k_{on}[R]}$$

Thus, the expression we have used to calculate the fractional flux through conformational selection is the following:

$$\frac{F_{CS}}{F_{CS} + F_{IF}} = \frac{k_U}{(k_f + k_U) + k_{on}[R]}$$

4.0 HIGHLY EFFICIENT COMPUTATION OF BASAL K_{ON} USING DIRECT SIMULATION OF PROTEIN-PROTEIN ASSOCIATION WITH FLEXIBLE MOLECULAR MODELS

The text in this chapter has been adapted from Ali S. Saglam and Lillian T. Chong, *J. Phys. Chem. B* **2016**, *120*, 117-122.

4.1 CHAPTER SUMMARY

An essential baseline for determining the extent to which electrostatic interactions enhance the kinetics of protein-protein association is the "basal" k_{on}, which is the rate constant for association in the absence of electrostatic interactions. However, since such association events are beyond the milliseconds timescale, it has not been practical to compute the basal k_{on} by directly simulating the association with flexible models. Here, we computed the basal k_{on} for barnase and barstar, two of the most rapidly associating proteins, using highly efficient, flexible molecular simulations. These simulations involved a) pseudo-atomic protein models that reproduce the molecular shapes, electrostatic, and diffusion properties of all-atom models, and b) application of the weighted ensemble path sampling strategy, which enhanced the efficiency of generating association events by >130-fold. We also examined the extent to which the computed basal k_{on} is affected by inclusion of intermolecular hydrodynamic interactions in the simulations.

4.2 INTRODUCTION

Of fundamental interest to biology is the extent to which electrostatic interactions enhance the rate of protein-protein association. An essential baseline for determining the magnitude of these rate enhancements is the "basal" k_{on}, which is the rate constant for association in the absence of electrostatic interactions.[118] In principle, the basal k_{on} should be measured in the same solvent environment using the hydrophobic isosteres - that is, hypothetical mutants with molecular shapes that are identical to those of the wild-type proteins, but are entirely uncharged. However, due to the difficulty of engineering hydrophobic isosteres, experimental studies have instead estimated the basal k_{on} by measuring the k_{on} for the wild-type proteins at various salt concentrations and extrapolating to the limit of infinite salt concentration where electrostatic interactions would be completely screened.[118]

An alternative approach is to construct the exact hydrophobic isosteres *in silico* by setting all partial charges of the wild-type proteins to zero and directly computing the basal k_{on} by simulating the association of the hydrophobic isosteres. Ideally, such simulations would involve the use of flexible molecular models in order to capture conformational changes during the association process. However, since the weak associations of completely hydrophobic proteins are beyond the milliseconds timescale,[118–124] it has only been feasible to directly compute the basal k_{on} using rigid, models with atomically detailed simulations.[119] Theoretical estimates of the basal k_{on} have also been made using spherical models with orientational constraints[120–123] and applications of transition-rate theory to rigid, atomistic models.[124]

Here, for the first time, we directly computed the basal k_{on} for a protein-protein association process using flexible models with molecular simulations. We focused on barnase and barstar, which are among the most rapidly associating proteins by virtue of their complementary electrostatic surfaces.[119] Our simulations employed flexible, pseudo-atomic protein models of barnase and barstar that were designed by Frembgen-Kesner and Elcock[114] to retain the molecular shapes, electrostatic potentials, and diffusion properties of the corresponding atomistic models at the experimental ionic strength (50 mM).[125] The same authors have demonstrated that the use of these models with standard "brute force" simulations can reproduce the experimental k_{on} values of both the wild-type and mutant protein pairs. How-

ever, they were unable to carry out such simulations to obtain a statistically robust estimate of the k_{on} for the hydrophobic isosteres (i.e. the basal k_{on}) due to the large computational cost.[114]

A critical feature of our study is the application of the weighted ensemble (WE) strategy[25] to enhance the sampling of rare events, e.g. the slow association of completely hydrophobic, uncharged proteins. Although the WE strategy has been previously applied to protein binding processes using Brownian dynamics (BD) simulations,[25,108] these studies were carried out without the inclusion of HIs) between, and within, the diffusing proteins. In the absence of explicit solvent, it has been demonstrated that the translational and rotational diffusion coefficients of flexible protein models are drastically underestimated unless intramolecular HIs are included in the simulations.[100] In addition, the neglect of *intermolecular* HIs in previous BD studies of protein binding processes[114 119,122,126] is likely to have contributed to their consistent overestimation of association rate constants.[114] Importantly, our simulations were validated by computing the k_{on} values for both wild-type barnase and its R59A mutant, which associates more slowly than wild-type barnase with barstar,[125] and comparing the computed values to experiment.

4.3 METHODS

4.3.1 The protein model and energy function

The wild-type and mutant pairs of barnase and barstar were represented using flexible, pseudo-atomic models developed by Frembgen-Kesner and Elcock.[114] Full details of these models are provided in ref[114]. Briefly, the generation of these models began with all-atom models of the wild-type proteins, which were based on the crystal structure of the barnase-barstar complex (PDB code: 1BRS);[127] the same models were used for both the unbound and bound states. Approximately one pseudo-atom was then used to represent every three amino acid residues (33 pseudo-atoms for the 110 residues of barnase and 27 pseudo-atoms for the 89 residues of barstar). For the wild-type proteins and R59A mutant barnase, the effective

charge method[126] was used to derive effective charges for the pseudo-atomic models such that the electrostatic potential of the corresponding all-atom model was reproduced. Electrostatic potentials were obtained by numerically solving the non-linear Poisson-Boltzmann equation under experimental conditions (pH 8, 25 °C, and ionic strength of 50 mM).[125] Pseudo-atoms were then positioned and sized to replicate the electron density envelope of the all-atom model. To generate models of the exact hydrophobic isosteres of barnase and barstar, we started with the pseudo-atomic models of the wild-type proteins and set all effective charges to zero.

The energy function consisted of a single intramolecular term involving flexible, harmonic bonds between the pseudo-atoms and intermolecular terms for electrostatic and non-electrostatic interactions. To maintain the molecular shapes of the proteins, three bonds per pseudo-atom were formed on average. All intermolecular electrostatic interactions between pseudo-atoms were calculated using the Debye-Hückel equation; intramolecular electrostatic interactions were omitted. Non-electrostatic interactions were calculated using a very weak Gō-type potential energy function with a shallow well depth ($\epsilon = 0.1$ kcal/mol). Thus, native contacts were only slightly rewarded by a weakly attractive Lennard-Jones-like potential and nonnative contacts were penalized by a purely repulsive potential.[98,99] The well-depth was kept at a minimal value in order to avoid implicitly double counting the attractive electrostatic interactions, which are assumed to be a primary driving force for the formation of the barnase-barstar complex.[125] Two pseudo-atoms were considered to form a native contact if any non-hydrogen atoms of the residues in the all-atom model are within 5.5 Å of each other in the crystal structure of the native complex, yielding a total of 34 intermolecular native contacts.

4.3.2 Weighted ensemble (WE) simulations

All simulations were carried out using the WE path sampling strategy,[25] as implemented in the WESTPA software package (https://westpa.github.io/westpa).[50] In this strategy, a large number of simulations, or trajectory "walkers", are started in parallel from the initial state and iteratively evaluated at fixed time intervals τ for resampling in which walkers

are either replicated or combined to maintain a similar number of walkers per bin along a progress coordinate towards the target state. Rigorous management of the statistical weights associated with each walker ensures that no bias is introduced into the dynamics.

In this study, the WE strategy was applied using steady-state simulations within the framework of the Northrup-Allison-McCammon (NAM) method.[104] This framework involves the definition of two concentric spherical surfaces with radii b and q that correspond to center-to-center separation distances for barnase and barstar. The inner sphere, or b surface, represents the initial, unbound state, and the outer sphere, or q surface, is an absorbing surface that is positioned at a much larger separation distance ($q \gg b$) to avoid wasting computational effort sampling the indefinite drifting apart of the proteins. Each WE simulation was started from 24 configurations of the unbound state in which barnase and barstar were randomly oriented at a center-to-center separation distance of b. A walker was continued until the pair of proteins either exceeded a separation distance q or satisfied the criterion for the target state for successful association, i.e. reaching a threshold value, Q_{rxn}, in the fraction of native intermolecular contacts, Q, that reproduces the experimental k_{on} for the wild-type proteins. Consistent with previous brute force simulations,[114] b and q were set to 100 and 500 Å, respectively. Upon reaching the q surface, a walker was "recycled" by starting a new walker from the unbound state with the same statistical weight thereby maintaining a steady state and enforcing a constant effective protein concentration (3.2 μM). Upon reaching a particular Q_{rxn} value, a walker was effectively recycled after completing the WE simulation by removing the walker and its replicas prior to calculating the k_{on}.

For each barnase-barstar pair, five independent WE simulations were performed with different initial random seeds for BD propagation. In each simulation, the configurational space of the protein pairs was divided into 760 bins along a progress coordinate that was intended to capture the slowest protein motions of the association process. We used a progress coordinate that consisted of three zones: a) a "far" zone involving the distance between barnase and barstar, b) an "intermediate" zone involving the RMS deviation of barstar from its bound-state position following alignment of barnase, and c) a "near" zone involving the same RMS deviation metric as in b) and the fraction of native contacts between barnase and barstar. Simulations were evaluated for resampling at fixed time intervals τ (or

iterations) of 2 ns to maintain 24 walkers per bin. Each simulation was carried out for 1000 iterations, or a molecular time of 2 μs (defined as Nτ where N is the number of iterations).

4.3.3 Propagation of dynamics

The dynamics of our WE simulations were propagated using the UIOWA-BD software,[100,106] which is the same BD engine that was used for the brute force simulations by Frembgen-Kesner and Elcock.[114] Consistent with these simulations, our WE simulations were performed at a constant temperature of 25 °C using a standard BD algorithm with the inclusion of hydrodynamic interactions (HIs) via calculation of the diffusion tensor using the equations of Rotne & Prager and Yamakawa;[128,129] the same values were used for the hydrodynamic radii of the pseudo-atoms to reproduce the translational diffusion coefficients of the corresponding all-atom protein models by the hydrodynamics program HYDROPRO;[113] and a time step of 0.25 ps was used throughout the simulations.

4.3.4 Calculation of k_{on} values

For each barnase-barstar pair, the k_{on} value was computed from each of five independent WE simulations using conformations that were sampled every 20 ps once a steady state was achieved (Figure S17, Supporting Information). These values were then averaged. All WE simulations were sufficiently long to yield relative percent uncertainties in the average k_{on} of <20% (Figure S18, Supporting Information). Uncertainties in the average k_{on} values were represented by calculating 95% confidence intervals. The k_{on} from each WE simulation was calculated using the NAM method according to the following equation:[104]

$$k_{on} = \frac{k_D(b)\beta}{[1-(1-\beta)k_D(b)/k_D(q)]} \tag{4.1}$$

where k(b) and k(q) are the diffusion rate constants for achieving separation distances of b and q, respectively, and β is the probability of successful collisions, i.e. that a simulation starting from the unbound state with a separation distance of b (100 Å) reaches the bound state before drifting apart to a separation distance of q (500 Å). Assuming that the motions of the binding partners are isotropic, k(b) and k(q) are given by the Smoluchowski result;

$k(r) = 4Dr$ where D is the relative translational diffusion coefficient of the two proteins (i.e. the sum of their corresponding diffusion coefficients). As done for the brute force simulations by Frembgen-Kesner and Elcock,[114] we used the estimate from HYDROPRO[113] for $D = 2.672 \times 10^2 \mathring{A}^2 ps^{-1}$. The β value was calculated using the following equation:

$$\beta = \frac{f_{SS}^{bind}}{f_{SS}^{bind} + f_{SS}^{qsurf}} \tag{4.2}$$

where f_{SS}^{bind} is the steady-state flux into the bound state and f_{SS}^{qsurf} is the steady-state flux into the q surface. As evident in the above equations, the influence of HIs is considered in our calculation of the probability of successful collisions (β), but only approximately on the diffusion of the two proteins by using the sum of their diffusion coefficients (D).[130]

4.3.5 Calculation of WE efficiency

For each barnase-barstar pair, we determined the efficiency of a single WE simulation relative to brute force simulation in computing the k_{on} for each of five independent WE simulations; these efficiencies were then averaged and uncertainties in the efficiencies were determined by calculating the 95% confidence intervals. The efficiency of each WE simulation was calculated using the following equation:

$$\text{Efficiency of WE} = \frac{t_{BF}}{t_{WE}} \tag{4.3}$$

where t_{BF} and t_{WE} are the wall-clock times required by brute force simulation and the WE simulation, respectively, to generate the same number of independent (uncorrelated) association events using the same computing resource (i.e. 256 CPU cores of 2.3 GHz AMD Interlagos processors). Association events were considered independent if, within the period between the event and one correlation time before the event, their corresponding trajectories did not share a common simulation segment. The correlation time was determined by monitoring autocorrelation of the flux into the bound state as a function of the lag time and identifying the first lag time that results in zero autocorrelation (within a 95% confidence interval; see Figure S17, Supporting Information). Since it was not practical to directly

obtain t_{BF} for all of the barnase-barstar pairs (i.e., the hydrophobic isosteres), we estimated t_{BF} in a consistent manner for each pair using the following equation:

$$t_{BF} = M_{BF}(\frac{0.02 \text{ days/trajectory/CPUcore}}{256 \text{ CPU core}}) \tag{4.4}$$

$$M_{BF} = \frac{\text{number of association events}}{\beta} \tag{4.5}$$

where M_{BF} is the number of trajectories in a brute force simulation to generate the same number of independent association events observed in a WE simulation - given that the brute force trajectories are terminated when the proteins either associate or reach a separation distance of q according to the NAM method; 0.02 days/trajectory/core is the average wall-clock time that would be required to complete a single brute force trajectory before the proteins reach a separation distance of q; and β (as defined above) is the probability calculated by WE for a single brute force trajectory to generate a successful association event before dissociating to a separation distance of q.

4.4 RESULTS

Our general strategy for computing k_{on} values from our simulations was to first identify a criterion for successful association that reproduces the experimental k_{on} for wild-type barnase and barstar. Next, we validated the simulations by using this criterion to calculate the k_{on} for R59A mutant barnase and wild-type barstar, which associates 9-fold more slowly than the wild-type proteins,[125] and comparing the calculated k_{on} to the experimental value. Finally, we used this criterion to estimate the basal k_{on}, i.e. the k_{on} for the hydrophobic isosteres in which all effective charges of the wild-type proteins are set to zero. Following the brute force simulations by Frembgen-Kesner and Elcock,[114] our criterion for successful association was to reach a threshold value, Q_{rxn}, in the fraction of native intermolecular contacts, Q; dynamics were propagated using the same BD engine with the inclusion of intramolecular HIs to achieve realistic diffusive properties of the individual proteins; and k_{on} values were calculated according to the NAM method (see Methods section).[104]

4.4.1 Validation of the simulation strategy

Figure 13 shows the computed k_{on} as a function of Q_{rxn} for all five independent WE simulations of each barnase-barstar pair. The experimental k_{on} for wild-type barnase and barstar ($2.86 \times 10^8 M^{-1}s^{-1}$)[125] was reproduced when using Q_{rxn} values of 0.27 and 0.56 for simulations with and without intermolecular HIs, respectively. These values differ slightly from those determined by Frembgen-Kesner and Elcock using brute force simulations and the same protein models (0.32 and 0.47, respectively)[114] due to more frequent monitoring of the reaction criterion (every 20 ps instead of 100 ps); thus, our WE simulations are less likely to have missed conformations that satisfy the reaction criterion. Importantly, using the Q_{rxn} values that we have identified, the computed k_{on} values for R59A barnase and wild-type barstar are in excellent agreement with experiment, regardless of whether or not intermolecular HIs were included (Figure 14; see also Table S3, Supporting Information). The reproduction of experimental k_{on} values for both wild-type and mutant pairs of barnase and barstar is consistent with results from brute force simulations,[114] providing validation of our WE simulation protocol. Relative to the basal k_{on}, our computed k_{on} values for wild-type barnase and barstar are 53- and 103-fold faster with and without intermolecular HIs, respectively. These rate enhancements are solely due to the electrostatic interactions between the wild-type proteins given the omission of intramolecular electrostatic interactions in our simulations.

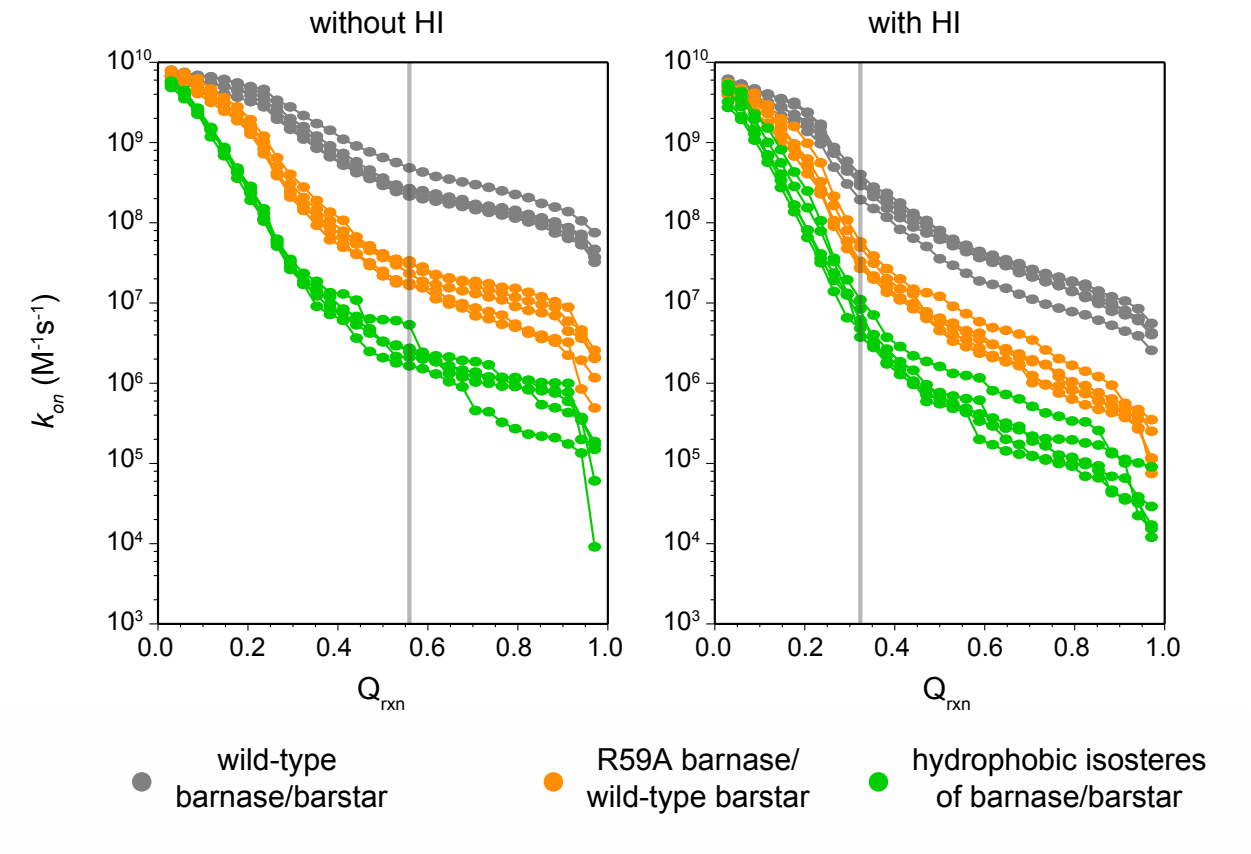

Figure 13: Computed k_{on} values for each barnase-barstar pair from each of five independent WE simulations as a function of the fraction of intermolecular native contacts Q_{rxn}. Results from simulations without and with the inclusion of intermolecular HI are shown in the left and right panels, respectively. The vertical gray line in each panel indicates the Q_{rxn} value that reproduces the experimental k_{on} for the wild-type pair for simulations without and with HI (0.56 and 0.27, respectively) and was used for calculating k_{on} values for the mutant pairs in that panel.

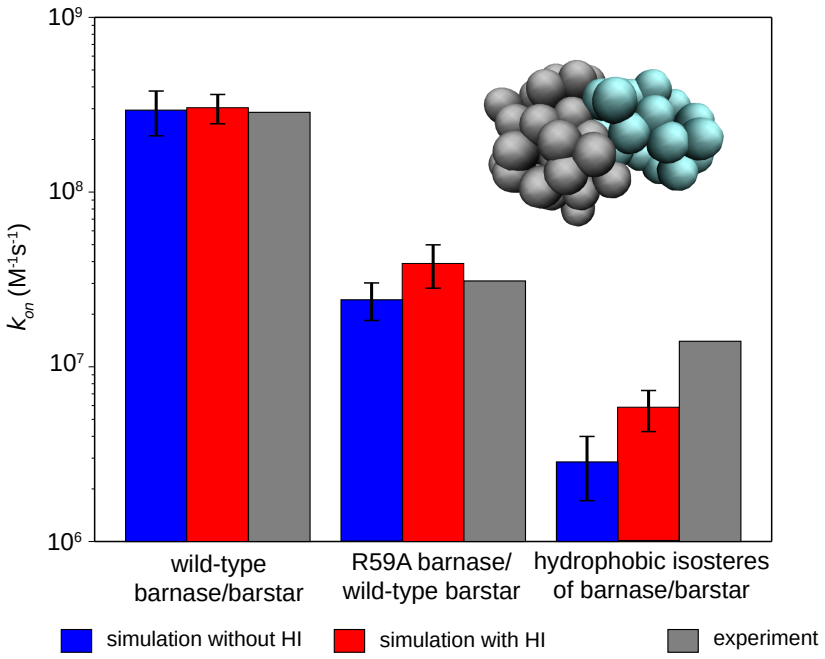

Figure 14: Comparison of computed and experimental[125] k_{on} values on a log scale. Reported values from simulation (with and without intermolecular HI) are averages from five independent WE simulations with the error bars representing 95% confidence intervals. The pseudo-atomic protein model of barnase (gray) and barstar (cyan) is shown in the upper right corner.

4.4.2 Estimation of the basal k_{on}

The basal k_{on} computed from our simulations with and without intermolecular HIs are (2.85 $\pm$ 1.30) $\times 10^6$ $M^{-1}s^{-1}$ and (5.79 $\pm$ 0.17) $\times 10^6$ $M^{-1}s^{-1}$, respectively. At the effective protein concentration maintained in our simulations (3.2 µM), these rate constants correspond to timescales beyond tens of milliseconds. Our computed basal k_{on} values are similar to those using less computationally intensive strategies; in particular, the use of spherical models with orientational constraints[120–123] has provided estimates in the range of 10^5-10^6 $M^{-1}s^{-1}$ and the use of rigid, atomistic models in either the application of transition-rate theory[124] or direct BD simulation of protein-protein association[119] has yielded estimates of $\sim 1 \times 10^6$

$M^{-1}s^{-1}$. The similarity of our estimates to these previous estimates suggests that flexible models may not be essential for obtaining realistic estimates of k_{on} values for proteins such as barnase and barstar that do not undergo significant conformational changes upon binding (the C_α RMS deviation between the crystal structures of the unbound[131,132] and bound[127] conformations is only 0.5Å for both barnase and barstar). However, it has not been possible to directly estimate the basal k_{on} with uncertainties of <100% using standard BD simulations with rigid, atomistic models since the association events were much slower in the absence of electrostatic forces.[119] On the other hand, our WE simulations with flexible molecular models enable significantly more precise calculations of the k_{on} (uncertainties of 22-46%) and could therefore be used for even more complicated binding processes, including ones that involve large conformational changes. Notably, our computed k_{on} values are significantly lower than that obtained by experiment from extrapolation to infinite salt concentration ($1.4 \times 10^7 M^{-1}s^{-1}$),[118] suggesting that the favorable electrostatic interactions between the proteins are not completely eliminated at high salt concentrations.

4.4.3 Effect of intermolecular HIs on the kinetics of association

Although the inclusion of intermolecular HIs has no effect on the ability of the simulation model to reproduce the effects of mutation on the k_{on}, for a fixed value of Q_{rxn}, the inclusion of intermolecular HIs significantly slows down the rate of association for all three pairs of the barnase-barstar system (Figures 14 and 15). Surprisingly, the extent to which k_{on} decreases is essentially the same for wild-type and R59A mutant pairs (e.g. by $\sim$ 5-fold at $Q_{rxn} = 0.27$). In contrast, the impact of intermolecular HIs in the brute force simulations by Frembgen-Kesner and Elcock was more pronounced for slower associating mutants of barnase such as R59A in which the electrostatic interactions with barstar are diminished.[114] Based on these results, it was predicted that the impact would be the most pronounced for the hydrophobic isosteres of barnase and barstar. However, the enhanced sampling provided by the WE strategy reveals no statistical difference between the impact of the intermolecular HIs on the k_{on} for the wild-type and R59A mutant pairs. For the hydrophobic isosteres, our results are inconclusive. Although it was possible to obtain statistically robust estimates of the basal k_{on}

–which was the primary goal of this work– our simulations did not reach the level of precision in the ratio of the k_{on} values with and without intermolecular HIs that would be required to determine the effect of HIs on the association kinetics relative to the wild-type pair (note the large confidence intervals in Figure 15). For future studies of this effect, significantly greater sampling using a larger number of simulations and/or longer simulations would be required to achieve a sufficient level of precision in the computed k_{on} values, particularly in the absence of intermolecular HIs.

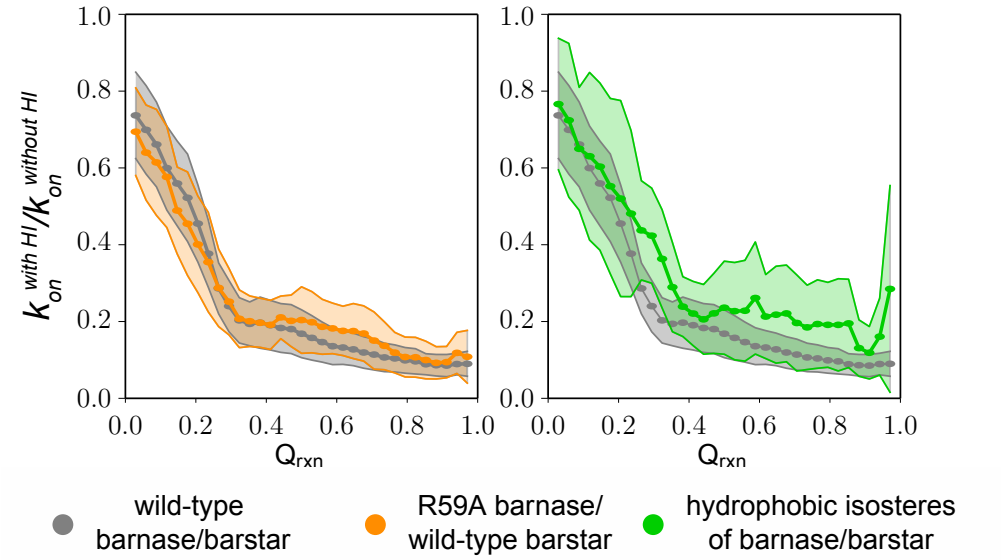

Figure 15: Ratio of association rate constants $k_{on}^{withHI}/k_{on}^{withoutHI}$ computed from simulations with and without intermolecular HIs as a function of the fraction of intermolecular contacts Q_{rxn}. The shaded regions represent 95% confidence intervals for averages (filled circles) from five independent WE simulations.

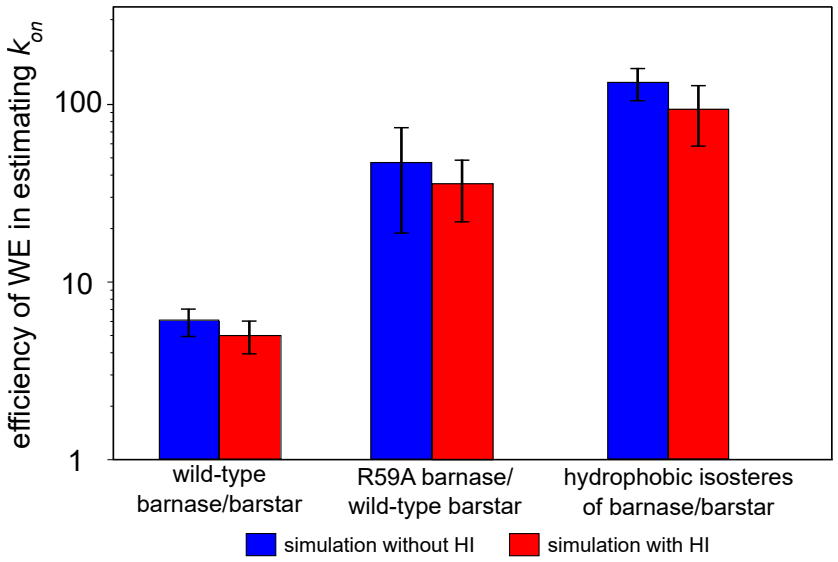

Figure 16: Average efficiencies of WE relative to brute force simulation in computing the k_{on}. Full details for estimating efficiencies are described in Methods. Uncertainties represent 95% confidence intervals for averages from five independent WE simulations.

4.4.4 Efficiency of WE simulation

Finally, it would not have been practical to obtain converged estimates of the basal k_{on} without the use of the WE strategy. In addition, a highly scalable, parallel implementation of this strategy was essential since it would have otherwise required > 2 years to carry out the simulations using a serial implementation. To determine the efficiency of parallelized WE vs. brute force simulation in estimating the k_{on}, we compared the wall-clock time that would be required of WE vs. brute force simulation (both using the NAM framework) to generate the same number of independent (uncorrelated) association events using the same computing resource (256 CPU cores of 2.3 GHz AMD Interlagos processors). Figure 16 shows the efficiencies of WE simulations relative to brute force simulations for each barnase-barstar pair (see also Table S4, Supporting Information). For the wild-type pair, a WE simulation was 6-fold more efficient than brute force simulation with the inclusion of intermolecular

HI. This efficiency increased to 46-fold for the R59A mutant pair and ultimately 131-fold for the hydrophobic isosteres. In the latter case, brute force simulation using the same flexible protein models would be highly impractical, requiring 386 days in wall-clock time to generate the same number of association events (> 1000) as a single WE simulation, which required only 3 days. The greater efficiency of WE observed for the slower processes (i.e. increasing with the barrier height) is consistent with previous WE studies of other rare events.[69,108,133,134]

4.5 CONCLUSIONS

In conclusion, we have directly computed the basal k_{on} for a protein-protein association process for the first time using flexible models with molecular simulations. In particular, we computed the basal k_{on} for the barnase-barstar system using highly efficient, flexible molecular simulations. Our computed basal k_{on} is significantly lower than that obtained by experiment from extrapolation to infinite salt concentration, suggesting that the electrostatic interactions are not completely eliminated at high salt concentrations. This result underscores the importance of directly computing the basal k_{on} using the true hydrophobic isosteres of the proteins under regular salt concentrations—a goal that can only be achieved by molecular simulation. Relative to our basal k_{on}, the electrostatic interactions of the wild-type proteins enhance the rate of association by > 130-fold. As demonstrated by Frembgen-Kesner and Elcock using brute force simulations,[114] the inclusion of intermolecular HIs significantly decreases the computed k_{on} values for both wild-type and mutant pairs. However, the extensive sampling provided by our WE simulations has revealed that the extent by which the k_{on} is reduced is the same for both the wild-type and R59A mutant pairs. For the hydrophobic isosteres, the relative extent to which the k_{on} was affected by the intermolecular HIs was inconclusive due to insufficient precision in the ratio of the k_{on} with and without intermolecular HI. Finally, our results demonstrate that WE simulations are orders of magnitude more efficient than brute force simulation in providing converged estimates of rate constants for the slow associations of proteins in the complete absence of electrostatic

interactions. The computation of such rate constants is otherwise impractical when using flexible protein models—even when these models are coarse-grained. Given its high efficiency, the simulation strategy used in this study would be useful for even more complicated systems, including those that undergo large conformational changes upon binding.

4.6 ACKNOWLEDGEMENTS

We thank Adrian Elcock for valuable discussions and making the UIOWA-BD software available. We also thank Alex DeGrave and A. Pratt for critical reading of the manuscript. Financial support was provided by NSF CAREER MCB0846216, NIH 1R01GM115805-01, and a DAAD graduate research grant. Computational resources were provided by NSF CNS-1229064 and the University of Pittsburgh's Center for Simulation and Modeling

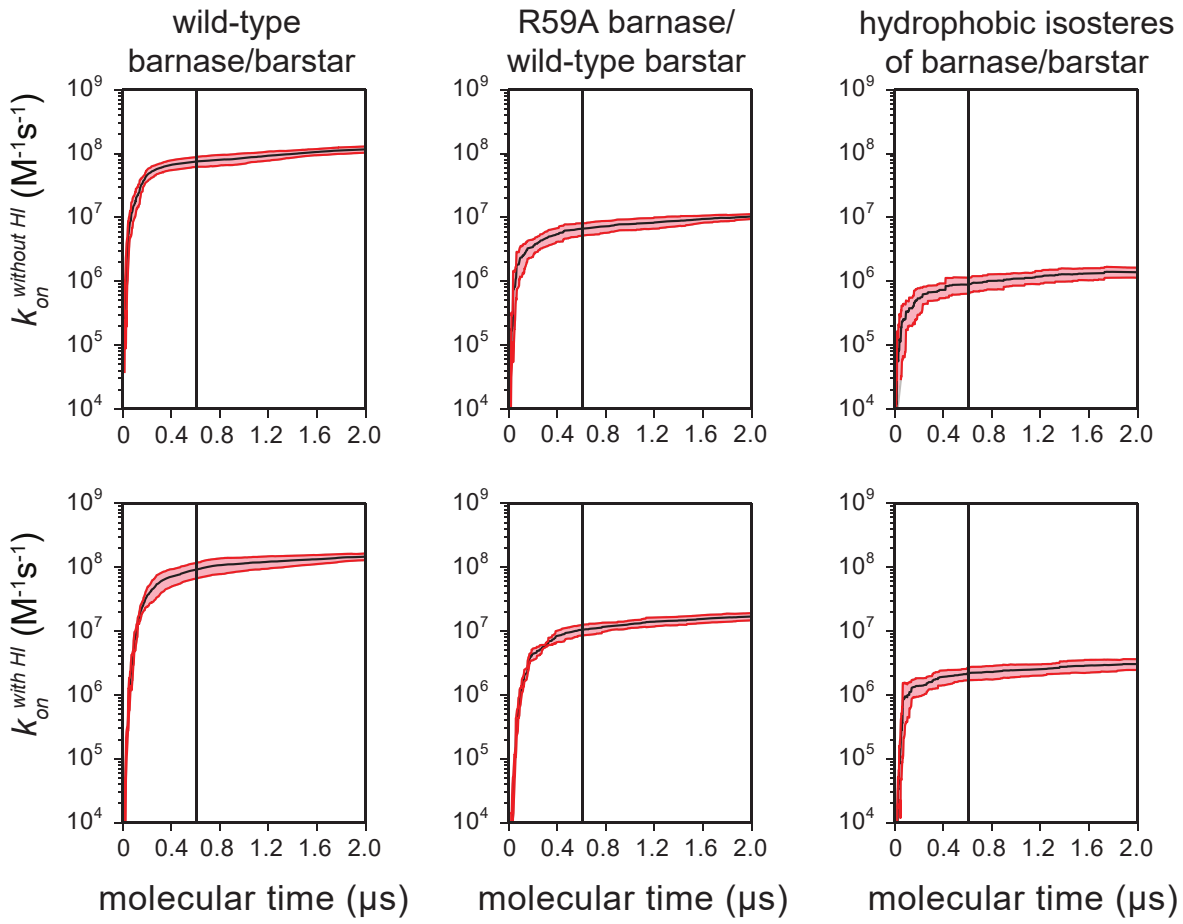

Figure 17: Average calculated k_{on} values for each barnase-barstar pair as a function of the molecular time from five independent WE simulations without and with the inclusion of intermolecular HI (top and bottom panels, respectively). The molecular time is defined as $N\tau$ where N is the number of WE iterations and τ is the fixed time interval of each iteration. Uncertainties (shaded in pink) are 95% confidence intervals. All subsequent analysis of the simulations was performed starting from a molecular time where an approximate steady state has been reached (vertical lines).

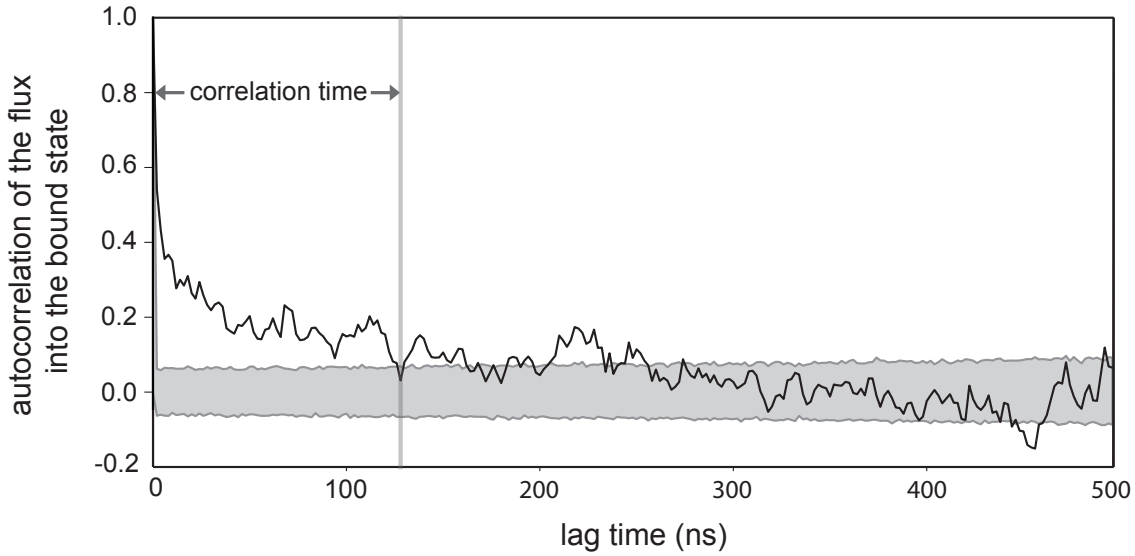

Figure 18: A representative autocorrelation of the flux into the bound state from a single WE simulation as a function of the lag time and its use in determining the number of independent association events. Data shown corresponds to a simulation involving the hydrophobic isosteres with the inclusion of intermolecular HI. The lag time is the frequency with which the flux into the bound state ($Q_{rxn} = 0.27$) was sampled and the correlation time (vertical gray line) is the time interval required to reach zero autocorrelation in the flux, i.e. the 95% confidence interval centered on zero (gray shaded region). Association events were considered independent if, within the period between the event and one correlation time before the event, their corresponding trajectories did not share a common simulation segment.

			$k_{on}/10^8(M^{-1}s^{-1})$
		simulation	experiment
	without HI	with HI	
wild-type barnase/barstar	2.94 ± 0.96	3.04 ± 0.66	2.86
R59A barnase/ wild-type barstar	0.243 ± 0.067	0.390 ± 0.124	0.31
hydrophobic isosteres of barnase/barstar	0.0285 ± 0.0130	0.0579 ± 0.0017	0.14

Table 3: Average calculated k_{on} values for each barnase-barstar pair from five independent WE simulations (with and without intermolecular HI) vs. experimental values. Uncertainties represent 95% confidence intervals.

	wild-type barnase/barstar		R59A barnase/ wild-type barstar		hydrophobi isostares of barnase/barstar	
	with HI	without HI	with HI	without HI	with HI	without HI
number events	1926 ± 70	2182 ± 168	1329 ± 62	958 ± 23	1096 ± 102	374 ± 18
t_{WE} (days)	2.1 ± 0.3	3.0 ± 0.6	2.0 ± 0.5	2.8 ± 1.0	3.1 ± 0.8	3.4 ± 0.5
t_{BF} (days)	13.6 ± 3.2	15.9 ± 3.0	77.0 ± 23.8	87.8 ± 22.0	386.0 ± 80.6	311.4 ± 103.6
$\beta/10^{-2}$	1.214 ± 0.261	1.182 ± 0.364	0.156 ± 0.049	0.095 ± 0.003	0.025 ± 0.007	0.011 ± 0.005
efficiency of WE	6 ± 1	5 ± 1	46 ± 27	35 ± 13	131 ± 26	92 ± 34

Table 4: Average efficiencies of weighted ensemble (WE) vs brute force (BF) simulation in estimating the k_{on} using five independent WE simulations. The efficiency of each WE simulation was estimated using t_{BF}/t_{WE} where t_{WE} is the wall-clock time required of the WE simulation and t_{BF} is the wall-clock time required of BF simulation to generate the same number of independent association events using the same computing resource (256 CPU cores of 2.3GHz AMD Interlagos processors). The latter was estimated using the probability β of capturing a successful association event over the course of the WE simulation (see Methods section). Averages for the free energy barrier to association, number of independent association events, t_{WE}, t_{BF}, and β are also provided. Uncertainties represent 95% confidence intervals.

5.0 PROTEIN-PROTEIN BINDING KINETICS AND CONTINUOUS PATHWAYS FROM ATOMISTIC SIMULATIONS IN EXPLICIT SOLVENT

5.1 CHAPTER SUMMARY

A complete characterization of a protein binding process requires the generation of atomically detailed pathways, which *are the mechanism* of the binding process with all of the key states, including states that are too transient to capture using experiments. Here we applied the weighted ensemble (WE) path sampling strategy to enable the atomistic simulation of protein-protein binding pathways in explicit solvent for the barnase/barstar system. Our WE simulation generated 203 continuous binding pathways and yielded a computed k_{on} that is in good agreement with experiment. Results reveal three residues in barnase that are kinetically important for binding the barstar ligand. In addition, partial desolvation of the proteins occurs late in the binding process during the rearrangement of the encounter complex to the bound state. Interfacial waters are crucial for forming the native bound structure and hydrogen bond bridging waters found in the crystal structure can be found in the bound state of the WE simulation.

5.2 INTRODUCTION

Protein-protein interactions enable essential biological functions such as signal transduction, cell metabolism, and muscle contraction. A complete understanding of the mechanisms of protein-protein binding processes, however, remains inaccessible to experimental studies, due to the difficulty in characterizing the transient states along the binding pathways. These

processes can be fully characterized using binding pathways with atomistic level detail which can be used to characterize not only the transient states but also the exact paths between states.

Alternatively, molecular dynamics (MD) simulations can provide atomistic pathways with high temporal resolution. However, due to the computational prohibitive timescales of protein binding processes, only a few atomistic simulations of protein-ligand binding processes have been reported. In addition, only one atomistic simulation of protein-protein binding process involving the barnase/barstar system has been generated; this simulation was carried out with explicit solvent and consisted of short, discontinuous trajectories, which were subsequently analyzed using a Markov State Model to compute rate constants for the long-timescale binding process of barnase and barstar.[135]

Here, we have applied the weighted ensemble (WE) path sampling strategy in conjunction with atomistic MD simulations to generate complete pathways for the protein-protein binding process of the barnase/barstar system in explicit water. The WE strategy can generate continuous trajectories and rigorous rate constants for any type of stochastic dynamics for a rare event (e.g., protein folding and binding). This strategy can be orders of magnitude more efficient than standard simulations in generating pathways and rate constants for rare events and has already enabled atomistic simulations of a protein-peptide binding process.[10] To our knowledge, our study provides the first continuous, atomistic pathways of a protein-protein association process with rigorous kinetics. The inclusion of explicit solvent has also enabled the characterization of the role of solvent in the mechanism of the binding process.

5.3 METHODS

Atomistic simulations of protein-protein binding pathways for barnase-barstar were enabled in this study by the application of the WE path sampling strategy.[25] This strategy involves carrying out a large number of trajectories in parallel, with each trajectory assigned a weight to properly represent the path ensemble. To control the distribution of trajectories, configurational space is divided into bins along a progress coordinate towards the target state and

trajectories are evaluated at fixed time intervals τ for resampling. The resampling procedure involves either the replication or combination of trajectories to maintain a specified number of target trajectories/bin while adjusting trajectory weights according to rigorous statistical rules such that no bias is introduced into the dynamics. In the present study, the WE strategy was applied under equilibrium conditions to permit the refinement of key states after completion of the simulation as well as the calculation of rate constants.[29]

5.3.1 Weighted Ensemble (WE) Simulations.

All WE simulations were carried out in explicit water using the open-source, highly scalable WESTPA software (https://westpa.github.io/westpa).[50] Prior to carrying out WE simulations of the protein-protein binding processes of the wild-type barnase-barstar system, representative unbound conformations of each binding partner were generated by running a separate equilibrium WE simulation starting from the conformation of that partner in the native, bound complex.

Equilibrium WE simulations of the isolated binding partners involved the use of a one-dimensional progress coordinate consisting of the heavy-atom RMSD of the protein from its conformation in the crystal structure of the complex. The progress coordinate was divided into 45 bins, with a fine bin spacing of 0.1-3.0 Å in the region corresponding to rearrangements of the encounter complex to the bound state and a coarser bin spacing of 0.5-10 Å. The simulations were carried out using a target number of 12 trajectories/bin for 1200 WE iterations with each iteration having a fixed τ of 5 ps, yielding a maximum trajectory length of 6 ns to achieve reasonable convergence of the probability distributions as a function of the progress coordinate.

Unbound states for the binding simulations were then generated by selecting conformations of each binding partner according to its probability from the last iteration of the WE simulation and randomly orienting the partners with respect to each other at a separation of 20 Å to yield 1728 possible pairs of unbound conformations of barnase and barstar. These pairs were then reduced to 100 pairs by assigning trajectories to appropriate bins along the minimum separation distance dimension of the progress coordinate that was used for the

binding simulation and combining trajectories with small weights according to the standard WE algorithm. The initial ensemble of unbound states for the subsequent binding simulation consisted of 16 copies of each of the 100 unbound states to yield a total of 1600 unbound states with the weights of these states appropriately renormalized.

To simulate the binding process, an equilibrium WE simulation of the binding process was started from the set of 1600 pre-equilibrated unbound states. A two-dimensional progress coordinate was used throughout the simulation, consisting of (i) the heavy-atom RMSD of Asp35 and Asp39 (the most buried barstar residues in the crystal structure) of barstar relative to the barnase-bound crystal pose following alignment on barnase, and (ii) the minimum separation distance between and two binding partners. The RMSD dimension of the progress coordinate was divided into 71 bins with a fine bin spacing from an RMSD of 0.5-10 Å to focus the sampling primarily in the region corresponding to the rearrangement of the encounter complex to the bound state, and a coarser bin spacing from an RMSD of 1-60 Å. As done in our previous study,[8] to ensure that conformations that are in contact are not combined with ones that are not during resampling, the distance dimension of the progress coordinate was divided into only two bins using the encounter complex region as a dividing point, with one bin for distances < 5 Å and the other bin for distances ≥ 5 Å. To make optimal use of a given number of available CPU cores, the total number of trajectory segments that were being carried out at a time was fixed at a constant number of 1600, adjusting the number of target trajectories in each bin as appropriate. On average, these adjustments resulted in ~ 22 target trajectories/bin.

For the binding process of barnase and barstar, the equilibrium WE simulation was carried out for 650 iterations, each with a fixed time interval of $\tau = 20$ ps to yield a maximum trajectory length of 13 ns and 18 µs of aggregate simulation time. After all existing initial states formed an encounter complex (by a maximum trajectory length of 6 ns), the sampling was focused on the encounter complex region. The simulation was sufficiently long to yield a steady value of the k_{on} (Fig. S28 Supporting Information).

5.3.2 Propagation of dynamics.

For the equilibration WE simulations, the dynamics were propagated using GROMACS 4.6.7 dynamics engine with the all-atom AMBER03* force field[136] and TIP3P water model[137]. Heavy-atom coordinates for initial models of the unbound proteins and native complex were taken from the crystal structure of the wild-type complex (PDB code: 1BRS[127]). Hydrogen atoms were added to each model using ionization states present in solution at pH 7. System was immersed in a sufficiently large dodecahedron box of explicit water molecules to provide a minimum 12 Å clearance between the solutes and box walls for the unbound states in which the binding partners were separated by 20 Å. A total of 31 Na^+ and 29 Cl^- ions were included to both neutralize the net charge of the protein system and yield an ionic strength of 50 mM, yielding $\sim 100,000$ atoms for the total system.

Prior to production simulations using the WE strategy, the systems were first subjected to energy minimization and then two stages of equilibrating the solvent while applying harmonic constraints to the proteins with a force constant of 10 kcal mol^{-1}• $Å^{-2}$. During the first stage, the system was equilibrated for 20 ps at constant temperature (25 °C) and volume. During the second stage, the system was equilibrated for 1 ns at constant temperature (25 °C) and pressure (1 atm). Since the WE strategy requires stochastic dynamics, the temperature was maintained using a stochastic velocity rescaling thermostat with a coupling constant of 0.1 ps; pressure was maintained using a weak Berendsen barostat with a coupling constant of 0.5 ps. Bonds involving hydrogens were constrained using the LINCS algorithm to enable a 2-fs time step. Van der Waals interactions were switched off smoothly between 8 and 9 Å along with the application of a long-range analytical dispersion correction to energy and pressure. Real-space electrostatic interactions were truncated at 10 Å. Long-range electrostatic interactions were calculated using particle mesh Ewald summation. Conformations were sampled every 20 ps for subsequent analysis.

5.3.3 State definitions.

Prior to the calculation of rate constants, definitions of the unbound state, encounter complex, and bound state were determined from the equilibrium WE simulation of the binding

process. The unbound state was defined as having a minimum separation distance of $\geq$ 20Å between the proteins. The metastable encounter complex was defined to include only complexes that had a sufficiently long survival time to proceed to the native complex, *i.e.* heavy-atom RMSD of $\geq$ 4 Å and $\leq$ 20 Å for Asp35 and Asp39 of barstar after alignment on barnase and a minimum separation distance of $\leq$ 3 Å between the proteins. The bound state was defined as having a heavy-atom RMSD $\leq$ 3.5 Å for Asp35 and Asp39 of barstar after alignment on barnase and a minimum separation distance of $\leq$ 3 Å between the proteins.

5.3.4 Calculation of rate constants.

Rate constants k_{ij} between states i and j along the binding processes were calculated using the following:

$$k_{ij,bimolecular} = (f_{ij}C_0)\left(\frac{1}{p_i C_0^2}\right) = \left(\frac{f_{ij}}{p_i}\right)\left(\frac{1}{C_0}\right) \tag{5.1}$$

$$k_{ij,unimolecular} = \frac{f_{ij}}{p_i} \tag{5.2}$$

where f_{ij} is the flux of probability carried by trajectories originating in state i and arriving in state j, p_i is the fraction of trajectories more recently in state i than in j, and C_0 is the reference concentration of the binding partners, calculated as $1/(N_A V)$ where N_A is Avogadro's number and V is the volume of the dodehedral box used for the simulations (956 Å^3). The C_0 for both binding simulations was 1.7 mM. The bimolecular form (equation (1)) was used for the rearrangement of the encounter complex to the bound state ($\mathrm{k_2}$) and the unimolecular form (equation (2)) was used for the formation of the encounter complex ($\mathrm{k_1}$). All reported uncertainties in rate constants represent 95% confidence intervals and were estimated using a Monte Carlo blocked bootstrapping technique. Rate constant $\mathrm{k_1}$ was calculated using the entire simulation while $\mathrm{k_2}$ and $\mathrm{k_{on}}$ were calculated using the latter half of the simulation that focused greater sampling on the rearrangement of the encounter complex to the bound state.

5.3.5 Calculation of pairwise residue contact maps:

To identify kinetically important residues, we analyzed the contacts formed in the encounter complex of the transition path ensemble (TPE) which consisted of only productive pathways each of which begins where the trajectory last exited the initial unbound state and ends where the trajectory first enters the bound state. This allows us to look at only the productive pathways and the contacts they form without being obscured by the unproductive pathways, as was shown to be effective in a recent study.[138] A contact was defined as having two heavy atoms within 4.5Å of each other and was calculated every τ, considering only the contacts between binding partners. The statistical weight of each conformation in TPE was defined as the sum of the weights of its successful child trajectories, similar to the aforementioned study.[138] The probability of contact formation of each residue i with residue j was calculated as the sum of the TPE probability of every trajectory where residue i and residue j are in contact.

5.3.6 Analysis of conformation space networks.

To visualize the various tracks of binding pathways, we generated conformational space networks as done by others using a WE-based strategy[139]. For each of the 203 binding events, the longest two pathways were selected and then altogether organized into 2000 clusters by applying the KCenters clustering algorithm with a Canberra distance metric as implemented in MSMBuilder;[140] the feature vector for the clustering consisted of the RMSD progress coordinate and minimum separation distance between the binding partners. Network graphs of the sampled conformational space were then generated using the Gephi 0.9.2 software package[141] and ForceAtlas 2 layout algorithm,[142] with each node represent a cluster center and the edges between nodes representing observed transitions between each cluster. The size of each node is proportional to the total weight of the conformations in the corresponding cluster and colored according to the weighted average of the property of interest over every conformation in that cluster. The committor probability for each cluster was calculated from the number of transitions between relevant nodes.

5.3.7 Detection of bridging water molecules between the proteins.

To detect interfacial bridging water molecules between proteins in the bound state, we calculated the probability that a water molecule forms hydrogen bonds with both proteins over conformations sampled every 20 ps. Probabilities were calculated from the WE simulation using trajectory weights that were normalized by the population of the bound state. Hydrogen bonds were identified using the MDAnalysis Python library[143,144] and defined as having a ≤ 3 Å distance between donor and acceptor atoms and a $\geq 120°$ donor-hydrogen-acceptor angle.

5.3.8 Monitoring protein desolvation and tryptophan burial during the binding process.

To monitor the desolvation of the two proteins during the binding process, we tracked the number of water molecules N_w within 6 Å of each protein to encompass the first two solvation shells. We then calculated the "percent solvation" by dividing the average number of waters in a particular conformation by the average number of waters observed in the unbound state. The percent burials of the barstar residues, Trp38 and Trp44, upon binding were calculated as (SASA in barstar)/(SASA in solution) x 100% where the SASA is the solvent accessible surface area that was calculated using the Shrake and Rupley algorithm[145] as implemented in MDTraj Python library.[146] Both analyses were performed every 20 ps on the same ensemble of successful binding pathways that was used for the conformational space networks (see above).

5.3.9 Calculation of conformational entropy per residue.

The conformational entropy of each residue was calculated using the following equation:

$$S_x = -R \sum_i p_x(i) \ln p_x(i) \tag{5.3}$$

where S_x is the entropy of residue x, R is the ideal gas constant, and $p_x(i)$ is the probability of observing a particular heavy-atom RMSD value of i for residue x among the distribution

of RMSD values corresponding to the conformations sampled. The RMSD for each residue was calculated by aligning on the α carbons of the corresponding protein.

5.4 RESULTS

A complete characterization of the mechanism of protein-protein binding requires analyzing the relevant kinetics as well as the ensemble of binding pathways. To generate a diverse set of binding pathways, our simulation protocol involved the following two features: (i) provided multiple chances for each pre-equilibrated unbound state to result in successful binding pathways by generating 16 copies of each unbound state to yield a total of 1600 initial states for the binding simulation, and (ii) reduced the likelihood of an initial state to be terminated via recombination during the early stages of the simulation by setting the total number of trajectories across all bins at a given iteration to the number of initial states (1600) thereby "front-loading" the simulation with a large number of trajectories in bins between the unbound state and encounter complex.

5.4.1 Mechanism of binding.

Our WE simulation was successful in generating a large ensemble of continuous atomistic pathways for barnase-barstar association in explicit solvent. A total of 203 independent binding pathways were generated within 30 days by carrying out an equilibrium WE simulation using 1600 CPU cores at a time on the XSEDE Stampede supercomputer with an aggregate simulation time of 18 µs and a maximum continuous trajectory length of 13 ns.

Results reveal that the binding process of this system involves a two-step process in which a metastable "encounter complex" intermediate (Fig. 19) is first formed, followed by rearrangement of this complex to the bound state. Approximately 81% of the aggregate simulation time resulted in diffusional collisions of the binding partners and 11 $\pm$ 5% of them were productive (*i.e.* eventually reaching the native complex). While only 5% of the aggregate simulation time yielded successful binding pathways, our simulation was partic-

ularly effective at generating productive encounter complexes, which resulted from 35% of the aggregate simulation time.

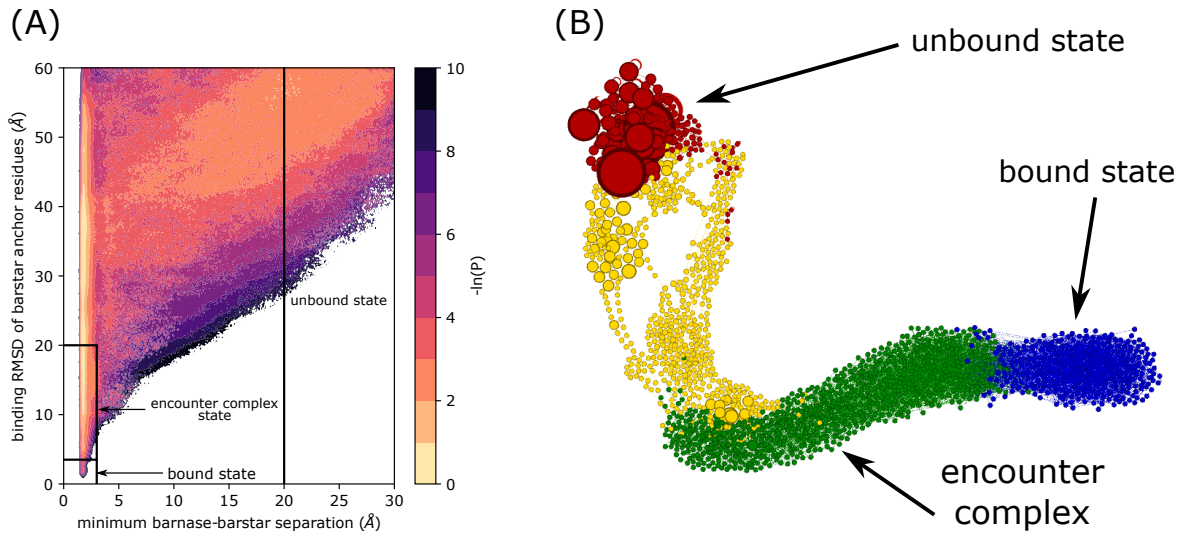

Figure 19: **(A)** Probability distribution of the protein-protein binding process as a function of the WE progress coordinate. The progress coordinate consisted of the heavy-atom RMSD of residues Arg35bs and Arg39bs of barstar after alignment of barnase from the crystal structure of barnase-barstar complex and minimum distance between barnase-barstar. Definitions of the encounter complex and bound state are delineated by the solid black lines. The color scale represents –RTln P where P is the pseudo-equilibrium probability density based on trajectory weights. **(B)** Reference conformational space network of barnase and barstar built from binding trajectories, highlighting the location of key states along the network.

Since our WE simulation was focused primarily on enhancing the sampling of association events, we did not observe a sufficient number of dissociation events to compute statistically robust rate constants for the unbinding direction and therefore focused exclusively on characterizing the kinetics in the association direction. As shown in Table 5, our computed rate constant k_{on} [(2.3 ± 1.0) x 10^8 $M^{-1}s^{-1}$] is in good agreement with experiment [(2.86 ± 0.7) x 10^8 $M^{-1}s^{-1}$]..[119] Given that the computed rate constant for the formation of the encounter complex k_1 [(1.8 ± 0.2) x 10^9 $M^{-1}s^{-1}$] is approximately equal to the k_{on} and that

process	rate constant	value
unbound state → encounter complex	$k_1(M^{-1}s^{-1})$	$1.8 \pm 0.2 \times 10^9$
encounter complex → bound state	$k_2(s^{-1})$	$2.7 \pm 0.5 \times 10^{10}$
unbound state → bound state	k_{on} $(M^{-1}s^{-1})$	$2.3 \pm 1.0 \times 10^8$
	experimental k_{on} $(M^{-1}s^{-1})$	$2.86 \pm 0.67 \times 10^8$

Table 5: Computed rate constants and 95% confidence intervals for the barnase-barstar binding process. Rate constant k_1 was calculated using the entire simulation and rate constants involving the bound state, k_{on} and k_2, were calculated using the second half of the simulation where the sampling was focused on the encounter and bound states and our rate constant estimates are converged (see Fig. 28).

the computed rate constant for the rearrangement of the encounter complex k_2 to the bound state is relatively fast [(2.7 ± 0.5) x 10^9 s^{-1}], the rate-limiting step is the diffusion-controlled formation of the encounter complex. The rate constant k_1 for this initial step is on the order of the Smoluchowski limit (∼ 5 x 10^9 $M^{-1}s^{-1}$) despite the orientational constraints due to electrostatic interactions between the proteins[118] and the ∼ 3× faster diffusion that results with the TIP3P water model.[137]

5.4.2 Diversity of binding pathways.

Five different tracks of complete binding pathways (I, II, III, IV, and V) were generated by our simulation with each track originating from a different pre-equilibrated unbound state and therefore not sharing any common trajectory segments with other binding tracks. As shown in Fig. 20 and Fig. 21, the initial unbound states of these binding tracks involve a variety of different relative orientations of the binding partners, including orientations in which the binding interface of barstar is facing the opposite side of barnase from the binding pocket (Track V). Thus, the binding tracks can be differentiated according to the extent to which the binding partners must rotate relative to each other to form productive encounter

complexes with Track III being the most direct track, requiring the least amount of relative rotations of the binding partners, and Track V being the most indirect track (Fig. 20), requiring the greatest amount of relative rotations of the binding partners. Despite the fact that Tracks I and II originated from unbound states with similar relative orientations, Track II involved a less direct route to forming the encounter complex (Fig. 21).

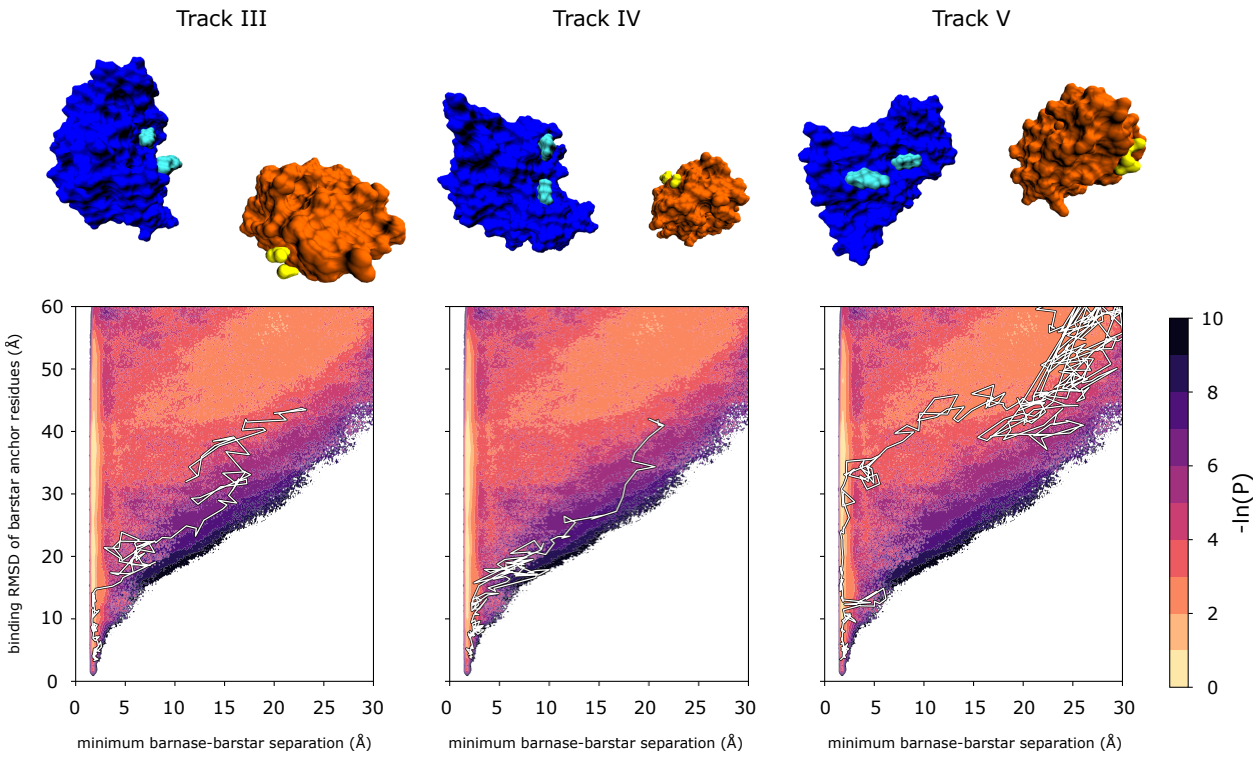

Figure 20: Diversity of starting structures that led to binding for the most direct and indirect tracks. **Top:** Surface representation of the starting structures, barnase is shown in blue, barstar is shown in orange, cyan residues are Lys27bn and Arg59bn which form the strongest interactions with the most buried residues in barstar binding helix, Asp35bs and Asp39bs shown in yellow. **Bottom**: A representative pathway from Tracks III-V overlayed on the probability distribution estimated from the WE simulation. Color map represents the $-\ln$ of the probability, the paths are shown in white, plotted every 20 ps.

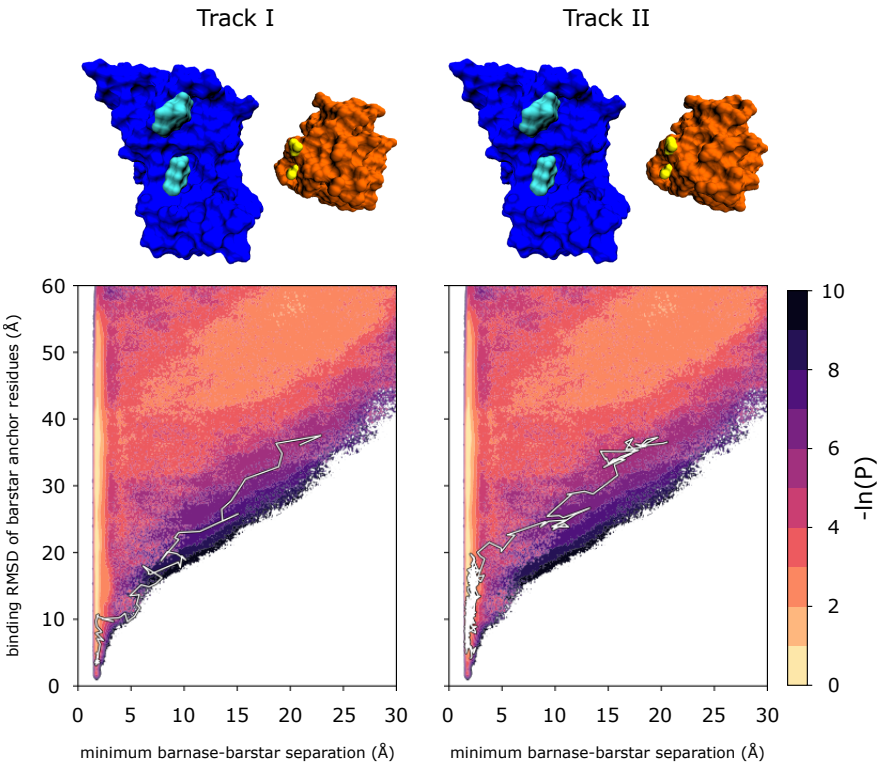

Figure 21: Similar starting structures leading to diverse pathways. **Top:** Surface representation of the starting structures, barnase is shown in blue, barstar is shown in orange, cyan residues are Lys27bn and Arg59bn which form the strongest interactions with the most buried residues in barstar binding helix, Asp35bs and Asp39bs shown in yellow. **Bottom:** A representative pathway from each track overlayed on the probability distribution estimated from the WE simulation. Color map represents the $-\ln$ of the probability, the paths are shown in white, plotted every 20 ps.

We have also tracked the percent burial of individual residues during the binding process in our simulations. Interestingly, the two interfacial Trp residues in barnase, Trp38 and Trp44, become buried upon forming the encounter complex with Trp44 becoming buried before Trp38. Thus, the detection of binding in stopped-flow Trp fluorescence experiments would include the formation of encounter complexes as well as the native complex. In addition, our results reveal that the barstar residues, Asp35bs and Asp39bs, that become the most buried in the bound state end up burying themselves earlier in the binding process

than other barstar residues with Asp35bs burying earlier than Asp39bs (Fig. 22).

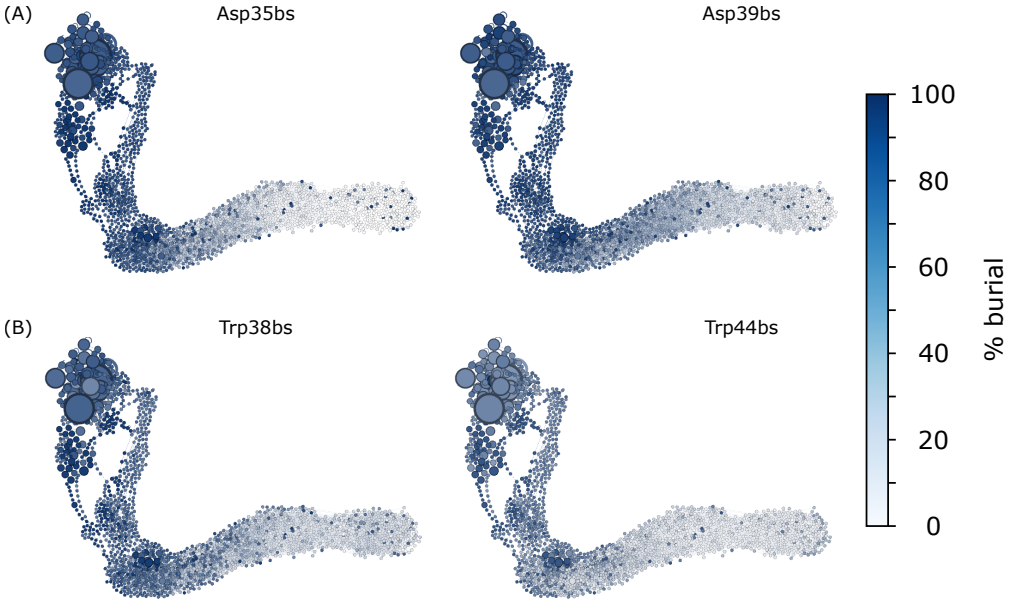

Figure 22: Conformation space networks showing the minimum percent burial of residues on barstar in each cluster. The color scale represents percent burial of a residue on barstar and sizes of nodes corresponds to the total weight of walkers in each cluster. See Fig. 19B for reference conformational space network.

Our WE simulation was successful in generating a diverse ensemble of encounter complexes that resulted from a variety of relative orientations of the two proteins in the unbound state, as illustrated by the cloud of "collision entry points" for barstar that is mapped onto the surface of a unit sphere centered on barnase (Fig. 23). These encounter complexes resulted from 1564 continuous pathways that originated from 6 of the 100 pre-equilibrated unbound states. A comparison of collision entry points of all pathways (Fig. 23B) and only productive pathways (Fig. 23C) shows that the productive collisions generally occurred near Arg59bn (front).

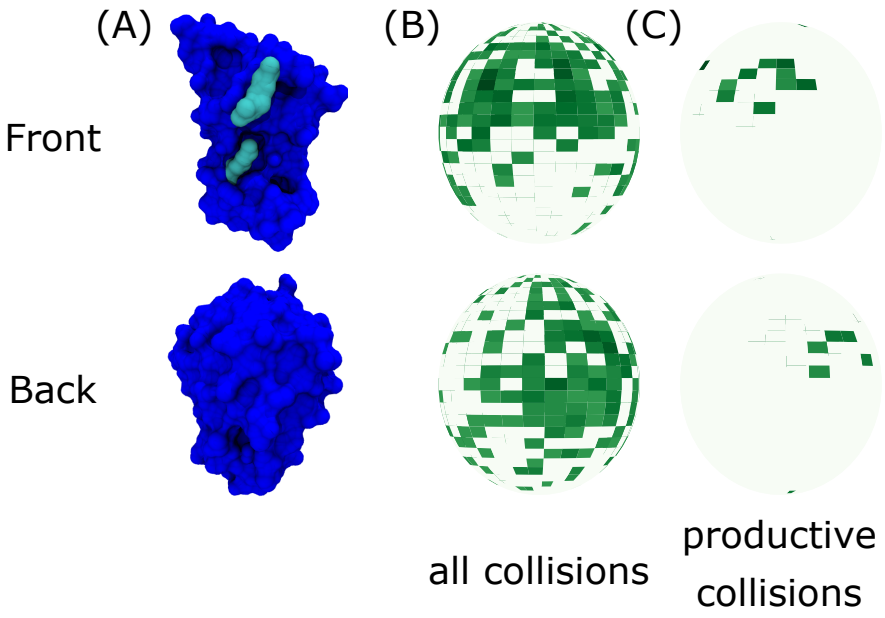

Figure 23: Cloud of collision entry points of barstar mapped onto a unit sphere centered on barnase. (A) Front and back orientations of barnase (blue) with binding interface residues Lys27bn and Arg59bn highlighted in cyan. These orientations are used to generate panels B and C in the corresponding rows. (B) Cloud of entry points for all diffusional collisions, and (C) cloud of entry points for productive collisions that form encounter complexes, which eventually rearrange to the native complex. Spheres are colored according to $-\ln p$ where p is the probability distribution of collision entry points projected onto the surface of unit sphere, ranging from least probable (white) to most probable (dark green).

A limitation of all rare-event sampling strategies is that these strategies may only capture the faster pathways depending on the maximum length of the trajectories and that indirect, slower pathways may be missed. For WE and related approaches, relevant free energy barriers to be surmounted may be orthogonal to the progress coordinate used to focus the sampling. In principle, however, if the progress coordinate captures the slowest relevant motion, then faster, correlated coordinates will also be captured. In the present study, the progress coordinate focuses the sampling of binding pathways in which the binding interfaces of the two proteins are pointing towards each other before diffusional collisions to form the encounter complex. Thus, the successful binding pathways involve unbound states in

which the binding interfaces of the two proteins were already oriented towards each other or unbound states in which the proteins end up rotating to into promising relative orientations. A future goal of great interest for WE methods development is to generate indirect pathways to binding such as those that involve rearrangement of the encounter complex to the native, bound state via "crawling" of the binding partners over each other's molecular surfaces. Promising strategies for achieving this goal are the use of progress coordinates that exhaustively cover the configurational space (*e.g.* Voronoi bins based on the pairwise RMSD between sampled conformations) and the improvement of schemes for replication and combination of trajectories to minimize the merging of very low-weight trajectories along indirect tracks to forming the bound state.

5.4.3 Kinetically important residues.

Based on our WE simulation, we identified kinetically important residues for the binding process by monitoring the frequency of intermolecular pairwise residue interactions formed by each interfacial residue in the encounter complexes (Fig. 24).

Our analysis revealed three barnase residues and three barstar residues that are involved in the formation of intermolecular contacts in the majority of encounter complexes: Arg59bn, His102bn, Ser38bn, Asp35bs, Gly43bs, and Trp44bs ("bn" for barnase and "bs" for barstar). The barnase residues, Arg59bn, His102bn, and Ser38bn, form interactions with the α-helix of barstar that lies at its binding interface (the binding helix) in 81%, 77%, and 68% of the encounter complexes, respectively. The barstar residues, Asp35bs, Gly43bs and Trp44bs, are located either on or near the binding helix, forming intermolecular contacts in 78%, 90% and 66% of the encounter complexes, respectively.

Our results are consistent with previous experimental and simulation studies.[147] Experimental studies have identified Lys27bn and Arg59bn as playing an important role in the association kinetics of barnase and barstar.[147] In addition, the kinetic importance of Ser38bn was also predicted by a recent simulation study[135] involving the construction of Markov State Models and the use of a different simulation model (AMBER ff99SB-ILDN and TIP3P).[137,148]

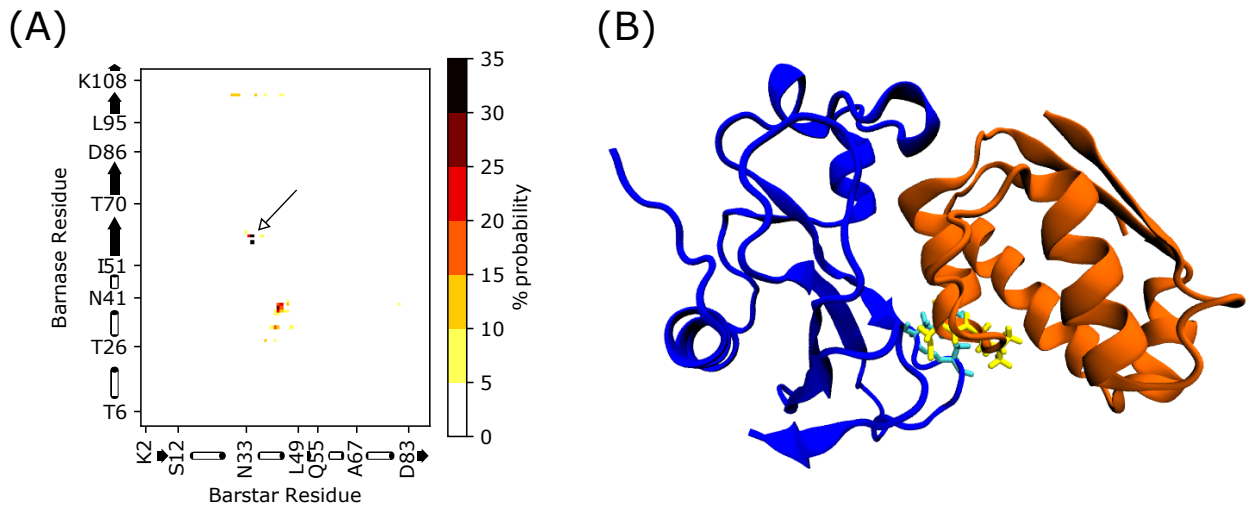

Figure 24: **(A)** Map of pairwise residue contacts formed in encounter complex ensemble generated by our WE simulation. The start of each secondary structure is shown in both axes where a beta sheet is shown as an arrow and an α-helix is shown with a cylinder. Kinetically most important set of contacts are highlighted with an empty-headed arrow. **(B)** Locations of the most kinetically important residue in the crystal structure of the native complex of barnase (blue) and barstar (orange). Arg59bn, His102bn and Ser38bn are shown in cyan and Asp35bs, Gly43, Trp44 are shown in yellow.

5.4.4 Changes in the conformational entropy of individual residues during the binding process.

To quantify the extent to which individual residues in barnase and barstar change in conformational flexibility, we calculated the conformational entropy of each residue according to the distributions of heavy-atom RMS deviations that residue after aligning on the C_α atoms of the corresponding protein in the crystal structure of the native complex.

As shown in Fig. 25, the largest loss of flexibility at the binding interface was observed in barstar binding helix, particularly residues Ala36bs to Leu41bs. At the binding interface

of barnase, residues Ser38bn and His102bn were the ones that lost the most flexibility upon binding. Interestingly, quite a few residues in both proteins that are not at the binding interface also lost flexibility upon binding. In particular, barstar residues Lys60bs and Asp83bs lost flexibility the most, followed by barnase residues on the third helix of barnase, Gly40bn, Leu42bn and Ala46bn as well as beta sheet residues Asn77bn and Thr79bn lost flexibility upon binding.

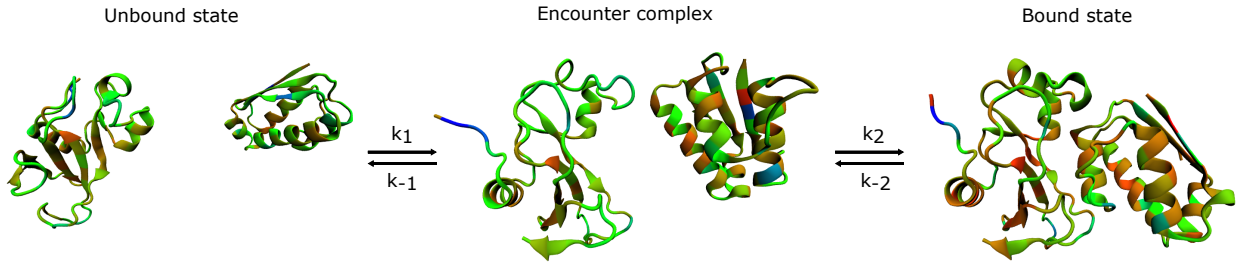

Figure 25: Per residue entropy calculated from per residue RMSD distributions shown as binding occurs. Proteins are shown in cartoon representation, entropy is shown using a red-green-blue color map and each residue is colored accordingly. In units of the gas constant R, red is 0.2 – 0.6, green is 0.6 – 1.4 and blue is 1.4 – 2.25.

5.4.5 Desolvation during the binding process.

While it is well-known that the desolvation of proteins occurs during protein-protein binding processes, it is not known *when* in the binding processes this desolvation occurs (*e.g.* upon forming the encounter complex and/or during rearrangement of the encounter complex to the bound state). To monitor the progress of desolvation during the barnase-barstar binding process in our explicit-solvent simulations, we calculated the percent solvation of each conformation relative to the unbound state, tracking the number of water molecules within 6 Å of each protein (see Methods). We then generated a conformational space network to visualize the various binding tracks and colored this network according to the minimum percent solvation thereby detecting *any* instance of desolvation. As shown in Fig. 26A, desolvation of the proteins occurs in the late stages of the binding process in our simulations. In partic-

ular, the two proteins undergo the greatest extent of desolvation during the rearrangement of the encounter complex to the native complex.

We also determined if a "drying effect" was occurring during the binding process in our simulations. As predicted by previous theoretical studies, the water molecules that occupy hydrophobic binding cavities may undergo drying effect, *i.e.* phase transition from a liquid to a gas phase[149–152](ref). This effect has been demonstrated by simulation studies involving the association of hydrophobic slabs[153] and for hydrophobic cavities of six proteins including Cox-2.[154,155] As shown in Fig. 26B, there are no large shifts in the density of the surrounding water molecules during the binding process in our simulations of the barnase/barstar system. Thus, no drying effect was detected in our simulations.

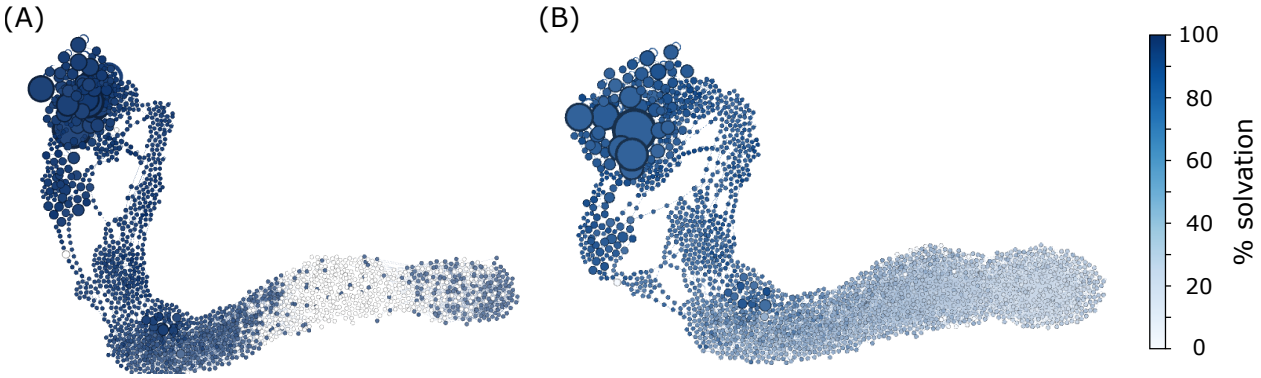

Figure 26: Conformational space networks of barnase and barstar built from binding trajectories, showing desolvation during binding. **(A)** Conformational space network colored by the minimum number of water molecules observed within 6Å of each protein in a given cluster of conformations. **(B)** Conformational space network colored by the average number of water molecules observed within 6Å of each protein in a given cluster of conformations. See Fig. 19B for reference conformational space network.

5.4.6 Interfacial, structural water molecules.

To determine the extent to which the positions of interfacial crystallographic water molecules are occupied in our simulation, we calculated the percent occupancy of each of the nine

Residues bridged	% occupancy in the bound state
Lys62bn/Tyr103bn – Asp35bs	4
Lys62bn/Asn58bn – Asp35bs	16
Arg59bn – Asp35bs	18
Glu73bn – Asp35bs	0
Ile55bn/Glu73bn -Trp38bs	16
Lys27bn/Glu73bn – Asp39bs	27
Arg83bn – Gly43bs	3
Ser38 – Val45bs	2
Ser38bn – Tyr47bs	2

Table 6: Percent occupancies of crystal water molecules that bridge hydrogen bonds between wild-type barnase and barstar (PDB code: 1BRS)[127] in the bound state sampled by WE simulation.

positions in the bound state ensemble. As shown in Table 6, these water molecules bridge hydrogen bonds between barnase and barstar, and all except one of the nine positions are occupied with four of these positions occupied >15% of the time. The occupancy of these positions of the crystal water molecules in the bound state is an encouraging validation of the ability of the force field and water model, particularly since the simulations were started from the unbound state.

In addition, our simulation identified water molecules in the bound-state ensemble that were not resolved in the crystal structure of the native complex and bridge hydrogen bonds between residues that were identified above as kinetically important (Fig. 27). One of these water molecules bridge hydrogen bonds between two barnase residues, Arg83bn and Lys27bn, and Asp39bs of barstar. The other water molecule bridges hydrogen bonds between Ser38bn of barnase and Trp44bs of barstar.

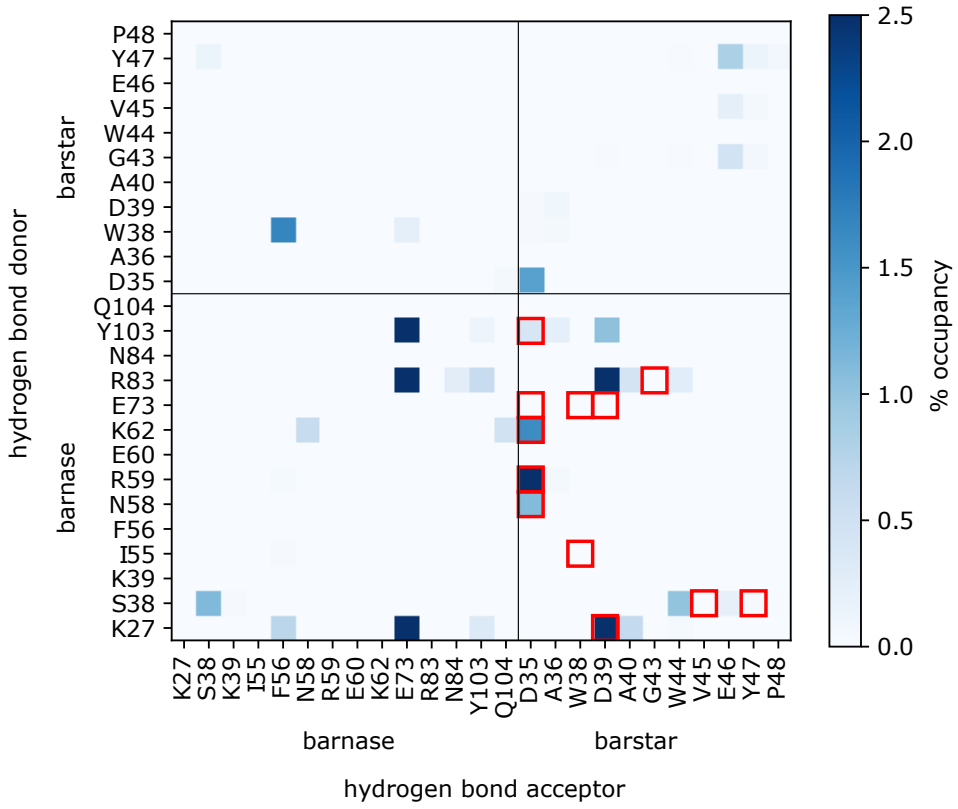

Figure 27: Occupancy map of waters that bridge residues through hydrogen bond formation. Both axes contain residues from both proteins and the black lines delineate the two proteins. Residues that are bridged by crystal waters are marked with red boxes. Occupancies are calculated from conformations in the bound state ensemble every 20 ps.

5.5 CONCLUSIONS

Starting from a large unbound state ensemble, we have generated 203 independent continuous binding pathways for the all-atom, explicit solvent barnase-barstar system. Our use of the WE strategy enabled the generation of continuous pathways for a protein-protein binding process in 30 days on a supercomputer. Binding pathways included indirectly oriented starting structures, improving on our previous study in terms of generating diverse binding

pathways.

We have directly computed the association rate constant, which was in good agreement with the experimental association rate constant. The diffusion-controlled formation of the encounter complex was the rate limiting step of the binding mechanism. Furthermore, we have shown that desolvation happens during rearrangement of the encounter complex into the bound state by direct analysis of the explicit water molecules. Despite observing desolvation in some binding pathways, the binding interface was still solvated in the bound state ensemble of our simulation, indicating that water molecules are important for binding of barnase and barstar, as predicted by previous studies. Hydrogen bridging water molecules that were present in the crystal structure were identified in the bound state ensemble, further underlining the importance for modelling explicit waters for protein-protein binding simulations.

5.6 SUPPORTING INFORMATION

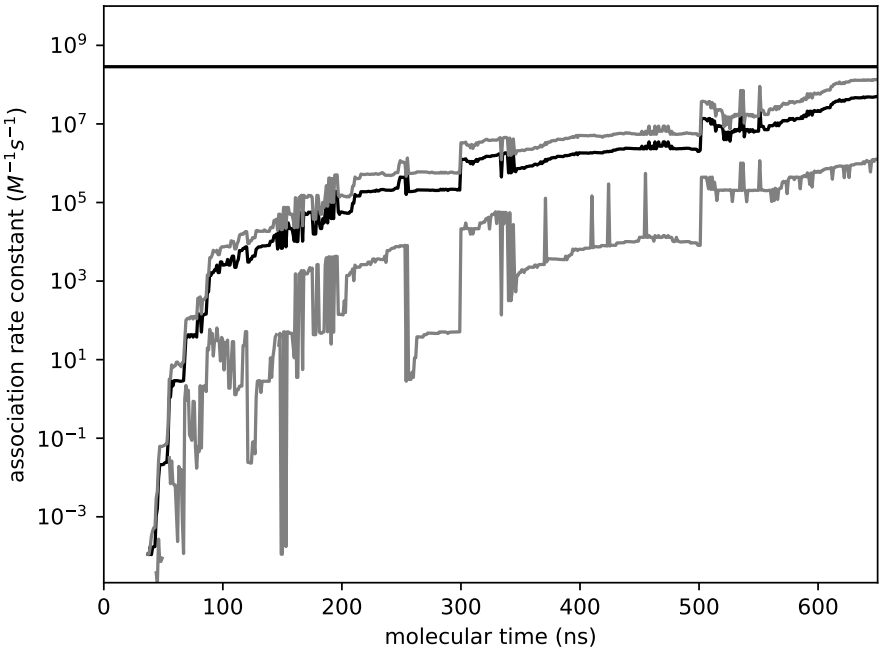

Figure 28: Evolution of calculated k_{on} values as a function of molecular time for barnase and barstar binding simulations. Gray lines show the 95% confidence interval. The molecular time is defined as Nτ where N is the number of WE iterations and τ is the fixed time interval of each iteration.

6.0 CONCLUSIONS AND FUTURE DIRECTIONS

Molecular dynamics simulations remain the only means to generate atomically detailed views of all of the protein conformational changes that occur during protein binding processes. As such, these simulations provide an ideal complement to experimental studies of these processes that in turn, provide important validation of the simulations. In this work, I have tackled a variety of research questions regarding the mechanisms of protein binding processes.

First, I have determined whether preorganization affects the binding kinetics of the intrinsically disordered p53 peptide to the MDM2 protein using flexible molecular models (Chapter 3). Using *in silico* modifications I have tuned the extent of preorganization in the the p53 peptide from fully disordered to fully preorganized, and directly simulated the association of each peptide variant with the MDM2 protein. Coupled with the weighted ensemble strategy, the resulting simulations yielded > 3000 binding events and statistically robust association rate constants for each p53 peptide variant. Not only was the association rate constant in reasonable agreement with experiment, the association rate constant of the fully preorganized and fully disordered variants were in agreement as well, indicating no kinetic advantage to being disordered for the p53 peptide. Further analysis indicates that the rearrangement step for the fully disordered peptide upon binding to MDM2 to be very rapid. Therefore, a potential future direction to this work involve simulating the association of larger and slower folding proteins with their binding partners using a similar methodology that directly address the question.

Secondly, I have directly computed the basal k_{on} for a protein-protein association process using using flexible molecular models in Chapter 4. Estimating the association rate constant of barnase and barstar in the absence of electrostatic interactions proved to be unfeasible

using standard simulations in previous studies. The application of the WE strategy enabled efficient sampling of binding between hydrophobic isosteres of barnase and barstar which resulted in a converged, computed basal k_{on}, revealing that the electrostatic interactions in the wild-type system enhance the association rate constant by $>$ 130-fold.

Finally, I have characterized the barnase barstar binding process using atomically detailed models with explicit water in Chapter 5. Starting from a large set of unbound structures, I have generated a diverse set of 203 continuous binding pathways. The diffusion-controlled formation of the encounter complex was the rate limiting step for this binding process and the directly computed k_{on} was in good agreement with the experimental association rate constant. Analysis of the explicit water molecules revealed that the binding interface is still solvated in the bound state, including hydrogen bond bridging water molecules that can be found in the crystal structure. Any desolvation happened during binding, happened late in the process, right before the formation of the binding complex.

Overall, results of this work show the feasibility of using molecular level simulations in generating of protein-protein binding pathways, enabled by the use of WE strategy. In particular, generating atomically detailed binding pathways for the protein-protein binding process including explicit waters have been shown to be feasible on standard computing clusters. Potential future directions for this work include further increasing the efficiency of generating a diverse set of pathways using history based replication and combination rules and focusing the sampling on indirect pathways where binding partners "crawl" over each others molecular surfaces.

APPENDIX A

"RULES OF THUMB" FOR RUNNING BINDING SIMULATIONS

During my graduate career, I have carried out hundreds of WE simulations of protein binding processes across different timescales and using a variety of different simulation models that have included residue-level as well as atomic levels of detail. As a result, I have gained a deeper understanding of choosing the optimal WE parameters for such simulations. My recommendations are summarized below.

A.1 PROGRESS COORDINATE

The WE strategy offers a great deal of flexibility in focusing the sampling on transitions between stable states. Typically, the sampling is focused using a progress coordinate and this progress coordinate can be modified "on the fly" during a simulation. This progress coordinate should capture the slowest relevant motion of the system.

For binding processes, an RMSD progress coordinate can be particularly sensitive in discriminating between non-native and native complexes, provided that the alignment involves just one of the binding partners thereby quantifying the relative orientations of the two binding partners. Furthermore, the sensitivity of the RMSD coordinate can be further improved by focusing on the minimal set of atoms that must "click" into place in the binding pocket to ensure that the remainder of the binding interface residues reach their intended positions in the native complex. From my experience, focusing on the "anchor" residues of

the protein ligand that become the most buried upon binding the protein receptor has been effective as it reduces the protein ligand to a "small-molecule" mimic. Furthermore, it has been important to include the distance between the binding partners as a dimension of the progress coordinate in order to separate states that are within van der Waals contact from states that are no in contact. This separation has been helpful in providing more thorough sampling of non-native complexes such as the encounter complex intermediate and greatly simplifies the analysis that involve encounter complexes (e.g. rate constant for encounter complex formation) after the simulation has been completed.

A.2 PLACEMENT OF BINS

An important rule of thumb for the placement of bins along a progress coordinate is to make sure to include one or more bins for every state of interest. As mentioned in the previous section, this ensures that transitions in and out of these states are better sampled, allowing for the characterization of kinetics at these regions of the configuration space. For binding, separating the structures that are in contact, but not bound, from the ones that are not in contact is important to ensure better sampling of rearrangement from the encounter complex to the bound state.

Furthermore, as shown in a previous study[138], the efficiency of the WE strategy can increase exponentially with the free energy barrier for the process of interest given optimal placement of bins. For example, the use of finer spacings between bins along the steeper parts of the barrier has been helpful in improving the efficiency of WE simulations. For protein binding processes, it has been helpful to use a finer bin spacing in the region of the progress coordinate that corresponds to the rearrangement of the encounter complex to the native complex. On the other hand, adding extra bins to regions corresponding to stable states or low barriers can lead to less efficient sampling. While the oversampling of these regions can reduce the efficiency of sampling binding events, the diversity of binding pathways may be greatly improved.

A.3 SIMULATION CONVERGENCE

As with any simulation, it is important to demonstrate the convergence of the simulation according to the computed properties of interest. For binding simulations, one would monitor the flux into the target bound state and hence the computed association rate constant as well as the evolution of the probability distribution as a function of the WE progress coordinates. It is important to note that, generally, state populations converge more slowly than rate constants and that it is possible to get converged rate constants even when state populations have not converged.

In the event that the progress coordinate is switched to a different one, it is important to note that the probability distribution over the new progress coordinate may be incorrect since the sampling was focused on the previous progress coordinate and the new progress coordinate might not be sufficiently sampled. Depending progress coordinates, it might be difficult to converge to the correct probability distribution once the progress coordinate has been switched. My suggestion in this scenario is to monitor the evolution of the probability distribution over the new progress coordinate and ensure that this distribution is converged before calculating any observables of interest.

As an additional consideration for analysis, storage is an important factor for large systems. In particular all-atom, explicit solvent simulations of protein-protein binding processes require a large amount of storage. Estimating the amount of storage space and ensuring there is enough storage ahead of time is important. Additionally, a good way to analyze these systems once the simulation is over is to save the protein coordinates every iteration in a separate file for every iteration. WESTPA provides tools to analyze data stored in this way in a parallelized fashion and the option to save the iteration data in separate files automatically is planned for an upcoming release.

APPENDIX B

SOFTWARE DEVELOPED

B.1 YAML INTERFACE FOR WESTPA PARAMETERS

Despite recent efforts in rare event sampling software to make these algorithms more accessible to a wider range of researchers, most rare event sampling methods remain difficult to use for non-expert users. An important part of developing a rare event sampling software that is accessible is to have simple user interfaces that allow the user to control the rare event sampling algorithm without requiring a lot of prior programming experience.

The weighted ensemble path sampling method has a flexible and open-source implementation in WESTPA, used throughout this work and developed by the Chong lab. I have developed an interface for WESTPA that allows non-programmer users to enter system parameters in a fashion that doesn't require extensive coding knowledge. This interface builds up on the previous interface for defining other WESTPA parameters and uses widely used YAML markup language. Prior to this development two files were required to fully setup WESTPA parameters for a simulation, one of which must be written in Python programming language. With this YAML interface, now the standard for defining WESTPA simulation parameters (such as progress coordinate dimensionality, data format of progress coordinate, progress coordinate binning and number of simulations per bin) it is possible to setup WESTPA simulation parameters with a single file that requires no prior programming knowledge to edit.

BIBLIOGRAPHY

[1] Shaw, D. E.; Deneroff, M. M.; Dror, R. O.; Kuskin, J. S.; Larson, R. H.; Salmon, J. K.; Young, C.; Batson, B.; Bowers, K. J.; Chao, J. C. Anton, a special-purpose machine for molecular dynamics simulation. *Commun. ACM* **2008**, *51*, 91–97.

[2] Stone, J. E.; Hardy, D. J.; Ufimtsev, I. S.; Schulten, K. GPU-accelerated molecular modeling coming of age. *J. Mol. Graph. Modell.* **2010**, *29*, 116–125.

[3] Le Grand, S.; Gotz, A. W.; Walker, R. C. SPFP: Speed without compromise - a mixed precision model for GPU accelerated molecular dynamics simulations. *Comput. Phys. Commun.* **2013**, *184*, 374–380.

[4] Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y.; Wriggers, W. Atomic-level characterization of the structural dynamics of proteins. *Science* **2010**, *330*, 341–346.

[5] Zhao, G.; Perilla, J. R.; Yufenyuy, E. L.; Meng, X.; Chen, B.; Ning, J.; Ahn, J.; Gronenborn, A. M.; Schulten, K.; Aiken, C.; Zhang, P. Mature HIV-1 capsid structure by cryo-electron microscopy and all-atom molecular dynamics. *Nature* **2013**, *497*, 643–646.

[6] Perilla, J. R.; Hadden, J. A.; Goh, B. C.; Mayne, C. G.; Schulten, K. All-atom molecular dynamics of virus capsids as drug targets. *J. Phys. Chem. Lett.* **2016**, *7*, 1836–1844.

[7] Frembgen-Kesner, T.; Elcock, A. Computer simulations of the bacterial cytoplasm. *Biophys. Rev.* **2013**, *5*, 109–119.

[8] Saglam, A. S.; Chong, L. T. Highly efficient computation of the basal kon using direct simulation of protein-protein association with flexible molecular models. *J. Phys. Chem. B* **2016**, *120*, 117–122.

[9] Dickson, A.; Mustoe, A. M.; Salmon, L.; Brooks, C. L. Efficient in silico exploration of RNA interhelical conformations using Euler angles and WExplore. *Nucleic Acids Research* **2014**, *42*, 12126–12137.

[10] Zwier, M. C.; Pratt, A. J.; Adelman, J. L.; Kaus, J. W.; Zuckerman, D. M.; Chong, L. T. Efficient atomistic simulation of pathways and calculation of rate constants for a protein-peptide binding process: Application to the MDM2 protein and an intrinsically disordered p53 peptide. *J. Phys. Chem. Lett.* **2016**, *7*, 3440–3445.

[11] van Erp, T. S.; Moroni, D.; Bolhuis, P. G. A novel path sampling method for the calculation of rate constants. *J. Chem. Phys.* **2003**, *118*, 7762.

[12] Bhatt, D.; Zuckerman, D. M. Beyond microscopic reversibility: Are observable nonequilibrium processes precisely reversible? *J. Chem. Theory Comput.* **2011**, *7*, 2520–2527.

[13] Dickson, A.; Warmflash, A.; Dinner, A. R. Separating forward and backward pathways in nonequilibrium umbrella sampling. *Journal of Chemical Physics* **2009**, *131*, 10.

[14] Weinan, E.; Vanden-Eijnden, E. In *Multiscale Modelling and Simulation*; Attinger, S., Koumoutsakos, P., Eds.; Springer Berlin Heidelberg: Berlin, Heidelberg, 2004; pp 35–68.

[15] Hummer, G. From transition paths to transition states and rate coefficients. *Journal of Chemical Physics* **2004**, *120*, 516–523.

[16] Suarez, E.; Adelman, J. L.; Zuckerman, D. M. Accurate Estimation of Protein Folding and Unfolding Times: Beyond Markov State Models. *Journal of Chemical Theory and Computation* **2016**, *12*, 3473–3481.

[17] Pratt, L. A statistical method for identifying transition-states in high dimensional problems. *J. Chem. Phys.* **1986**, *85*, 5045–5048.

[18] Bolhuis, P. G.; Dellago, C.; Chandler, D. Sampling ensembles of deterministic transition pathways. *Faraday Discussions* **1998**, *110*, 421–436.

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

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

[21] Woolf, T. B. Path corrected functionals of stochastic trajectories: towards relative free energy and reaction coordinate calculations. *Chemical Physics Letters* **1998**, *289*, 433–441.

[22] Ottinger, H. C. VARIANCE REDUCED BROWNIAN DYNAMICS SIMULATIONS. *Macromolecules* **1994**, *27*, 3415–3423.

[23] Kloeden, P.; Platen, E. *Numerical Solution of Stochastic Differential Equations*; Springer-Verlag, 1992.

[24] Zuckerman, D. M.; Woolf, T. B. Dynamic reaction paths and rates through importance-sampled stochastic dynamics. *Journal of Chemical Physics* **1999**, *111*, 9475–9484.

[25] Huber, G. A.; Kim, S. Weighted-Ensemble Brownian dynamics simulations of protein association reactions. *Biophys. J.* **1996**, *70*, 97–110.

[26] Kahn, H.; Harris, T. E. Estimation of particle transmission by random sampling.

[27] Zhang, B. W.; Jasnow, D.; Zuckerman, D. M. The "weighted ensemble" path sampling method is statistically exact for a broad class of stochastic processes and binning procedures. *J. Chem. Phys.* **2010**, *132*, 054107.

[28] Bhatt, D.; Zhang, B.; Zuckerman, D. Steady-state simulations using weighted ensemble path sampling. *J. Chem. Phys.* **2010**, *133*, 014110.

[29] Suarez, E.; Lettieri, S.; Zwier, M. C.; Stringer, C. A.; Subramanian, S. R.; Chong, L. T.; Zuckerman, D. M. Simultaneous computation of dynamical and equilibrium information using a weighted ensemble of trajectories. *J. Chem. Theory Comput.* **2014**, *10*, 2658–2667.

[30] Cerou, F. Adaptive multilevel splitting for rare event analysis. *Stochastic Analysis and Applications* **2007**, *25*, 417–443.

[31] Zimmerman, M. I.; Bowman, G. R. FAST Conformational Searches by Balancing Exploration/Exploitation Trade-Offs. *J. Chem. Theory Comput.* **2015**, *11*, 5747–5757.

[32] Preto, J.; Clementi, C. Fast recovery of free energy landscapes via diffusion-map-directed molecular dynamics. *Phys. Chem. Chem.Phys.* **2014**, *16*, 19181–19191.

[33] Becker, N. B.; Allen, R. J.; ten Wolde, P. R. Non-stationary forward flux sampling. *Journal of Chemical Physics* **2012**, *136*, 18.

[34] Dickson, A.; Maienschein-Cline, M.; Tovo-Dwyer, A.; Hammond, J. R.; Dinner, A. R. Flow-Dependent Unfolding and Refolding of an RNA by Nonequilibrium Umbrella Sampling. *J. Chem. Theory Comput.* **2011**, *7*, 2710–2720.

[35] Wales, D. J. Discrete path sampling. *Molecular Physics* **2002**, *100*, 3285–3305.

[36] Carr, J. M.; Wales, D. J. Folding pathways and rates for the three-stranded beta-sheet peptide Beta3s using discrete path sampling. *Journal of Physical Chemistry B* **2008**, *112*, 8760–8769.

[37] Fackovec, B.; Vanden-Eijnden, E.; Wales, D. J. Markov state modeling and dynamical coarse-graining via discrete relaxation path sampling. *Journal of Chemical Physics* **2015**, *143*, 13.

[38] Barkema, G. T.; Mousseau, N. Event-based relaxation of continuous disordered systems. *Physical Review Letters* **1996**, *77*, 4358–4361.

[39] St-Pierre, J. F.; Mousseau, N.; Derreumaux, P. The complex folding pathways of protein A suggest a multiple-funnelled energy landscape. *Journal of Chemical Physics* **2008**, *128*, 8.

[40] Du, W. N.; Bolhuis, P. G. Adaptive single replica multiple state transition interface sampling. *Journal of Chemical Physics* **2013**, *139*, 11.

[41] Swenson, D. W. H.; Bolhuis, P. G. A replica exchange transition interface sampling method with multiple interface sets for investigating networks of rare events. *Journal of Chemical Physics* **2014**, *141*, 11.

[42] Allen, R. J.; Warren, P. B.; ten Wolde, P. R. Sampling rare switching events in biochemical networks. *Physical Review Letters* **2005**, *94*, 4.

[43] Glowacki, D. R.; Paci, E.; Shalashilin, D. V. Boxed Molecular Dynamics: A Simple and General Technique for Accelerating Rare Event Kinetics and Mapping Free Energy in Large Molecular Systems. *Journal of Physical Chemistry B* **2009**, *113*, 16603–16611.

[44] Warmflash, A.; Bhimalapuram, P.; Dinner, A. R. Umbrella sampling for nonequilibrium processes. *Journal of Chemical Physics* **2007**, *127*, 8.

[45] Vanden-Eijnden, E.; Venturoli, M. Exact rate calculations by trajectory parallelization and tilting. *J. Chem. Phys.* **2009**, *131*, 7.

[46] Dickson, A.; Dinner, A. R. In *Annual Review of Physical Chemistry, Vol 61*; Leone, S. R., Cremer, P. S., Groves, J. T., Johnson, M. A., Richmond, G., Eds.; Annual Review of Physical Chemistry; Annual Reviews: Palo Alto, 2010; Vol. 61; pp 441–459.

[47] Faradjian, A.; Elber, R. Computing time scales from reaction coordinates by milestoning. *J. Chem. Phys.* **2004**, *120*, 10880–10889.

[48] Majek, P.; Elber, R. Milestoning without a reaction coordinate. *J. Chem. Theory Comput.* **2010**, *6*, 1805–1817.

[49] Bello-Rivas, J. M.; Elber, R. Exact milestoning. *Journal of Chemical Physics* **2015**, *142*, 19.

[50] Zwier, M. C.; Adelman, J. L.; Kaus, J. W.; Pratt, A. J.; Wong, K. F.; Rego, N. B.; Suarez, E.; Lettieri, S.; Wang, D. W.; Grabe, M.; Zuckerman, D. M.; Chong, L. T. WESTPA: An interoperable, highly scalable software package for weighted ensemble simulation and analysis. *J. Chem. Theory Comput.* **2015**, *11*, 800–809.

[51] Abdul-Wahid, B.; Feng, H.; Rajan, D.; Costaouec, R.; Darve, E.; Thain, D.; Izaguirre, J. A. AWE-WQ: Fast-forwarding molecular dynamics using the accelerated weighted ensemble. *J. Chem. Inf. Model.* **2014**,

[52] Kratzer, K.; Berryman, J. T.; Taudt, A.; Zeman, J.; Arnold, A. The Flexible Rare Event Sampling Harness System (FRESHS). *Computer Physics Communications* **2014**, *185*, 1875–1885.

[53] Elber, R.; Roitberg, A.; Simmerling, C.; Goldstein, R.; Li, H. Y.; Verkhivker, G.; Keasar, C.; Zhang, J.; Ulitsky, A. Moil: A program for simulations of macromolecules. *Comput. Phys. Commun.* **1995**, *91*, 159–189.

[54] Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM - a program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **1983**, *4*, 187–217.

[55] Kirmizialtin, S.; Johnson, K. A.; Elber, R. Enzyme Selectivity of HIV Reverse Transcriptase: Conformations, Ligands, and Free Energy Partition. *Journal of Physical Chemistry B* **2015**, *119*, 11513–11526.

[56] Kirmizialtin, S.; Nguyen, V.; Johnson, K.; Elber, R. How conformational dynamics of DNA polymerase select correct substrates: Experiments and simulations. *Structure* **2012**, *20*, 618–627.

[57] Adelman, J. L.; Scarbrough, A.; Zwier, M. C.; Bhatt, D.; Chong, L. T.; Zuckerman, D. M.; Grabe, M. Simulations of the alternating access mechanism of the sodium symporter Mhp1. *Biophys. J.* **2011**, *101*, 2399–2407.

[58] Stelzl, L. S.; Fowler, P. W.; Sansom, M. S. P.; Beckstein, O. Flexible Gates Generate Occluded Intermediates in the Transport Cycle of LacY. *Journal of Molecular Biology* **2014**, *426*, 735–751.

[59] Adelman, J. L.; Grabe, M. Simulating current-voltage relationships for a narrow ion channel using the weighted ensemble method. *J. Chem. Theory Comput.* **2015**, *11*, 1907–1918.

[60] Du, W. N.; Bolhuis, P. G. Sampling the equilibrium kinetic network of Trp-cage in explicit solvent. *Journal of Chemical Physics* **2014**, *140*, 17.

[61] Velez-Vega, C.; Borrero, E. E.; Escobedo, F. A. Kinetics and mechanism of the unfolding native-to-loop transition of Trp-cage in explicit solvent via optimized forward flux sampling simulations. *J. Chem. Phys.* **2010**, *133*, 105103.

[62] Juraszek, J.; Bolhuis, P. G. Rate constant and reaction coordinate of trp-cage folding in explicit water. *Biophys. J.* **2008**, *95*, 4246–4257.

[63] Du, W. N.; Bolhuis, P. G. Equilibrium Kinetic Network of the Villin Headpiece in Implicit Solvent. *Biophysical Journal* **2015**, *108*, 368–378.

[64] Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E. How fast-folding proteins fold. *Science* **2011**, *334*, 517–520.

[65] Piana, S.; Lindorff-Larsen, K.; Shaw, D. E. Atomic-level description of ubiquitin folding. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 5915–5920.

[66] Dickson, A.; Lotz, S. D. Ligand release pathways obtained with WExplore: Residence times and mechanisms. *J. Phys. Chem. B* **2016**, *120*, 5377–5385.

[67] Teo, I.; Mayne, C. G.; Schulten, K.; Lelievre, T. Adaptive multilevel splitting method for molecular dynamics calculation of benzamidine-trypsin dissociation time. *J. Chem. Theory Comput.* **2016**, *12*, 2983–2989.

[68] Votapka, L. W.; Amaro, R. E. Multiscale Estimation of Binding Kinetics Using Brownian Dynamics, Molecular Dynamics and Milestoning. *Plos Computational Biology* **2015**, *11*, 24.

[69] Adelman, J. L.; Grabe, M. Simulating rare events using a weighted ensemble-based string method. *J. Chem. Phys.* **2013**, *138*, 044105.

[70] Dickson, A.; Brooks, C. L. WExplore: Hierarchical Exploration of High-Dimensional Spaces Using the Weighted Ensemble Algorithm. *J. Phys. Chem. B* **2014**, *118*, 3532–3542.

[71] Morelli, M. J.; ten Wolde, P. R.; Allen, R. J. DNA looping provides stability and robustness to the bacteriophage lambda switch. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 8101–8106.

[72] Donovan, R. M.; Sedgewick, A. J.; Faeder, J. R.; Zuckerman, D. M. Efficient stochastic simulation of chemical kinetics networks using a weighted ensemble of trajectories. *J. Chem. Phys.* **2013**, *139*, 115105.

[73] Donovan, R. M.; Tapia, J.-J.; Sullivan, D. P.; Faeder, J. R.; Murphy, R. F.; Dittrich, M.; Zuckerman, D. M. Unbiased rare event sampling in spatial stochastic systems biology models using a weighted ensemble of trajectories. *PLoS Comput. Biol.* **2016**, *12*, e1004611.

[74] Tse, M. J.; Chu, B. K.; Roy, M.; Read, E. L. DNA-Binding Kinetics Determines the Mechanism of Noise-Induced Switching in Gene Networks. *Biophysical Journal* **2015**, *109*, 1746–1757.

[75] Tse, M. J.; Chu, B. K.; Read, E. L. Mapping Epigenetic Landscapes of Gene Regulatory Networks by Adaptive Weighted Ensemble Sampling. *Biophysical Journal* **2016**, *110*, 494A–495A.

[76] Daigle, B. J.; Roh, M. K.; Gillespie, D. T.; Petzold, L. R. Automated estimation of rare event probabilities in biochemical systems. *Journal of Chemical Physics* **2011**, *134*, 13.

[77] Roh, M. K.; Daigle, B. J.; Gillespie, D. T.; Petzold, L. R. State-dependent doubly weighted stochastic simulation algorithm for automatic characterization of stochastic biochemical rare events. *Journal of Chemical Physics* **2011**, *135*, 11.

[78] Fink, A. Natively unfolded proteins. *Curr. Opin. Struct. Biol.* **2005**, *15*, 35–41.

[79] Wright, P. E.; Dyson, H. J. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol.* **1999**, *293*, 321–31.

[80] Shoemaker, B. A.; Portman, J. J.; Wolynes, P. G. Speeding molecular recognition by using the folding funnel: the fly-casting mechanism. *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 8868–73.

[81] Zhou, H. X.; Pang, X.; Lu, C. Rate constants and mechanisms of intrinsically disordered proteins binding to structured targets. *Phys. Chem. Chem. Phys.* **2012**, *14*, 10466–10476.

[82] Iesmantavicius, V.; Dogan, J.; Jemth, P.; Teilum, K.; Kjaergaard, M. Helical propensity in an intrinsically disordered protein accelerates ligand binding. *Angew. Chem. Int. Ed.* **2014**, *53*, 1548–1551.

[83] Papadakos, G.; Sharma, A.; Lancaster, L. E.; Bowen, R.; Kaminska, R.; Leech, A. P.; Walker, D.; Redfield, C.; Kleanthous, C. Consequences of inducing intrinsic disorder in a high-affinity protein-protein interaction. *J. Am. Chem. Soc.* **2015**, *137*, 5252–5255.

[84] Gianni, S.; Morrone, A.; Giri, R.; Brunori, M. A folding-after-binding mechanism describes the recognition between the transactivation domain of c-Myb and the KIX domain of the CREB-binding protein. *Biochem. Biophys. Res. Commun.* **2012**, *428*, 205–209.

[85] Rogers, J. M.; Wong, C. T.; Clarke, J. Coupled folding and binding of the disordered protein PUMA does not require particular residual structure. *J. Am. Chem. Soc.* **2014**, *136*, 5197–5200.

[86] Kohn, J.; Plaxco, K. Engineering a signal transduction mechanism for protein-based biosensors. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 10841–10845.

[87] Meisner, W.; Sosnick, T. Fast folding of a helical protein initiated by the collision of unstructured chains. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 13478–13482.

[88] Wang, J.; Lu, Q.; Lu, H. P. Single-molecule dynamics reveals cooperative binding-folding in protein recognition. *PLoS Comput. Biol.* **2006**, *2*, e78.

[89] Chen, J. Intrinsically disordered p53 extreme C-terminus binds to S100B(beta) through "fly-casting". *J. Am. Chem. Soc.* **2009**, *131*, 2088–2089.

[90] Dickson, A.; Ahlstrom, L. S.; Brooks, C. L. Coupled folding and binding with 2D window-exchange umbrella sampling. *J. Comp. Chem.* **2016**, *37*, 587–594.

[91] Turjanski, A.; Gutkind, J.; Best, R.; Hummer, G. Binding-induced folding of a natively unstructured transcription factor. *PLoS Comput. Biol.* **2008**, *4*, e1000060.

[92] Huang, Y.; Liu, Z. Kinetic advantage of intrinsically disordered proteins in coupled folding-binding process: A critical assessment of the "fly-casting" mechanism. *J. Mol. Biol.* **2009**, *393*, 1143–1159.

[93] Zhou, G.; Pantelopulos, G. A.; Mukherjee, S.; Voelz, V. A. Bridging Microscopic and Macroscopic Mechanisms of p53-MDM2 Binding with Kinetic Network Models. *Biophysical Journal* **2017**, *113*, 785–793.

[94] Daniels, K. G.; Tonthat, N. K.; McClure, D. R.; Chang, Y. C.; Liu, X.; Schumacher, M. A.; Fierke, C. A.; Schmidler, S. C.; Oas, T. G. Ligand concentration regulates the pathways of coupled protein folding and binding. *J. Am. Chem. Soc.* **2014**, *136*, 822–825.

[95] Hammes, G. G.; Chang, Y. C.; Oas, T. G. Conformational selection or induced fit: A flux description of reaction mechanism. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 13737–13741.

[96] Oldfield, C. J.; Cheng, Y.; Cortese, M. S.; Romero, P.; Uversky, V. N.; Dunker, A. K. Coupled folding and binding with alpha-helix-forming molecular recognition elements. *Biochemistry* **2005**, *44*, 12454–12470.

[97] Greenblatt, M. S.; Bennett, W. P.; Hollstein, M.; Harris, C. C. Mutations in the p53 tumor suppressor gene: Clues to cancer etiology and molecular pathogenesis. *Cancer Res.* **1994**, *54*, 4855–4878.

[98] Go, N. Theoretical studies of protein folding. *Annu. Rev. Biophys. Bioeng.* **1983**, *12*, 183–210.

[99] Takada, S. Gō-ing for the prediction of protein folding mechanisms. *Proc. Natl. Acad. Sci. USA* **1999**, *96*, 11698–11700.

[100] Frembgen-Kesner, T.; Elcock, A. Striking effects of hydrodynamic interactions on the simulated diffusion and folding of proteins. *J. Chem. Theory Comput.* **2009**, *5*, 242–256.

[101] Zuckerman, D. M.; Chong, L. T. Weighted ensemble simulation: Review of methodology, applications, and software. *Ann. Rev. Biophys.* **2017**, *46*, 43–57.

[102] Kussie, P.; Gorina, S.; Marechal, V.; Elenbaas, B.; Moreau, J.; Levine, A. J.; Pavletich, N. Structure of the MDM2 oncoprotein bound to the p53 tumor suppressor transactivation domain. *Science* **1996**, *274*, 948–953.

[103] Huang, Y.; Liu, Z. Nonnative interactions in coupled folding and binding processes of intrinsically disordered proteins. *PLoS One* **2010**, *5*, e15375.

[104] Northrup, S.; Allison, S.; McCammon, J. Brownian dynamics simulation of diffusion-influenced bimolecular reactions. *J. Chem. Phys.* **1984**, *80*, 1517–1524.

[105] Ermak, D.; McCammon, J. Brownian dynamics with hydrodynamic interactions. *J. Chem. Phys.* **1978**, *69*, 1352–1360.

[106] Elcock, A. Molecular simulations of cotranslational protein folding: Fragment stabilities, folding cooperativity, and trapping in the ribosome. *PLoS Comput. Biol.* **2006**, *2*, e98.

[107] Hess, B.; Bekker, H.; Berendsen, H.; Fraaije, J. LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem.* **1997**, *18*, 1463–1472.

[108] Rojnuckarin, A.; Livesay, D. R.; Subramaniam, S. Bimolecular reaction simulation using Weighted Ensemble Brownian dynamics and the University of Houston Brownian Dynamics program. *Biophys. J.* **2000**, *79*, 686–693.

[109] Kiefhaber, T.; Bachmann, A.; Jensen, K. S. Dynamics and mechanisms of coupled protein folding and binding reactions. *Curr. Opin. Struct. Biol.* **2012**, *22*, 21–29.

[110] Schon, O.; Friedler, A.; Bycroft, M.; Freund, S.; Fersht, A. Molecular mechanism of the interaction between MDM2 and p53. *J. Mol. Biol.* **2002**, *323*, 491–501.

[111] Clementi, C.; Nymeyer, H.; Onuchic, J. Topological and energetic factors: What determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? An investigation for small globular proteins. *J. Mol. Biol.* **2000**, *298*, 937–953.

[112] Koga, N.; Takada, S. Roles of native topology and chain-length scaling in protein folding: A simulation study with a Gō-like model. *J. Mol. Biol.* **2001**, *313*, 171–180.

[113] Garcia de la Torre, J.; Huertas, M. L.; Carrasco, B. Calculation of hydrodynamic properties of globular proteins from their atomic-level structure. *Biophys. J.* **2000**, *78*, 719–730.

[114] Frembgen-Kesner, T.; Elcock, A. Absolute protein-protein association rate constants from flexible coarse-grained Brownian dynamics simulations: The role of intermolecular hydrodynamic interactions in barnase-barstar association. *Biophys. J.* **2010**, *99*, L75–L77.

[115] Chodera, J. D.; Singhal, N.; Pande, V. S.; Dill, K. A.; Swope, W. C. Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics. *The Journal of Chemical Physics* **2007**, *126*, 155101.

[116] Noé, F.; Horenko, I.; Schütte, C.; Smith, J. C. Hierarchical analysis of conformational dynamics in biomolecules: Transition networks of metastable states. *The Journal of Chemical Physics* **2007**, *126*, 155102.

[117] Liu, J.; Faeder, J. R.; Camacho, C. J. Toward a quantitative theory of intrinsically disordered proteins and their function. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 19819–19823.

[118] Schreiber, G.; Haran, G.; Zhou, H. Fundamental Aspects of Protein - Protein Association Kinetics. *Chem. Rev.* **2009**, *109*, 839–860.

[119] Gabdoulline, R.; Wade, R. Simulation of the diffusional association of barnase and barstar. *Biophys. J.* **1997**, *72*, 1917–29.

[120] Northrup, S.; Erickson, H. Kinetics of protein-protein association explained by Brownian dynamics computer simulation. *Proc. Natl. Acad. Sci. USA* **1992**, *89*, 3338–3342.

[121] Zhou, H. Enhancement of Protein-Protein Association Rate by Interaction Potential : Accuracy of Prediction Based on Local Boltzmann Factor. *Biophys. J.* **1997**, *73*, 2441–2445.

[122] Camacho, C.; Kimura, S.; DeLisi, C.; Vajda, S. Kinetics of desolvation-mediated protein-protein binding. *Biophys. J.* **2000**, *78*, 1094–105.

[123] Schlosshauer, M.; Baker, D. Realistic protein–protein association rates from a simple diffusional model neglecting long-range interactions, free energy barriers, and landscape ruggedness. *Prot. Sci.* **2004**, *13*, 1660–1669.

[124] Alsallaq, R.; Zhou, H. Prediction of protein-protein association rates from a transition-state theory. *Structure (London, England : 1993)* **2007**, *15*, 215–24.

[125] Schreiber, G.; Fersht, A. Rapid, electrostatically assisted association of proteins. *Nat. Struct. Biol.* **1996**, *3*, 427–431.

[126] Gabdoulline, R. R.; Wade, R. C. Effective charges for macromolecules in solvent. *J. Phys. Chem.* **1996**, *100*, 3868–3878.

[127] Buckle, A.; Schreiber, G.; Fersht, A. Protein-protein recognition: crystal structural analysis of a barnase-barstar complex at 2.0-A resolution. *Biochemistry* **1994**, *33*, 8878–8889.

[128] Rotne, J.; Prager, S. Variational treatment of hydrodynamic interaction in polymers. *J. Chem. Phys.* **1969**, *50*, 4831–4837.

[129] Yamakawa, H. Transport properties of polymer chains in dilute solution - hydrodynamic interaction. *J. Chem. Phys.* **1970**, *53*, 436–443.

[130] Dlugosz, M.; Antosiewicz, J.; Zielinkski, P.; Trylska, J. Contributions of far-field hydrodynamic interactions to the kinetics of electrostatically driven molecular association. *J. Phys. Chem. B* **2012**, *116*, 5437–5447.

[131] Martin, C.; Richard, V.; Salem, M.; Hartley, R.; Mauguen, Y. Refinement and structural analysis of barnase at 1.5 Å resolution. *Acta Crystallogr. Sect. D* **1999**, *D55*, 386–398.

[132] Ratnaparkhi, G. S.; Ramachandran, S.; Udgaonkar, J. B.; Varadarajan, R. Discrepancies between the NMR and X-ray structures of uncomplexed barstar: analysis suggests that packing densities of protein structures determined by NMR are unreliable. *Biochemistry* **1998**, *37*, 6958–6966.

[133] Zhang, B.; Jasnow, D.; Zuckerman, D. Efficient and verified simulation of a path ensemble for conformational change in a united-residue model of calmodulin. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 18043–18048.

[134] Zwier, M. C.; Kaus, J. W.; Chong, L. T. Efficient explicit-solvent molecular dynamics simulations of molecular association kinetics: Methane-methane, Na+/Cl-, methane/benzene, and K+/18-crown-6 ether. *J. Chem. Theory Comput.* **2011**, *7*, 1189–1197.

[135] Plattner, N.; Doerr, S.; De Fabritiis, G.; Noe, F. Complete protein-protein association kinetics in atomic detail revealed by molecular dynamics simulations and Markov modelling. *Nature Chemistry* **2017**, *9*, 1005–1011.

[136] Best, R.; Hummer, G. Optimized molecular dynamics force fields applied to helix-coil transition of polypeptides. *J. Phys. Chem. B* **2009**, *113*, 9004–9015.

[137] Jorgensen, W.; Chandrasekhar, J.; Madura, J.; Impey, R.; Klein, M. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926–935.

[138] DeGrave, A. J.; Ha, J. H.; Loh, S. N.; Chong, L. T. Large enhancement of response times of a protein conformational switch by computational design. *Nature Communications* **2018**, *9*, 9.

[139] Dickson, A.; Lotz, S. D. Multiple ligand unbinding pathways and ligand-induced destabilization revealed by WExplore. *Biophys. J.* **2017**, *620*, 620–629.

[140] Beauchamp, K. A.; Bowman, G. R.; Lane, T. J.; Maibaum, L.; Haque, I. S.; Pande, V. S. MSMBuilder2: Modeling Conformational Dynamics on the Picosecond to Millisecond Scale. *Journal of Chemical Theory and Computation* **2011**, *7*, 3412–3419.

[141] Bastian, M.; Heymann, S.; Jacomy, M. Gephi: An Open Source Software for Exploring and Manipulating Networks. **2009**,

[142] Jacomy, M.; Venturini, T.; Heymann, S.; Bastian, M. ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. *Plos One* **2014**, *9*, 12.

[143] Michaud-Agrawal, N.; Denning, E. J.; Woolf, T. B.; Beckstein, O. Software News and Updates MDAnalysis: A Toolkit for the Analysis of Molecular Dynamics Simulations. *Journal of Computational Chemistry* **2011**, *32*, 2319–2327.

[144] Richard J. Gowers,; Max Linke,; Jonathan Barnoud,; Tyler J. E. Reddy,; Manuel N. Melo,; Sean L. Seyler,; Jan Domański,; David L. Dotson,; Sébastien Buchoux,; Ian M. Kenney,; Oliver Beckstein, MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations. **2016**, 98 – 105.

[145] Shrake, A.; Rupley, J. A. ENVIRONMENT AND EXPOSURE TO SOLVENT OF PROTEIN ATOMS - LYSOZYME AND INSULIN. *Journal of Molecular Biology* **1973**, *79*, 351–371.

[146] McGibbon, R. T.; Beauchamp, K. A.; Harrigan, M. P.; Klein, C.; Swails, J. M.; Hernández, C. X.; Schwantes, C. R.; Wang, L.-P.; Lane, T. J.; Pande, V. S. MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophysical Journal* **2015**, *109*, 1528 – 1532.

[147] Schreiber, G.; Fersht, A. Interaction of barnase with its polypeptide inhibitor barstar studied by protein engineering. *Biochemistry* **1993**, *32*, 5145–5150.

[148] Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins* **2010**, *78*, 1950–1958.

[149] Huang, D. M.; Chandler, D. Cavity formation and the drying transition in the Lennard-Jones fluid. *Physical Review E* **2000**, *61*, 1501–1506.

[150] Huang, D. M.; Chandler, D. The hydrophobic effect and the influence of solute-solvent attractions. *Journal of Physical Chemistry B* **2002**, *106*, 2047–2053.

[151] Huang, X.; Margulis, C. J.; Berne, B. J. Dewetting-induced collapse of hydrophobic particles. *Proceedings of the National Academy of Sciences of the United States of America* **2003**, *100*, 11953–11958.

[152] Berne, B. J.; Weeks, J. D.; Zhou, R. H. *Annual Review of Physical Chemistry*; Annual Review of Physical Chemistry; Annual Reviews: Palo Alto, 2009; Vol. 60; pp 85–103.

[153] Zhang, X. Y.; Zhu, Y. X.; Granick, S. Hydrophobicity at a Janus interface. *Science* **2002**, *295*, 663–666.

[154] Young, T.; Abel, R.; Kim, B.; Berne, B. J.; Friesner, R. A. Motifs for molecular recognition exploiting hydrophobic enclosure in protein-ligand binding. *Proceedings of the National Academy of Sciences of the United States of America* **2007**, *104*, 808–813.

[155] Young, T.; Hua, L.; Huang, X. H.; Abel, R.; Friesner, R.; Berne, B. J. Dewetting transitions in protein cavities. *Proteins-Structure Function and Bioinformatics* **2010**, *78*, 1856–1869.