EFFICIENT COMPUTER SIMULATIONS OF PROTEIN-PEPTIDE BINDING USING WEIGHTED ENSEMBLE SAMPLING

by

Matthew C. Zwier

B.S., Hope College, 2003

M.S., University of Illinois at Urbana-Champaign, 2006

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

2013

UNIVERSITY OF PITTSBURGH

DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

Matthew C. Zwier

It was defended on

August 6, 2013

and approved by

Lillian T. Chong, Ph. D., Associate Professor of Chemistry

Rob A. Coalson, Ph. D., Professor of Chemistry and Professor of Physics

Kenneth D. Jordan, Ph. D., Professor of Chemistry

Daniel M. Zuckerman, Ph. D., Associate Professor of Computational and Systems Biology

Dissertation Director: Lillian T. Chong, Ph. D., Associate Professor of Chemistry

EFFICIENT COMPUTER SIMULATIONS OF PROTEIN-PEPTIDE BINDING USING WEIGHTED ENSEMBLE SAMPLING

Matthew C. Zwier, PhD

University of Pittsburgh, 2013

Molecular dynamics simulations can, in principle, provide detailed views of protein-protein association processes. However, these processes frequently occur on timescales inaccessible on current computing resources. These are not particularly slow processes, but rather they are rare — fast but infrequent. The weighted ensemble (WE) sampling approach provides a way to exploit this separation of timescales and focus computing power efficiently on rare events. In this work, it is demonstrated that WE sampling can be used to efficiently obtain kinetic rate constants, pathways, and energy landscapes of molecular association processes. Chapter 1 of this dissertation further discusses the need for enhanced sampling techniques like the WE approach. In Chapter 2, WE sampling is used to study the kinetics of association of four model molecular recognition systems (methane/methane, Na^+/Cl^-, methane/benzene, and K^+/18-crown-6 ether) using molecular dynamics (MD) simulations in explicit water. WE sampling reproduces straightforward "brute force" results while increasing the efficiency of sampling by up to three orders of magnitude. Importantly, the efficiency of WE simulation increases with increasing complexity of the systems under consideration. In Chapter 3, weighted ensemble Brownian dynamics (BD) simulations are used to explore the association between a 13-residue fragment of the p53 tumor suppressor and the MDM2 oncoprotein. The association rates obtained compare favorably with experiment. By directly comparing both flexible and pre-organized variants of p53, it is shown that the "fly-casting" effect, by which natively unstructured proteins may increase their association rates, is not significant in MDM2-p53 peptide binding. Including hydrodynamic interactions in the simulation model dramatically alters the association rate, indicating that the detailed motion of solvent may have substantial effects on the kinetics of protein-protein association. In Chapter 4, an all-atom molecular dynamics simulation of p53-MDM2 binding is described. We obtain an association rate that agrees with the experimental value. The free energy landscape of binding is "funnel-like", downhill after the initial encounter between p53 and MDM2. Together, the studies described here establish that WE sampling is highly effective in simulating rare molecular association events.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

A dissertation is the capstone achievement of a graduate career, demonstrating that the writer is capable of independent thought and action. A dissertation is therefore something of a paradox, as I could not be writing this without having depended and continuing to depend on the support and assistance of so many.

I thank my advisor, Dr. Chong, for her consistent and consistently graceful support, for knowing, usually before I did, when I needed time to explore and when I needed firm direction or correction. I thank my committee, for their input and guidance. I thank Dr. Toby Chapman, who at an ACS job-hunting meeting suggested I consider joining the department at the University of Pittsburgh. I thank the members of the Chong group and Dr. Joshua Adelman, for inspiration, commiseration, and frequent raucous laughter. I thank my parents, who gave their time and energy so that my wife could complete her own studies while allowing our daughter to spend all her days surrounded by family. I thank my daughter Claire Lucia, whose clear light has shone so brightly in my life these four years. To my sons, Joshua Cecil and Benjamin Jude: you won't remember this time in words or images, but I won't ever forget our late night couch time. I thank my child Olivia or Oliver for watching over our family from above for the last year and a half, and Sister Marie Jude whose prayers, I'm convinced, assisted me in finishing this dissertation.

Above all, I thank my wife Karen. Thank you for being my companion. Thank you for your support, your sacrifices, your stability. Thank you for needing the same of me; I pray I will be these ever more for you.

1.0 THE NEED FOR ENHANCED SAMPLING

This chapter is based on a review article previously published as: Zwier, M. C., and Chong, L. T. (2010) Reaching biological timescales with all-atom molecular dynamics simulations. *Curr Opin Pharmacol 10*, 745–752

1.1 INTRODUCTION

Many biological processes — including enzyme catalysis, signal transduction, and protein-protein binding — involve protein motions that occur on multiple timescales.[1] As illustrated in Figure 1.1, these motions include ps-ns dynamics of side chains, ns-μs relative motions of protein domains, and μs-ms allosteric transitions.[2] Furthermore, the shorter timescale dynamics can influence and be influenced by longer-timescale motions.[3,4] The flexibility of proteins and the associated ensemble of alternate conformational states are important for many pharmaceutically-relevant species.[3,5,6]

Although X-ray crystallography or NMR spectroscopy can provide ensemble-averaged structures of certain conformational states, they cannot always characterize short-lived or unstructured species, such as transient binding pockets,[8] alternative conformations of active sites,[9] or proteins with intrinsically disordered regions,[10,11] and it may be precisely those species that could lead to new classes of pharmaceuticals.[6,12] Molecular dynamics (MD) simulations can complement experiments by providing the time resolution and atomic detail necessary for monitoring the step-by-step progression of conformational changes (e.g. the opening and closing of active-site protein "flaps"). Given sufficient computational resources, such simulations can span multiple timescales, revealing how fast, local fluctuations (ps-ns) might facilitate slower, functionally-relevant collective motions of the protein ($\geq \mu$s), providing detailed views of the mechanisms of conformational transitions. However, typical computing resources limit these simulations, which ideally include explicit water molecules, to the nanosecond timescale. Thus, direct "brute force" simulations — simply running simulations for a sufficiently long time (*i.e.* many times longer than the slowest event of interest) — have limited use in capturing biologically-relevant

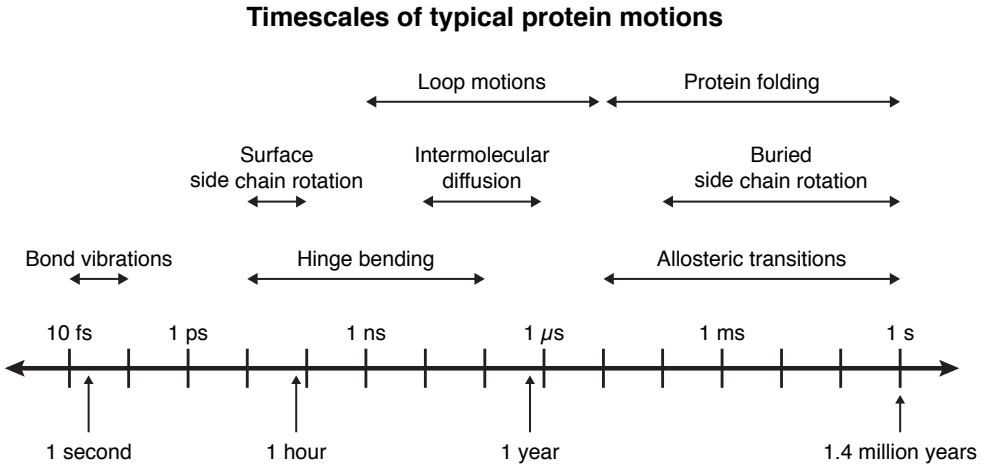


Figure 1.1: Timescales of typical protein motions and estimated computational time to simulate them. Motions and their corresponding timescales are indicated above the axis. Below the axis is a rough estimate of the amount of "wallclock" time required to perform a molecular dynamics (MD) simulation of a typically-sized protein-protein complex solvated in explicit water (~45,000 atoms) on a typical (2.6 GHz dual-core) desktop computer. To capture the fastest motions of proteins (*i.e.* bond vibrations), MD simulations must employ femtosecond time steps; a large number of simulation steps are therefore required to reach biological timescales, making MD simulations of protein systems very computationally expensive. For the computer described above, tens of nanoseconds of dynamics are accessible within weeks, but to reach the millisecond timescale would require millennia. (Timescales are from Refs. 2 and 7, except for intermolecular diffusion, which is derived from Ref. 7 assuming 1 mM concentration of the diffusing species.)

motions (*e.g.* induced-fit binding[13]). As illustrated in Figure 1.1, many biologically-relevant motions (fs-ns) are readily accessible to modern computers, but many motions which may be of interest (μs and beyond) are far out of reach.

The desire to access biological timescales with MD simulations has driven the development of innovative enhanced sampling techniques. These techniques invariably increase computational throughput at the cost of introducing additional assumptions, *e.g.* the system under study is strictly at equilibrium, or that initial and final states of a transition can be unambiguously identified and rigorously defined. Common to all these enhanced-sampling techniques is the assumption of a separation of timescales, where a process has a long characteristic time not because the transitions involved are slow, but rather because the transitions are rare, with long waiting times between otherwise fast events. It is the elimination of this waiting time that allows these techniques to access biologically-relevant timescales.

Here, we discuss recent developments in both brute force simulation and enhanced sampling techniques. Due to space constraints, we have restricted our discussion to methods which involve motion on a single, unmodified free energy surface.[a] The distinct advantage of such approaches is that the dynamics of the system under study are completely unperturbed, and furthermore, realistic kinetic information may be readily extracted from simulations; this information (e.g. kinetic rates and timescales of motion) provides another means of validating simulation with experiment. On the other hand, this admittedly limited scope precludes us from discussing in detail other promising enhanced sampling methods,[2,14] including those that under some circumstances can yield realistic kinetic rates.[15,16] In preparing this review, we found it helpful to summarize the similarities and differences among each of the many techniques discussed below. As shown in Figure 1.2, the techniques discussed here can be grouped according to whether they provide continuous, atomically-detailed pathways of transitions between two states or not, and also according to the amount of *a priori* information required to construct a simulation. Generally, methods that require more information about a system involve more assumptions, but can generate pathways (or trajectories) of the biological event of interest with greater efficiency.

[a]The free energy surface of a chemical system — free energy as a function of atomic coordinates — completely defines its dynamics; it depends on the microscopic interactions between atoms and macroscopic thermodynamic variables such as temperature, volume, and pressure.

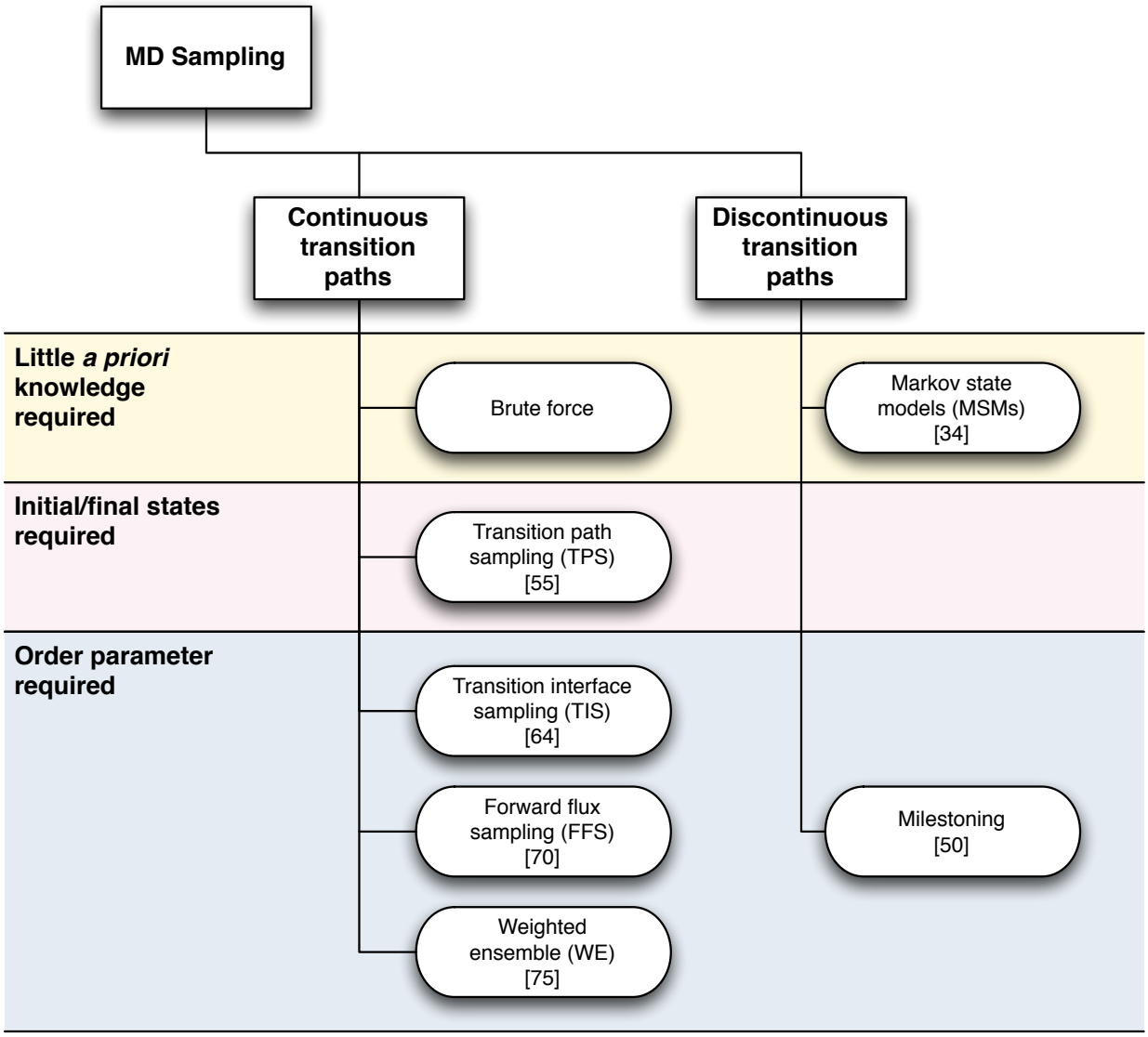


Figure 1.2: Techniques capable of accessing biomolecular timescales. When feasible, large-scale brute force dynamics with explicit consideration of solvent provide the greatest possible detail with the fewest possible assumptions. Where such simulations are not possible, other enhanced sampling techniques can be used to obtain kinetic information, transition paths, or both. In general, where increased *a priori* knowledge (yellow, red, and blue regions, top to bottom) is required to run a simulation, greater efficiency gains compared to brute force dynamics are possible; the price of increased throughput is often a greater number of assumptions about or restrictions on the system being simulated. For convenience, references are provided in which each enhanced sampling method was first presented for simulation of biochemical systems. In the case of Markov state models, the first discussion of using multiple shorter-timescale simulations to reach longer-timescale kinetics of biological systems is cited.

1.2 BRUTE FORCE DYNAMICS

When feasible, brute force simulations provide the greatest possible detail with the fewest possible assumptions relative to other MD-based sampling techniques. As highlighted in Figure 1.2, very little *a priori* knowledge is required to run a brute force dynamics simulation — generally only a force field and a representative initial structure. These are not trivial concerns, particularly as the force field completely defines the thermodynamics of a system,[17] and slight imbalances in force field parameterization can significantly alter the results of simulations,[18] especially for long time scales;[19] however, agreement with experiment is generally acceptable for the biological questions being addressed.[17,20]

While the cost of running these brute force simulations is high, computing resources continue to grow in size and power. Continued optimization of MD software, coupled with the decreasing cost of commodity hardware — particularly multi-core processors — has played a key role in reaching microsecond timescales for typical biological systems (*e.g.* $\sim 10^4 - 10^5$ atoms), timescales that were inaccessible only five years ago. Furthermore, recently-developed hardware specialized for performing MD simulations is poised to generate microseconds of dynamics per day on similarly-sized biological systems.[20,21]

Though the overall computing landscape has not changed dramatically in the last year, the use of general-purpose graphics processing units (GPGPUs) in the field of molecular dynamics continues to grow (*cf., e.g.,* 22–26, AMBER 11,[27] GROMACS,[28] OpenMM[29]). While impressive gains in throughput are possible with GPGPUs for certain calculations involved in an MD simulation,[30] the small on-board memory and high cost of communication with GPGPUs has severely limited the ability of GPGPUs to accelerate MD calculations effectively.[25] This communications problem must be overcome before GPGPU-accelerated MD calculations can hope to supplant traditional large-scale parallel MD calculations, which continue to provide access to long timescales[31] and large systems.[32]

1.3 KINETIC "GLUE"

The high computational cost of accessing biological timescales with brute force simulations (see Figure 1.1) has led to the development of several methods that attempt to obtain long-timescale information by "gluing together" shorter-timescale simulations. Two examples with proven applicability to biological systems are Markov state models and Milestoning.

1.3.1 Markov state models

Markov state models (MSMs) — discrete-state kinetic models — seek to describe the behavior of a system in terms of a finite number of metastable (relatively long-lived) states and the rates of transitions between them.[33–40] These transition networks are generally constructed by grouping many conformations from multiple, relatively short brute force simulations such that conformational transitions within states (groups) are common but transitions between states are rare.[34,36] The requirements for this statewise decomposition of conformational space can be expressed several ways: (A) the timescales for intrastate transitions are short but the timescales for interstate transitions are long; (B) the probability of having moved from some state i at time t to some other state j at time $t + dt$ depends only on the lag time dt; or (C) the probability of moving from state i to state j does not depend on how the system came to be in state i. It is the first property (A, the separation of timescales) that enables MSMs to capture long-timescale kinetic information from short-timescale dynamics, and it is the lattermost property (C, "memorylessness") that is perhaps most familiar as the defining property of a Markov process, hence the name Markov state model.

Construction of a Markov state model reduces many individual conformations and the detailed dynamics connecting them into a discrete set of states, their relative probabilities, and a matrix of (lag-time-dependent) transition probabilities between each pair of states.[34] Each state represents a distribution of quickly-interconverting conformations, which may (but need not) correspond to experimentally well-defined populations, such as those that might be identified by NMR spectroscopy. From the transition probability matrix and state populations, quantities such as the overall transition rate between two states, the set of paths connecting them, and the contribution of each path to the overall rate can be calculated.[38,41] Thus, MSMs present a coarse-grained view of both the conformational space of the system and the dynamics within it. MSMs have been used primarily to study protein folding mechanisms, showing good agreement with experiment in both structural information and folding rates.[35,36,39,40,42–47] However, these models are generally applicable to the problems of identifying kinetically-distinct states of proteins under given conditions (*e.g.* temperature, ionic strength) and determining the kinetics of slow conformational transitions. The particular strength of MSMs may in fact be their ability to indicate native state ensembles of structures, providing information complementary to that provided by X-ray crystallography and NMR spectroscopy experiments.

It should be noted that short trajectories suitable for MSM construction may be generated with replica exchange molecular dynamics (REMD),[48] a popular method for enhancing sampling of con-

formational space.[49] Many copies ("replicas") of a system are simulated in parallel at multiple temperatures, and configurations are occasionally swapped between temperatures (generally according to a Boltzmann criterion). Although this technique does not generally permit the extraction of reaction rates, Markov state models can be constructed from the brief trajectory segments between replica exchanges, and reaction rates may then be determined from the MSMs.[15]

1.3.2 Milestoning

The Milestoning approach developed by Elber and co-workers also uses kinetic information from shorter-timescale simulations to infer long-timescale kinetics.[50,51] Unlike MSMs, from which definitions of states may be obtained, Milestoning requires *a priori* definitions of initial and final states and a one-dimensional order parameter[b] that specifies "how far along" a simulation is in a transition between the initial and final states. This order parameter is divided by a number of surfaces ("milestones") and equilibrated ensembles of simulations are prepared at each milestone. In a second simulation phase, the constraint holding simulations to milestone surfaces is removed, and as simulations reach neighboring milestones, the time required by each simulation to reach a neighboring milestone (in either a forward or backward direction) is recorded. This simulation between milestones — rather than between initial and final states — effectively eliminates the waiting time that would otherwise be sampled by brute force simulations.

The central assumption of Milestoning is that all degrees of freedom other than the order parameter relax completely between subsequent milestones. Under this assumption, the "incubation times" between milestones, obtained as described above, may be transformed into the global first-passage time distribution, the probability distribution of times required to reach the final state from the initial state.[51] When a single timescale dominates a system, the first-passage time distribution is exponential and the reaction rate is simply the inverse of the mean first passage time, but in a system where multiple timescales are important, the first passage time distribution is capable of describing the resulting non-exponential behavior,[52] as has been demonstrated explicitly for Milestoning simulations.[50,51] This added flexibility reflects the fact that the central assumption of Milestoning (complete relaxation along all non-order-parameter coordinates) is less restrictive than that of Markov state models (where complete relaxation is assumed in all degrees of freedom within each state). However, the

[b] This order parameter (also called a "progress coordinate" by some practitioners) is a scalar value which varies continuously and (generally) monotonically between particular values at the initial and final states. It may, but does not necessarily, reflect a formal reaction coordinate.

cost of this increased detail in the determination of kinetics is reduced detail in determination of conformational states; initial and final states, the order parameter, and adequate milestones must be known prior to a Milestoning simulation of a system. The utility of the Milestoning approach is demonstrated well by a recent study involving the recovery stroke of myosin (a millisecond process), which provided experimentally-testable mechanistic insights and a rate consistent with experiment.[53]

1.4 PATH SAMPLING TECHNIQUES

Path sampling approaches seek to determine the detailed dynamics of pathways between well-defined metastable states. These techniques are complementary to MSMs and Milestoning, which can provide definitions of metastable states and detailed kinetics of transitions between well-defined metastable states, respectively. The most widely used methods in recent years are transition path sampling (TPS) and its variants, such as transition interface sampling (TIS); forward flux sampling (FFS); and weighted ensemble (WE) sampling.

1.4.1 Transition path sampling

Transition path sampling (TPS), which is based on early work by Pratt,[54] was first presented more than a decade ago[55] and has subsequently been extensively employed, refined, and reviewed.[56–59] TPS is a Monte Carlo sampling of MD-simulated paths between initial and final states, which (as highlighted in Figure 1.2) must be known *a priori*. Each path is typically generated by randomly selecting a segment of the previously-sampled path, perturbing its coordinates and/or momenta, and then "shooting off" MD trajectories both forward and backward in time from the perturbed segment;[57] thus, this scheme requires the dynamics of the system to be invariant under time reversal, *i.e.* the system must be at equilibrium.[58] The resulting set of paths between the initial and final states and their relative probabilities together provide a detailed picture of how transitions between the initial and final states progress. TPS does not directly provide kinetic information; rather, a subsequent (computationally-expensive) umbrella sampling calculation is required,[57] a limitation which led directly to the development of transition interface sampling (discussed below). As with all path sampling methods, the presence of long-lived intermediate states[60] or multiple distinct transition pathways separated by substantial free energy barriers[61] may severely reduce the effectiveness of TPS. Nonetheless, TPS is capable of describing rare

transitions in biological systems. Recently, TPS was used to determine the pathways of conformational change in the activation of photoactive yellow protein (PYP), predicting experimentally-detectable intermediates and suggesting experiments which can be used to validate the TPS results.[62] In a striking combination of a number of enhanced sampling techniques, Juraszek and Bolhuis used transition path sampling to determine the pathways of conformational change in folding and unfolding mechanisms of formin binding protein 28 (FBP28) and then map the free energy landscape of the protein; the computed unfolding barrier is in agreement with experiment.[63]

1.4.2 Transition interface sampling

The high computational cost of obtaining kinetic information with TPS inspired the development of transition interface sampling (TIS) and several variants thereof.[64–66] Transition interface sampling, along with FFS and WE, partitions an order parameter connecting initial and final states with several dividing surfaces ("interfaces"); this again represents an increase in the amount of information required to start a simulation (see Figure 1.2). In TIS, a Monte Carlo procedure very similar to TPS — including forward and backward shooting of MD trajectories — is used to sample paths between each pair of adjacent interfaces; the reaction rate is then determined by the flux out of the initial state and the conditional probabilities of reaching each interface in turn.[67] In this way, the paths and transition rate between initial and final states are determined simultaneously. Interface-based sampling methods like TIS may suffer greatly in efficiency if significant free-energy barriers exist between interfaces, particularly if the barriers must be surmounted in order to reach the next interface.[68]

1.4.3 Forward flux sampling

The forward flux sampling (FFS) method of Allen *et al.* was presented as an alternative to TPS and TIS without the requirement of microscopic reversibility, thus allowing path-sampling studies of nonequilibrium systems.[58,59,69–71] Like TIS, FFS requires well-defined initial and final states, an order parameter describing the transition between them, and partitioning of the order parameter by interfaces. Rather than using Monte Carlo techniques to sample transition paths, MD simulations — propagating forward in time only — are used to determine the paths between interfaces. When a dynamics trajectory reaches an interface, its coordinates and momenta at the interface are saved, then used to start a number of new simulations from the interface. The reaction rate is calculated in terms of a set of conditional crossing probabilities, and transition paths between initial and final states may be obtained by tracing completed

paths from the final state back to the initial state.[70] As in TIS and WE, high barriers between interfaces may cause a sharp drop in sampling efficiency, as simulations can progress to the next interface only rarely.[68] A particularly interesting feature of FFS is the existence of well-defined estimates for computational efficiency as functions of FFS simulation parameters (*e.g.* the number of interfaces), allowing for selection of efficient parameters.[72,73] FFS has been used primarily in simplified models of various systems (*cf.* Refs. 58 and 59), but it has also been applied to an all-atom folding simulation of the trp-cage mini-protein.[74]

1.4.4 Weighted ensemble sampling

The weighted ensemble (WE) sampling technique is conceptually quite similar to FFS, though it predates FFS by nearly a decade.[75] Originally conceived to accelerate sampling in Brownian dynamics simulations,[75–77] weighted ensemble sampling is asymptotically correct for a much broader class of stochastic simulations, including MD simulations.[78] WE sampling uses independent simulations, each carrying a statistical weight, to explore conformational space. Like TIS and FFS, WE sampling requires definitions of initial and final states, an order parameter, and the partitioning of space along the order parameter into bins. Simulations are propagated for a fixed time period, after which a statistically-rigorous reweighting procedure is used to keep the number of simulations in each bin constant without altering the total probability in each bin. Thus, as unoccupied bins become populated, more simulations are created with which to explore that region of phase space, and as simulations cross backwards into previously-traversed bins, they will likely be eliminated, reducing oversampling. As simulations reach the destination state, their probability weights are recycled to the initial state, establishing a steady-state flow of probability from the initial state to the final state. The resulting transition paths are continuous, and the macroscopic reaction rate is obtained simultaneously as the net flow of probability into the destination state.[75] WE sampling has a theoretical and algorithmic framework that naturally supports more than one order parameter, making it an attractive option for sampling rare events in systems that cannot be described with a single order parameter.[68] Achievement of a steady-state probability flow from the initial state to the final state may be accelerated using concepts developed from non-equilibrium umbrella sampling[79] (see Section A.3), partially ameliorating the difficulty shared by WE, TIS, and FFS of surmounting barriers between interfaces (*i.e.* within bins).[68]

WE sampling in the context of a residue-based Monte Carlo simulation has recently been used to study the kinetics and conformational transitions between the *apo* and *holo* forms of calmodulin, show-

ing excellent agreement and efficiency gains compared to brute force Monte Carlo sampling.[80] Our own group has recently determined that WE sampling in conjunction with MD simulations achieves high efficiency in modeling simple molecular association events (methane/methane, methane/benzene, Na^+/Cl^-, and 18-crown-6/K^+) in explicit water (see Chapter 2 and Reference 81), indicating that this approach is a promising one for studying protein/small-molecule and protein/protein interactions. Applications of the WE approach to the study of protein-peptide association kinetics and energy landscapes are discussed in Chapters 3 and 4.

1.5 CONCLUSIONS

Conformational changes in biologically-relevant systems span an enormous range of time scales, from picosecond dynamics of side chains through microsecond or slower dynamics of coordinated conformational transitions. All-atom MD simulations have typically been limited by computing power to microseconds of simulation time or less. Even so, with increasing computing power, brute force MD simulations continue to provide detailed views on biologically-relevant conformational transitions. Additionally, a number of enhanced sampling techniques have matured to the point of reaching biological timescales with MD simulations. The most promising avenue for exploration of the dynamics and kinetics of pharmacologically-relevant systems appears not to be any single MD sampling technique, but combinations of techniques that, when used together, yield far more information than any technique alone (*e.g.* Ref. 63). With advances in simulation approaches and computing power, MD simulation is becoming increasingly useful in providing detailed structural and mechanistic insight into biologically-relevant events that are of pharmaceutical interest.

1.6 ACKNOWLEDGEMENTS

This work was supported by an NSF CAREER award (MCB-0845216) to Lillian T. Chong and a University of Pittsburgh Arts & Sciences Fellowship to MCZ. We are grateful to John Chodera and Dan Zuckerman for insightful discussions and constructive comments on the manuscript.

1.7 APPENDIX: REFERENCE NOTES

The following notes highlight specific references used throughout this chapter.

* = of particular interest. ** = of outstanding interest.

15* Markov models are used to extract realistic kinetic information from replica exchange molecular dynamics simulations, which are generally unable to provide such information.

16* A unique study of how exact, unperturbed thermodynamic averages and (in some cases) realistic macroscopic reaction rates can be extracted from perturbed dynamics.

21* A lucid introduction to high-performance computing as applied to biomolecular systems.

32* A detailed discussion of the technical complexities involved in scaling molecular dynamics simulations to large core counts.

44** The folding of a small protein (the PinWW domain) is studied using short MD trajectories in explicit solvent, Markov state models, and transition path theory; folding occurs along multiple pathways, and therefore the authors argue that a probabilistic view of protein folding mechanisms is necessary.

51** An exceptionally clear and thorough description of the Milestoning procedure, with a superior balance of detailed theory and straightforward explanation, a forthright discussion of the strengths and weaknesses of the Milestoning method, and detailed comparison to similar methods.

58* A concise review of enhanced sampling techniques, including all those discussed here, focusing on biological applications.

59* Another review of enhanced sampling techniques, with a particular focus on forward flux sampling (FFS).

63** A model study demonstrating the great utility of combining multiple enhanced sampling techniques. The authors use three techniques, each according to its own particular strength, to elucidate the folding mechanism of a small protein (the FBP28 WW domain).

80* The first application of weighted ensemble sampling to large-scale conformational transitions of a protein switch (apo and holo forms of calmodulin).

2.0 THE EFFICIENCY OF THE WEIGHTED ENSEMBLE APPROACH FOR MOLECULAR ASSOCIATION SIMULATIONS

This chapter is based on a research article previously published as: Zwier, M. C.; Kaus, J. W.; Chong, L. T. *J Chem Theory Comput* 2011, *7*, 1189–1197.

2.1 INTRODUCTION

Proteins bind their partners in a highly specific manner. Understanding the mechanisms of these binding events is not only fundamentally interesting, but could also impact fields such as protein engineering, host-guest chemistry, and drug discovery. Atomistic molecular dynamics (MD) simulations can potentially offer the most detailed views of molecular recognition events, especially when performed with explicit solvent. However, only up to a microsecond of simulation is practical on typical computing resources, while protein binding events require microseconds and beyond.[1] It is therefore computationally prohibitive to capture these events by sufficiently long "brute force" simulations. Fortunately, the long timescales required for protein binding events are not necessarily because the actual events take a long time; instead, the events may be fast but infrequent, separated by long waiting times.

Path sampling approaches[50,55,64–66,69,70,75,82] aim to capture rare events by minimizing the simulation of long waiting times between events.[83] Weighted ensemble sampling[75] is one such approach which is rigorously correct for any type of stochastic simulation,[78] easily parallelized, and simultaneously provides both transition paths and their associated kinetics.[75] Weighted ensemble sampling has been applied to Brownian dynamics simulations of protein-protein binding,[75] protein-substrate binding,[77] protein folding,[76] Monte Carlo simulations of large-scale conformational transitions in the molecular switches calmodulin[80] and adenylate kinase,[84] and molecular dynamics simulations of alanine dipeptide in implicit solvent.[68]

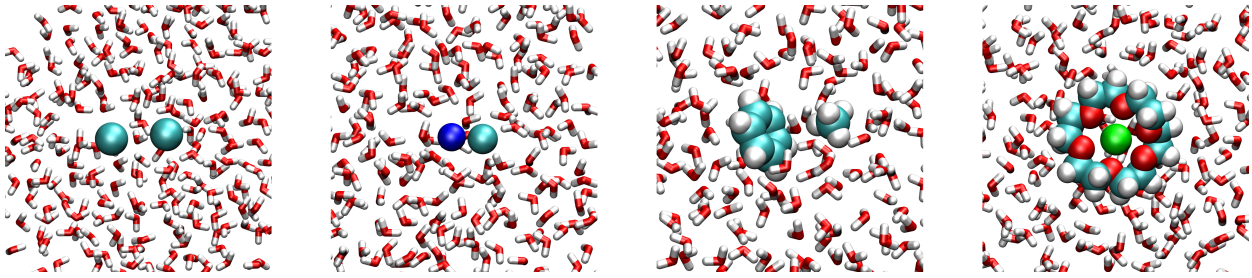


Figure 2.1: Molecular recognition systems of this study. From left to right, two methane molecules; Na^+/Cl^-; benzene and methane; and K^+ ion with 18-crown-6 ether. All systems were immersed in explicit water molecules. (Prepared with VMD.[105])

We apply the weighted ensemble path sampling approach with explicit-solvent MD simulations. Our goal is to determine the efficiency of the weighted ensemble approach relative to brute force simulation in sampling molecular associations for a range of well-studied systems: methane/methane,[85–90] Na^+/Cl^-,[91–100] methane/benzene,[101,102] and K^+/18-crown-6 ether[103,104] (Figure 2.1). These systems were chosen because of their small size and relatively low barriers to association ($\sim 2\ k_B T$); combined, these features make feasible the simulation of association events by brute force, providing us with opportunities to evaluate not only the efficiency of the weighted ensemble approach, but its validity as well.

2.2 THEORY

2.2.1 Overview of weighted ensemble sampling

Weighted ensemble sampling uses "statistical ratcheting" to efficiently sample rare events using stochastic simulations.[68,75,80,83] To monitor the progress of these simulations toward the rare event of interest (*i.e.* molecular association), a progress coordinate between the source (A, unbound) and destination (B, bound) states is defined by one or more order parameters; this progress coordinate is then divided into bins. A number of simulations are started in the unbound state A, which are then propagated for a fixed time τ. After this propagation time, if a simulation has progressed into a bin closer to the destination state B, its current state is used to start replicas of that simulation; these replicas diverge due to

the stochastic nature of the underlying dynamics. Alternatively, if the simulation has regressed toward the source state A, it is effectively terminated. This resampling procedure[78] involving the replication of productive simulations and termination of unproductive simulations is repeated at fixed intervals (τ, 2τ, 3τ, and so on) until the desired number of rare events (crossings into state B) is sampled. Once a simulation reaches the destination state B, it is removed from the destination state B and "recycled" as a new simulation starting from the source state A. As this propagation and resampling procedure is repeated, the transition path ensemble — an ensemble of continuous trajectories between the source and destination states — is generated. As shown in Figure 2.2, some common history is shared among this ensemble of trajectories, and each trajectory has a maximum length τN_τ after N_τ iterations of propagation and resampling. When the trajectories are generated using molecular dynamics simulations, a stochastic thermostat is required to allow for divergence of trajectories after resampling.

To maintain correct statistics and kinetics of the transition paths, each simulation is assigned an appropriate statistical weight. When simulations are replicated, their statistical weights are split; when simulations are terminated for regressing toward the source state A, their statistical weights are merged into existing replicas; and when simulations are terminated for reaching the destination state B, their statistical weights are removed from the destination state B and assigned to newly-created replicas in the initial state A.

2.2.2 Rate constants

Weighted ensemble sampling yields kinetic information as a simulation progresses. After steady-state probability recycling is attained, the rate constant k is given by the average probability current I_B into the destination state B:[52,75,80]

$$k = \langle I_B \rangle \tag{2.2.1}$$

where the angle brackets indicate a time average. Because the recycling procedure described above eliminates all probability from the destination state B at each resampling, the probability current I_B may be approximated as

$$I_B \approx \frac{P_B(\tau N_\tau)}{\tau} \tag{2.2.2}$$

where τ is the weighted ensemble propagation/resampling timestep and $P_B(\tau N_\tau)$ is probability contained in the destination state at time τN_τ (weighted ensemble iteration N_τ) immediately prior to recycling. Since $P_B(\tau N_\tau)$ must be monitored in order to ensure probability conservation during a weighted ensemble simulation, the rate constant k is obtained "for free."

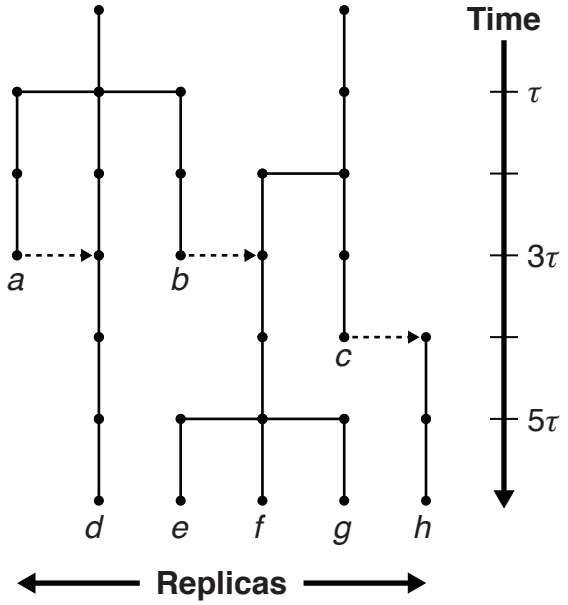


Figure 2.2: Schematic diagram of weighted ensemble molecular dynamics trajectories. Replication and termination of simulations occurs at intervals of τ, the propagation/resampling time. Trajectories a and b are terminated at $t = 3\tau$ and trajectory c reaches the destination state at $t = 4\tau$, at which time its statistical weight is assigned to a newly-created replica which traces out trajectory h; the dotted arrows indicate a transfer of statistical weight but not history. Trajectories are replicated at at $t = \tau$, $t = 2\tau$, and $t = 5\tau$. Note that trajectories e, f, and g share common history but are independent from trajectories d and h, which themselves are mutually independent.

The partially shared history of weighted ensemble trajectories results in highly correlated probability current measurements; that is, $I_B(\tau N_\tau)$ and $I_B(\tau[N_\tau - 1])$ are not statistically independent. The time average $\langle I_B \rangle$ may be computed in the usual way, but the associated confidence interval (encompassing the statistical error in the rate constant) must be computed with a method that accounts for the time correlation within I_B, such as Monte Carlo bootstrapping.[75,106,107]

On the other hand, the quantity accessible from brute force dynamics is not the probability current into the destination state, but is rather a set of elapsed times between completed transition events. That is, brute force simulation does not yield the rate constant directly, but rather the first passage time distribution. For transitions dominated by a single timescale, this distribution is exponential, and the rate constant is simply the inverse of the mean first passage time $\langle t_{\text{fp}} \rangle$:[52]

$$k = \langle t_{\text{fp}} \rangle^{-1} \tag{2.2.3}$$

It should be noted that these two methods for determining the rate constant k are alternative mathematical descriptions of the same underlying physical principles (for an extensive discussion, see Ref. 52). Thus, rate constants obtained from brute force and weighted ensemble simulations may be directly compared, given that the same model was used for propagating dynamics in both cases and that the confidence interval for the rate constant is calculated correctly for the weighted ensemble simulation.

2.2.3 Transition event durations

If a system exhibits rare but fast events, then the transition event duration t_{ed} — the amount of time it takes a transition to complete once it starts — is much less than the mean first passage time $\langle t_{\text{fp}} \rangle$ (which includes the waiting time between rare events):

$$t_{\text{ed}} \ll \langle t_{\text{fp}} \rangle$$

The cumulative probability distribution of t_{ed}, $F(t_{\text{ed}})$, is at least approximately an indicator of the extent of sampling of the transition pathways. Distinct pathways will have associated characteristic transition durations,[108] and as transition pathways are sampled, $F(t_{\text{ed}})$ will become better-resolved. Thus, the self-convergence of $F(t_{\text{ed}})$ is a strong indicator that the transition path ensemble has been adequately explored.

The cumulative transition event duration distribution $F(t_{\text{ed}})$ is built up directly from simulation trajectories simply by noting the time elapsed between exiting the source state A and entering the destination state B. In the brute force case, a set of event durations is transformed directly into a cumulative

distribution function in the usual manner (by counting the number of t_{ed} less than a specified value):

$$F(t_{\mathrm{ed}}) = \frac{1}{N}\sum_i h(t_{\mathrm{ed}(i)}) \tag{2.2.4}$$

where i indexes transitions, N is the number of transition events observed, $t_{\mathrm{ed}(i)}$ is the duration of transition event i, and h is an indicator function satisfying

$$h(t_{\mathrm{ed}(i)}) = \begin{cases} 1 & \text{if } t_{\mathrm{ed}(i)} \leq t_{\mathrm{ed}} \\ 0 & \text{otherwise} \end{cases}$$

Weighted ensemble simulations, on the other hand, yield not the set of event durations $\{t_{\mathrm{ed}}\}$ but a set $\{(t_{\mathrm{ed}}, w)\}$ of transition event durations and corresponding terminal statistical weights. These terminal weights partially encode the probability of arriving at the final state and so a weighted variation of Equation 2.2.4 must be used:

$$F(t_{\mathrm{ed}}) = \frac{\sum_i w_i h(t_{\mathrm{ed}(i)})}{\sum_i w_i} \tag{2.2.5}$$

There are several advantages to describing the transition event duration distribution as an (empirical) cumulative distribution function. First, rigorous confidence bands may be assigned to empirical distribution functions,[109,110] allowing one to assign error bars to the entire t_{ed} distribution and facilitating the comparison of simulation results. Second, the number of points N_e in a realization of $F(t_{\mathrm{ed}})$ is equal to the number of unique transition event durations sampled, and as such can be considered a statistical sample size for the purposes of quantifying sampling, even in the weighted ensemble case. For this same reason, Equation 2.2.5, despite being cast in a weighted form, describes a formal empirical distribution function and is therefore an unbiased estimator of the true cumulative distribution function.[110]

2.2.4 Relative efficiency of weighted ensemble simulations

Any meaningful metric for comparing the relative efficiencies of weighted ensemble and brute force simulations must account for not only the computational expense of obtaining an estimate on a quantity such as the rate constant, but also the uncertainty of that estimate. In other words, an efficiency metric must take error bars into account. For a given quantity like the reaction rate k, we define the efficiency of weighted ensemble sampling relative to brute force as

$$S = \frac{t_{(\mathrm{WE})}}{t_{\mathrm{eff}}} \tag{2.2.6}$$

where $t_{(\mathrm{WE})}$ is the aggregate weighted ensemble simulation time (not overcounting shared history) and t_{eff} is the effective amount of brute force simulation time that would be required to obtain an estimate

with the same size error bar as that obtained from a weighted ensemble simulation. Consideration of the error structure of brute force simulations and application of Equation 2.2.6 gives the following efficiency metrics S_k and S_{ed} for sampling of the association rate constant k and t_{ed} distribution, respectively:

$$S_k = \frac{t_{(\text{BF})}}{t_{(\text{WE})}} \left(\frac{\Delta k^*_{(\text{BF})}}{\Delta k^*_{(\text{WE})}} \right)^2 \tag{2.2.7}$$

$$S_{\text{ed}} = \frac{t_{(\text{BF})}}{t_{(\text{WE})}} \left(\frac{N_{e(\text{WE})}}{N_{e(\text{BF})}} \right) \tag{2.2.8}$$

where t represents total simulation time, Δk^* is the width of the 95% confidence interval on the rate constant k relative to the time average $\langle k \rangle$, and N_e is the number of unique time values in the empirical distribution function $F(t_{\text{ed}})$; the subscripts (BF) and (WE) represent values from brute force and weighted ensemble simulations, respectively. Detailed derivations of Equations 2.2.7 and 2.2.8 are provided in Sections 2.7.2 and 2.7.3.

2.3 METHODS

2.3.1 Model systems

Four systems were used to test the feasibility of using weighted ensemble sampling with explicit-solvent MD simulations to study molecular association events. These systems all possess simple, one-dimensional progress coordinates by which it is possible to unambiguously define "how close to binding" a simulation is at any point in time. All systems were immersed in boxes of explicit water molecules. The model systems in order of progressively more challenging features are described below.

Methane/methane. This system is a simple example of a hydrophobic interaction. The natural progress coordinate of this system is simply the center-to-center distance between the two methane molecules.

Na^+/Cl^-. This system is a simple example of an electrostatic interaction. The natural progress coordinate of this system is the center-to-center distance between the two ions.

Methane/benzene. Like the methane/methane system, methane/benzene is a model of hydrophobic interactions, but unlike the previous two systems, it does not exhibit an effective spherical symmetry. However, our brute force simulations of this system revealed that the condensed-phase bound state involves precession of the methane molecule about the surface of the benzene ring. Therefore, de-

spite the broken spherical symmetry, the natural progress coordinate for this system is effectively one-dimensional and was taken to be the distance between the methane carbon and the center of mass of the benzene carbon atoms.

K^+/18-crown-6 ether. This system is a simple example of the binding of a (trivially) rigid substrate (K^+) by a flexible partner (18-crown-6 ether). Like methane/benzene, this system does not exhibit effective spherical symmetry. However, both simulation[103,104] and X-ray crystallography[111] have indicated that the bound structure consists of the K^+ ion co-planar with the crown ether oxygen atoms. The natural progress coordinate for this system is therefore the distance between the K^+ ion and the center of mass of the ether oxygen atoms.

2.3.2 Simulation details

Both brute force and weighted ensemble simulations were performed using the GROMACS 4.0.5 software package.[28] Production dynamics (both brute force and weighted ensemble) were propagated in the canonical (NVT) ensemble at 300 K using a Langevin thermostat[112] (coupling time 1 ps). Van der Waals interactions were switched off smoothly between 8 and 9 Å; to account for the truncation of the van der Waals interactions, a long-range analytical dispersion correction[113] was applied to energy and pressure. Real-space electrostatic interactions were truncated at 10 Å. Long range electrostatic interactions were calculated using particle mesh Ewald[114] (PME) summation. Bonds to hydrogen atoms were constrained to their equilibrium lengths using LINCS,[115] allowing for a 2 fs integration timestep.

Each model system was constructed in its unbound state and solvated in a dodecahedral periodic box with a minimum 12-Å clearance between the solutes and the box walls. Following a 1000-step steepest-descent energy minimization, each system was subjected to 20 ps of NVT thermal equilibration followed by 1 ns of constant-pressure (NPT) density equilibration using a weak isotropic Berendsen barostat[116] (reference pressure 1 bar, coupling time 5 ps, and compressibility 4.5×10^{-5} bar^{-1}). In both equilibration stages, all heavy atoms were restrained using a harmonic potential. The resulting equilibrated systems were used as starting points for both brute force and weighted ensemble MD simulations. The initial pair separations were 10, 10, 17, and 15 Å for methane/methane, Na^+/Cl^-, methane/benzene, and K^+/18-crown-6 ether, respectively. The GROMOS 45A3 united-atom force field[117] and SPC/E[118] water model were used for methane/methane and Na^+/Cl^-, while the OPLS-AA/L force field[119] and the TIP3P[120] water model were used for methane/benzene and K^+/18-crown-6 ether. Atom type assignments for K^+/18-crown-6 ether are provided in Figure 2.4.

2.3.3 Brute force dynamics propagation

Brute force simulations for all model systems were started from the endpoints of their respective second-stage (density) equilibration runs. Each simulation was continued until a sufficient number of transition events were observed, with solute positions recorded every 10 fs. The methane/methane and methane/benzene systems were both run as single 1-μs trajectories. Na^+/Cl^- and K^+/18-crown-6 ether required multiple independent trajectories to observe a sufficient number of transition events; ten independent 1-μs trajectories were run for Na^+/Cl^-, and 100 independent 100-ns trajectories were run for K^+/18-crown-6 ether.

2.3.4 Determination of bound and unbound states

The analysis of brute force trajectories and the construction of weighted ensemble simulations require unambiguous definitions of bound and unbound states for each system. Because all four model systems possess one-dimensional progress coordinates, the same protocol for determining these states was applied to all four model systems. Pairwise condensed-phase interactions can be described by the potential of mean force (PMF) $u(r)$, the free energy of the system as a function of pair separation r.[121] Taking the zero of energy to be the non-interacting limit, for constant-volume systems $u(r)$ is given by the following:[90]

$$u(r)/k_BT = -\left(\ln\frac{P(r)}{r^2} - \ln\frac{P(r_0)}{r_0^2}\right) \tag{2.3.1}$$

where $P(r)$ is the probability of observing the system at a pair separation r, r_0 is the shortest distance at which the pair is effectively non-interacting ($du/dr \approx 0$ for all $r > r_0$), and the factors of r^2 arise from the transformation between the Cartesian coordinates of the MD simulation and the spherical polar coordinates in which $u(r)$ is expressed. For each model system, the PMF $u(r)$ was determined using Equation 2.3.1 with pairwise distance probabilities $P(r)$ taken from the brute force trajectories. The unbound state A was defined as $A = \{r : r \geq r_0\}$, where (as above) r_0 is the shortest distance at which the pair is effectively non-interacting. This definition ensures that binding events observed in brute force simulations are very nearly statistically independent. The bound state B was readily identified as being near the global minimum of $u(r)$, and defined as $B = \{r : r < r_B\}$, where r_B is the separation at which the global minimum well of $u(r)$ becomes concave up; that is, B is the basin of attraction of the global minimum of $u(r)$. The remainder of progress coordinate space defines a transition region $T = \{r : r_B \leq r < r_0\}$ wherein the partners are interacting but not definitively bound. PMF curves for each system are provided in Figures 2.6 – 2.9.

2.3.5 Determination of weighted ensemble simulation parameters

In addition to definitions of bound and unbound states, a weighted ensemble simulation requires selection of optimal bin sizes, numbers of replicas per bin, and propagation/resampling interval τ. In making these selections, the extent of sampling should be maximized (generally meaning more bins and more replicas per bin) while minimizing the overall computational cost (generally meaning fewer bins and fewer replicas per bin).

For all four model systems, the potential of mean force was used to determine a bin spacing aimed at maximizing the "ratcheting" effect of the weighted ensemble approach. Where the PMF was changing rapidly with respect to pair separation, bin boundaries were chosen such that the crossing of a bin does not require climbing more than $\sim k_B T$ in energy as indicated by the appropriate PMF. This ensures that the system can move about the progress coordinate with relative ease. Conversely, in the region where the PMF is slowly-varying, a constant spacing of bins was adopted. The propagation period τ was then chosen so that the RMS change in pair separation over a time τ was approximately equal to the width of the bins in the slowly-varying region of the PMF. This resulted in bins of width $\sim 0.1 - 1.0$ Å. Initial tests indicated that 50 replicas per bin yielded sufficiently precise values for the rate constant k at a reasonable computational cost, so this value was used for all four model systems. Detailed listings of the resulting bin boundaries are provided in Figures 2.6 – 2.9, and the remaining weighted ensemble sampling parameters are summarized in Table 2.4.

2.3.6 Weighted ensemble dynamics propagation

Weighted ensemble dynamics runs used exactly the same simulation parameters (force field, thermostat parameters, box volume, *etc.*) as those of the corresponding brute force simulations. As with the brute force simulations, the initial atomic coordinates and velocities were taken from the end of the equilibration phase for each model system. The weighted ensemble sampling algorithm was implemented in an in-house computer code as described above. Replicas were propagated in parallel on 32 – 96 CPU cores, requiring a few days to simulate each model system. Both the rate constant k and the (cumulative) transition event duration distribution $F(t_{\mathrm{ed}})$ were monitored every 50 or 100 τ, and the weighted ensemble simulation was terminated when k was constant within uncertainty and $F(t_{\mathrm{ed}})$ had converged to within 95% confidence and remained at that level, as determined by a two-sided Kolmogorov-Smirnov test[110] (a standard test of the statistical equivalence of two empirical distribution functions). Though resampling was performed with a period of τ, all analysis of the simulations was conducted at a time res-

olution of 10 fs (the period with which solute positions were recorded during the underlying dynamics simulations). The resulting aggregate simulation times for each system are presented in Table 2.5.

2.4 RESULTS AND DISCUSSION

The purpose of this study was to determine the efficiency of weighted ensemble sampling relative to brute force sampling for association events in four molecular recognition systems. As described above, both the association rate constant k and the transition event duration distribution $F(t_{ed})$ can be used to quantify sampling of the transition path ensemble. We compare the efficiency and accuracy of weighted ensemble simulations relative to brute force simulations in terms of both rate constants and transition event distributions.

2.4.1 Rate constants

The rate constant (k) values for brute force and weighted ensemble simulations were separately converged to within statistical uncertainty. As shown in Table 2.1, the weighted ensemble simulations are in qualitative agreement with brute force simulations for all systems; quantitative agreement was achieved for Na^+/Cl^- and methane/benzene. The relative efficiency S_k of weighted ensemble sampling of the rate constant was modest (1.4-fold) for Na^+/Cl^-, greater than five-fold for the diffusive systems (methane/methane and methane/benzene), and 300-fold for the most complex system, K^+/18-crown-6 ether.

It is not surprising that the rate constant obtained by weighted ensemble sampling for K^+/18-crown-6 ether does not agree with the brute force simulation, as the brute force $F(t_{ed})$ did not converge; it is less clear why the rate constants for methane/methane are not in agreement. One possibility is that either the brute force or the weighted ensemble simulation did not sample the full set of waiting times between rare events. The waiting time t_w between subsequent $A \rightarrow B$ transition events relates the first passage time t_{fp} and the transition event duration t_{ed} according to

$$t_{fp} = t_{ed} + t_w$$

In all cases (including that in which t_{ed} and t_w are not statistically independent),

$$\langle t_{fp} \rangle = \langle t_{ed} \rangle + \langle t_w \rangle$$

where the angle brackets denote the expectation (mean) value. Since $\langle t_{\mathrm{ed}} \rangle \ll \langle t_{\mathrm{fp}} \rangle$ for all four systems considered here, the discrepancy between brute force and weighted ensemble simulations in mean waiting time $\langle t_{\mathrm{w}} \rangle$ accounts almost completely for the discrepancy in rate constants between simulation techniques (see Table 2.2). It is likely that the overestimated brute force waiting time for K^+/18-crown-6 ether is due to poor convergence of the brute force simulation. Similarly, it seems likely that the methane/methane brute force simulation underestimated t_{w} for that system. In both of these cases, the efficiencies presented in Table 2.1 represent lower bounds, as they assume complete convergence of the brute force simulations.

Implicit in the foregoing analysis is the assumption that the first passage time distribution is exponential:

$$f(t_{\mathrm{fp}}) = k \exp(-k t_{\mathrm{fp}}) \tag{2.4.1}$$

$$F(t_{\mathrm{fp}}) = 1 - \exp(-k t_{\mathrm{fp}}) \tag{2.4.2}$$

where k is the rate constant, $f(t_{\mathrm{fp}})$ is the probability density of the first passage time distribution, and $F(t_{\mathrm{fp}})$ is its cumulative distribution function. An exponential first passage time distribution would occur in a system possessing (effectively) a single barrier of constant height. In this case, the rate constant k is equal to the inverse mean first passage time [cf. Equation 2.2.3]. If the (cumulative) first passage time distribution $F(t_{\mathrm{fp}})$ is not exponential, then the inverse mean first passage time is at best an approximation to the true rate constant; conversely, the weighted ensemble approach samples k directly, and so it can be expected to recover the correct rate constant (within the bounds of statistical uncertainty) regardless of whether the underlying physical mechanisms lead to an exponential first passage time distribution. For three of the four model systems (Na^+/Cl^-, methane/benzene, and K^+/18-crown-6), the first passage time distributions obtained from brute force simulations conform to Equation 2.4.2 to within 95% confidence (see Figure 2.5). For methane/methane, however, the first passage time distribution deviates from the expected exponential distribution for $t_{\mathrm{fp}} \lesssim 300$ ps. This offers an alternative explanation for why the rate constant values obtained for methane/methane differ between brute force and weighted ensemble simulations: because the first passage time distribution $F(t_{\mathrm{fp}})$ is not exponential, the rate constant k obtained from the brute force first passage time distribution as $\langle t_{\mathrm{fp}} \rangle^{-1}$ may in fact be inaccurate.

Table 2.1: Brute force (BF) and weighted ensemble (WE) aggregate simulation times t, rate constants (k), and relative sampling efficiencies (S_k) for the four model systems.

System	t_{BF}	t_{WE}	$k_{\mathrm{BF}}\,(\mathrm{ps}^{-1})$	$k_{\mathrm{WE}}\,(\mathrm{ps}^{-1})$	S_k
Methane/methane	1 μs	299 ns	$1.91 \pm 0.10 \times 10^{-3}$	$1.61 \pm 0.06 \times 10^{-3}$	7.0
Na^+/Cl^-	10 μs	3.86 μs	$1.86 \pm 0.09 \times 10^{-4}$	$1.82 \pm 0.11 \times 10^{-4}$	1.4
Methane/benzene	1 μs	369 ns	$8.6 \pm 0.7 \times 10^{-4}$	$7.7 \pm 0.3 \times 10^{-4}$	8.7
K^+/18-Crown-6	10 μs	322 ns	$2.1 \pm 0.3 \times 10^{-5}$	$4.8 \pm 0.2 \times 10^{-5}$	300

Aggregate simulation times correspond to the combined length of all trajectories (either brute force or weighted ensemble) for each system, without overcounting common history in the case of weighted ensemble simulations. Uncertainties on the rate constants represent 95% confidence intervals. Relative efficiencies were calculated using Equation 2.2.7.

Table 2.2: Ratios of rate constants k and average waiting times $\langle t_{\mathrm{w}} \rangle$ for brute force (BF) and weighted ensemble (WE) simulations.

System	$k_{(\mathrm{WE})} / k_{(\mathrm{BF})}$	$\langle t_{\mathrm{w}} \rangle_{(\mathrm{BF})} / \langle t_{\mathrm{w}} \rangle_{(\mathrm{WE})}$
Methane/methane	0.842	0.841
Na^+/Cl^-	0.977	0.977
Methane/benzene	0.827	0.822
K^+/18-Crown-6	1.93	1.94

2.4.2 Transition event duration distributions

In general, the weighted ensemble simulations were as good or better than brute force simulations in generating well-resolved (cumulative) transition event duration distributions $F(t_{ed})$. As shown in Figure 2.3, $F(t_{ed})$ was well-resolved by both brute force and weighted ensemble simulations for all systems except K^+/18-crown-6 ether, for which brute force sampling was not capable of providing a converged $F(t_{ed})$ distribution. The resolution of distributions from weighted ensemble simulations far exceeds that of distributions obtained from brute force simulations, as demonstrated in the increased number N_e of unique transition durations sampled (see Table 2.3). Further, pathways generated by weighted ensemble sampling and having different transition event durations were indeed noticeably different from each other (see Supporting Information). These are strong indications that the weighted ensemble algorithm effectively enhances sampling of the transition path ensemble. The relative efficiency S_{ed} of sampling $F(t_{ed})$ increased with the complexity of the molecular recognition system, ranging from one to three orders of magnitude. The 1100-fold relative efficiency of weighted ensemble sampling for K^+/18-crown-6 ether is a conservative estimate, as the referenced brute force simulation had not even reached convergence with respect to $F(t_{ed})$.

As shown in Tables 2.1 and 2.3, $S_k < S_{ed}$ in all four cases. This is partly a consequence of our definitions of the efficiency metrics S_k and S_{ed} (see above and Sections 2.7.2, 2.7.3, and 2.7.4), but also reflects that the rate constant k is generally more difficult to sample than the set of transition event durations $\{t_{ed}\}$. In particular, convergence of the rate constant k requires sampling of *all* important pathways as well as a steady state flow of probability through them.

2.4.3 How much sampling is required?

As evident for K^+/18-crown-6 ether, the most complex system of this study, it is not always possible to obtain converged brute force simulations of molecular association events. In such cases, how does one know if the weighted ensemble approach has achieved sufficient sampling? One can, at least, gauge the self-convergence of the association rate constants k and the transition event duration distributions $F(t_{ed})$ obtained from the weighted ensemble simulations. However, self-convergence of these metrics does not guarantee that the simulation has converged to the true value of k or $F(t_{ed})$.

As an illustration, consider the convergence of $F(t_{ed})$, the probability distribution of the event duration times t_{ed}. Even if two transition event distributions $F_{\tau(1)}(t_{ed})$ and $F_{\tau(2)}(t_{ed})$ obtained by time points $N_{\tau(1)}$ and $N_{\tau(2)} > N_{\tau(1)}$ in a weighted ensemble simulation are statistically equivalent, this does not nec-

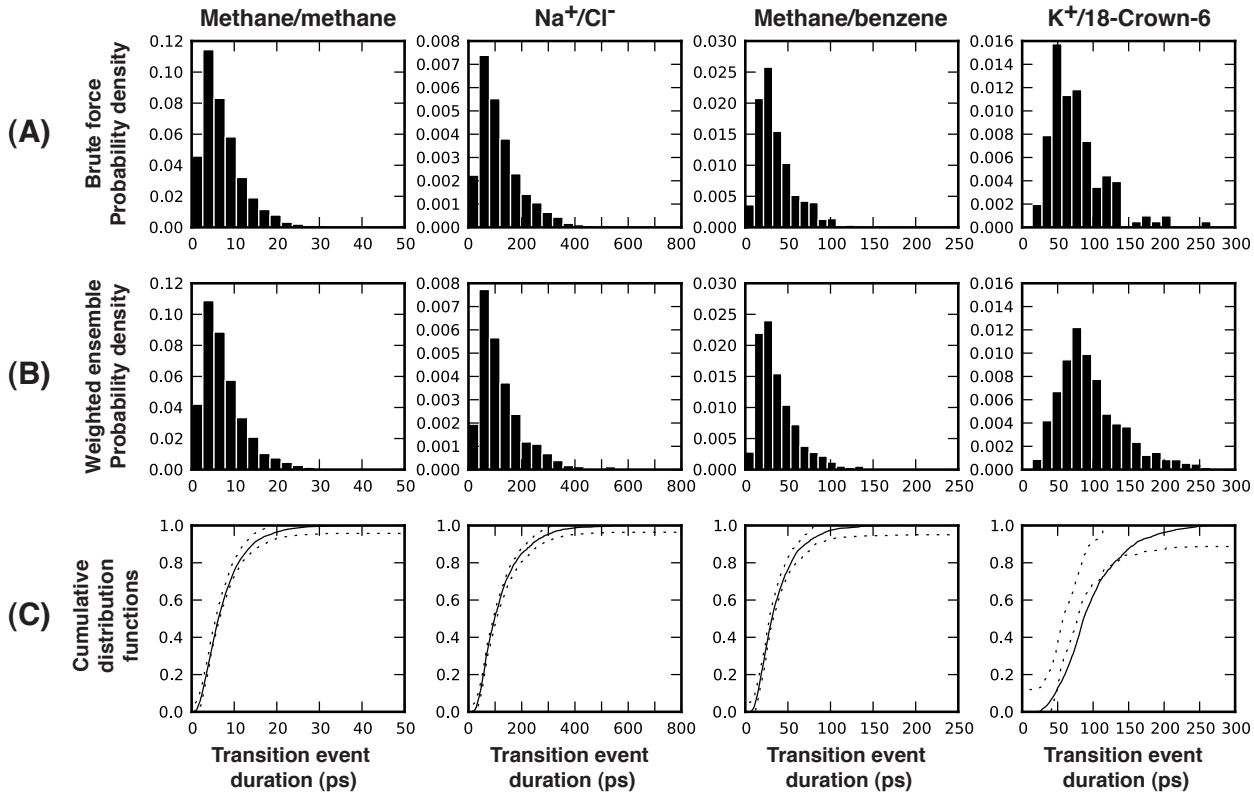


Figure 2.3: Transition event duration distributions obtained from (A) brute force and (B) weighted ensemble simulations. The cumulative distribution function (CDF) of the transition event duration probability for each model system is shown in (C); the brute-force CDF is plotted as a 95% confidence interval with dotted lines, and the solid line is the CDF obtained from the weighted ensemble simulation.

essarily indicate asymptotic convergence on the true transition event duration distribution. Because a weighted ensemble simulation of length N_τ iterations contains only trajectories of maximum length τN_τ, then a the statistical equivalence of $F_{\tau(1)}(t_{\mathrm{ed}})$ and $F_{\tau(2)}(t_{\mathrm{ed}})$ does not indicate that the entire event duration distribution has been adequately sampled, merely that all pathways taking time $t \leq \tau N_{\tau(1)}$ to traverse have been adequately sampled. Thus, for a weighted ensemble simulation of length τN_τ, one must ultimately decide whether data obtained for time scales less than τN_τ are sufficient to provide insights into the systems under study.

2.4.4 How does one choose optimal weighted ensemble parameters?

Efficient use of weighted ensemble sampling involves finding the optimal balance between computational expense and level of sampling. A poor choice of progress coordinate bins can easily lead to oversampling relatively unimportant regions of phase space. A large number of replicas not only aids rapid exploration of phase space, but also determines the precision of probability current value and thus kinetic information; however, the total computational cost of weighted ensemble scales approximately linearly with the maximum number of system replicas. A short propagation/resampling period τ allows many opportunities for replicas to split and explore newly-visited regions of phase space and for replicas to merge to avoid oversampling regions of phase space, but ultimately may not allow sufficient divergence of trajectories to allow for efficient exploration of phase space.

Integral to the construction of a weighted ensemble simulation is the choice of a progress coordinate that is sufficiently sensitive to quantify "how far along" the reaction is. Any number of relatively low-cost enhanced-sampling or energy landscape smoothing techniques[2,14,122] might be employed to guide the choice of a progress coordinate, including metadynamics;[123,124] targeted,[125] steered,[126] or accelerated[127] molecular dynamics; or the recently-developed orthogonal space random walk method.[128] A number of short brute force simulations may be required to determine the average time evolution of the progress coordinate, which in turn determines the most efficient choices of bin spacing and the propagation/reweighting period τ. Finally, it may be necessary to adjust these parameters "on the fly" during a simulation, especially for large systems with complex, rough energy landscapes (*i.e.* proteins) where long-lived intermediate states may be encountered in the course of a simulation. A discussion of how to select these parameters in practice is included in Section A.1.

The complexities and advantages of actively adjusting the numbers of bins, their boundaries, and the number of replicas in each bin has been discussed in detail;[75] such schemes could be used to detect replicas that "stall" in certain progress coordinate bins and adjust the weighted ensemble simulation to compensate. These schemes would not be able to cope effectively with systems possessing intermediate states with lifetimes comparable to the mean first passage time; such systems do not exhibit the separation of timescales which weighted ensemble sampling is designed to exploit. However, using ideas developed from nonequilibrium umbrella sampling, it is possible to reweight phase space density analytically in order to accelerate the attainment of steady-state probability recycling;[68] this would in turn accelerate the determination of the rate constant in systems with $t_{\mathrm{ed}} \approx t_{\mathrm{fp}}$ at the possible expense of efficient sampling of the transition path ensemble.

Finally, it should be noted that the weighted ensemble approach is but one instance of a class of "interface-based" enhanced sampling techniques which share a number of strengths and potential weaknesses;[58,59,83] other such techniques include transition interface sampling (TIS) and variants,[64–66] forward flux sampling (FFS),[69,70] and Milestoning.[50] (See Chapter 1 for a brief overview of these methods.) All of the methods in this class are rare event sampling methods that divide phase space along distinct interfaces, and each method is capable of providing realistic kinetic rates. Provided a well-chosen progress coordinate, these methods are equivalent in principle with respect to the information which can be obtained from them and the efficiency with which that information is obtained, at least for equilibrium systems. Among these methods, however, the weighted ensemble approach is uniquely flexible; in particular, sampling can be maximized while minimizing computational cost both by dividing phase space according to arbitrary boundaries in any number of dimensions, and by adjusting the level of sampling within each region (by adjusting the number of simulation replicas within a bin). The cost of this flexibility, however, is the complexity of determining efficient choices for parameters such as the progress coordinate, bin boundaries, and the number of replicas per bin. In situations where a reasonable progress coordinate cannot be determined, a method not dependent on a progress coordinate (such as transition path sampling[55,56,60] or a recently-developed variation of Milestoning[129]) may be necessary. Similarly, if efficient choices for simulation parameters (such as bin boundaries and the number of replicas per bin) cannot be made in advance and adjustment of these parameters during a simulation is impractical, then a method like FFS (for which analytical expressions for efficiency as a function of simulation parameters exist[72,73]) may be a better choice.

2.4.5 Why are efficiencies what they are?

The efficiency of a weighted ensemble simulation is largely determined by weighted ensemble simulation parameters, particularly the propagation/resampling period τ, the choice of progress coordinate(s), and the locations of bin boundaries.[108] For some systems, brute force simulation is already highly efficient at sampling the molecular association events; this is confirmed by the modestly increased weighted ensemble sampling efficiencies (S_k and S_{ed}) for methane/methane, Na^+/Cl^-, and methane/benzene. However, the fact that the weighted ensemble approach increases rather than decreases efficiency indicates that even in such cases, the weighted ensemble technique is capable of accelerating sampling of both k and $F(t_{\mathrm{ed}})$. On the other hand, the very high relative efficiency of sampling in K^+/18-crown-6 ether is particularly encouraging. Despite the small size of the system, brute force MD was incapable of effective sampling of rate constants and transition event duration distributions for K^+/18-crown-6 ether, almost certainly due to the high (approximately 14 k_BT, 8.3 kcal/mol) barrier to dissociation. Weighted ensemble sampling was able to obtain self-converged values of both the rate constant k and the transition event duration distribution $F(t_{\mathrm{ed}})$. This is primarily because probability recycling completely circumvents the necessity to climb the 14 k_BT dissociation barrier in order to observe another binding event.

These results point encouragingly to the ability to simulate protein-protein binding events with weighted ensemble molecular dynamics. With well-chosen bin boundaries, the weighted ensemble technique should increase sampling efficiency exponentially with increasing barrier heights. This is because placing bin boundaries sufficiently close to each other effectively linearizes the probability of crossing a number of bins in succession, rather than surmounting a barrier in one step with a probability which decreases exponentially with barrier height.[51] As a concrete example, the barrier to association in a diffusion-limited protein-protein system is approximately 10 k_BT (roughly five times that of the model systems). If this exponential efficiency scaling holds, then one can expect about 20,000-fold improvement in sampling for such a system. In other words, if a given computational resource is otherwise capable of generating 500 ps per calendar day (a substantial but accessible level of computational power), this efficiency gain corresponds to reaching a timescale of about 1 ms in 100 days, compared to the 50 ns that would otherwise be possible in the same amount of time. However, since protein-protein binding pathways involve significant metastable intermediate states (*e.g.* encounter complexes[130]), it is possible for a simulation to "stall" in such a state. As discussed above, several techniques exist which may

partially ameliorate this difficulty, but in the end, a number of simulations connecting the intermediate states may be necessary to fully explore binding events in such systems.

2.5 CONCLUSIONS

We have applied the weighted ensemble path sampling approach to molecular dynamics simulations in explicit solvent, enabling the detailed sampling of rare molecular association events. We have compared the efficiency of weighted ensemble sampling relative to brute force sampling in simulating association events of methane/methane, Na^+/Cl^-, methane/benzene, and K^+/18-crown-6 ether. Relative to brute force simulation, weighted ensemble sampling of these four systems confirms that the weighted ensemble approach reproduces or even improves sampling of both the rate constant k and the distribution of transition event durations. This improvement is on the order of 300 and 1100-fold, respectively, for a system exhibiting significant conformational flexibility (K^+ binding with 18-crown-6 ether). We expect efficiency gains to grow with increasing barriers to association. However, the existence of significant metastable intermediate states may hinder sampling in such systems, requiring the use of various enhancements to the weighted ensemble method in order to explore binding events in such systems. Nonetheless, these results indicate that weighted ensemble sampling in conjunction with MD simulations is likely to allow for the effective determination of transition paths and rate constants for protein binding events.

2.6 ACKNOWLEDGEMENTS

We thank Dan Zuckerman and Divesh Bhatt (U. Pitt. Dept. of Computational Biology), Bin Zhang (U. Michigan Dept. of Chemistry), Michael Grabe and Josh Adelman (U. Pitt. Dept. of Biological Sciences), Gary Huber (UCSD Dept. of Bioengineering), and Karen Zwier and Jonathan Livengood (U. Pitt. Dept. of History and Philosophy of Science) for helpful discussion; we also thank Xianghong Qi for initial efforts. This work was supported by NSF CAREER award MCB-0845216 to LTC, a University of Pittsburgh Arts & Sciences Fellowship to MCZ, and a University of Pittsburgh Brackenridge Fellowship (underwritten by the United States Steel Foundation) to JWK.

2.7 SUPPORTING INFORMATION

2.7.1 Description of K^+/18-Crown-6 Ether Binding Pathways

To further explore the distinct pathways of binding that result in a broad distribution of event duration times t_{ed}, the pathways of binding for the K^+/18-crown-6 ether system were examined. Five trajectories were taken from the weighted ensemble simulation, corresponding to the minimum (Movie S1), first quartile (Movie S2), median, third quartile, and maximum event duration times. In the shortest trajectory, the K^+ ion binds quickly to the oxygens of the ether (Figure 2.10A) and spends a very short time bound to one or two oxygens before it fully binds to all six oxygens (Figure 2.11A) . In the other four trajectories, the K^+ ion spends variable amounts of time at certain distances away from the crown ether (Figure 2.10B), suggesting that the K^+ ion may be penetrating solvation shells prior to binding to the ether oxygens. Upon approaching the ether, the K^+ ion binds to one or two of the ether oxygens (Figure 2.11B), occasionally moving to another oxygen (or pair of oxygens) before binding fully. Thus, the amount of time needed to penetrate each solvation shell and the amount of time spent bound to only one or two oxygens appears to be the primary differences between trajectories having different event duration times t_{ed}.

2.7.2 Derivation of the Relative Efficiency Metric S_k

Our derivation loosely follows that of Huber and Kim in their original discussion of the efficiency of weighted ensemble sampling.[75] Assuming that binding events in brute force simulations are independent, the width ΔA of a confidence interval on any time-averaged quantity of interest A depends on the number N_{obs} of transitions observed as

$$\Delta A^* \propto N_{obs}^{-1/2}$$

where

$$\Delta A^* = \left| \frac{\Delta A}{\langle A \rangle} \right|$$

is the width ΔA of the confidence interval relative to the mean value $\langle A \rangle$. Since the rate constant k is simply the number of transitions observed per unit simulation time, $N_{obs} \approx kt$, where t is the total amount of time simulated; thus,

$$\Delta A^* \propto t^{-1/2} \tag{2.7.1}$$

Table 2.3: Number of unique transition durations N_e and relative efficiency S_{ed} of sampling of the transition event duration distribution for brute force (BF) and weighted ensemble (WE) simulations. Relative efficiency was calculated using Equation 2.2.8.

System	$N_{e(\text{BF})}$	$N_{e(\text{WE})}$	S_{ed}
Methane/methane	1021	2304	7.5
Na^+/Cl^-	1415	8780	16
Methane/benzene	750	5485	20
K^+/18-Crown-6	145	5007	1100

Table 2.4: Weighted ensemble simulation parameters. r_B is the maximum separation of the bound state, r_0 is the minimum separation of the unbound state, L is the width of uniformly-spaced bins in slowly-varying regions of the PMF (see "Methods"), N_b is the total number of bins used in the simulation, and τ is the dynamics propagation period.

System	r_B (Å)	r_0 (Å)	L (Å)	N_b	τ (ps)
Methane/methane	4.00	10.00	0.2	13	0.5
Na^+/Cl^-	2.80	14.98	1.0	22	5.0
Methane/benzene	5.65	17.00	0.8	15	1.0
K^+/18-Crown-6	0.20	11.60	1.0	14	0.5

This allows us to construct an efficiency metric by asking how long a brute force simulation is required to obtain the same confidence interval as is obtained from a weighted ensemble simulation. Constructing an equality of proportions using Equation 2.7.1,

$$\frac{t_{\text{eff}}^{-1/2}}{t_{(\text{BF})}^{-1/2}} = \frac{\Delta A^*_{(\text{WE})}}{\Delta A^*_{(\text{BF})}}$$

where t_{eff} is the effective brute force time required to obtain a confidence interval of relative width $\Delta A^*_{(\text{WE})}$ (as obtained from a weighted ensemble simulation, as by block averaging or bootstrapping) and $\Delta A^*_{(\text{BF})}$ is the relative width of the confidence interval as obtained from a brute force simulation of length $t_{(\text{BF})}$. This leads immediately to

$$t_{\text{eff}} = t_{(\text{BF})} \left(\frac{\Delta A^*_{(\text{BF})}}{\Delta A^*_{(\text{WE})}} \right)^2$$

We define efficiency as the reciprocal of simulation time ("faster is better"), and thus the efficiency S of weighted ensemble sampling relative to brute force is the ratio of weighted ensemble and effective brute force reciprocal simulation times:

$$S = \frac{t_{(\text{WE})}^{-1}}{t_{\text{eff}}^{-1}} = \frac{t_{\text{eff}}}{t_{(\text{WE})}} \tag{2.7.2}$$

Thus,

$$\begin{aligned} S &= \frac{t_{\text{eff}}}{t_{(\text{WE})}} \\ &= \frac{t_{(\text{BF})}}{t_{(\text{WE})}} \left(\frac{\Delta A^*_{(\text{BF})}}{\Delta A^*_{(\text{WE})}} \right)^2 \end{aligned} \tag{2.7.3}$$

where $t_{(\text{WE})}$ is the total dynamics time of the weighted ensemble simulation without overcounting shared history.

Substituting the reaction rate k and the relative width of its confidence interval Δk^* for A and ΔA^* for both brute force (BF) and weighted ensemble (WE) results in the following expression for the relative efficiency S_k of sampling the rate constant:

$$S_k = \frac{t_{(\text{BF})}}{t_{(\text{WE})}} \left(\frac{\Delta k^*_{(\text{BF})}}{\Delta k^*_{(\text{WE})}} \right)^2 \tag{2.2.7}$$

Note that efficiency S_k increases with decreasing weighted ensemble simulation time $t_{(\text{WE})}$ or confidence interval width $\Delta k^*_{(\text{WE})}$.

2.7.3 Derivation of the Relative Efficiency Metric S_{ed}

Again, we assume that transitions between the initial and destination states are independent events occurring with an average rate k (the rate constant), and thus in a brute force simulation of length t we

expect to see $N = kt$ transition events. Here, k is the (unknown) *true* rate constant, not the rate constant determined by either brute force or weighted ensemble sampling.

We then ask, given a brute force simulation with transition event duration (t_{ed}) effective sample size $N_{e(\text{BF})}$ and a weighted ensemble simulation with t_{ed} effective sample size $N_{e(\text{WE})}$, we ask how long a brute force simulation would be required to produce an effective sample size of $N_{e(\text{WE})}$. Proceeding as above and constructing an equality of proportions from $N = kt$:

$$\frac{N_{e(\text{BF})}}{N_{e(\text{WE})}} = \frac{k\,t_{(\text{BF})}}{k\,t_{\text{eff}}}$$

where t_{eff} is the effective brute force simulation time required to produce $N_{e(\text{WE})}$ transition events. Solving for t_{eff} gives

$$t_{\text{eff}} = t_{(\text{BF})}\frac{N_{e(\text{WE})}}{N_{e(\text{BF})}}$$

and inserting in Equation 2.7.2 gives

$$\begin{aligned} S_{\text{ed}} &= \frac{t_{\text{eff}}}{t_{(\text{WE})}} \\ &= \frac{t_{(\text{BF})}}{t_{(\text{WE})}}\left(\frac{N_{e(\text{WE})}}{N_{e(\text{BF})}}\right) \end{aligned} \tag{2.2.8}$$

Note that S_{ed} increases with decreasing weighted ensemble simulation time $t_{(\text{WE})}$ and increasing weighted ensemble effective sample size $N_{e(\text{WE})}$.

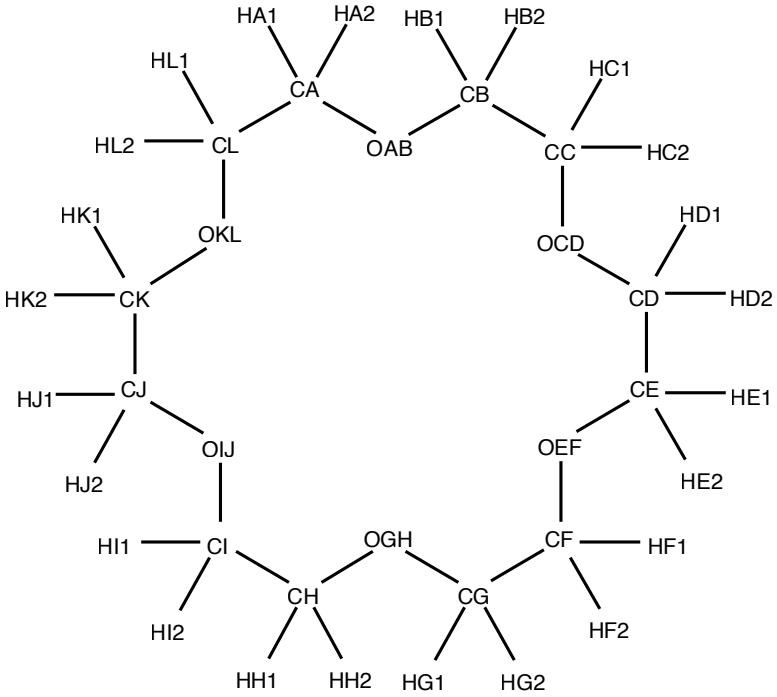
As it is possible to assign a confidence band to an empirical distribution function,[109] it is possible to follow the derivation of S_k in a step-by-step manner, replacing the width of the confidence interval by the width of the confidence band about the empirical distribution function $F(t_{\text{ed}})$. The resulting expression is identical to that of Equation 2.2.8.

2.7.4 Why is $S_k < S_{\text{ed}}$?

The definitions of Equations 2.2.7 and 2.2.8 result in $S_k < S_{\text{ed}}$, in practice. For equal numbers of transition events, the ratio $\Delta k^*_{(\text{BF})}/\Delta k^*_{(\text{WE})} < 1$ in Equation 2.2.7, since the time correlation in k increases the statistical uncertainty in k. Conversely, for equal amounts of aggregate dynamics time, the ratio $N_{e(\text{WE})}/N_{e(\text{BF})} > 1$ in Equation 2.2.8, since the weighted ensemble algorithm effectively eliminates the waiting time between events that brute force simulation is required to sample. The power of 2 in Equation 2.2.7 further reduces S_k relative to S_{ed}.

Table 2.5: Total simulation time required for convergence of weighted ensemble sampling. τ is the dynamics propagation period, N_τ is the number of weighted ensemble iterations necessary to achieve and sustain a 95% confidence level on the transition event duration distribution $F(t_{\text{ed}})$, and τN_τ is the maximum continuous trajectory length supported by the WE simulation. The aggregate time corresponds to the combined length of all trajectories without overcounting common history.

System	τ (ps)	N_τ	τN_τ (ps)	Aggregate time
Methane/methane	0.5	1000	500	299 ns
Na^+/Cl^-	5.0	800	4000	3.86 μs
Methane/benzene	1.0	500	500	369 ns
K^+/18-Crown-6	0.5	1000	500	322 ns



Atom	Type	Atom	Type	Atom	Type	Atom	Type
CA	182	CD	182	HG1	185	HJ1	185
HA1	185	HD1	185	HG2	185	HJ2	185
HA2	185	HD2	185	OGH	180	CK	182
OAB	180	CE	182	CH	182	HK1	185
CB	182	HE1	185	HH1	185	HK2	185
HB1	185	HE2	185	HH2	185	OKL	180
HB2	185	OEF	180	CI	182	CL	182
CC	182	CF	182	HI1	185	HL1	185
HC1	185	HF1	185	HI2	185	HL2	185
HC2	185	HF2	185	OIJ	180	K^+	408
OCD	180	CG	182	CJ	182		

Figure 2.4: OPLS/AA atom type assignments for 18-crown-6 ether.

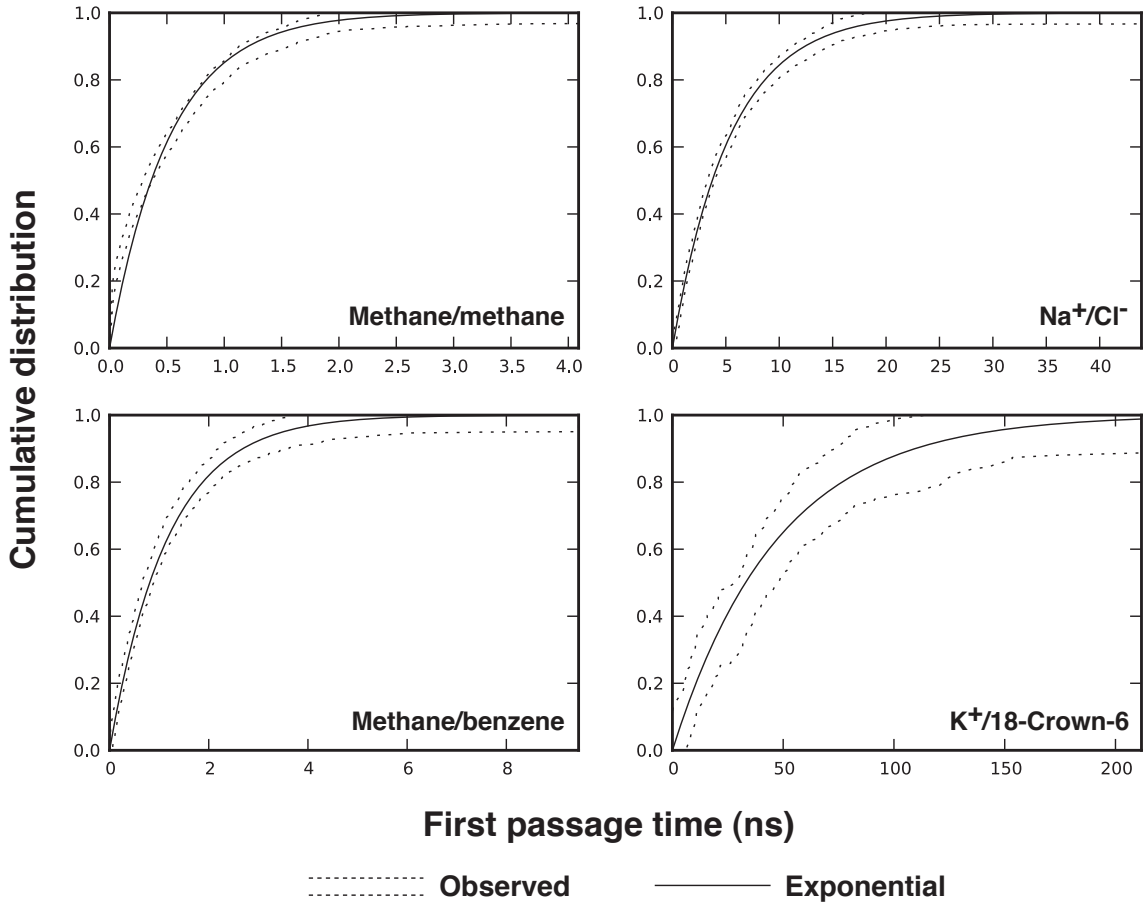
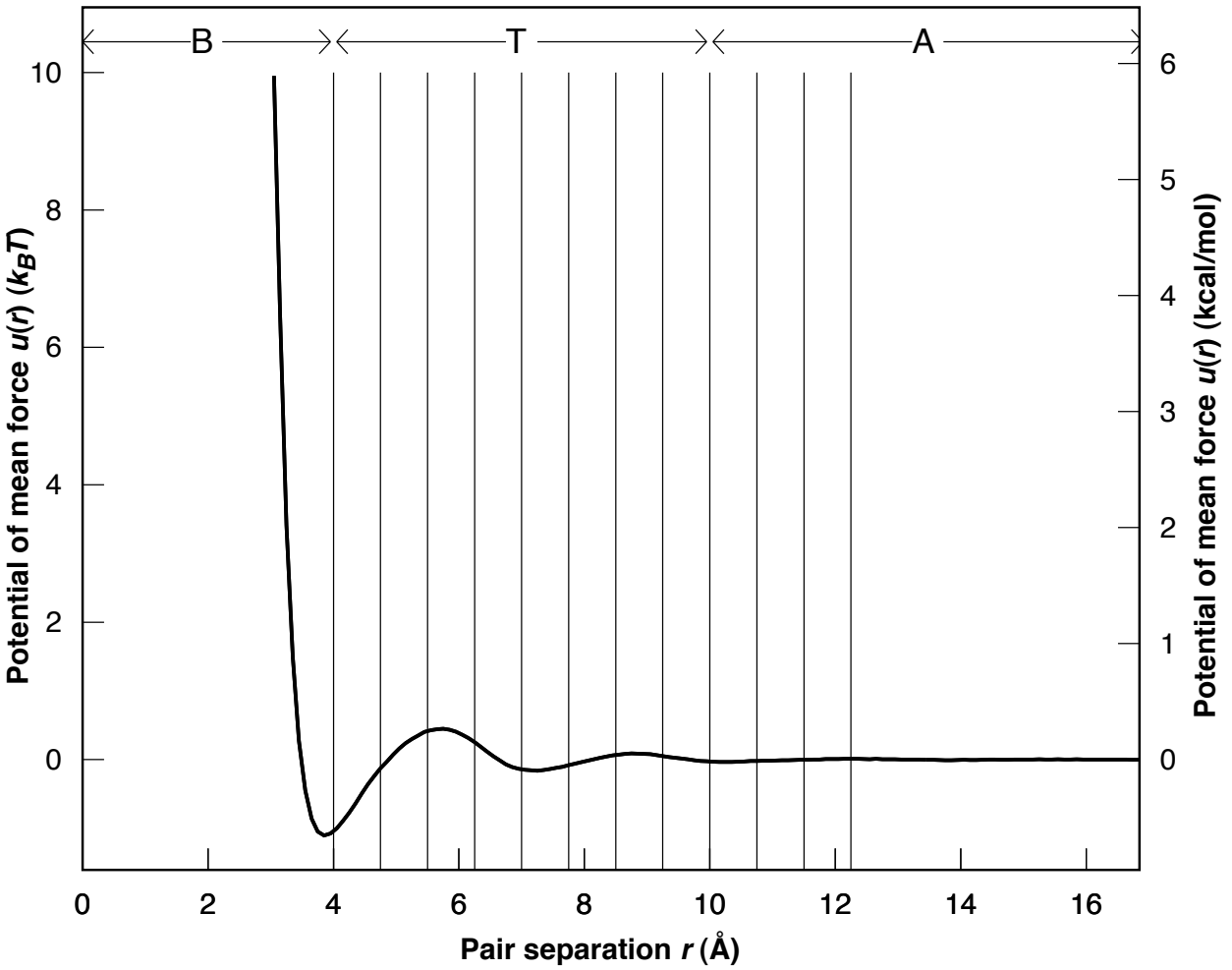


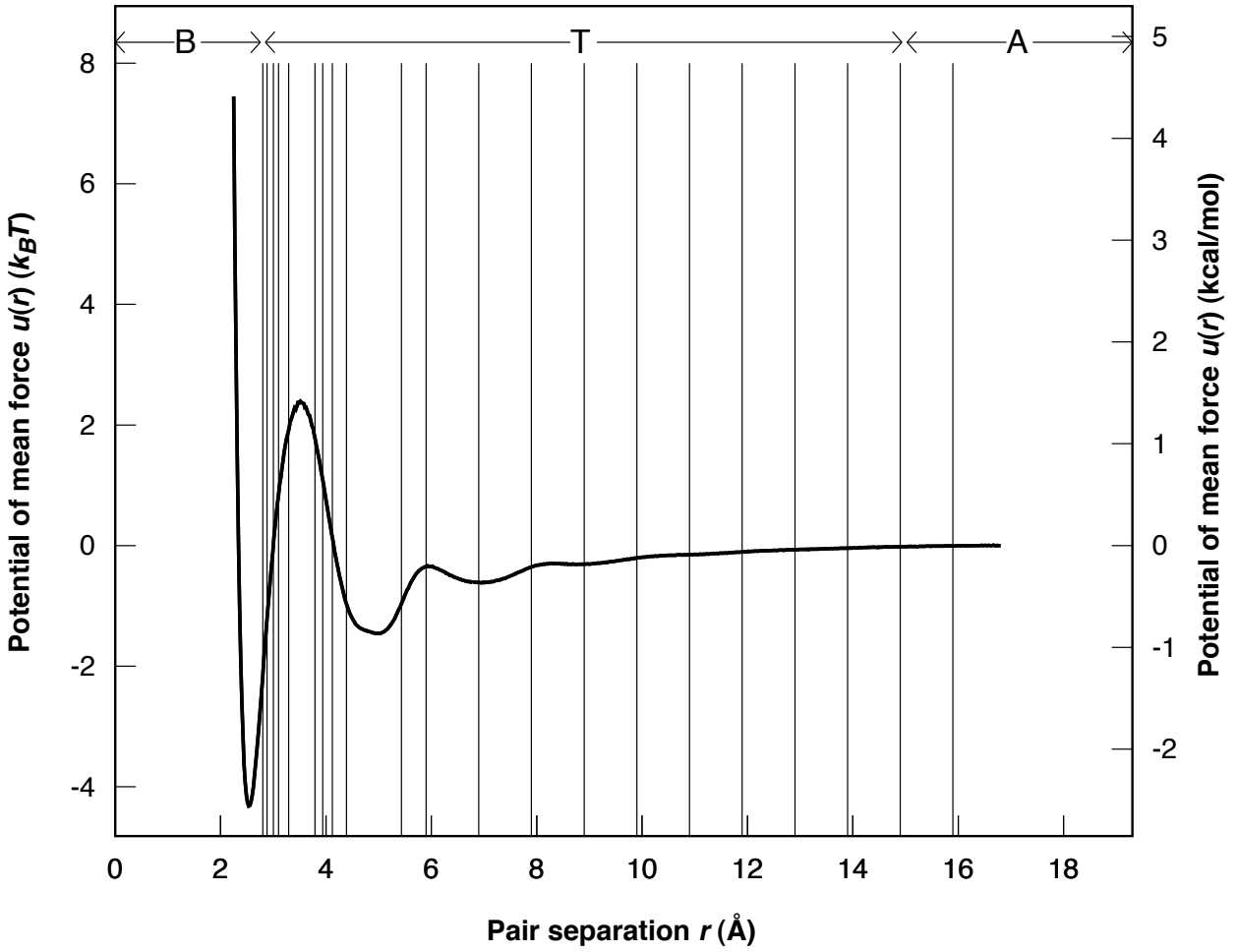
Figure 2.5: First passage time distributions $F(t_{\mathrm{fp}})$ from brute force simulations. The observed distributions are shown as a 95% confidence interval bounded by dotted lines. Exponential distributions of first passage times $\left[F(t_{\mathrm{fp}}) = 1 - \exp(-k\,t_{\mathrm{fp}})\right]$ with rate constant $k = \langle t_{\mathrm{fp}} \rangle^{-1}$ are shown with solid lines.



Bin boundaries r and potential of mean force $u(r)$ at each boundary:

r/Å	$u(r)/k_BT$	r/Å	$u(r)/k_BT$
4.00	-1.04	8.50	0.07
4.75	-0.13	9.25	0.05
5.50	0.43	10.00	-0.03
6.25	0.25	10.75	-0.02
7.00	-0.14	11.50	0.00
7.75	-0.08	12.25	0.00

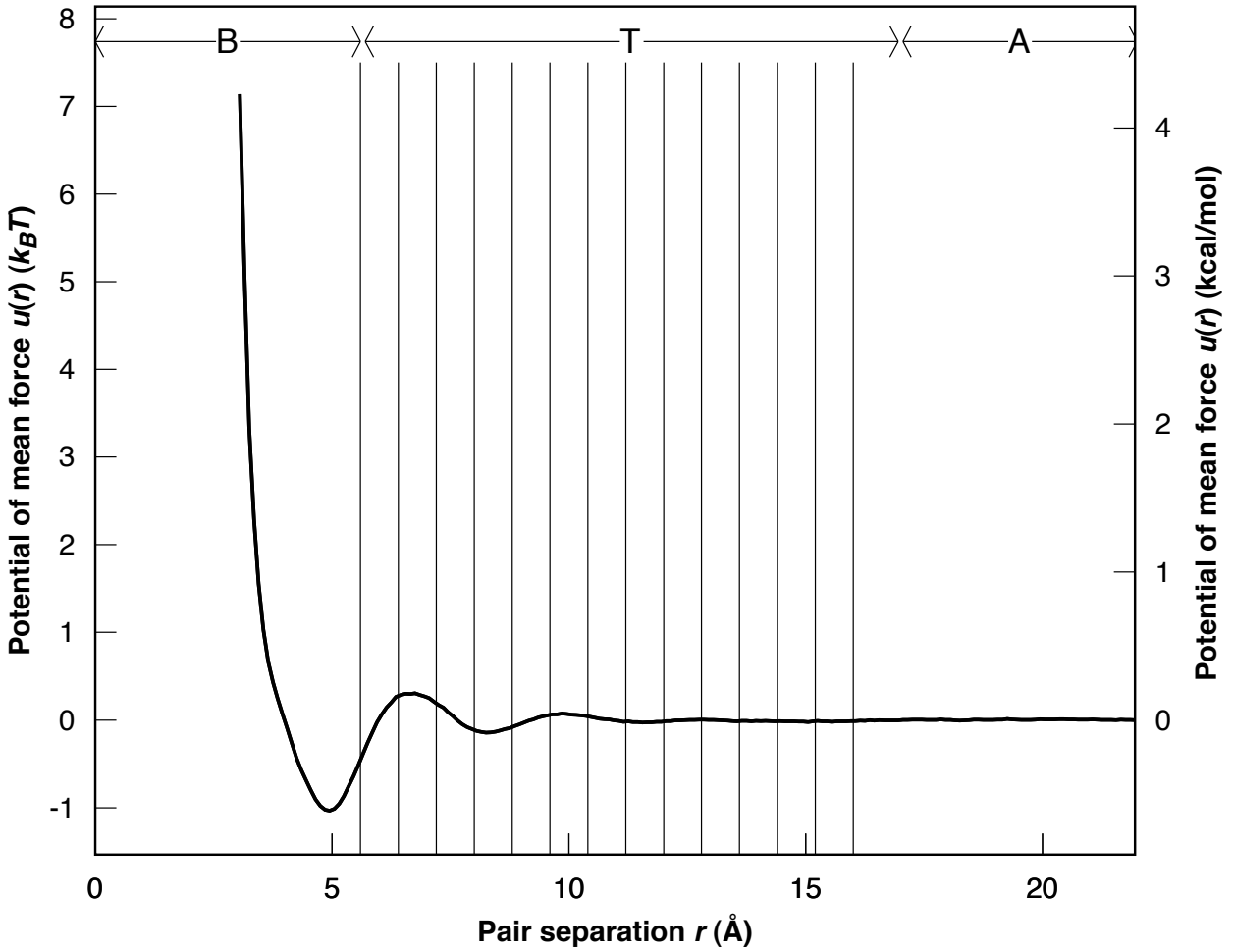
Figure 2.6: Potential of mean force $u(r)$ for methane/methane associations, given in units of k_BT (left axis) and kcal/mol (right axis). The unbound state A, transition region T, and bound state B are marked with horizontal arrows. Bin boundaries are marked with vertical lines and tabulated.



Bin boundaries r and potential of mean force $u(r)$ at each boundary:

r/Å	$u(r)/k_BT$	r/Å	$u(r)/k_BT$
2.80	-2.21	6.90	-0.61
2.88	-1.21	7.90	-0.36
3.00	-0.09	8.90	-0.31
3.10	0.85	9.90	-0.20
3.29	1.93	10.90	-0.15
3.79	1.76	11.90	-0.10
3.94	1.03	12.90	-0.07
4.12	0.12	13.90	-0.04
4.39	-0.96	14.90	-0.02
5.43	-0.97	15.90	-0.01
5.90	-0.35		

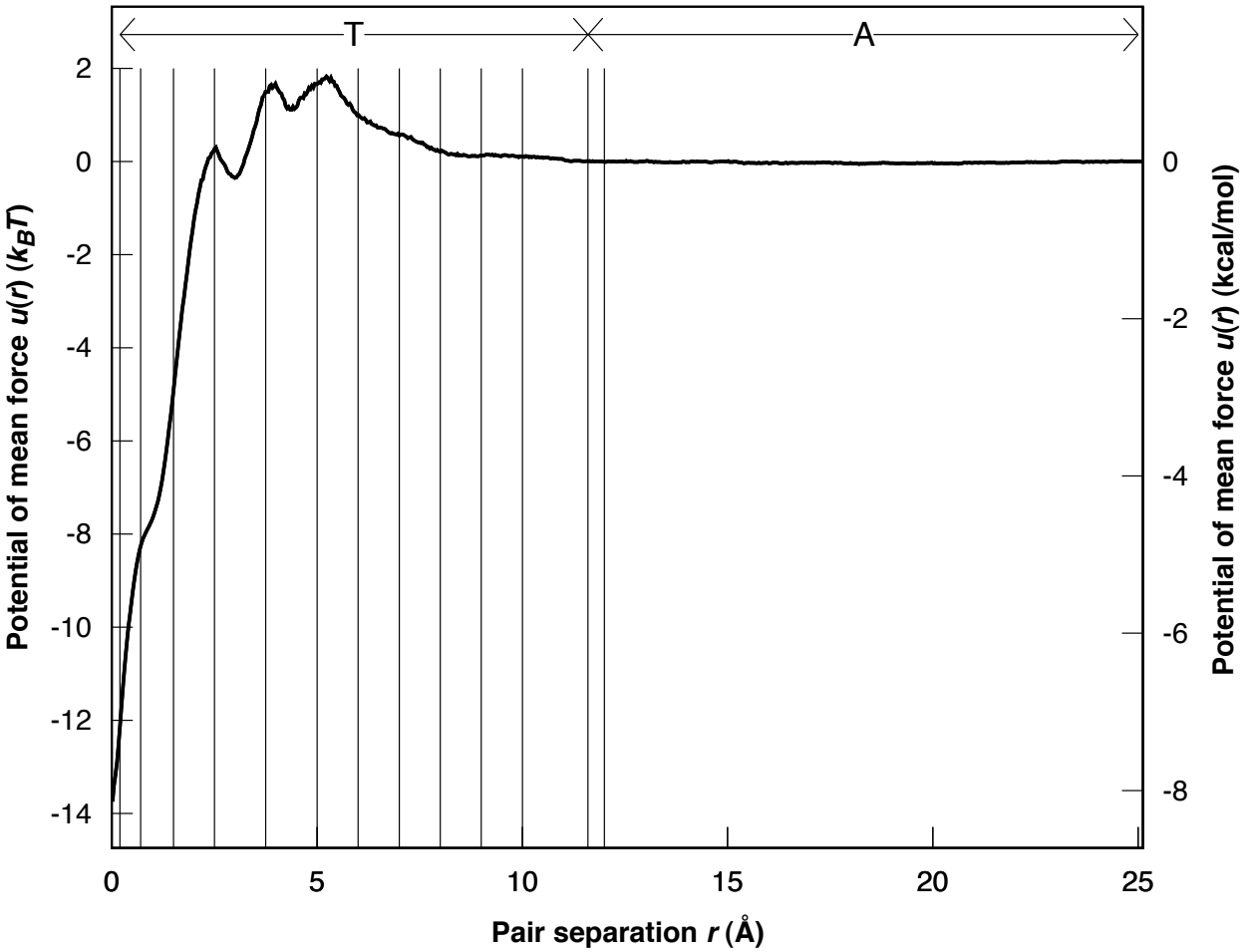
Figure 2.7: Potential of mean force $u(r)$ for Na^+/Cl^- associations, given in units of k_BT (left axis) and kcal/mol (right axis). The unbound state A, transition region T, and bound state B are marked with horizontal arrows. Bin boundaries are marked with vertical lines and tabulated.



Bin boundaries r and potential of mean force $u(r)$ at each boundary:

r/Å	$u(r)/k_BT$	r/Å	$u(r)/k_BT$
5.6	-0.45	11.2	-0.02
6.4	0.27	12.0	-0.02
7.2	0.19	12.8	0.00
8.0	-0.11	13.6	-0.01
8.8	-0.08	14.4	-0.01
9.6	0.06	15.v2	-0.01
10.4	0.04	16.0	-0.01

Figure 2.8: Potential of mean force $u(r)$ for methane/benzene associations, given in units of k_BT (left axis) and kcal/mol (right axis). The unbound state A, transition region T, and bound state B are marked with horizontal arrows. Bin boundaries are marked with vertical lines and tabulated.



Bin boundaries r and potential of mean force $u(r)$ at each boundary:

r/Å	$u(r)/k_BT$	r/Å	$u(r)/k_BT$
0.2	-12.15	7.0	0.56
0.7	-8.30	8.0	0.20
1.5	-4.87	9.0	0.13
2.5	0.27	10.0	0.10
3.8	1.49	11.6	0.01
5.0	1.67	12.0	-0.01
6.0	1.00		

Figure 2.9: Potential of mean force $u(r)$ for K^+/18-crown-6 ether associations, given in units of k_BT (left axis) and kcal/mol (right axis). The unbound state A and transition region T are marked with horizontal arrows. The bound state $B = \{r : r < 0.20\ \text{Å}\}$ is not labeled. Bin boundaries are marked with vertical lines and tabulated.

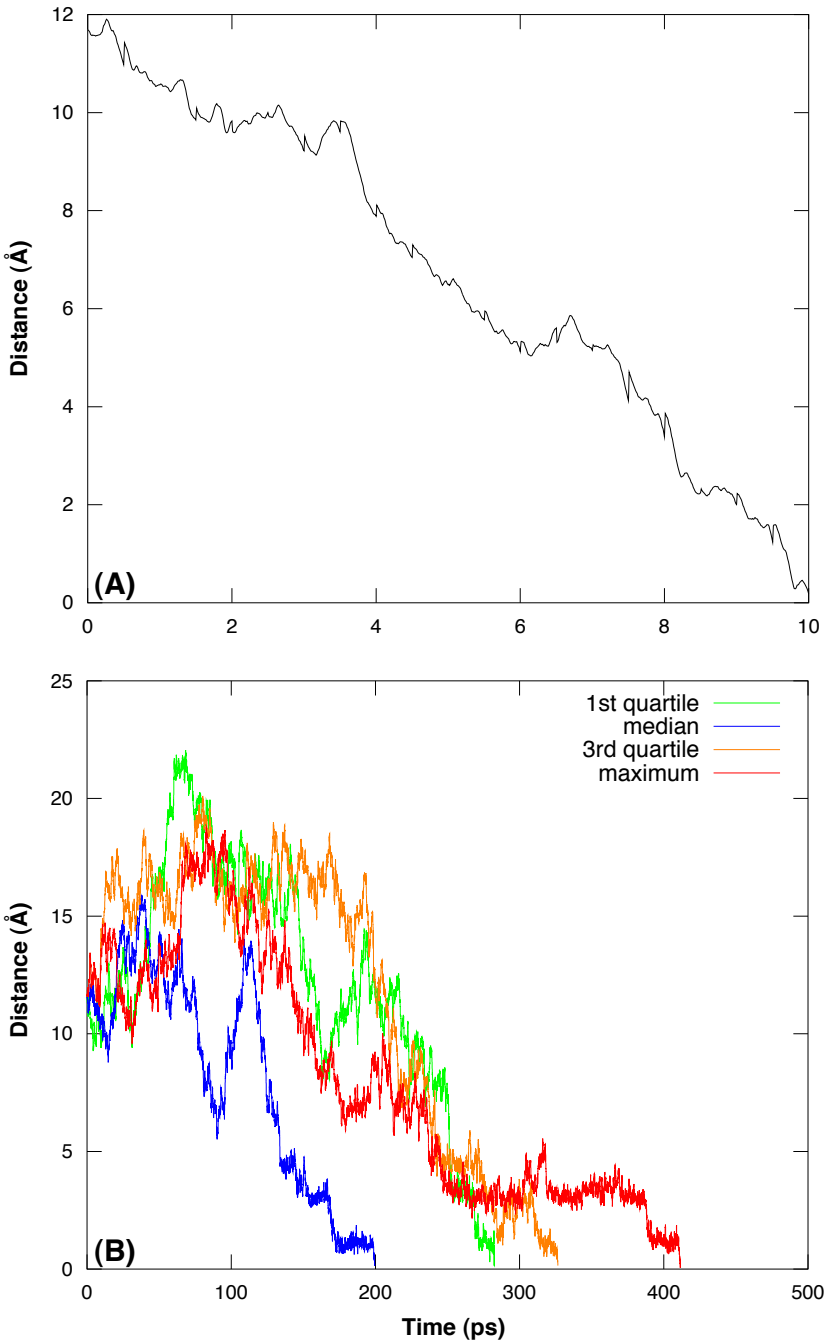


Figure 2.10: Distance between the center of mass of the 18-crown-6 oxygen atoms and the K^+ ion for (A) the shortest weighted ensemble trajectory and (B) for weighted ensemble trajectories with first-, second- (median), third-, and fourth-quartile (maximum) transition event duration times.

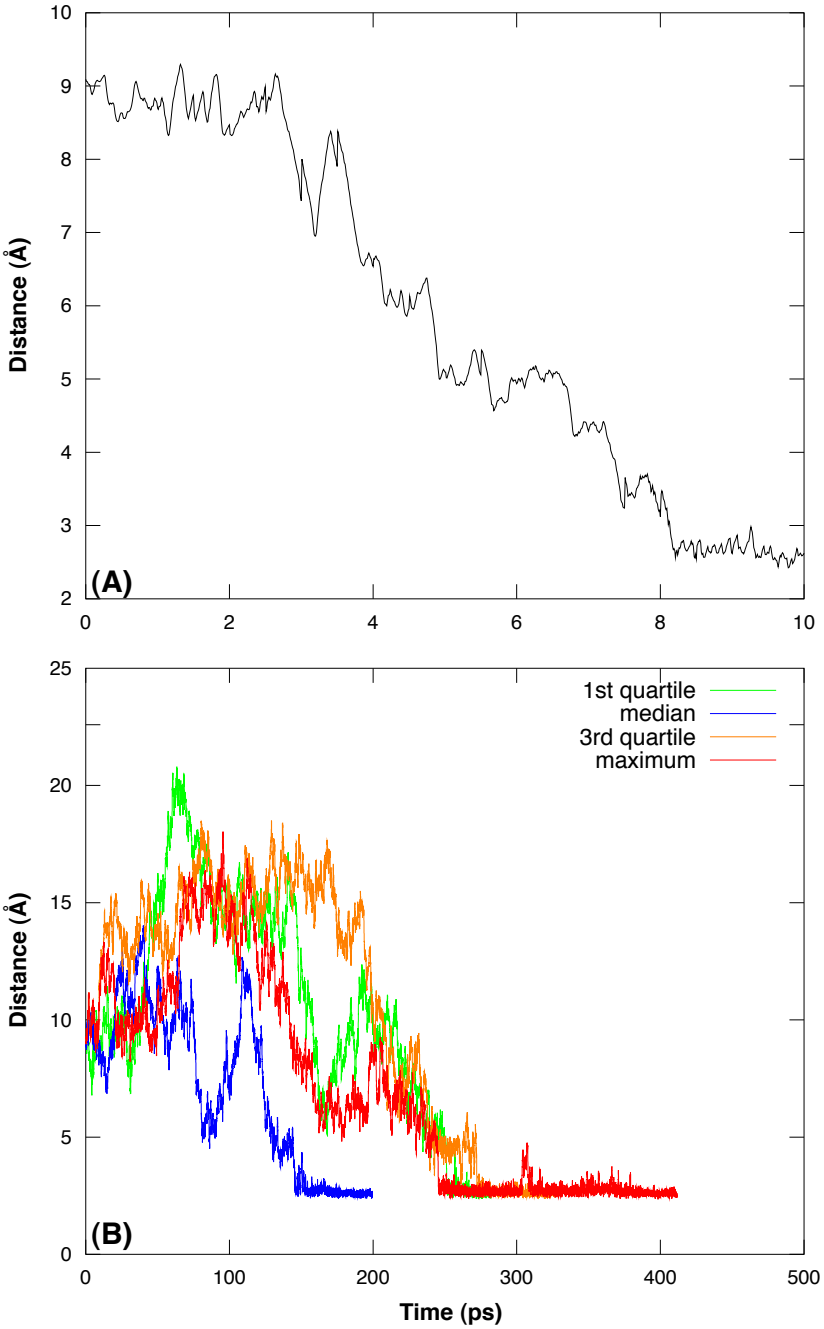


Figure 2.11: Minimum distance between the K^+ and any oxygen of 18-crown-6 for (A) the shortest weighted ensemble trajectory and (B) for weighted ensemble trajectories with first-, second- (median), third-, and fourth-quartile (maximum) transition event duration times.

3.0 COARSE-GRAINED SIMULATIONS OF PROTEIN-PEPTIDE ASSOCIATIONS

The work described in this chapter was performed in equal-authorship collaboration with David Wang, who performed the simulations themselves and compiled an initial draft of the manuscript which formed the basis of this chapter.

3.1 INTRODUCTION

The kinetics of molecular recognition events are of long-standing interest to a variety of fields, including host-guest detection, drug design, and protein engineering. Among these fields, a prevailing assumption is that the preorganization of the unbound ligand conformation to its receptor-bound conformation will result in faster association rates than more flexible unbound conformations. This assumption, however, may be challenged by the prevalence of intrinsically disordered proteins,[131] many of which adopt well-defined structures only upon binding their partners. In particular, a theoretical study has suggested that disorder provides a kinetic advantage through a "fly-casting" mechanism.[132] In this mechanism, it is hypothesized that a flexible, disordered protein has a larger "capture" radius than a preorganized, well-folded, protein, enabling it to bind its partner weakly, but more quickly at relatively long distances. The protein then folds as it "reels in" its partner, coupling its folding to the binding process.

To directly test the fly-casting hypothesis, one must compare the association rates of the disordered and exact, preorganized versions of the protein to the same partner. While a number of experimental studies have explored this hypothesis,[133–145] none provide definitive proof since it is not possible to create the preorganized analog of the intrinsically disordered protein without significantly altering its chemical structure. As a result, many have studied the coupled folding and binding process of intrinsically disordered proteins using molecular simulations with minimal models that provide residue-level detail (*i.e.* C_α models).[146–148] Only one of these studies has compared the binding kinetics of the disordered and preorganized versions of a protein, focusing on a classic system that has been proposed to bind via fly-casting: the phosphorylated KID domain and its partner protein KIX. Results from this study

suggest that the disordered KID domain does indeed have a kinetic advantage over its fully preorganized analog (~2.5x faster). However, this advantage is not attributed to a larger capture radius, as proposed in the fly-casting mechanism, but rather to a fewer number of intermolecular collisions that are required to form the intended, native complex.

Here, we use molecular simulations to directly compare the association rates for another classic system: a peptide fragment (residues 17-29) of the transactivation domain of the p53 tumor suppressor and its cellular inhibitor, the MDM2 oncoprotein. Upon binding MDM2, the intrinsically disordered p53 peptide adopts an alpha-helical conformation. The mechanism of MDM2-p53 binding has also been of great biomedical interest since many cancers have been linked to overexpression of MDM2, which inactivates p53.[149] Our study is novel in several respects. First, we employ a more detailed protein model than previous simulation studies using the usual, minimal C_α model, but with the addition of coarse-grained side chains. This level of detail was required since the binding-induced folding of the disordered peptide into an alpha-helix occurs only when steric effects of side chains are introduced. Second, we examine the kinetic effects of including hydrodynamic interactions (HI) between the protein residues in the simulations since these interactions for the disordered and preorganized p53 peptides are potentially different and could thereby result in different kinetics of binding to the MDM protein. Last, but not least, we apply the "weighted ensemble" path sampling approach[75] in conjunction with our molecular simulations to efficiently generate an extensive ensemble of binding pathways and rigorously compute association rate constants. In this study, all weighted ensemble simulations were performed using the open-source, high-performance WESTPA (Weighted Ensemble Simulation Toolkit with Parallelization and Analysis) software, which exhibits nearly perfect scaling out to thousands of CPUs (see Section A.4).

3.2 METHODS

3.2.1 The Protein Model

All proteins were represented by a residue-level model with coarse-grained side chains in which each amino acid consists of one C_α pseudo-atom and up to three side chain pseudo-atoms.[150] Using this coarse-grained model, the MDM2-p53 peptide complex consists of 262 pseudo-atoms, which is ~32% of the system size for the corresponding all-atom model (820 atoms). Coordinates for the MDM2-p53 complex were taken from the X-ray crystal structure (PDB code: 1YCR).[151] A Gō-type potential energy func-

tion[152,153] governs the conformational dynamics of the protein model. Bonded interactions between atoms are modeled by standard molecular mechanics terms:

$$E_{\text{bonded}} = \sum_{\text{bonds}} k_{\text{bond}}(r - r_{\text{eq}})^2 + \sum_{\text{angles}} k_{\text{angle}}(\theta - \theta_{\text{eq}})^2 + \sum_{\text{dihedrals}} V_1[1 + \cos(\phi - \phi_1)] + V_3[1 + \cos(3\phi - \phi_3)] \tag{3.2.1}$$

in which r, θ, ϕ are pseudo-bond lengths, pseudo-angles, and pseudo-dihedrals, respectively; V_1 and V_3 are potential barriers for the dihedral terms. 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/radian, respectively.

Nonbonded interactions between residues separated by four or more pseudo-bonds were modeled in one of two ways, depending on whether or not the residues form (native) contacts in the native, bound state. A native contact was defined as two heavy atoms located within 5.5 Å of each other in the crystal structure of the protein complex. Native contact interactions were modeled using a Lennard-Jones-like potential:

$$E_{ij}^{\text{native}} = \epsilon^{\text{native}} \left[5 \left(\frac{\sigma_{ij}^{\text{native}}}{r_{ij}} \right)^{12} - 6 \left(\frac{\sigma_{ij}^{\text{native}}}{r_{ij}} \right)^{10} \right] \tag{3.2.2}$$

in which ϵ^{native} is the energy well depth for the native interaction, r represents interatomic distance during simulation, and σ^{native} represents the corresponding distance in the crystal structure. Non-native interactions were modeled using a purely repulsive potential:

$$E_{ij}^{\text{native}} = \epsilon^{\text{non-native}} \left(\frac{\sigma_{ij}^{\text{non-native}}}{r_{ij}} \right)^{12} \tag{3.2.3}$$

in which $\sigma_{ij}^{\text{non-native}}$ and $\epsilon^{\text{non-native}}$ are set to 4.0 Å and 0.60 kcal/mol, respectively. The number of intramolecular native contacts for p53 and MDM2 are 63 and 884, respectively. In addition, there are 174 intermolecular native contacts between p53 and MDM2.

3.2.2 Brownian dynamics (BD) simulations

All simulations were performed using a standard Brownian dynamics algorithm,[154] as implemented in the UIOWA-BD software.[150,155] For simulations that do not include hydrodynamic interactions (HI), the only nonzero elements in the diffusion tensor are the diagonal elements where pseudo-atom $i = j$, which were calculated with the Stokes-Einstein relation $D_{ii} = k_B T / 6\pi\eta_s a$, where η_s is the solvent viscosity (set to 0.89 cP to represent water at 25 °C) and a is the hydrodynamic radius of the pseudo-atom. For simulations that include HI, the off-diagonal elements ($i \neq j$) are also nonzero, hydrodynamically coupling

pseudo-atoms i and j. The elements of the D_{ij} submatrices were evaluated as described by Frembgen-Kesner and Elcock.[150] We used a hydrodynamic radius of 3.5 Å, which has been found to be the optimal value for reproducing translational diffusion coefficients of all-atom protein models for the coarse-grained model used in our study.[150] A time step of 50 fs was used, constraining pseudo-bonds between residues to their native bond lengths using the LINCS algorithm.[115]

3.2.3 Parameterization of the model

To model the degree to which the MDM2 protein and p53 peptide are folded, we tuned the potential well depth ϵ^{native} accordingly. For the fully flexible, disordered version of the p53 peptide, we use an ϵ^{native} value of 0.05 kcal/mol, which resulted in a low average fraction of native contacts (0.14 ± 0.05; uncertainties represent one standard deviation) over five independent, standard Brownian dynamics simulations of 10 μs each. For the preorganized version of the p53 peptide, we used an ϵ^{native} value of 1.7 kcal/mol to keep the peptide folded even in the unbound state (average fraction of native contacts: 0.995 ± 0.0096). An ϵ^{native} value of 0.6 kcal/mol was chosen for the MDM2 protein to ensure that it remains folded throughout 10-μs simulations. Finally, the ϵ^{native} value for native residue-residue contacts between MDM2 and p53 was tuned to the minimum value (6.0 kcal/mol) that would enable the fully disordered peptide to adopt the expected helical conformation when bound to MDM2 and remain in this state throughout 10 μs of standard Brownian dynamics simulations. This value was set to be the same for the disordered and preorganized versions of the p53 peptide, assuming that their MDM2 binding affinities are the same. This assumption is supported by the fact that the distributions of the fraction of native contacts between MDM2 and the p53 peptide observed from the 10-μs simulations are virtually indistinguishable from one another for the disordered and preorganized versions of the peptide (Figure 3.1).

3.2.4 Weighted ensemble simulations

To efficiently simulate protein binding events, we used the weighted ensemble (WE) path sampling approach[75] in conjunction with Brownian dynamics (BD) propagation. In the WE approach, a progress coordinate is set up between the initial and target states (*e.g.* unbound and bound states, respectively). The progress coordinate is then divided into bins with the goal of populating each bin with N simulations, or "walkers," which are each assigned a statistical weight. Starting with N walkers in the initial state, the dynamics of each walker are simultaneously propagated and occasionally coupled by replica-

tion and combination events at short time intervals τ based on their progress towards the target state, splitting and combining the statistical weights, respectively, such that no bias is introduced into the dynamics.[78,81] Any walkers that reach the target state are terminated, and their probabilities are recycled into new initial states, thereby maintaining steady-state conditions. Further, any walkers where the separation between p53 and MDM2 exceeded 50 Å were similarly terminated and recycled into new initial states, in order to maintain a constant effective concentration of 3.2 mM.

The simulations were started from 3000 unbound states in which the p53 peptide and MDM2 protein were separated by b = 35 Å with randomly selected orientations. The unbound conformations of each binding partner were randomly selected from five 10-μs standard Brownian dynamics simulations of that binding partner. A one-dimensional progress coordinate was used for WE sampling, consisting of the root-mean-square deviation (RMSD) of C_α atoms in the p53 peptide after alignment of MDM2. This progress coordinate was partitioned with a bin spacing of 4 Å. The WE propagation/resampling interval τ was set to 100 ps, which allowed for at least one walker to traverse a bin after propagation for a time τ; the number of walkers per bin was enforced to be 48. Trajectories reaching either the bound state or a "drift" state in which p53 and MDM2 are at least 50!Å apart were recycled to the initial state with randomly chosen orientations and conformations. The bound state was defined as having a C_α RMSD of p53 within one standard deviation of the average value obtained from simulations of the bound state (0.098 ± 0.018 Å). Simulations were run for 2000 – 3000 iterations, requiring ~6 days using 112 cores at a time on 2.66 GHz Intel Xeon quad-core processors (325 ns/day/core). This number of iterations was sufficient for achieving convergence of the computed association rate constants (Figure 3.4). To generate free energy landscapes of the binding process, each of the two simulations with HI was extended an additional 500 iterations with no recycling at the bound state (while maintaining recycling at the outer sphere) to generate state populations under pseudo-equilibrium conditions.

3.2.5 Calculation of rate constants

All rate constants were calculated using the first 2000 iterations (for simulations with HI) or 3000 iterations (for simulations without HI) of the weighted ensemble simulations in which the dynamics were propagated under steady-state conditions. To calculate a rate constant k_{ij} for transitions from state i to j (*i.e.* rate constants of transitions from the unbound state to the bound state k_{on}, unbound state to encounter complex k_1, encounter complex to unbound state k_{-1}, and encounter complex to bound state k_2), we monitor the flux of probability f_{ij} carried by WE walkers from state i to state j per unit time, and

normalize by the fraction of probability p_i most recently in state i:[156]

$$k_{ij} = \frac{f_{ij}}{p_i} \tag{3.2.4}$$

To provide internal validation of our computed association ("on") rate constants k_{on}, we also computed k_{on} using a hybrid approach which combines probability fluxes from WE simulation with the method of Northrup, Allison, and McCammon for calculating association rates with BD simulations (the "NAM approach").[157] Applying the NAM approach involves the creation of two concentric spheres in which MDM2 is positioned at the center. The radius b of the inner sphere, on which p53 is initially positioned, is chosen sufficiently large so that the motions of the binding partners are isotropic; here, $b = 35$ Å. In addition to being terminated at the bound state, walkers are also terminated at the outer sphere of radius $q = 50$ Å (the "truncation sphere"), which marks the boundary between the simulation region and a region where relative diffusion of p53 and MDM2 is considered analytically rather than by simulation. The association rate is then given by

$$k_{\text{on}} = \frac{k_D(b)\beta}{1-(1-\beta)k_D(b)/k_D(q)} \tag{3.2.5}$$

in which $k_D(r)$ is the diffusion-limited rate constant for the two partners achieving a separation r from an infinite separation and β is the probability that a simulation starting from the unbound state with a separation of b (35 Å) reaches the bound state before drifting apart to a separation of q (50 Å). Assuming that the motions of the two partners are isotropic for $r > b$ leads to the Smoluchowski result $k_D(r) = 4\pi D r$, in which D represents the relative translational diffusion coefficient of the two partners (see Section A.1.2 for a derivation of the relative translational diffusion coefficient in terms of individual translational diffusion coefficients). Under this assumption, Equation 3.2.5 reduces to

$$k_{\text{on}} = \frac{4\pi D b \beta}{1-(1-\beta)b/q} \tag{3.2.6}$$

The value for D was computed by evaluating the diffusion coefficient of each binding partner from five standard 10-μs BD simulations of each partner (a total of 100,000 conformations, sampled every 100 ps). The β value was estimated using the following equation:[77]

$$\beta = \frac{f_{\text{bind}}}{f_{\text{bind}} + f_{\text{drift}}} \tag{3.2.7}$$

where f_{bind} is the steady-state flux into the bound state and f_{drift} is the steady-state flux into the drift ($q > 50$ Å) state in the WE simulation.

Uncertainties in the rate constants were computed using a block bootstrapping method on the steady state fluxes. This method of error analysis is appropriate for a time-correlated data set with an unknown

distribution. From our dataset of fluxes, we created 10,000 new datasets of the same size as the initial sample. Each new dataset was generated by randomly selecting, with replacement, fluxes from the original dataset. A distribution of means was generated from the new datasets, from which variability was used to calculate 95% confidence intervals. Because many segments that reach the target states in our simulations share common history and are time correlated, the assumption of independent events is only valid when the data is sampled at a frequency at which there is no autocorrelation. Therefore, the autocorrelation function was calculated for flux data every iteration of weighted ensemble, which provides an estimate of the correlation time in number of iterations. Bootstrapping is then applied to the subset of flux information separated by the correlation time. Finally, error analysis and association rate calculations were performed only after an approximate steady state was achieved (Figure 3.4).

3.2.6 Calculation of percentages of productive collisions

The percentage of successful collisions was calculated by the ratio of the steady-state flux into the bound state and into a collision state defined by a minimum distance of 5 Å.

3.2.7 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:

$$R = \begin{bmatrix} \sum x_n^2 & \sum x_n y_n & \sum x_n z_n \\ \sum y_n x_n & \sum y_n^2 & \sum y_n z_n \\ \sum z_n x_n & \sum z_n y_n & \sum z_n^2 \end{bmatrix} \tag{3.2.8}$$

where R is the gyration tensor and (x_n, y_n, z_n) are the coordinates of the nth psudoatom assuming the center of geometry is located at the origin. The eigenvalues of 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:

$$R_M = 2\sqrt{\lambda_3} \tag{3.2.9}$$

3.3 RESULTS AND DISCUSSION

3.3.1 Modeling the binding-induced folding of p53

In order to simulate the MDM2-binding kinetics of the disordered and preorganized versions of the p53 peptide, it is essential to ensure that both versions of the peptide can fold to a similar extent upon binding the MDM2 protein. In particular, simulations of the native MDM2-p53 complex should be sufficiently detailed such that the fully flexible, disordered version of the peptide will remain folded. This was not possible using a minimal C_α protein model; however, the addition of coarse-grained side chains (as described in Methods) accomplished this goal, with the average fraction of native contacts of the bound peptide being 0.80 ± 0.06 (uncertainties represent one standard deviation; see Figure 3.1) over 10 μs of standard molecular simulations that included HI between protein residues. The steric effects of the side chains are therefore required for modeling the coupled folding and binding process of the fully disordered peptide.

3.3.2 Binding pathways and free energy landscape

With the aid of the WE approach, extensive sampling of MDM2-p53 binding pathways was obtained using molecular simulations, yielding 33,393 and 55,842 binding events, respectively, for the disordered and preorganized versions of the p53 peptide when HI were included in the simulations (Table 3.2). These binding events resulted from a large, diverse ensemble of 3000 unbound states in which the MDM2 protein and p53 peptide were randomly oriented. While many of these binding pathways share common histories, a large number of the pathways are fully independent: 399 and 629 pathways for the disordered and preorganized peptides, respectively.

Interestingly, the free energy landscapes of binding corresponding to these coarse-grained simulations (Figure 3.2) are consistent with those from atomistic simulations (Chapter 4), which reveal a "funnel-like" landscape near the binding site once the encounter complex is formed. This similarity in free energy landscapes suggests that the MDM2-p53 binding process may indeed involve a minimally frustrated free energy landscape that is driven by short-range interactions, as exemplified by the Gō-like potential employed here.

3.3.3 Binding kinetics

As shown in Table 3.1, the computed rate constants for MDM2-p53 associations are essentially the same for the disordered and preorganized versions of the p53 peptide, regardless of whether or not hydrodynamic interactions (HI) were included. In addition, the association rate constants computed using Equation 3.2.4 are consistent with those computed using the hybrid WE/NAM approach of Equation 3.2.6, providing internal validation of the computed rates.

With the inclusion of HI, the resulting association rate constants are ~3-fold faster than those from simulation without HI. The increased rate is likely due to the faster relative diffusion coefficient of the binding partners ($\sim 3.9 \times 10^{-6}$ cm^2 s^{-1} *vs.* $\sim 0.5 \times 10^{-6}$ cm^2 s^{-1}), which is in good agreement with that predicted for the corresponding all-atom protein models by the hydrodynamics program HYDROPRO[158] (3.6×10^{-6} cm^2 s^{-1}), confirming the effectiveness of using appropriate hydrodynamic radii to reproduce translational diffusion coefficients of the corresponding all-atom models.[150] At first glance, the increased association rate that results when HI is included may appear to be at odds with the findings from a recent study involving coarse-grained simulations of associations between the proteins barnase and barstar,[159] in which it was found that the inclusion of intermolecular HI in the simulations decreases the association rate relative to the rate computed from simulations with no HI. Our results are not in conflict with these results, since our comparison involves the inclusion of *full* HI (both intramolecular and intermolecular) relative to the case without HI.

Interestingly, percentages of productive collisions for disordered and preorganized peptides are similar (see Table 3.1). This result is in contrast to a previous simulation study in which the disordered version of the KID domain (which at ~28 residues is about twice as long as the p53 peptide considered here) appears to require a fewer number of intermolecular collisions than the ordered version to form the native complex.[148]

3.3.4 Does MDM2-p53 binding involve a "fly-casting" effect?

In our molecular simulations, which employ a Gō-type potential, the fully disordered version of the p53 peptide folds only upon binding the MDM2 protein (*i.e.* folding begins only when ~40% of the native MDM2-p53 contacts are formed; see Figure 3.5). We have therefore created the best possible scenario for observing a fly-casting effect in our simulations. Even so, we observe no significant difference in the overall association rate constants for the disordered vs. preorganized versions of the p53 peptide. Therefore, our results do not reveal an overall kinetic advantage of the disordered version of the p53 peptide

in binding MDM2. These results are consistent with the fact that the rate of binding a high-affinity peptide by a truncated, unfolded version of the Fyn SH3 domain (lacking four carboxyl-terminal residues) is essentially the same as that of the full-length, folded version of the domain.[160] In addition, it is worth noting that the kinetic advantage predicted by the original study that proposed the fly-casting mechanism is only up to 1.6-fold for the binding of a single arc repressor molecule to a DNA site.[132] This modest difference in rates is consistent with a previous simulation study comparing the kinetics of binding the KIX protein by the disordered *vs.* preorganized version of the phosphorylated, KID domain.[148] While it is reported that the binding by the disordered version of the domain is ~2.5-fold faster, the domain exists in a much greater range of conformations (*i.e.* forming ~30 – 90% of intramolecular native contacts) than the preorganized version (60 – 90%) for in the bound state defined for the rate calculations (forming 80% of intermolecular native contacts). The difference in the rates would be expected to be less than 2.5-fold if the definition of the bound state required the same extent of folding in the KID domain.

Furthermore, the disordered version of the p53 peptide does not appear to have a significantly larger capture radius than its preorganized analog (Figure 3.3). As monitored by the maximum principal axis radius (R_M), the most probable capture radius of the disordered peptide is only slightly greater (by 1.3 Å, ~8.5%) than that of its preorganized analogue. While the disordered peptide achieves a maximum radius of 6.8 Å (~41%) greater than that of the preorganized peptide, it also assumes more contracted conformations, resulting in an average only slightly different from the preorganized case. One potential explanation for the absence of "fly-casting" may be that the p53 peptide is not sufficiently long to exhibit a significantly larger capture radius in its unfolded state vs. folded state. However, according to the simulation study involving the much longer, disordered KIX domain (28 residues), any kinetic advantage resulting from the increased capture radius is negated by the effects of a slower diffusion coefficient, which is inversely proportional to the capture radius.[148]

3.3.5 Efficiency of weighted ensemble simulations

The computation of well-converged association rate constants requires the generation of a large number of binding pathways, which was greatly facilitated by the use of the WE approach. We estimate the efficiency of WE simulation over standard "brute force" simulations by considering the number of CPU hours required to generate the same number of binding events as WE simulation using brute force simulations (on the same computer resource). We approximate binding as a single-exponential process such that the number of binding events can be estimated by Nkt, where N is the number of brute force simu-

lations, k is the association rate constant, and t is the length of a simulation. If each simulation is ~2 μs in length (an amount that can be computed in about a week on a single CPU), then ~500 of these simulations would need to be run to generate the 629 binding pathways (each from a distinct unbound state) that resulted from our WE simulations with HI for the disordered p53 peptide. These ~500 simulations would require ~82,500 CPU hours. Our WE simulations of MDM2-p53 binding required only ~8600 CPU hours, representing a ~10-fold gain in efficiency over brute force simulation.

3.4 CONCLUSIONS

We have directly tested the "fly-casting" binding mechanism for a classic system involving an intrinsically disordered peptide: a p53 peptide and its partner protein, MDM2. In particular, we compared the MDM2-association rates of the fully disordered peptide with its exact, preorganized analog using molecular simulations. By applying the weighted ensemble path sampling approach, we have simulated, with high efficiency, an extensive ensemble of binding pathways, with $> 30,000$ binding events among hundreds of distinct pathways. Based on the association rate constants computed from these simulations, our weighted ensemble simulations are at least ~10 times more efficient than standard ("brute force") simulations.

The rate constants are essentially the same for both the disordered and preorganized versions of the peptide, indicating that the flexibility of the peptide has no impact on its kinetics of binding. Fly-casting is therefore not a significant effect for the MDM2-p53 peptide system, even though the ideal scenario for this effect was modeled using a Gō-type potential that ensured folding of the peptide only upon binding MDM2. Our results also do not support the conventional assumption that preorganized ligands should lead to faster binding than more flexible ligands for the MDM2-p53 peptide system. It appears that the fully flexible, "disordered" version of the p53 peptide may already preorganize the most critical residues for binding (*i.e.* residues that are the most deeply buried upon binding MDM2, forming the greatest number of residue-residue contacts between MDM2 and p53) by simply linking them together with the peptide backbone.

Finally, we investigated the effects of hydrodynamic interactions (HI) on the association rate constants. The inclusion of HI results in faster association rate constants due to faster translational diffusion coefficients of the binding partners.

Given the general features of our protein model, all of the above conclusions are likely to be relevant to any protein-peptide system in which long-range interactions are not critical and the peptide is a similar length (~13 residues) and intrinsically disordered, adopting an alpha helix only upon binding the partner protein.

3.5 ACKNOWLEDGEMENTS

This work was supported in part by NSF CAREER Award MCB-0845216 to Lillian T. Chong; University of Pittsburgh Arts & Sciences and Mellon Fellowships to M.C.Z.; and the James V. Harrison Fund and University Honors College Brackenridge Research Fellowships to David W. Wang. We are also grateful for NSF XSEDE allocation TG-MCB100109 and the University of Pittsburgh's Center for Simulation and Modeling for use of its Linux cluster. In addition we would like to thank Dan Zuckerman and David Swigon (University of Pittsburgh), Kevin Plaxco (University of California at Santa Barbara), and Dmitri Makarov (University of Texas at Austin) for insightful conversation. We also thank Adrian Elcock (University of Iowa) for making the UIOWA-BD software available.

Table 3.1: Computed rate constants and percentage of productive collisions from weighted ensemble simulation of associations of the disordered and preorganized versions of the p53 peptide with the MDM2 protein, with or without hydrodynamic interactions (HI). The relative diffusion coefficients of the p53 peptide and MDM2 protein are also reported, computed based on five 10-μs simulations of each binding partner. Uncertainties represent 95% confidence intervals.

	With HI		Without HI	
	Disordered	Preorganized	Disordered	Preorganized
WE k_{on} (10^8 $\text{M}^{-1}\,\text{s}^{-1}$)	1.8 ± 0.2	2.3 ± 0.2	0.42 ± 0.05	0.44 ± 0.04
Hybrid WE/NAM k_{on} (10^8 $\text{M}^{-1}\,\text{s}^{-1}$)	2.6 ± 0.3	2.6 ± 0.2	0.46 ± 0.05	0.44 ± 0.05
k_1 (10^{11} $\text{M}^{-1}\,\text{s}^{-1}$)	2.9 ± 0.2	2.5 ± 0.2	3.1 ± 0.8	2.1 ± 0.4
k_{-1} (10^9 s^{-1})	1.4 ± 0.3	2.8 ± 0.2	6.0 ± 2.6	17.5 ± 1.2
k_2 (10^6 s^{-1})	5.8 ± 0.7	7.4 ± 0.6	1.3 ± 0.2	1.4 ± 0.1
% productive collisions	1.05 ± 0.025	1.18 ± 0.15	0.16 ± 0.07	0.22 ± 0.06
D (10^{-9} $\text{cm}^2\,\text{s}^{-1}$)	3.91 ± 0.02	3.98 ± 0.02	0.495 ± 0.002	0.495 ± 0.002

Table 3.2: Extent of sampling of binding events.

	With HI		Without HI	
	Disordered	Preorganized	Disordered	Preorganized
Number of binding events	33,393	55,842	43,214	102,938
Number of distinct binding pathways	399	629	95	116

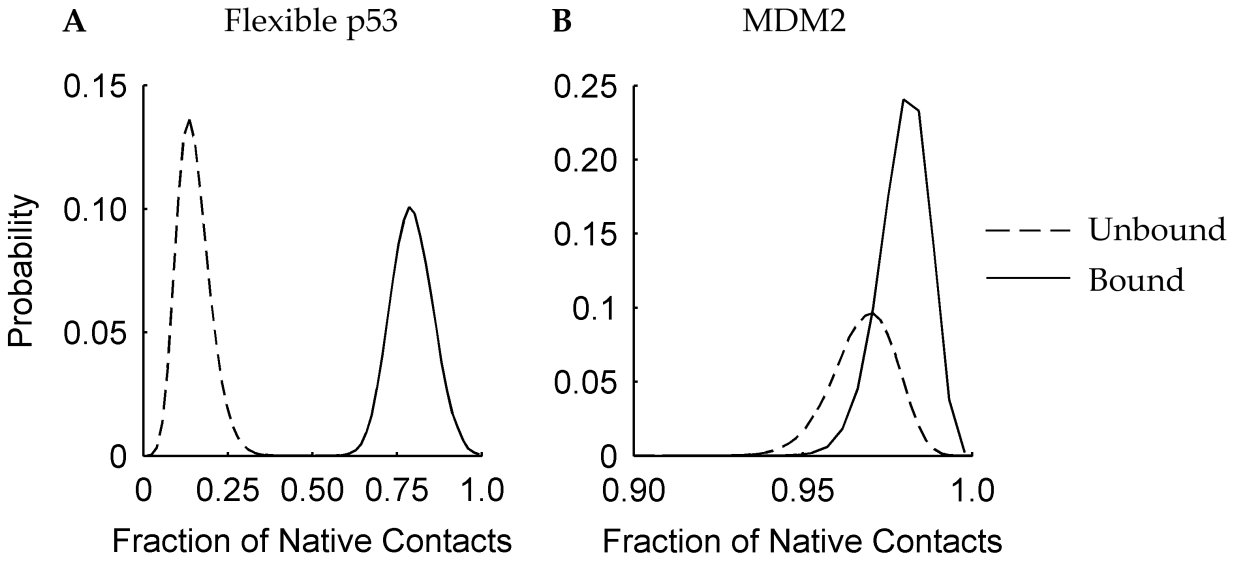


Figure 3.1: Shifts in the probability distributions of the fraction of native contacts of each binding partner from the unbound to bound state. (A) The disordered p53 peptide exhibits dramatic conformational change between the unbound (dashed line) and bound (solid line) states. The preorganized peptide was highly folded even in the unbound state (0.995 ± 0.0096) (B) MDM2 becomes slightly more folded upon binding MDM2. All distributions were generated from five 10 μs runs of standard Brownian dynamics simulations. Conformations were taken every 100 ps.

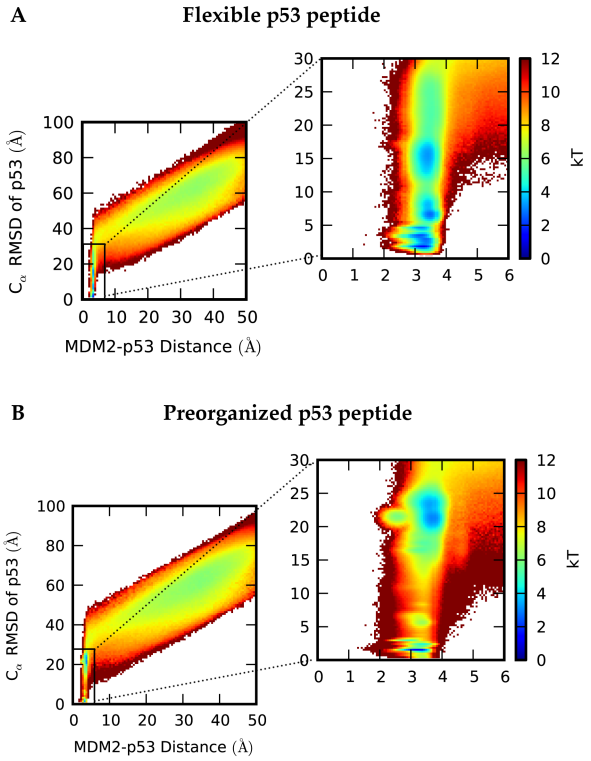


Figure 3.2: Free energy landscapes as a function of the C_α RMSD of the p53 peptide after alignment of the MDM2 protein (from the crystal structure of the bound state[151]) vs. the minimum distance between the MDM2 protein and p53 peptide. (A) The diffusion of the disordered p53 peptide is nearly barrier-free. After colliding with MDM2, it climbs a small, but significant, kinetic barrier for rearrangement into the bound state. (B) The preorganized p53 peptide exhibits similar barrier-free diffusion. In contrast to disordered p53, it must overcome a large barrier for rearrangement. Data was taken from all trajectories in weighted ensemble simulations, where conformations were sampled every 1 ps. Energies were calculated as the negative log of the probability in each bin. Probabilities were determined based on the weights of each walker in weighted ensemble.

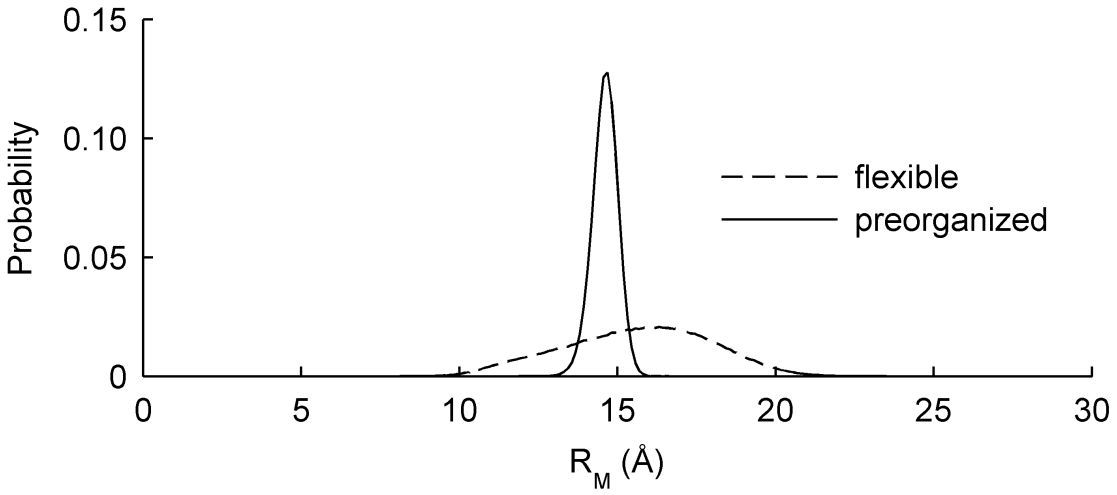


Figure 3.3: Distributions of the "capture" radius, as measured by the maximum principal axis radius R_M, for both the disordered (dashed line) and preorganized (solid line) p53 peptides. Distributions are taken from 100 μs of molecular simulations of unbound p53, where conformations are taken every 100 ps.

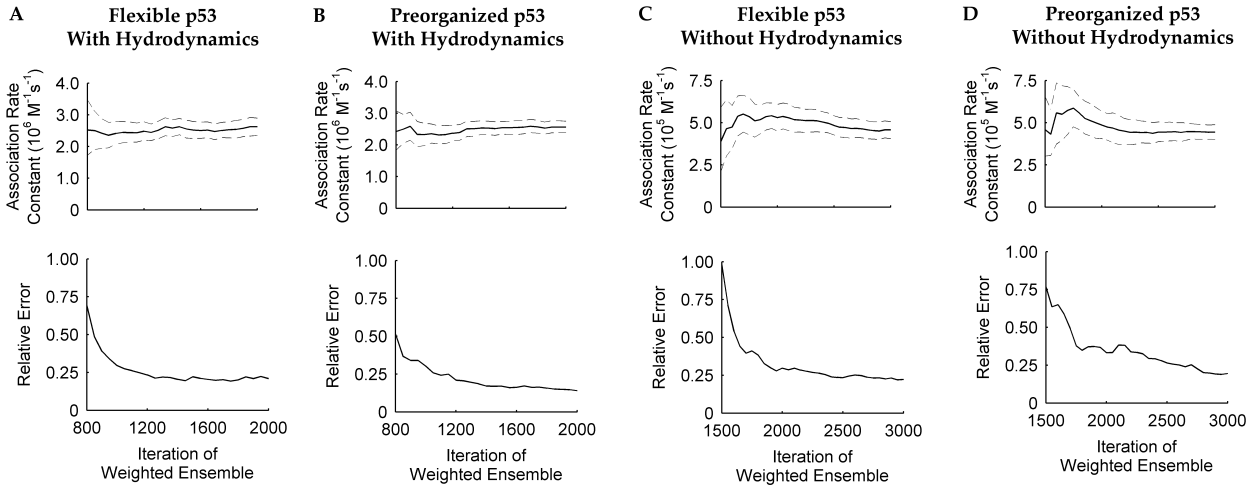


Figure 3.4: Evolution of MDM2-p53 association rates. The dashed lines indicate the 95% confidence intervals.

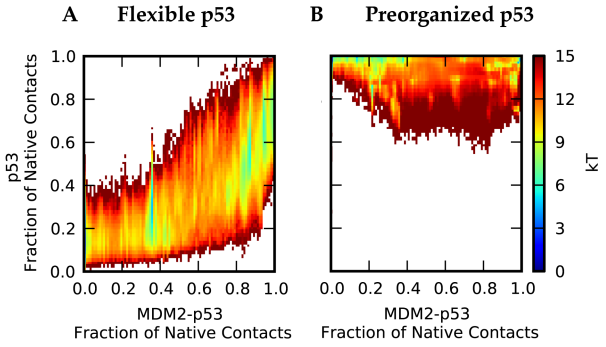


Figure 3.5: Free energy landscapes as a function of the fraction of native contacts formed within the p53 peptide vs. between the MDM2 protein and p53 peptide. Conformations were sampled every 1 ps. (A) The flexible p53 peptide remains folds only after a significant portion of the peptide is bound to MDM2. (B) The preorganized p53 peptide undergoes a small degree of unfolding upon binding before reassuming its native, bound conformation.

4.0 ATOMISTIC SIMULATIONS OF PROTEIN-PEPTIDE ASSOCIATIONS

4.1 INTRODUCTION

The complete characterization of protein binding processes has remained elusive to laboratory experiments due to the fleeting nature of transition and intermediate states along the pathways. Molecular dynamics (MD) simulations can, in principle, provide a "microscope" for viewing these critical biological processes in full atomic detail and high temporal resolution, but are computationally demanding. While advances in computer hardware and software have led to notable successes in simulating protein binding events with either small-molecule inhibitors [161–164] or peptides, [165] such "brute force" simulations — simply running the simulations long enough to capture at least one binding event — are not practical on typical computing resources. Some efforts have therefore been made to estimate rate constants for long-timescale processes from relatively short, discontinuous simulations [163,166] with the construction of Markov state models. [167] However, it is not clear that sufficient sampling has been achieved in these studies to characterize the free energy landscape of binding.

A promising, alternative strategy for efficiently capturing long-timescale processes is to focus computing effort on the "rare", or infrequent, functional motions rather than the stable states (e.g. protein conformational changes upon binding and not on the unbound or bound states) without altering the underlying dynamics. [83] One such strategy is the "weighted ensemble" (WE) path sampling approach, [75] which can generate continuous trajectories and rate constants for the rare event of interest in a rigorous manner for any type of stochastic dynamics. [78] The WE approach has been used in conjunction with MD simulations of molecular associations, [81] Brownian dynamics simulations of protein binding [75,77] and protein folding [76] and in coarse-grained simulations of large conformational transitions between alternate folded states of proteins. [80,84,168] We have demonstrated that the WE approach can be orders of magnitude more efficient than standard simulations — in terms of total computing time — in generating pathways and rate constants of rare events for benchmark systems. [80,81,84,169] To efficiently simulate com-

plex biological processes that involve metastable intermediates (*e.g.* protein folding and binding), however, it is necessary to apply the WE approach with a reweighting procedure for either non-equilibrium[68] or equilibrium conditions.[156] The latter enables the simultaneous generation of equilibrium and non-equilibrium observables, including state populations and rate constants; importantly, states do not have to be defined in advance to compute rate constants between them. Other rare-events methods (*cf.* Refs. 50,58,59,71,170,171), to our knowledge, have not been applied to protein binding.

Here, we apply the WE approach with an equilibrium reweighting procedure[156] (see Section A.3) to characterize the free energy landscape and kinetics of binding between a peptide fragment of the p53 tumor suppressor and its cellular inhibitor, the MDM2 oncoprotein. The MDM2-p53 peptide complex is a classic system for studying protein-peptide binding in which the intrinsically disordered p53 peptide[172] adopts a helical conformation upon binding to a well-defined hydrophobic pocket of MDM2.[151] Since many cancers are related to inactivation of p53 due to overexpression of MDM2,[149] there is great interest in characterizing the MDM2-p53 binding process in order to determine ways to disrupt the binding.[173] Previous MD simulations of MDM2-p53 peptide binding have focused on solely the unbound and/or bound states,[174–182] and the final approach of p53 to MDM2 from very short (~3–5 Å) distances.[183,184] In addition, Brownian dynamics simulations have been used to generate diffusional collisions between rigid models of the MDM2 protein and p53 peptide to form transient "encounter complex" intermediates.[185]

In this work, we use a standard, fully flexible, all-atom protein model with GB/SA implicit solvent[186,187] to simulate the complete MDM2-p53 binding process, including transitions from the metastable encounter complex to the native complex. More than 2000 continuous binding pathways have been generated starting from an ensemble of >1500 unbound states in which the binding partners are separated by 30 Å at random orientations. Our results demonstrate that the WE approach can simultaneously and efficiently generate both free energy landscapes and rigorous rate constants for a protein binding process.

4.2 RESULTS AND DISCUSSION

We have used the weighted ensemble (WE) path sampling approach[75,156] with an equilibrium reweighting procedure[156] to generate pathways, free energy landscapes, and rate constants for associations and dissociations between an N-terminal p53 peptide (residues 17 – 29) and the MDM2 oncoprotein (resi-

dues 25 – 109). This approach couples thousands of parallel trajectories to ensure uniform coverage of a pre-selected progress coordinate without any bias in the dynamics of the simulation.[75,80] Consistent with previous simulation studies of MDM2-p53 interactions,[178–181,183,184] our model of MDM2 is a truncated version of the protein (residues 25-109) that lacks a mobile N-terminal region that is unresolved in the crystal structure[151] and functions as a "lid" over the p53 binding cleft.[188]

4.2.1 Free energy landscape of binding

Starting from >1500 well-equilibrated unbound states (see Figure 4.4 and "Methods" for details), our WE simulation our WE simulation resulted in a thorough exploration of the space around the MDM2 protein by the p53 peptide (Figure 4.5). The simulation generated a total of 296 continuous binding pathways for the p53 peptide and MDM2 protein and was completed in ~15 days using ~3,500 CPU cores at a time on the XSEDE Stampede supercomputer (aggregate simulation time of ~120 μs). Progress toward binding was monitored with an RMSD-based progress coordinate and the bound state definition was refined prior to kinetics analysis (see "Methods" and Figure 4.7). Two thirds (~80 μs) of our ~120 μs of aggregate dynamics are involved in binding pathways, with ~10% of the diffusional collisions between the p53 peptide and MDM2 protein being productive (*i.e.* resulting in the native complex). All ~1500 of our initial unbound configurations of p53 and MDM2 contributed to binding, including non-helical p53 conformations (Figure 4.4) and configurations where p53 did not have a direct approach to the binding pocket. The free energy landscape of the simulation as a whole (Figure 4.1A and B) and the portion of the free energy landscape explored by trajectories contributing probability flux to binding events (Figure 4.1C) are quite similar, indicating that we have effectively sampled the portion of the free energy landscape relevant to binding.

The free energy landscape (Figure 4.1A and B) of binding between the p53 peptide and MDM2 protein appears to be "funnel-like", involving multiple binding pathways going down a free energy gradient. In particular, after diffusive collisions of p53 and MDM2 form a metastable "encounter complex' intermediate, rearrangement of this encounter complex to the bound state is largely downhill. The idea of a funnel near the protein binding site has been suggested by others[189] within the context of lattice-model simulations[190] and docking studies[191–193] to rationalize the fact that protein-protein associations occur at rates which are $> 10^3$ times faster than would be expected from the collision frequencies of spherical particles with specific docking constraints (~10^3 $M^{-1}\,s^{-1}$).[194] Importantly, the existence of a binding funnel has been recognized as a requirement for robust docking of small-molecule inhibitors to protein

drug targets.[195] Our atomistic simulations of protein binding pathways provide direct confirmation of the binding funnel for a classic protein-peptide interaction.

4.2.2 Binding affinity and rate constants

Although the durations of successful binding trajectories range from ~2.5 ns to ~20 ns, the computed association ("on") rate constant is $k_{on} = 7.1 \times 10^7$ $M^{-1} s^{-1}$ (with a 95% confidence interval of $k_{on} = (3.5, 11.5) \times 10^7$ $M^{-1} s^{-1}$), suggesting that the protein binding process involves true rare events, *i.e.* the actual duration of the binding transition itself is orders of magnitude less than its mean passage time. This k_{on} value is ~10x slower than the only experimentally measured k_{on} for the MDM2-p53 complex, which involves the full-length MDM2 protein (9.2×10^6 $M^{-1} s^{-1}$).[196] As evident in Figure 4.2, the rate-limiting step for MDM2-p53 binding (in the absence of the flexible MDM2 lid) is the formation of the transient encounter complex, which is diffusion-controlled ($> 10^5$ $M^{-1} s^{-1}$).[197] This step is then followed by relatively rapid conformational rearrangements of the encounter complex to the native complex (for a representative pathway, see Figure 4.3).

The faster k_{on} computed for MDM2-p53 binding with the truncated MDM2 protein relative to the experimental k_{on} with the full-length protein likely results from the constitutively open binding cleft in truncated MDM2 in contrast to the ~10% open population in full-length MDM2.[198] Based on NMR studies, it has been demonstrated that the open state is in a slow dynamic equilibrium (>10-ms timescale) with the closed state and that the only significant structural differences between these states are in the flexible MDM2 lid region.[198] Furthermore, these studies have shown that the open state gives rise to a set of resonances that is nearly identical to the single set of resonances observed for p53-bound MDM2, reflecting a shift of the conformational equilibrium towards the open state upon binding the p53 peptide. If one assumes that binding only occurs when the lid is open (*i.e.* 10% of the time), our measured rate decreases by an order of magnitude, to $k_{on} = \sim 7 \times 10^6$ $M^{-1} s^{-1}$, which coincides with the experimental value. It should also be noted that our computed k_{on} is likely faster than experimental measurements due to our use of an implicit solvent model, since the frictional collision frequency is uniform throughout this model. In particular, it has been proposed that the friction between a ligand and the solvent increases as the ligand approaches a hydrophobic binding pocket, slowing down its rate of association.[199]

In principle, an equilibrium WE simulation can simultaneously provide both association and dissociation ("on" and "off") rates. However, since our WE setup (*e.g.* progress coordinate, scheme for combining trajectories, *etc.*) was focused on sampling association events, we did not observe a sufficient

number of dissociation events to directly compute the k_{off} value. Instead, we estimated k_{off} as $k_{on}K_D$ (using our computed k_{on} of $\sim 7 \times 10^7$ $M^{-1}s^{-1}$) and computed K_D directly from the populations of the unbound and bound states sampled by our simulation. The computed K_D = 60 nM is within a factor of 2.5 of the experimentally measured value [121 nM[200]] for association of the p53 peptide to the same truncated MDM2 protein in our simulation. The resulting k_{off} = 4.3 s^{-1}, which is within ~2-fold of the experimentally measured value for dissociation of the p53 peptide from the full-length MDM2 protein (2 s^{-1}).[196]

4.2.3 The role of p53 residue F19

To further characterize the encounter complex, a free energy landscape was generated as a function of the heavy-atom RMSD of three key "anchor" residues of the p53 peptide from its MDM2-bound crystallographic pose after alignment of *(a)* MDM2 and *(b)* itself. The anchor residues — F19, W23, and L26 — are the residues of the p53 peptide that become the most buried upon binding MDM2. These residues have been found by experimental mutagenesis studies to be crucial for binding.[196] About 70% of the conformations in the encounter complex state feature burial of the p53 residue F19 at the p53-MDM2 interface (Figure 4.8 for details), suggesting that F19 may be a kinetically important residue for MDM2-p53 binding. This result is consistent with experiments which have found that an F19A mutation of p53 abolishes MDM2-p53 binding.[201] In addition, simulations involving the final approach of the F19A p53 peptide have found that the peptide is unable to anchor into the binding cleft, sliding on the MDM2 surface in a nonspecific manner instead.[202] Such deeply buried residues have been proposed to function as "anchors" that smooth out the binding process by avoiding kinetically costly structural rearrangements.[203]

4.2.4 Efficiency of WE simulation

To estimate the efficiency of the WE sampling relative to brute force simulation, we focus on the amount of computing time required on the same computing resource (XSEDE's Stampede) to estimate the rate constant for the longest timescale process, *i.e.* dissociation events, which occur > 10^4-fold more slowly than the association events. Based on our estimated k_{off}, the mean first passage time for unbinding is ~0.2 s. To generate just a single dissociation event, $\sim 7 \times 10^8$ CPU hours would be required for a brute force simulation; using 3500 CPU cores at a time, the simulations would require ~26 years of wall-clock time to complete. The estimation of k_{off} is clearly not tractable in this case, particularly since multiple dissociation events would be needed. Our WE simulation required only ~1 million CPU hours and

provided reasonable estimates of the K_D, k_{on}, and k_{off} as well as a well-resolved free energy landscape. In estimating the k_{off} value alone (as well as K_D from which k_{off} is computed), our WE simulation is ~1000-fold more efficient than brute force simulation. Since our simulation also generates a free energy landscape of binding, which would not be well-resolved by the brute force simulations, this efficiency is reasonably conservative.

4.3 CONCLUSIONS

We have used the WE path sampling approach[75] with an equilibrium reweighting procedure[156] to simulate the binding between an N-terminal peptide fragment of the p53 tumor suppressor with the MDM2 oncoprotein in atomic detail. Our simulation generated about 300 continuous binding pathways, resulting in computed k_{on}, k_{off}, and K_D estimates consistent with experiment. This encouraging agreement suggests that current simulation models such as the one used in this study [AMBER ff99SB-ILDN[204] with GB/SA implicit solvent[186,187]] are capable of identifying the native, bound state of the MDM2-p53 system and providing realistic kinetics as well as thermodynamics.

The resulting free energy landscape provides the first confirmation (to our knowledge) from unbiased, atomistic simulation that protein binding pathways can follow a "funnel-like" landscape, which has long been an important assumption of molecular docking strategies for virtual screening and drug discovery. In the absence of the N-terminal flexible lid in MDM2, MDM2-p53 binding is diffusion limited due to the rate-limiting formation of a metastable, encounter complex state, which subsequently undergoes rapid rearrangement to the native, bound state. A key feature of the encounter complex state is the anchoring of the p53 residue, F19, into the binding cleft of MDM2, suggesting that F19 may be a kinetically important residue for MDM2-p53 binding.

Our WE simulations simultaneously provided pathways and equilibrium state populations in atomic detail with rigorous rate constants for a complex biological process, *i.e.* protein-peptide binding, in ~15 days on a supercomputer, which is orders of magnitude more efficient than brute force simulations. The results suggest that rare-event strategies like WE could become an important technique for modern biophysics.

4.4 METHODS

4.4.1 Weighted ensemble (WE) simulation

We simulated the association between the MDM2 protein (residues 25 – 109) and p53 peptide (residues 17 – 29) using the WE approach[75] with an equilibrium reweighting procedure,[156] as implemented in the open-source, high-performance WESTPA (Weighted Ensemble Simulation Toolkit with Parallelization and Analysis) software (see Section A.4). In this approach, a large number of simulations, or "walkers", are started in parallel from the initial, unbound state and evaluated for replication or combination every τ (according to the standard WE algorithm[75]) to maintain the desired number of walkers per bin along a progress coordinate towards the target bound state; in our simulations, we used a τ value of 50 ps. All WE simulations were performed by applying the equilibrium reweighting procedure[156] at regular intervals of τ for the first half of the simulation to accelerate convergence in the sampling of thermodynamics and kinetics of association. This procedure (outlined in Section A.3) uses the local convergence of kinetics to properly redistribute weight across the entire progress coordinate space, and is required for accurate and efficient equilibrium WE simulations in the presence of metastable intermediate states.[68,156] As a test of simulation convergence, no equilibrium reweighting was applied in the second half of the trajectory to ensure that the results remain unchanged in this part of the trajectory.

To extensively sample the unbound conformations of each binding partner, we performed a separate WE simulation (with equilibrium reweighting) of each binding partner starting from its coordinates in the crystal structure of the native MDM2-p53 complex (PDB code: 1YCR)[151] (the MDM2 protein was capped with an acetyl group, with a charged C-terminus; the p53 peptide was capped with acetyl and NH_2 groups). These simulations involved the use of a one-dimensional progress coordinate consisting of a heavy-atom RMSD of protein/peptide from its conformation in the native complex, increasing towards 10 Å. The coordinate was partitioned with a bin spacing of 0.1 Å; a total of 32 walkers per bin was enforced throughout the simulations. Initial, 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 30 Å to yield ~6.2 million possible pairs of unbound conformations of MDM2 and the p53 peptide. These millions of walkers in the initial, unbound state were then reduced to ~1500 walkers by assigning the walkers to appropriate bins along the two-dimensional progress coordinate intended for the binding simulation and combining walkers with small weights according to the standard WE algorithm[75] to yield

8 walkers per occupied bin. This progress coordinate consisted of decreasing heavy-atom RMSDs of the p53 peptide relative to its MDM2-bound crystallographic pose[151] following alignment on *(a)* MDM2 (to monitor the extent of binding) and *(b)* itself (to monitor the extent of preorganization of the peptide for binding). The coordinate was partitioned with a bin spacing of 0.2 – 2 Å in MDM2-aligned RMSD and 0.5 Å in p53-aligned RMSD. Using this two-dimensional progress coordinate, a total of ~400 iterations were performed to generate binding pathways, with a maximum trajectory length of ~20 ns. After ~200 WE iterations (about 57 μs of aggregate simulation time), both the energy landscape in the progress coordinate (Figure 4.9) and the association rate (Figure 4.10) were reasonably converged. All analysis was performed using the latter half of the simulation with conformations sampled every picosecond.

4.4.2 Propagation of dynamics

Dynamics in the WE simulation were propagated using the GROMACS MD engine (version 4.5.3)[28] along with the AMBER ff99SB-ILDN force field[204] and a generalized Born/surface area (GB/SA) implicit solvent model[186,187] (Born radii calculated according to the OBC method[205]). The GROMACS 4.5.3 source code was modified to include the electrostatic effects of a uniform monovalent salt concentration on the free energy of solvation (as employed in GB/SA calculations) according to a Debye-Hückel screening expression.[205,206] Consistent with stopped-flow kinetics experiments,[196] the ionic strength was set to 150 mM. To maintain a constant temperature of 25° C, a Langevin thermostat[112] was used with a water-like collision frequency of 58 ps^{-1} [calculated from the self-diffusion coefficient of water[207]]. The use of a stochastic thermostat is required for WE sampling with MD simulation since the dynamics of the simulation "walkers" must diverge when the walkers are replicated. To enforce a constant effective protein/peptide concentration (3.2 mM) while maintaining equilibrium conditions, the momenta of both the MDM2 protein and p53 peptide in the trajectories were reversed whenever the minimum distance between the MDM2 protein and p53 peptide was greater than 50 Å and the center-of-mass distance between the protein and peptide was increasing. Bonds to hydrogen atoms were constrained to their equilibrium lengths with the LINCS algorithm,[115] which permitted the use of a 2 fs integration time step. Nonbonded interactions were truncated at 16 Å.

4.4.3 State definitions

The definition of the bound state (Figure 4.7, A and B) was refined based on the energy landscape obtained from the WE simulation and confirmed by a separate control simulation that was started from the

bound state (Figure S2); the average heavy-atom RMSD of the overall bound state from the corresponding crystal structure[151] was 2.5 ± 0.5Å (the uncertainty represents an approximate 95% confidence interval). Consistent with previous Brownian dynamics simulations of protein-protein association,[208] the encounter complex through which all binding pathways pass was defined as a specific complex (with at least one intermolecular native contact and a certain extent of binding by the p53 anchor residues) as delineated in Figure 4.7C. The unbound state was defined as a p53-MDM2 separation of > 20 Å. Rate constants for state-to-state transitions were computed as described below and were not particularly sensitive to the choice of state boundaries.

4.4.4 Rate calculations

The rate constant k_{ij} between states i and j is computed using the following:

$$k_{ij} = \frac{f_{ij}}{p_i} C_{\text{eff}}^{-1}$$

where f_{ij} is the flux of probability carried by walkers originating in state i and arriving in state j and C_{eff} is the effective concentration of binding partners, calculated as $C_{\text{eff}} = 1/[(4/3)\pi r^3 N_A]$ where r = 50 Å is the radius of the simulation region and N_A is Avogadro's number. (In these simulations, C_{eff} = 3.2 mM.) Normalization by p_i amounts to a separation of equilibrium fluxes into multiple steady-state fluxes, and is what allows us to extract rate constants corresponding to steady state experiments from equilibrium data.[209] The conditional flux f_{ij} from state i to state j is evaluated by tracing the continuous trajectories generated by the WE approach and noting when transitions from state i to state j occur; if such a transition occurs any time within iteration N_i of WE sampling, then that transition generates a contribution w/τ to the conditional flux $f_{ij}(N_i)$ from state i to state j arriving within iteration N_i, where w is the weight of the walker at the time of the transition. These flux values may be correlated in time (N_i) so, as done by others,[75] uncertainties in the rate constants k_{ij} were computed using a Monte Carlo bootstrapping strategy.[106] In particular, bootstrapping was used to first determine the correlation time t_c of f_{ij}, representing the maximum lag time for which the autocorrelation of flux was statistically significant. The f_{ij} measurements were then averaged in blocks of length t_c, and the averages of the blocks are used as input for another bootstrap to determine the 95% confidence interval of the mean flux. This procedure reduces to block averaging[210] in the limit of large effective sample size, but will properly account for non-normality of the sampling distribution of the mean at smaller effective sample sizes.

4.4.5 Dissociation constant calculation

We calculate the dissociation constant K_D as

$$\begin{aligned} K_D &= \frac{[\text{unbound state}]^2}{[\text{bound state}]} \\ &= \frac{[P_{\text{unbound}}]^2}{[P_{\text{bound}}]} C_{\text{eff}} \end{aligned}$$

where P_{unbound} and P_{bound} are the probabilities of being in the unbound or bound states and C_{eff} is the effective concentration, calculated as described above. Estimating P_{unbound} and P_{bound} from our simulation requires consideration of the Jacobian contribution to the volume of configuration space subtended by each state. As the Jacobian of the bound state described above is difficult to determine, we instead use the distance r of the most buried atom of p53 W23 to its position in the minimized crystal structure. We then create a histogram of r from the positions of all walkers in the latter half of the WE simulation, then divide the histogram by r^2 to account for the Jacobian term in going between the three-dimensional Cartesian coordinates of W23 position and the one-dimensional distance r to its position in the minimized crystal structure. We then normalize the histogram and integrate over the regions corresponding to the bound state ($r < 1.0$ Å) and the unbound state (25 Å $< r <$ 60 Å) to yield P_{bound} and P_{unbound} respectively. These states are determined by inspection of the potential of mean force along r (Figure 4.11) and are conservative (as small as possible for the bound state and as large as possible for the unbound state). The values of K_D and k_{off} thus obtained are conservative and are not particularly sensitive to the particular boundaries chosen.

4.5 ACKNOWLEDGEMENTS

This work was supported by NSF CAREER grant MCB-0845216 to Lillian T. Chong, University of Pittsburgh Arts & Sciences and Mellon Fellowships to MCZ, NIH grant T32-DK061296 to Joshua L. Adelman, NSF grants MCB-0643456 and MCB-1119091 to Daniel M. Zuckerman, and NSF XSEDE allocation TG-MCB100109 to Lillian T. Chong. We thank Karl Debiec and Ali Sinan Saglam for constructive discussions on data storage and analysis strategies and Adam Pratt for discussion and assistance in data analysis for Figure 4.1.

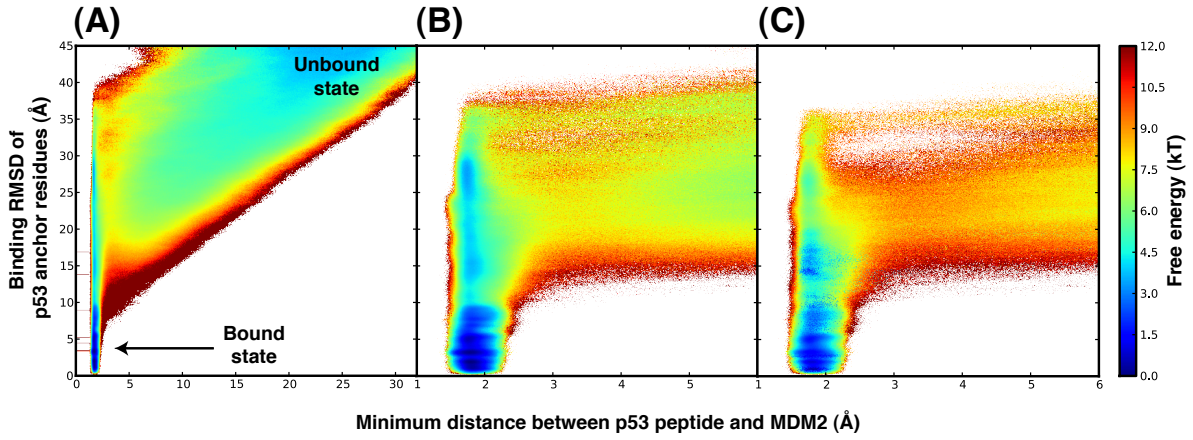


Figure 4.1: Landscapes for free energy of p53-MDM2 binding. The landscape of binding is funnel-like (A), as any trajectory that surmounts the barrier to initial association faces a downhill rearrangement (B) to the bound state. A representative continuous binding pathway is superimposed on the free energy landscape in (B) (see also Movie S1). Fully two-thirds of our 120 μs of dynamics are involved in binding, and the energy landscape obtained from trajectories contributing flux to binding (C) matches that of all trajectories (B). The binding RMSD of p53 anchor residues was defined as the heavy atom RMSD of the three key hydrophobic residues of p53 that are deeply buried upon binding the MDM2 protein — F19, W23, L26 — evaluated after alignment of the heavy atoms of MDM2.

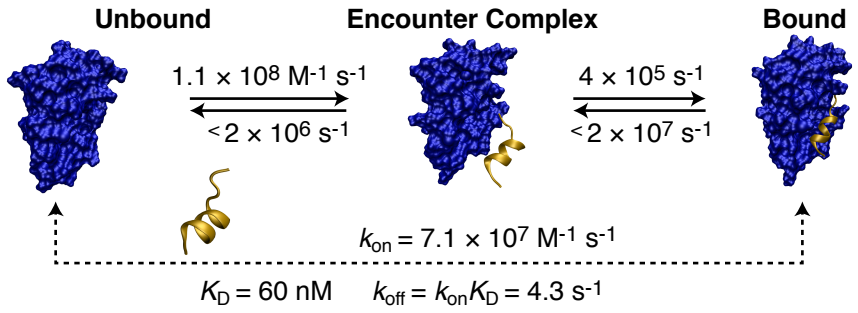


Figure 4.2: Kinetic mechanism for binding of p53 to MDM2 binding. Diffusion-limited collisions result in the formation of a metastable encounter complex (center), which rapidly interconverts with the bound state (right). Measuring the populations of the bound and unbound states allows us to determine a K_D, which in turn allows us to estimate the overall k_{off}.

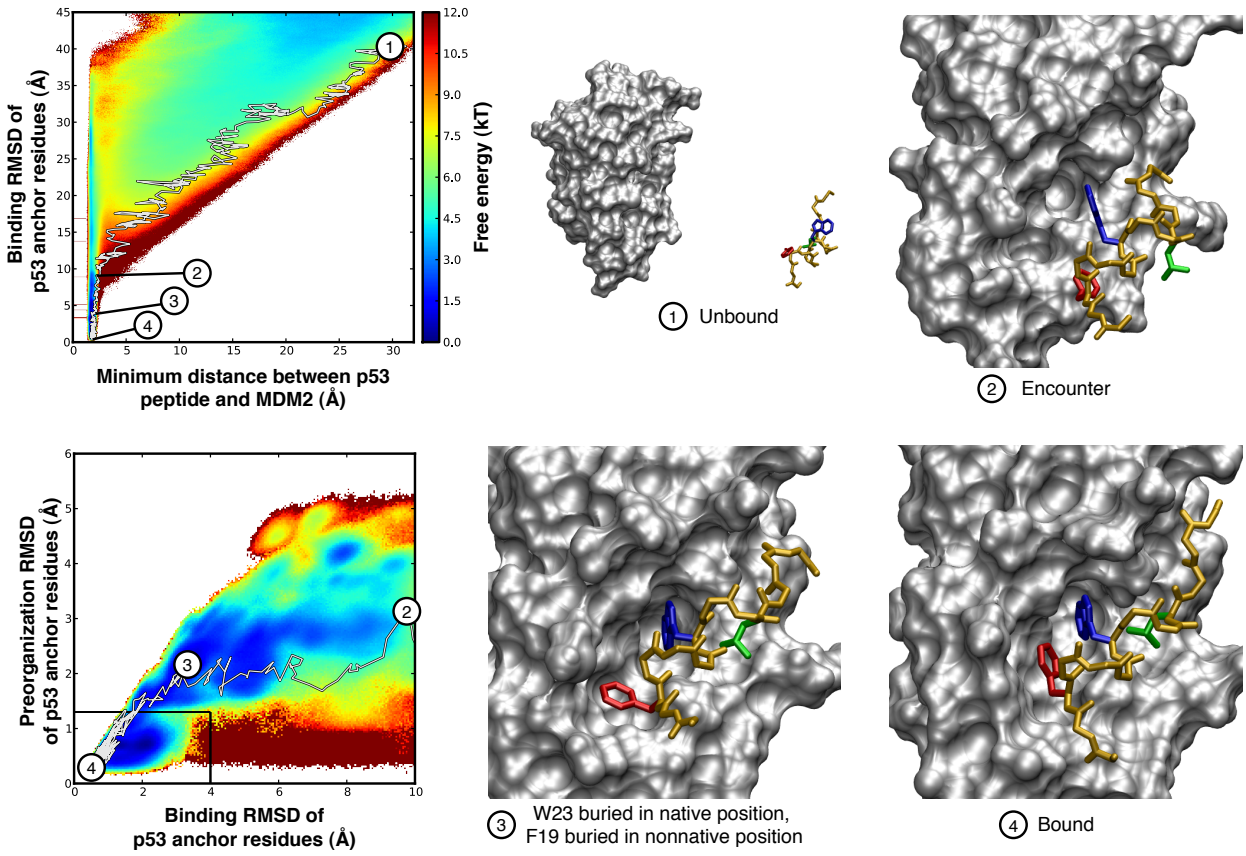


Figure 4.3: Continuous p53-MDM2 binding pathway obtained from WE simulation, superimposed on the free energy landscapes of binding: binding RMSD vs. minimum p53-MDM2 separation (upper left) and preorganization RMSD of the p53 peptide vs. binding RMSD (lower left). The preorganization RMSD reflects the similarity of the p53 peptide to its conformation in the minimized crystal structure, and was calculated as the heavy-atom RMSD of the three key hydrophobic residues (F19, W23, and L26) of p53 after alignment on all heavy atoms of p53. At the beginning of the WE simulation **(1)**, the p53 peptide (gold) is separated from MDM2 (silver) by 30 Å. After diffusing to an encounter complex **(2)**, p53 residue W23 (blue) buries in its native position **(3)** while F19 (red) is buried in a non-native position. The p53 peptide then rearranges to the bound state **(4)**, with F19 (red), W23 (blue), and L26 (green) buried in their native, bound-state positions. This trajectory is also illustrated in Movie S1.

Table 4.1: Rate constants and 95% confidence intervals for p53-MDM2 association.

Initial state	Final state	Rate	Lower bound	Upper bound	Unit
Unbound	Bound	7.1	3.5	11.5	$\times 10^7$ $M^{-1}s^{-1}$
Unbound	Encounter	1.1	0.9	1.3	$\times 10^8$ $M^{-1}s^{-1}$
Encounter	Bound	3.8	2.0	5.9	$\times 10^5$ s^{-1}
Bound	Encounter	8×10^4	0	2×10^7	s^{-1}
Encounter	Unbound	8×10^3	0	2×10^6	s^{-1}

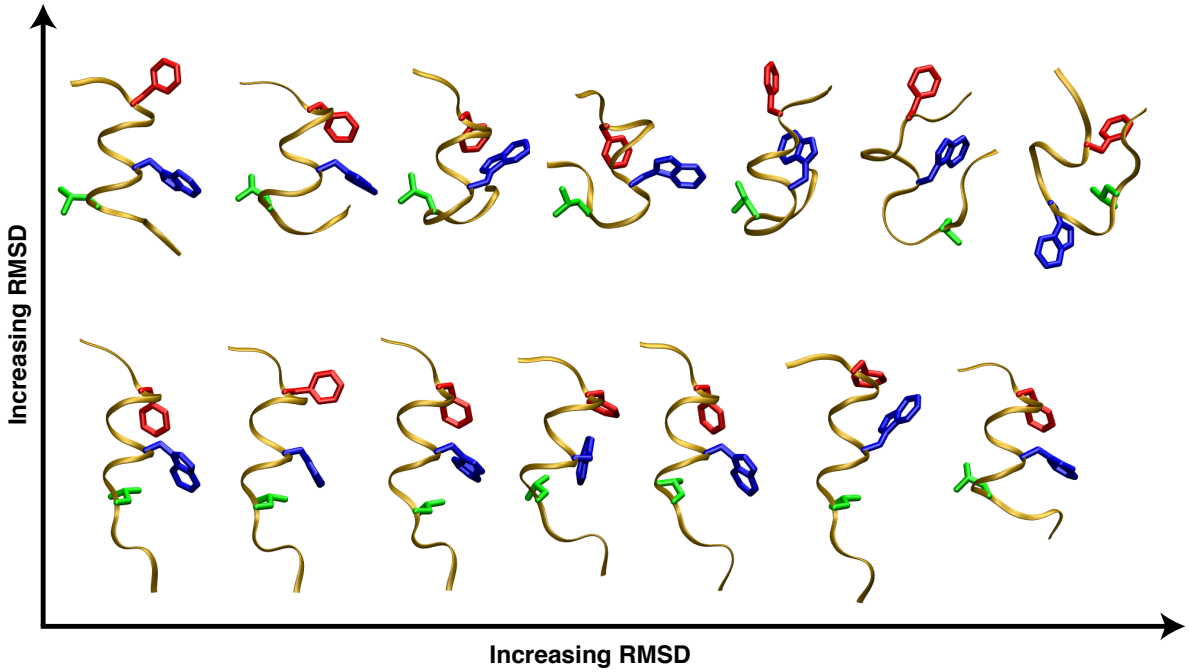


Figure 4.4: Diversity of p53 initial conformations. The heavy atom RMSD of p53 (after alignment on the heavy atoms of p53) increases both left to right and bottom to top. The minimized crystal structure conformation is at the lower left.

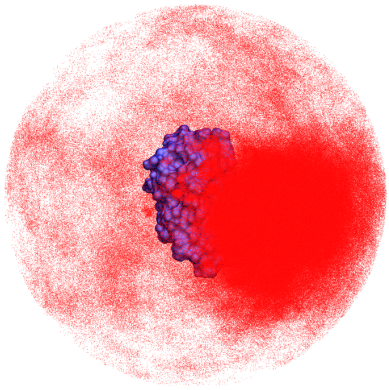


Figure 4.5: Locations visited by p53 center-of-mass (red) relative to MDM2 (blue) over the course of the WE simulation (about 2 million conformations, sampled every 50 ps).

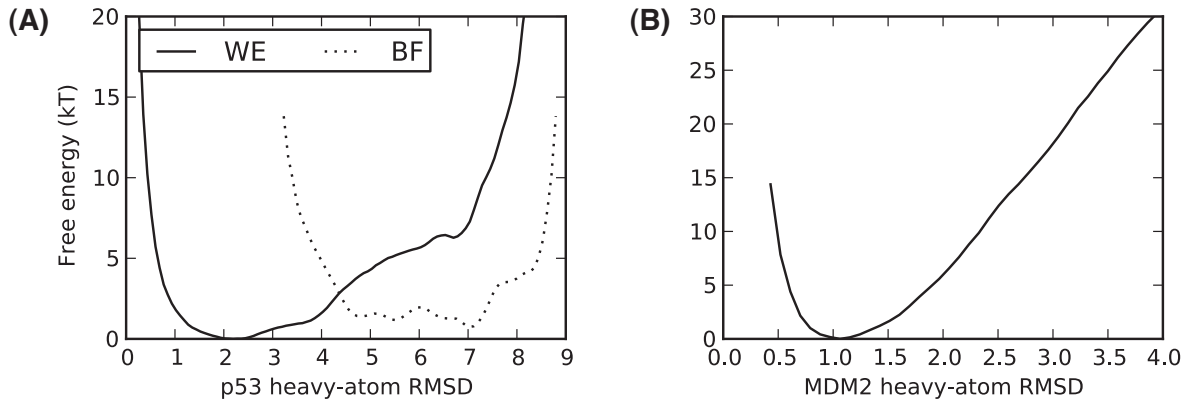


Figure 4.6: Sampling of initial states for unbound (A) p53 and (B) MDM2. For p53 (A), WE (solid line) samples unbound p53 conformations substantially better than 10 μs of brute force sampling (dotted line).

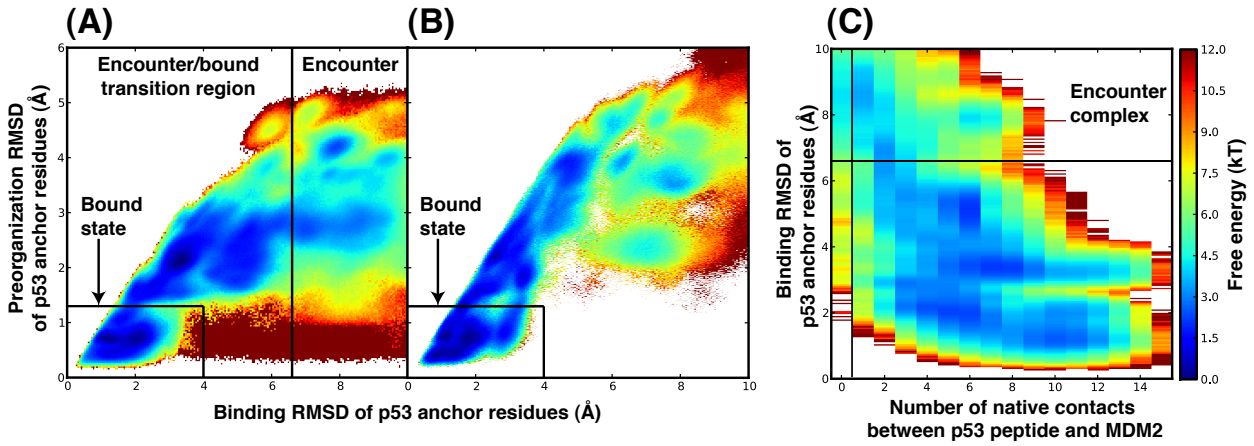


Figure 4.7: State definitions for p53-MDM2 binding as refined from WE simulations. The bound state identified in the binding simulation (A) was confirmed by a separate control simulation which yielded a similar energy landscape (B) as that obtained from the binding simulation. The encounter complex (C) was defined as having at least one native p53-MDM2 residue-residue contact. Binding and preorganization RMSDs were defined as the heavy-atom RMSDs of p53 relative to the minimized crystal structure after alignment on MDM2 or p53, respectively.

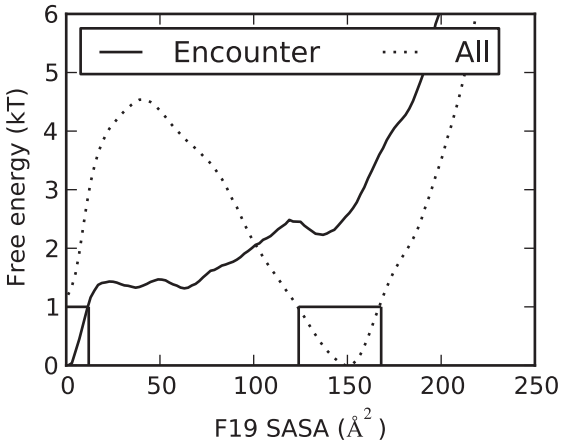


Figure 4.8: Burial of p53 residue F19 in the encounter complex. The buried (SASA < 12 Å^2, boxed at left) and unburied (124 Å^2 < SASA < 168 Å^2, boxed at center) states were defined as the range of SASA values within kT of the minima of the PMFs of F19 burial in the encounter complex and entire ensemble, respectively. Integration of the encounter complex PMF over these regions gives the relative populations of buried and unburied F19.

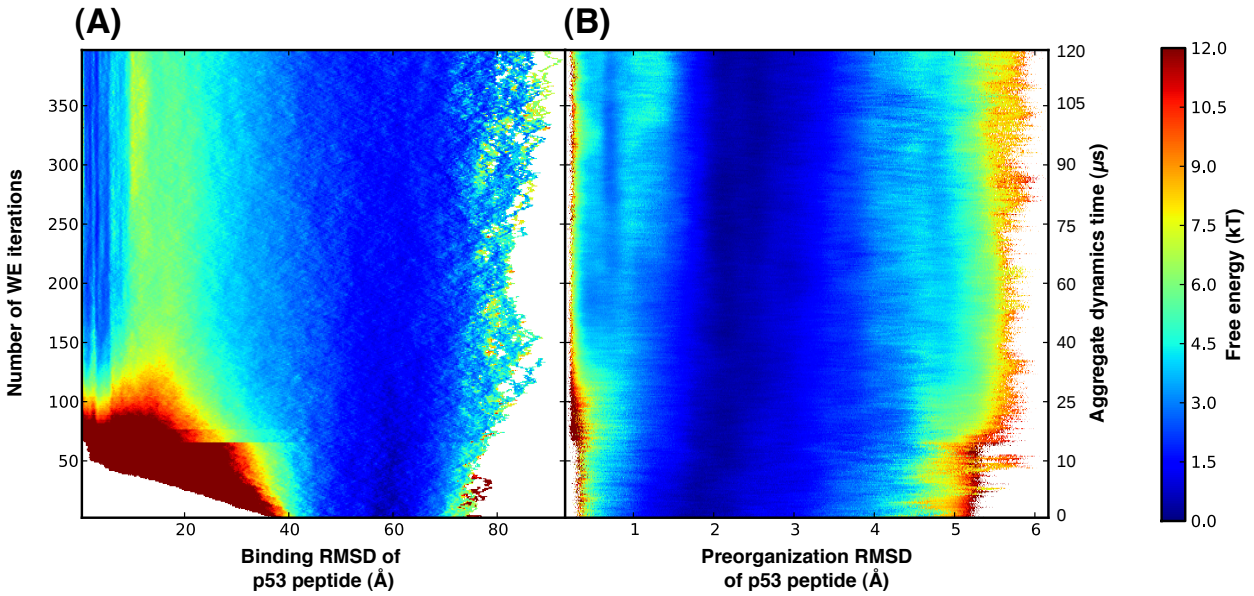


Figure 4.9: Evolution of the probability distribution of progress coordinate values for *(A)* the "binding" RMSD (heavy-atom RMSD of the p53 peptide after alignment on MDM2 heavy atoms) and *(B)* the "pre-organization" RMSD (heavy-atom RMSD of p53 after alignment on itself). Both probability distributions were relatively stable at 200 iterations of WE resampling. For reference, 200 iterations of WE corresponds to about 58 μsof total sampling, and 396 iterations (the end of the WE simulation) corresponds to about 120 μs.

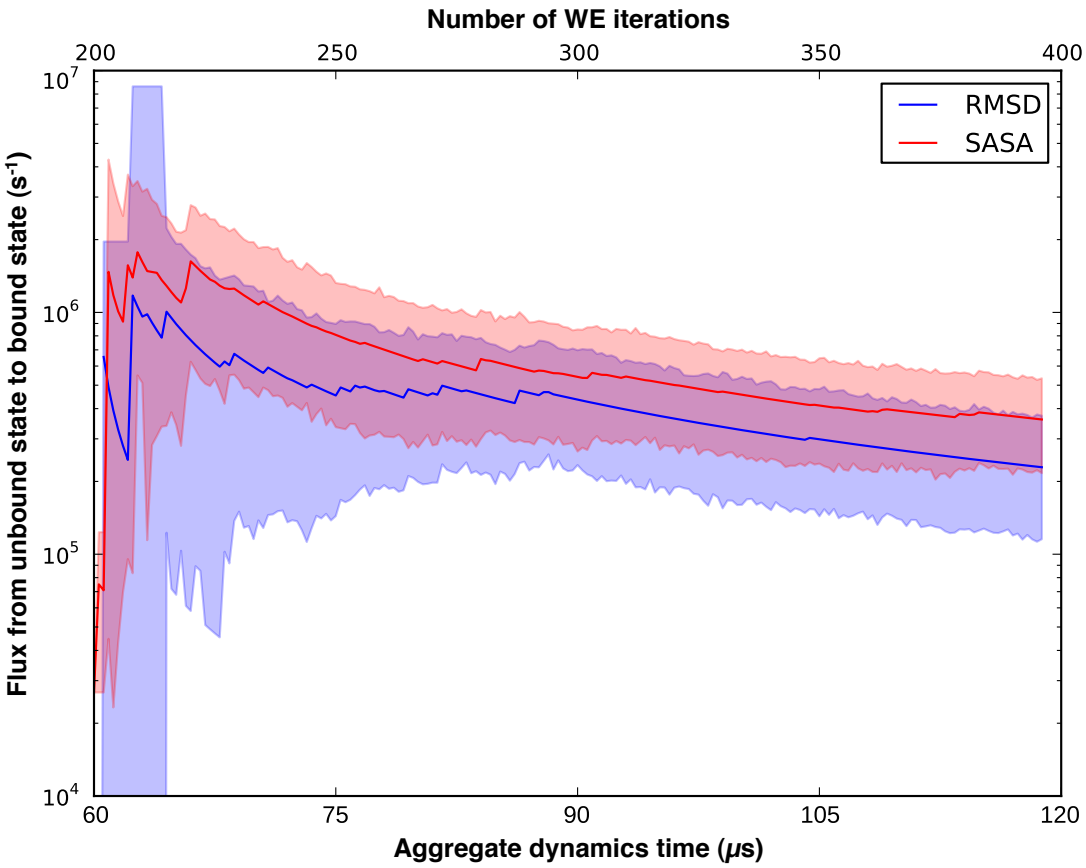


Figure 4.10: Evolution of flux into the bound state, plotted for both the RMSD-based bound state definition described in the main text (blue), as well as for a definition based on the burial of the 3 key hydrophobic residues (red). In the latter case, the bound state was defined as F19 burial > 25 Å^2, W23 burial > 50 Å^2, and L26 burial > 50 Å^2. The shaded regions represent 95% confidence intervals as determined by blocked bootstrapping; overlap indicates the agreement between the binding rates as determined by the two bound state definitions. Though the rate as determined from the RMSD-based bound state definition is slightly decreasing, the rate as determined from the burial-based bond state definition has stabilized, as is evident from the constant lower bound of the 95% confidence band.

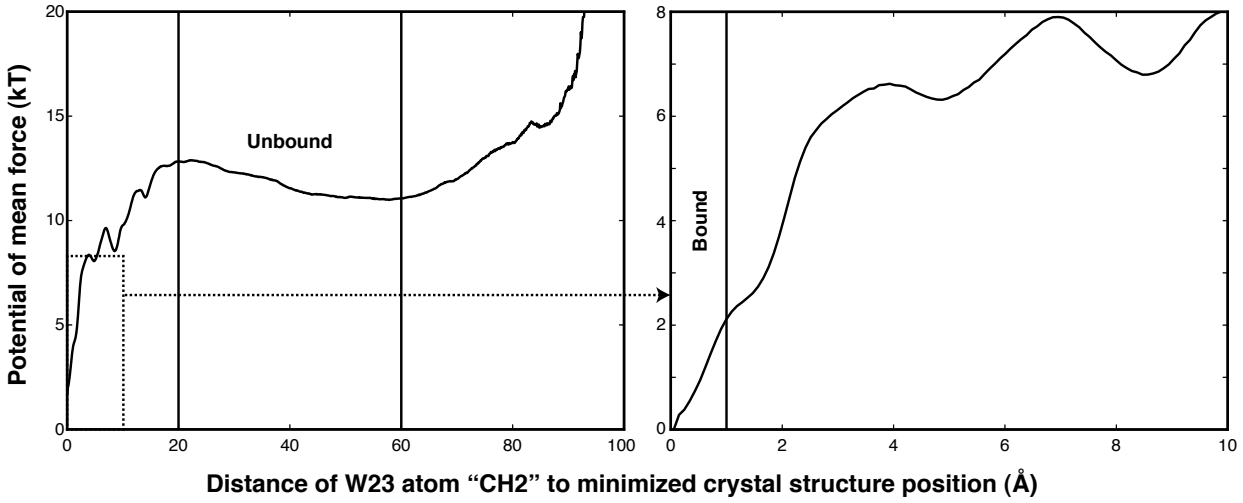


Figure 4.11: Potential of mean force of most buried atom of W23 to its minimized crystal structure position. The vertical lines mark the boundaries used to define the bound and unbound states for the K_D analysis. These boundaries were determined by inspection of the PMFs and chosen to minimize the extent of the bound state and maximize the extent of the unbound state, providing a conservative estimate of K_D. The K_D thus obtained was not particularly sensitive to the choice of these boundaries.

5.0 CONCLUSIONS AND FUTURE DIRECTIONS

As discussed in Chapter 1, there remains a need for enhanced sampling approaches to reach long biologically interesting timescales with molecular dynamics simulations. Despite the steady increase of computer power, the advent of GPU-augmented computing, and even the creation of a supercomputer hand-tuned for MD (Anton[211]), interesting biological processes beyond the millisecond timescale are not accessible with molecular dynamics simulations (*cf.* Figure 1.1). The longest, most detailed brute force simulations of microsecond-scale protein folding events performed to date have obtained only about ten folding events for each of the twelve small, fast-folding proteins studied.[212] Similarly, the longest, most detailed brute force simulation of protein-small molecule associations obtained less than five binding events for each of two drug-protein systems.[162]

We showed in Chapter 2 that weighted ensemble molecular dynamics simulations can increase the efficiency of sampling rare molecular association events by orders of magnitude. The degree to which weighted ensemble simulations are accelerated relative to brute force depends on the heights of barriers. Low-barrier systems see little acceleration relative to brute force simulation, but higher-barrier systems see substantial improvement (two to three orders of magnitude, as a lower bound). Weighted ensemble simulations reproduced brute force results in all cases, providing explicit confirmation that the formal exactness of WE[78] translates effectively to actual simulations.

Encouraged by these results, in Chapter 3 we extended our use of weighted ensemble sampling to a more complicated system: the association of a 13-residue fragment of p53 to MDM2 using a coarse-grained model. The association rate for p53-MDM2 binding agreed very well with the experimental value, indicating the utility of relatively simple structure-based potentials in examining the kinetics of protein-peptide associations. We directly tested the "fly-casting" mechanism by simulating association between MDM2 and both highly flexible and rigidly preorganized p53. The "fly-casting" mechanism hypothesizes that intrinsically disordered binding partners have a kinetic advantage owing to larger capture radii and subsequent coupled folding and binding.[132] We observed no significant difference in associ-

ation rates between MDM2 and preorganized or unstructured p53, demonstrating that the fly-casting mechanism is not a significant effect in binding between MDM2 and the p53 peptide considered here. Hydrodynamic interactions increase the computed association rates by a factor of ~4 – 5, due to the increased remarkably, mostly due to changing the translational diffusion coefficients that dictate the rate at which diffusional encounters between p53 and MDM2 occur.

All-atom simulations of p53-MDM2 binding were considered in Chapter 4. Using all-atom molecular dynamics simulations in GB/SA implicit solvent, we obtained ~300 binding events and the association rate and free energy landscape of p53-MDM2 binding. Our definition of the bound state was obtained from the simulation itself, rather than imposed from the beginning of the simulation. The association rate we obtain agrees with the experimental value when considering that we omitted the flexible "lid" that controls access to the MDM2 binding site. The free energy landscape of binding is "funnel-like"; once p53 encounters MDM2, it is largely a downhill transition into the bound state. These results would have been impossible to obtain with brute-force sampling, demonstrating that the WE approach does indeed extend computer simulations of biologically interesting processes to otherwise inaccessible timescales.

We obtain good agreement between our simulations of p53-MDM2 binding (k_{on} ranging from $\sim 4 \times 10^7$ $\text{M}^{-1}\,\text{s}^{-1}$ to $\sim 4 \times 10^8$ $\text{M}^{-1}\,\text{s}^{-1}$) and the experimental association rate ($\sim 1 \times 10^7$ $\text{M}^{-1}\,\text{s}^{-1}$). This agreement is strengthened when considering that the models considered here lack the flexible "lid" over the MDM2 binding cleft present in the experimental study, and we consequently may expect to obtain somewhat higher rates (up to ten-fold higher; see Chapter 4) by simulation than obtained in experiment. Further, the highest association rates were obtained with the Gō-like potential of Chapter 3 (particularly when including hydrodynamic interactions) which is not surprising given that Gō-type potentials are known to accelerate protein dynamics due to their neglect of non-native interactions.[213–215] Taken together, these results indicate that the models we use are capable of modeling the kinetics of biomolecular associations accurately. These results are likely to be applicable to the entire class of systems involving the burial of an intrinsically-disordered protein in a hydrophobic pocket.

In developing these simulations, we have extended WE sampling to equilibrium conditions, which opens up the possibility of determining (for example) protein-protein binding and unbinding rates, pathways, and equilibrium constants simultaneously. Further, this lifts the original requirement for WE sampling that the "target" state (B for $A \rightarrow B$ transitions) be known, and known well, in advance. We have also directly obtained an equilibrium constant for a binding process, which is particularly exciting, as as equilibrium constants are key values in describing molecular association in chemical, biological,

and pharmacological contexts. However, in these simulations, we have obtained no continuous bound-to-unbound trajectories in our atomistic p53-MDM2 simulation, which points to the necessity of further developing equilibrium WE simulations so that equilibrium WE simulations reach their potential for providing not only kinetics, energy landscapes, and equilibrium constants, but also the continuous pathways that make WE sampling unique among enhanced sampling approaches.

Perhaps the greatest difficulty in performing WE simulations is selecting an effective progress coordinate. There are preliminary indications that frequent resampling (short τ) WE simulations can *decrease* sampling in degrees of freedom not spanned by the progress coordinate, relative to brute force trajectories of length similar to the maximum trajectory length of a WE simulation.[a] This is perhaps not surprising, as accelerating sampling along a progress coordinate necessarily leaves other degrees of freedom to be sampled without acceleration — that is, essentially by brute force. Each WE dynamics segment then amounts to a τ-length brute force simulation of degrees of freedom not explicitly considered by progress coordinates. These degrees of freedom may prove unexpectedly critical in determining the kinetics of rare events.[216] Therefore, the progress coordinate that most effectively drives transitions from some state A to some state B may not be the most effective in determining the macroscopic rate k_{AB} of $A \rightarrow B$ transitions. This issue is not unique to WE simulations, but rather plagues any enhanced sampling technique that relies on a progress coordinate, including other path sampling approaches and Markov state models. This being the case, in practice we obtain encouraging agreement between simulations and experiment and between between simulations with different models. The coarse-grained models of Chapter 3, the atomistic model of Chapter 4, and the experimental association rate of p53-MDM2 are all in reasonable agreement. Further, alternative bound state definitions (RMSD and solvent accessible surface area) for the atomistic simulation of p53-MDM2 binding give the same association rate. This indicates that though the imposition of a progress coordinate *may* subtly complicate sampling of rare processes, it does not *necessarily* do so.

After finding an effective progress coordinate, appropriate values for the propagation/resampling time τ and locations of bin boundaries must be chosen. In all the studies contained herein, we have estimated τ and bin spacing using models for the relative diffusion of the two binding parters. Beyond this estimation, progress toward binding was ensured simply by doubling the density of bin boundaries in regions of progress coordinate space where the WE simulation appeared to stall. (A detailed description of this procedure and the reasoning behind it is presented in Section A.1.) There has been little detailed consideration of the effects of different progress coordinates, τ, or bin boundary locations on the effi-

[a] Daniel M. Zuckerman, personal communication

ciency of WE simulations. There is a distinct need for this information, as it will allow construction of WE simulations that use computer resources even more efficiently than they already do.

Finally, we have demonstrated that WE sampling not only effectively samples rare events like protein-peptide associations, but also that these simulations are practical on current computing resources. In order to perform these simulations, we developed our own WE software (WESTPA – the Weighted Ensemble Sampling Toolkit with facilities for Parallelization and Analysis), described in Section A.4. As has been demonstrated by its use in the studies discussed herein, our software scales from the desktop through many thousands of cores, efficiently exploiting the loosely-coupled nature of WE simulation trajectories. WESTPA can interface with any stochastic simulation software and places particular emphasis on making the massive amounts of data generated in WE simulations easily available for analysis. To our knowledge, no other enhanced sampling method has such extensive software support. We have released this software under an open-source license to enable others to use WE sampling for simulation of rare events. This software will be described in detail in a forthcoming paper.

Overall, our results demonstrate that the WE approach can simultaneously and efficiently generate both free energy landscapes and rigorous rate constants for rare chemical events, particularly the p53-MDM2 binding process. These simulations would not have been possible with brute force simulations, and point encouragingly to the utility of WE sampling in accessing otherwise inaccessible timescales. The efficiency and flexibility of WE sampling, combined with the availability of user-friendly software, makes the weighted ensemble approach likely to become a mainstay of computer simulation of rare events in the coming years.

APPENDIX

CONSTRUCTING AND RUNNING WEIGHTED ENSEMBLE SIMULATIONS

A.1 ESTIMATING WE SIMULATION PARAMETERS

The efficiency of weighted ensemble sampling depends on achieving a balance between good statistics (maximizing the number of bin-to-bin and state-to-state transitions) and minimizing the total dynamics time of the simulation. As is usually the case, these aims are somewhat at odds. The number of bin-to-bin transitions can be maximized by increasing the number of walkers per bin, decreasing the size of bins (and thus increasing the number of bins), or increasing the propagation/resampling period τ. The total dynamics time can be minimized by reducing the number of total walkers in use, thus requiring reducing the number of walkers per bin and/or the number of bins, or shortening τ. Thus, the size of bins, the number of walkers per bin, and τ are mutually dependent, and all three affect the cost of the WE simulation and the level of sampling attained. The following procedure has proven useful in both coarse-grained and atomistic WE simulations, leading to reasonable sampling in reasonable computer time. This is not proposed as an *optimal* procedure, but rather a *sufficient* one; a detailed study on the effects of these simulation parameters on non-trivial systems is likely necessary to achieve the most efficient possible sampling.

The procedure outlined below rests on a few observations:

1. In strongly-interacting regions of configuration space, the evolution of dynamics will be dominated by (free) energy barriers.
2. In non-interacting regions of configuration space, or regions of the energy landscape that are in some sense broad and flat, the evolution of dynamics will be dominated by diffusive mechanisms.

The goal is to place bin boundaries so that the probability of progressing from one bin to the next is locally constant. That is, the probability of traversing some number of bins M in N_τ iterations (corresponding to length $t = N_\tau \tau$ trajectory fragments) is linear in N_τ. This "linearizes" both diffusion (since the probability of traversing a distance R will increase linearly with time t rather than with $\sqrt{t}$) and barrier climbing (since the probability of climbing a barrier of height E is linear with time t rather than with $\exp[-E/kT]$). Incidentally, such a binning scheme will ensure a quadratic speedup of WE over brute force in diffusive regimes and an exponential speedup of WE over brute force in regimes dominated by high barriers.

On the surface it appears that item (1) above requires intimate knowledge of the potential of mean force of the system to be studied. In practice, it does not. Because we do not know the energy landscape in advance, we take advantage of the flexibility of the WE approach and simply double the density of bins when we notice that transitions from one bin to another in that region of configuration space are too rare, and repeat the process until transitions are sufficiently probable. Such an adjustment requires pausing a WE simulation, but not restarting it from the beginning; rather, the bin boundaries can be adjusted and the simulation continued. (This was the spirit of the on-the-fly binning approach presented by Huber and Kim in the paper introducing WE.[75]) Increasing bin density is superior to increasing the number of walkers or the length τ of propagation, as increasing bin density will bring sampling closer to local linearity, but increasing the number of walkers or increasing τ amounts to imposing a greater degree of sampling locally by brute force.

Similarly, item (2) requires knowledge of the diffusive characteristics of the system under study. This is generally not a problem, as the diffusive behavior of non-interacting species is well-understood, or at least can be modeled sufficiently well for our purposes. In flat but interacting regions of configuration space (say, flat but rough regions near protein binding or folding basins), we fall back to the procedure of doubling bin density until transitions are not too rare.

Of course, to do this, we need a working definition of how many bin-to-bin transitions are reasonable. We will thus require *a priori* the probability α that a walker traverses a bin in a single iteration of length τ. This can be accomplished by placing $N_B \approx \alpha^{-1}$ walkers in a bin and then tuning the size of the bin appropriately, which amounts to asserting that one walker out of N_B, on average, will traverse a bin in one iteration of WE propagation and resampling. One can also view α as the fraction of walkers that traverse a bin in one iteration. It is important that walkers *traverse* bins, rather than merely *exit* bins (possibly through the same boundary they entered from), as this helps to ensure that the energy landscape within a bin is approximately flat.

Given the fraction α of walkers that we desire to traverse a bin, we may use diffusion arguments to determine proper values for the propagation/resampling period τ and the width of bins R. This will require a diffusion coefficient D for motion along a certain progress coordinate (or a single dimension of a multi-dimensional progress coordinate). For the binding simulations considered in this work, the progress coordinate was either separation or an RMSD metric that behaves as separation at long distances, and thus the diffusion coefficient involved was that describing relative translational diffusion of the two binding partners. (See Section A.1.2 for a derivation of the relative diffusion coefficient $D = D_1 + D_2$ of two partners in terms of their individual translational diffusion coefficients D_1 and D_2.)

A.1.1 The Relationships Among Diffusion Coefficients, Propagation Time, Bin Width, and Replicas Per Bin

Given an appropriate diffusion coefficient D, we seek values of the propagation/resampling period τ and the width R of bins so that on average a fraction α of walkers in a bin traverse a bin within τ. Consider, for simplicity, the one-dimensional case, where diffusion is governed by

$$\frac{\partial P}{\partial t} = D\frac{\partial^2 P}{\partial r^2}, \tag{A.1.1}$$

whose fundamental solution is

$$P(r,\tau;D) = (4\pi D\tau)^{-1/2}\exp\left(-r^2/4D\tau\right) \tag{A.1.2}$$

The probability $\alpha = 1/N_B$ that one of N_B replicas will move at least a distance R along the relative coordinate r in a time τ is

$$\alpha = P(r > R,\tau;D) = \int_R^\infty (4\pi D\tau)^{-1/2}\exp\left(-\frac{r^2}{4D\tau}\right)dr;$$

note that here R represents the distance traveled (*i.e.* the bin width), not the coordinate of the collective motion of the two particles. Letting $\sigma = \sqrt{4D\tau}$, $x = r/\sigma$, $dx = dr/\sigma$, and $X = R/\sigma$, we have

$$\begin{aligned}
\alpha &= \int_X^\infty \frac{1}{\sqrt{\pi}}\exp\left(-x^2\right)dx \\
&= \frac{1}{2}\int_X^\infty \frac{2}{\sqrt{\pi}}\exp\left(-x^2\right)dx \\
&= \frac{1-\mathrm{erf}(X)}{2} \\
&= \frac{1-\mathrm{erf}(R/\sigma)}{2} \\
&= \frac{1-\mathrm{erf}\left(R/\sqrt{4D\tau}\right)}{2}.
\end{aligned} \tag{A.1.3}$$

Equation A.1.3 may then immediately be solved, giving

$$\tau = \frac{1}{4D}\left(\frac{R}{\text{erf}^{-1}(1-2\alpha)}\right)^2 \tag{A.1.4}$$

where $\text{erf}^{-1}(x)$ is the inverse error function satisfying $\text{erf}^{-1}(\text{erf}(x)) = x$. Equation A.1.4 states that for two particles with relative diffusion coefficient D, there is a probability α that they will move at least an amount R in relative displacement in a time τ. A typical use of Eq. A.1.4 would be to set $R = 1$ Å (representing bin boundaries placed every 1 Å) and $\alpha = 1/N_B$, and thereby obtain the optimal value of the propagation period τ to ensure that (on average) one out of N_B replicas per bin will advance 1 Å to the next bin in a single weighted ensemble iteration.

Equation A.1.4 provides a convenient, largely conservative estimate[a] for τ given a bin width R and probability α, but is not exact for systems diffusing in three dimensions; rather, it overestimates τ by about 50% at the $\alpha = 0.04$ level. The reduction to a one-dimensional problem amounts to assuming that relative motion is always either perfectly aligned or opposed, which is not true of three-dimensional isotropic diffusion. Extending this result to three dimensions gives

$$\alpha = \int_R^\infty \int_0^\pi \int_0^\pi \frac{1}{\sigma^3\pi^{3/2}} \exp\left(-\frac{r^2}{\sigma^2}\right) r^2 \sin\theta \, dr \, d\theta \, d\phi \tag{A.1.5}$$

$$\begin{aligned} &= \int_R^\infty \frac{1}{\sigma\sqrt{\pi}} \left(\frac{r}{\sigma}\right)^2 \exp\left(-\frac{r^2}{\sigma^2}\right) \\ &= \int_X^\infty \frac{1}{\sqrt{\pi}} x^2 \exp\left(-x^2\right) dx \\ &= \frac{1-\text{erf}(X)}{4} + \frac{X}{2\sqrt{\pi}} \exp\left(-X^2\right) \end{aligned} \tag{A.1.6}$$

where the integration in Eq. A.1.5 is over a hemisphere to represent that we wish to consider one direction of change in relative separation (*i.e.* forward progress) only, and $X = R/\sigma = R/\sqrt{4D\tau}$ as before. Equation A.1.6 is easily solved for τ with a numerical root finding algorithm.

A.1.2 The Relative Diffusion of Two Particles

For a binding simulation, we are interested in the diffusion constant D that defines the *relative* motion of the binding partners at long distances. The diffusion of two particles of diffusion coefficients D_1 and D_2 relative to each other obey a diffusion equation with effective diffusion coefficient $D = D_1 + D_2$. To see this, consider one-dimensional diffusion, which obeys the familiar relation

$$\frac{\partial P}{\partial t} = D\frac{\partial^2 P}{\partial x^2}, \tag{A.1.7}$$

[a] Specifically, this estimate is conservative in predicting τ for $R/\sigma \approx 0.6$ or higher.

where P is probability (*i.e.* a suitably normalized concentration or number density), D is the diffusion constant, and $x(t)$ is displacement at time t relative to time $t = 0$ and position $x = 0$. For two non-interacting particles with diffusion coefficients D_1 and D_2 and displacements x_1 and x_2 relative to their positions at time $t = 0$, this becomes

$$\frac{\partial P}{\partial t} = D_1 \frac{\partial^2 P}{\partial x_1^2} + D_2 \frac{\partial^2 P}{\partial x_2^2}. \tag{A.1.8}$$

Let the relative displacement $r = x_1 - x_2$, and also let $R = (D_2 x_1 + D_1 x_2)/(D_1 + D_2)$. Using the chain rule to rewrite $\partial P/\partial t$ in terms of $\partial^2 P/\partial r^2$ and $\partial^2 P/\partial R^2$, we have

$$\frac{\partial P}{\partial t} = \frac{D_1 D_2}{D_1 + D_2} \frac{\partial^2 P}{\partial R^2} + (D_1 + D_2) \frac{\partial^2 P}{\partial r^2}, \tag{A.1.9}$$

which indicates that diffusion in the relative coordinate r obeys a diffusion equation with coefficient $D_1 + D_2$ and the combined coordinate R (the "center of diffusion" in a sense analogous to center of mass) obeys a diffusion equation with coefficient $(D_1 D_2)/(D_1 + D_2)$ (the "reduced diffusion coefficient," in analogy to reduced mass).

The three-dimensional isotropic case is completely analogous, after replacing displacements with displacement vectors and each second spatial derivative $\partial^2 P/\partial q^2$ with the operator ∇_q^2 (for each coordinate $q \in \{x_1, x_2, r, R\}$), where ∇_q is the gradient along the vector $\vec{q}$. The diffusion equation in absolute coordinates $\vec{x}_1$ and $\vec{x}_2$ is

$$\frac{\partial P}{\partial t} = D_1 \nabla_{x_1}^2 P + D_2 \nabla_{x_2}^2 P \tag{A.1.10}$$

and in "relative" coordinates $\vec{r}$ and $\vec{R}$ is

$$\frac{\partial P}{\partial t} = \frac{D_1 D_2}{D_1 + D_2} \nabla_R^2 P + (D_1 + D_2) \nabla_r^2 P \tag{A.1.11}$$

The relative diffusion coefficient $D_1 + D_2$ may be obtained by sufficiently long molecular dynamics simulations of the individual particles; by consideration of the interaction of solvent with the (rigid) crystal structure of each particle (as in HYDROPRO[158]); or simply by assuming each particle to be approximately spherical, with radius equal to the radius of gyration R_g. In this lattermost case, the Stokes and Einstein relations lead immediately to

$$D = \frac{k_B T}{6\pi\eta R_g}, \tag{A.1.12}$$

where η is the solvent viscosity, T is absolute temperature, and k_B is Boltzmann's constant. Performing this calculation for two species of radii $R_{g(1)}$ and $R_{g(2)}$, we have the following for the relative diffusion coefficient D:

$$D = D_1 + D_2 = \frac{k_B T}{6\pi\eta} \left(\frac{1}{R_{g(1)}} + \frac{1}{R_{g(2)}} \right) \tag{A.1.13}$$

This spherical approximation has been surprisingly effective in practice.

A.1.3 Summary

In summary, to construct a bin space for a weighted ensemble simulation:

1. Choose a value α which defines the fraction of walkers in a bin which traverse a bin along one progress coordinate direction in one iteration. A reasonable balance of sampling and computational cost seems to be $\alpha \approx 0.1$, or about ten walkers per bin.
2. For large separations (or small interactions), choose a bin spacing R and a propagation resampling period τ so that a fraction α of walkers in a bin traverse that bin in τ. This can be accomplished using the diffusion arguments presented above.
3. Run the WE simulation, and check that about α^{-1} walkers progress from bin to bin each iteration. This is trivial for empty bins, as this amounts to one more bin being populated in each dimension with each new iteration.
4. If progress stalls, double the number of bins (halve the bin spacing) in the region where the simulation has stalled.
5. Continue the WE simulation, repeating the subdivision of bins as necessary. Failure to progress in the face of repeated increases in bin density may indicate a poor progress coordinate.

A.2 CALCULATION OF LANGEVIN THERMOSTAT COLLISION FREQUENCY FOR IMPLICIT SOLVENT SIMULATIONS

Implicit solvent simulations using a Langevin thermostat are propagated using the Langevin equation, where the acceleration of a given particle at a given timestep is

$$\frac{d^2\vec{r}}{dt^2} = \frac{1}{m}\vec{F} - \gamma\frac{d\vec{r}}{dt} + \frac{1}{m}\delta\vec{F} \tag{A.2.1}$$

where $\vec{r}$ is the position of the particle, m is its mass, $\vec{F} = -\nabla U$ is the force acting on the particle, $\delta\vec{F}$ is a stochastic force, and γ is a damping constant.[217] The damping constant γ represents drag on the particle due to the surrounding fluid. Invoking the imagery of frequent collisions with the surrounding medium imparting drag, γ (with units of [1/time]) is sometimes called the "collision frequency". Physically, the damping constant γ determines the terminal drift velocity of the particle under the influence of an applied force, as can be seen by setting $d^2\vec{r}/dt^2 = \delta\vec{F} = 0$ in Equation A.2.1 and noting

$$\vec{F} = m\gamma\frac{d\vec{r}}{dt} \tag{A.2.2}$$

or, in magnitude,

$$\frac{F}{v} = m\gamma = \zeta \tag{A.2.3}$$

where $v = |d\vec{r}/dt|$ is the velocity and ζ is called the viscous friction coefficient. The viscous friction coefficient ζ and the diffusion coefficient D are linked by the Einstein relation

$$\zeta D = kT \tag{A.2.4}$$

Substituting $\zeta = m\gamma$ into Equation A.2.4 allows one to obtain γ in terms of the diffusion coefficient D (which in general is temperature dependent):

$$\gamma = \frac{kT}{mD} \tag{A.2.5}$$

The self-diffusion coefficient of water at 298 K is $D = 2.299 \times 10^{-9}$ m^2/s (see Ref. 207), giving $\gamma = 5.98 \times 10^{13}$ s^{-1}, or 59.8 ps^{-1}. The substantial variation in γ seen in the wild for implicit solvent simulations likely derives from the estimation of D (or ζ) using various methods like Stokes' Law ($\zeta = 6\pi\eta R$ for a sphere of radius R in a medium of viscosity η) rather than experimentally-determined D or ζ values.

A.3 REWEIGHTING IN WEIGHTED ENSEMBLE SIMULATIONS

Though weighted ensemble simulations can be highly effective in accelerating sampling of kinetics and thermodynamics of rare events, the "equilibration" period necessary for the initial state of the WE simulation to relax to a state representative of the energy landscape can be quite long, particularly when long-lived metastable intermediates are present. For example, in a binding simulation, the initial state of a WE simulation might be such that all walkers are in the unbound state (however diverse the configurations of that state may be), which is not at all representative of the binding landscape as a whole; it then takes time for the entire landscape between the unbound state and the bound state to become populated, and for the populations of various regions to relax to their steady-state (or equilibrium) values. One can expect, however, that the rates for probability flows in and out of bins are determined almost as soon as the bins are occupied (assuming that the bins are sufficiently small). Knowing this, it becomes possible to use *local* bin-to-bin rate information to extrapolate the *global* population distribution.[68,156]

We begin with the mathematical expression of a steady-state population distribution:

$$0 = \frac{dP_i}{dt} = \sum_j k_{ji} P_j - \sum_j k_{ij} P_i \tag{A.3.1}$$

which expresses that bin i is at a steady state (the rate of change in population in bin i is zero), and that steady state must be determined by the total probability flow into the bin from all other bins ($\sum_j k_{ji} P_j$) and the total probability flow out of the bin to all other bins ($\sum_j k_{ij} P_i$). (Here, k_{ij} represents the flow from bin i to bin j.) This forms a system of algebraic equations in the populations P_i of bins and the rate matrix k_{ij}.

Given an estimate of the rate matrix k_{ij}, solving this system of equations for the populations P_i of bins gives the steady-state populations for all bins. (In practice, the system of equations A.3.1 is solved in the least-squares sense, as noise in both the rate matrix k_{ij} and populations P_i can interfere with an exact solution.) One can then uniformly scale the probabilities of walkers in bin i so that the total weight in bin i equals the P_i obtained from the steady state equations. That is, if a bin i contains weight P_i^*, and solving the steady-state equations indicates that bin i should contain probability P_i, one multiplies the weight of each walker in the bin by the same factor P_i / P_i^*. Equation A.3.1, and therefore this procedure, applies equally to both the "classic" steady-state formulation of WE and equilibrium WE (without sources and sinks), as equilibrium is itself a steady state.[209]

This procedure is exact in the limit of a precisely-known rate matrix k_{ij} with an infinite number of bins. These conditions are never actually met, so in practice this procedure provides an estimate of the steady-state probability distribution. If bins are constructed so that the population within a bin equilibrates within a single propagation/resampling period τ (which occurs in the limit of small bins, long τ, and/or the absence of substantial barriers within a bin), then this estimate rapidly converges to the true steady-state population distribution of the system when applied repeatedly over the coarse of a running WE simulation. Because the conditions under which convergence to the correct distribution is guaranteed are impossible to meet exactly, it is wise to use reweighting periodically until the population distribution seems to have converged, then cease this reweighting and observe whether the population distribution relaxes substantially as the simulation progresses. In particular, the presence of substantial barriers within bins can result in consistently incorrect rate matrix elements k_{ij}, which in turn can cause this reweighting procedure to adjust the landscape to a steady but incorrect population distribution.[b] Using reweighting only for a portion of the simulation may allow (and in fact has allowed[c]) for the detection of this condition.

[b] Daniel M. Zuckerman, personal communication.
[c] Joshua L. Adelman, personal communication.

A.4 THE WESTPA SOFTWARE PACKAGE

In order to perform the simulations described in this dissertation, software for performing and analyzing weighted ensemble simulations was necessary, and was therefore constructed over the course of about four years. The resulting software package is called WESTPA (the Weighted Ensemble Sampling Toolkit with facilities for Parallelization and Analysis), and has been released as open-source software for the benefit of the simulation community. WESTPA has the following features:

Accessible, modular and extensible WESTPA is written in the Python programming language[d], allowing WESTPA to leverage the diverse, high performance, and rapidly expanding Python scientific computing ecosystem. Efficiency-critical routines are written in C *via* the excellent Cython extension language for Python.[e] WESTPA currently runs on any Unix-like OS that is natively supported by Python, including Linux and Mac OS X. The design of the WESTPA code is highly modular, allowing relatively easy customization of how simulations are performed (Figure A1). WESTPA provides a number of hooks that allow plug-ins (small pieces of software written by the user) to perform custom processing at several points of the main simulation loop without the need to modify the WESTPA code. For more complex customizations, most components of the WESTPA software can be swapped for custom versions at runtime.

Compatible with existing dynamics software As the weighted ensemble approach is rigorously correct for any stochastic simulation,[78] WESTPA is designed to interface with any existing simulation package. This includes traditional stand-alone molecular dynamics packages like GROMACS,[218] AMBER,[219] or NAMD.[220] For increased flexibility and efficiency, WESTPA is also capable of interfacing with simulation toolkits like OpenMM[29] or fully user-programmed routines via a relatively simple interface; a customized dynamics propagator need only define three relatively simple Python functions ("propagate dynamics", "generate initial state for a new trajectory", and "get progress coordinate"). Notably, because propagation of dynamics is handled by programs or routines external to WESTPA, any optimizations that are already in place for propagation of dynamics (such as use of optimized linear algebra libraries or offloading of computational work to GPUs or other coprocessors) are automatically used by WESTPA simulations. Instructions for interfacing with various software packages are available through the WESTPA software web site.[f]

[d] http://www.python.org
[e] http://www.cython.org
[f] https://chong.chem.pitt.edu/WESTPA/

Flexible sampling The weighted ensemble approach is uniquely flexible among enhanced sampling techniques. It naturally supports multidimensional progress coordinates,[83] may be used to perform simulations under equilibrium conditions or generalized steady states (with sinks and sources for trajectories), and reweighting techniques may be used to accelerate convergence in both thermodynamic and kinetic observables.[68,156] Binning in WESTPA is highly flexible. Progress coordinates may be single- or multidimensional. These progress coordinates may be divided into bins by boundaries on grids, Voronoi cells, or any user-defined function that can map progress coordinate values to integers (bin numbers). The target number of walkers can vary from bin to bin, thus allowing a direct specification of how much computational power to devote to each region of progress coordinate space. Binning, including the ideal number of walkers in each bin, may change at any point in the simulation without needing to discard existing simulation data.

Efficient, flexible data storage. Weighted ensemble simulations can produce prodigious amounts of data. Foremost, different trajectories share history. For maximum efficiency, this shared history is stored only once, along with a directed graph describing the connectivity of trajectory segments. In addition to the (required) progress coordinate, arbitrary datasets associated with each trajectory segment may be calculated and stored during the simulation for efficient access during subsequent analysis. These datasets may be stored at an arbitrarily high time resolution. Trajectories themselves (*e.g.* the time-ordered configurations of molecular dynamics trajectories) are typically stored in the native format of the underlying dynamics engine, which is usually highly optimized for space or speed (for instance, XTC files for GROMACS, NetCDF files for AMBER, or DCD files for NAMD).

WESTPA stores all data in the cross-platform, language-neutral, highly efficient HDF5 file format.[g] The HDF5 library provides optimized storage and retrieval of numerical data. The file format is cross-platform and the library is language-neutral, meaning that the data produced by WESTPA may be accessed from any programming language which has HDF5 bindings (Python,[h] C, C++, Fortran, Java, and MATLAB, among others) on any machine for which the HDF5 library can be compiled (currently, all major personal and scientific computing platforms).

Parallelization and scaling. The weighted ensemble approach is highly parallelizable, as trajectories are very loosely coupled; the propagation of N walkers can always be distributed over up to N cores. As shown in Figure A2, essentially perfect scaling has been achieved over thousands of cores. For large-scale simulations (*e.g.* condensed-phase molecular dynamics simulations), performance is generally

[g] http://www.hdfgroup.org
[h] http://www.h5py.org

not limited by the speed of communication between processors, but rather how quickly external programs (*e.g.* for propagating dynamics or extracting progress coordinates) can be spawned or how quickly analysis tools can read data from disk.

WESTPA includes facilities to distribute tasks such as dynamics propagation or subsequent analysis over multiple cores within a node and multiple nodes within a network. Though designed with weighed ensemble simulations in mind, these facilities are completely general, and may be used for writing parallel programs in Python independent of the rest of the WESTPA framework. This parallelization can use a number of low-level communications methods, including threads, processes, custom TCP/IP communication, or MPI. The custom TCP/IP communication is notable for being operative even where MPI implementations do not allow repeated execution of child processes (as in external dynamics engines), and for being capable of supporting a fan-out communication pattern for reducing network load.

Analysis tools. WESTPA includes a suite of tools to analyze weighted ensemble simulations. Each tool focuses on performing one task (*e.g.* "assign walkers to bins","calculate the probability distribution of progress coordinate values", or "plot the time evolution of a probability distribution"), so that tools can be chained together to perform complicated analysis tasks. Most tools can run on multiple cores to increase analysis throughput. Intermediate data is stored in the flexible and efficient HDF5 file format. Because all data is stored in HDF5 format, analyses not directly addressed by the tools packaged with WESTPA are relatively simple to implement — there is no need to "decode" data in order to process it. That said, common analysis tasks are packaged as a programming framework for re-use in customized analysis scripts.

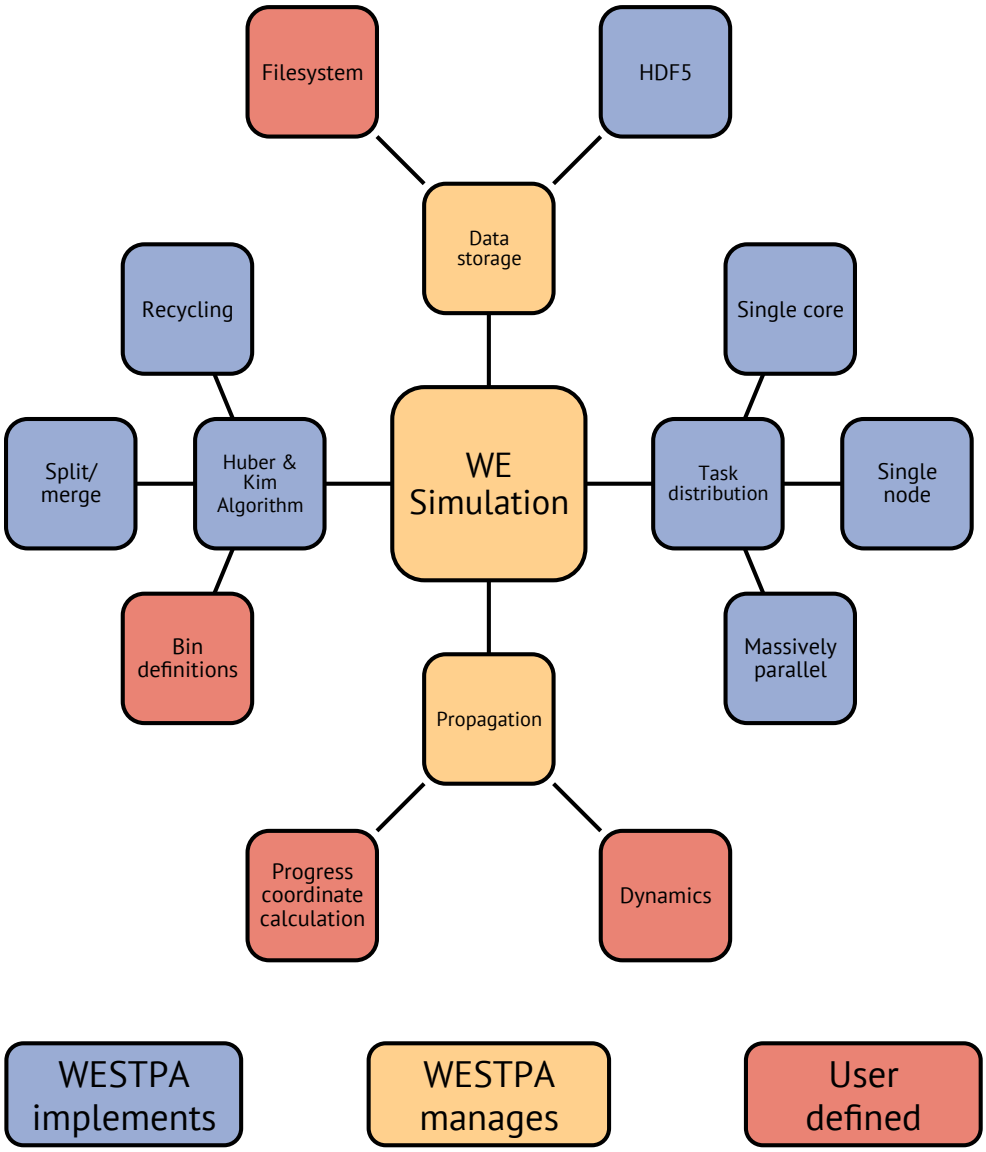


Figure A1: Logical structure of the WESTPA software package. components shown in blue are implemented in WESTPA and may be replaced, if desired, by custom code. Components shown in yellow indicate functionality that WESTPA coordinates, which can be customized by configuration options, plug-ins, or custom code. Components shown in red must be provided by the user. In particular, dynamics propagation and progress coordinate calculation must be provided by an external executable or a custom Python (or Python-accessible routine), and the presence of a filesystem shared among multiple nodes of a cluster is currently presumed.

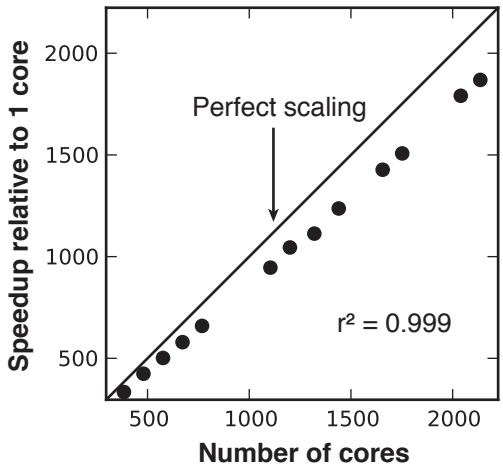


Figure A2: Scaling of WESTPA simulations. WESTPA scales perfectly with fixed overhead to thousands of cores. These data were taken from a production protein-peptide binding simulation.

BIBLIOGRAPHY

[1] Henzler-Wildman, K. A., and Kern, D. (2007) Dynamic personalities of proteins. *Nature 450*, 964–972.

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

[3] Hammes-Schiffer, S., and Benkovic, S. J. (2006) Relating protein motion to catalysis. *Annu Rev Biochem 75*, 519–541.

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

[5] Teague, S. J. (2003) Implications of protein flexibility for drug discovery. *Nat Rev Drug Discov 2*, 527–541.

[6] Lee, G. M., and Craik, C. S. (2009) Trapping moving targets with small molecules. *Science 324*, 213–215.

[7] Fersht, A. R. *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding*; W. H. Freeman and Company, 1999.

[8] Wells, J. A., and McClendon, C. L. (2007) Reaching for high-hanging fruit in drug discovery at protein-protein interfaces. *Nature 450*, 1001–1009.

[9] Hornak, V., and Simmerling, C. (2007) Targeting structural flexibility in HIV-1 protease inhibitor binding. *Drug discovery today 12*, 132–138.

[10] Dyson, H. J., and Wright, P. E. (2005) Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol 6*, 197–208.

[11] Uversky, V. N., Oldfield, C. J., and Dunker, A. K. (2008) Intrinsically disordered proteins in human diseases: introducing the D2 concept. *Annu Rev Biophys 37*, 215–246.

[12] Boehr, D. D., Nussinov, R., and Wright, P. E. (2009) The role of dynamic conformational ensembles in biomolecular recognition. *Nat Chem Biol 5*, 789–796.

[13] Chong, L. T., Snow, C. D., Rhee, Y. M., and Pande, V. S. (2005) Dimerization of the p53 oligomerization domain: identification of a folding nucleus by molecular dynamics simulations. *Journal of Molecular Biology 345*, 869–878.

[14] Elber, R. (2005) Long-timescale simulation methods. *Curr Opin Struct Biol 15*, 151–156.

[15] Buchete, N.-v., and Hummer, G. (2008) Peptide folding kinetics from replica exchange molecular dynamics. *Phys Rev E 77*, 1–4.

[16] Xin, Y., Doshi, U., and Hamelberg, D. (2010) Examining the limits of time reweighting and Kramers' rate theory to obtain correct kinetics from accelerated molecular dynamics. *J Chem Phys 132*, 224101.

[17] MacKerell, A. D. (2004) Empirical force fields for biological macromolecules: overview and issues. *J Comput Chem 25*, 1584–1604.

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

[19] Freddolino, P. L., Park, S., Roux, B., and Schulten, K. (2009) Force field bias in protein folding simulations. *Biophys J 96*, 3772–3780.

[20] Klepeis, J. L., Lindorff-Larsen, K., Dror, R. O., and Shaw, D. E. (2009) Long-timescale molecular dynamics simulations of protein structure and function. *Curr Opin Struct Biol 19*, 120–127.

[21] Pitera, J. W. (2009) Current developments in and importance of high-performance computing in drug discovery. *Curr Opin Drug Discov Devel 12*, 388–396.

[22] Eastman, P., and Pande, V. S. (2010) Efficient nonbonded interactions for molecular dynamics on a graphics processing unit. *J Comput Chem 31*, 1268–1272.

[23] Larsson, P., and Lindahl, E. (2010) A high-performance parallel-generalized born implementation enabled by tabulated interaction rescaling. *J Comput Chem* n/a–n/a.

[24] Harvey, M. J., Giupponi, G., and De Fabritiis, G. (2009) ACEMD: Accelerating Biomolecular Dynamics in the Microsecond Time Scale. *J Chem Theory Comput 5*, 1632–1639.

[25] Phillips, J. C., Stone, J. E., and Schulten, K. Adapting a message-driven parallel application to GPU-accelerated clusters. 2008.

[26] Harvey, M. J., and De Fabritiis, G. (2009) An Implementation of the Smooth Particle Mesh Ewald Method on GPU Hardware. *J Chem Theory Comput 5*, 2371–2377.

[27] Case, D. A. et al. (2010) AMBER 11.

[28] Hess, B., Kutzner, C., van der Spoel, D., and Lindahl, E. (2008) GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J Chem Theory Comput 4*, 435–447.

[29] Eastman, P. et al. (2013) OpenMM 4: A Reusable, Extensible, Hardware Independent Library for High Performance Molecular Simulation. *J Chem Theory Comput 9*, 461–469.

[30] Friedrichs, M. S., Eastman, P., Vaidyanathan, V., Houston, M., Legrand, S., Beberg, A. L., Ensign, D. L., Bruns, C. M., and Pande, V. S. (2009) Accelerating molecular dynamic simulation on graphics processing units. *J Comput Chem 30*, 864–872.

[31] Freddolino, P. L., Liu, F., Gruebele, M., and Schulten, K. (2008) Ten-microsecond molecular dynamics simulation of a fast-folding WW domain. *Biophys J 94*, L75–7.

[32] Schulz, R., Lindner, B., Petridis, L., and Smith, J. C. (2009) Scaling of Multimillion-Atom Biological Molecular Dynamics Simulation on a Petascale Supercomputer. *J Chem Theory Comput 5*, 2798–2808.

[33] Ozkan, S. B., Dill, K. A., and Bahar, I. (2003) Computing the transition state populations in simple protein models. *Biopolymers 68*, 35–46.

[34] Chodera, J. D., Swope, W. C., Pitera, J. W., and Dill, K. A. (2006) Long-Time Protein Folding Dynamics from Short-Time Molecular Dynamics Simulations. *Multiscale Model Simul 5*, 1214.

[35] Noé, F., Horenko, I., Schütte, C., and Smith, J. C. (2007) Hierarchical analysis of conformational dynamics in biomolecules: transition networks of metastable states. *J Chem Phys 126*, 155102.

[36] Chodera, J. D., Singhal, N., Pande, V. S., Dill, K. A., and Swope, W. C. (2007) Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics. *J Chem Phys 126*, 155101.

[37] Buchete, N.-v., and Hummer, G. (2008) Coarse master equations for peptide folding dynamics. *J Phys Chem B 112*, 6057–6069.

[38] Metzner, P., Schütte, C., and Vanden-Eijnden, E. (2009) Transition Path Theory for Markov Jump Processes. *Multiscale Model Simul 7*, 1192.

[39] Noé, F., and Fischer, S. (2008) Transition networks for modeling the kinetics of conformational change in macromolecules. *Curr Opin Struct Biol 18*, 154–162.

[40] Bowman, G. R., Beauchamp, K. A., Boxer, G., and Pande, V. S. (2009) Progress and challenges in the automated construction of Markov state models for full protein systems. *J Chem Phys 131*, 124101.

[41] Sriraman, S., Kevrekidis, I. G., and Hummer, G. (2005) Coarse master equation from Bayesian analysis of replica molecular dynamics simulations. *J Phys Chem B 109*, 6479–6484.

[42] Huang, X., Bowman, G. R., Bacallado, S., and Pande, V. S. (2009) Rapid equilibrium sampling initiated from nonequilibrium data. *Proc Nat Acad Sci USA 106*, 19765–19769.

[43] Yang, S., Banavali, N. K., and Roux, B. (2009) Mapping the conformational transition in Src activation by cumulating the information from multiple molecular dynamics trajectories. *Proc Nat Acad Sci USA 106*, 3776–3781.

[44] Noé, F., Schütte, C., Vanden-Eijnden, E., Reich, L., and Weikl, T. R. (2009) Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. *Proc Nat Acad Sci USA 106*, 19011–19016.

[45] Voelz, V. A., Bowman, G. R., Beauchamp, K. A., and Pande, V. S. (2010) Molecular simulation of ab initio protein folding for a millisecond folder NTL9(1-39). *J Am Chem Soc 132*, 1526–1528.

[46] Bowman, G. R., and Pande, V. S. (2010) Protein folded states are kinetic hubs. *Proc Nat Acad Sci USA 2010*, 1–6.

[47] Bowman, G. R., Huang, X., and Pande, V. S. (2010) Network models for molecular kinetics and their initial applications to human health. *Cell Research* 1–9.

[48] Hansmann, U. (1997) Parallel tempering algorithm for conformational studies of biological molecules. *Chem Phys Lett 281*, 140–150.

[49] Mitsutake, A., Sugita, Y., and Okamoto, Y. (2001) Generalized-ensemble algorithms for molecular simulations of biopolymers. *Biopolymers 60*, 96–123.

[50] Faradjian, A. K., and Elber, R. (2004) Computing time scales from reaction coordinates by milestoning. *J Chem Phys 120*, 10880–10889.

[51] West, A. M. A., Elber, R., and Shalloway, D. (2007) Extending molecular dynamics time scales with milestoning: example of complex kinetics in a solvated peptide. *J Chem Phys 126*, 145104.

[52] Hänggi, P., Talkner, P., and Borkovec, M. (1990) Reaction-rate theory: fifty years after Kramers. *Rev Mod Phys 62*, 251–341.

[53] Elber, R., and West, A. M. A. (2010) Atomically detailed simulation of the recovery stroke in myosin by Milestoning. *Proc Nat Acad Sci USA 107*, 5001–5005.

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

[55] Dellago, C., Bolhuis, P. G., Csajka, F., and Chandler, D. (1998) Transition path sampling and the calculation of rate constants. *J Chem Phys 108*, 1964.

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

[57] Dellago, C., and Bolhuis, P. G. (2007) Transition Path Sampling Simulations of Biological Systems. *Topics in Current Chemistry 268*, 291–317.

[58] Escobedo, F. A., Borrero, E. E., and Araque, J. C. (2009) Transition path sampling and forward flux sampling. Applications to biological systems. *Journal of Physics: Condensed Matter 21*, 333101.

[59] Allen, R. J., Valeriani, C., and Rein ten Wolde, P. (2009) Forward flux sampling for rare event simulations. *Journal of Physics: Condensed Matter 21*, 463102.

[60] Grünwald, M., Dellago, C., and Geissler, P. L. (2008) Precision shooting: Sampling long transition pathways. *J Chem Phys 129*, 194101.

[61] Zhang, B. W., Jasnow, D., and Zuckerman, D. M. (2009) Weighted Ensemble Path Sampling for Multiple Reaction Channels. *arXiv preprint arXiv:0902.2772*

[62] Vreede, J., Juraszek, J., and Bolhuis, P. G. (2010) Predicting the reaction coordinates of millisecond light-induced conformational changes in photoactive yellow protein. *Proc Nat Acad Sci USA 107*, 2397–2402.

[63] Juraszek, J., and Bolhuis, P. G. (2010) (Un)Folding mechanisms of the FBP28 WW domain in explicit solvent revealed by multiple rare event simulation methods. *Biophys J 98*, 646–656.

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

[65] Moroni, D., Bolhuis, P. G., and van Erp, T. S. (2004) Rate constants for diffusive processes by partial path sampling. *J Chem Phys 120*, 4055–4065.

[66] van Erp, T. S. (2007) Reaction Rate Calculation by Parallel Path Swapping. *Phys Rev Lett 98*, 1–4.

[67] Vanerp, T., and Bolhuis, P. G. (2005) Elaborating transition interface sampling methods. *Journal of Computational Physics 205*, 157–181.

[68] Bhatt, D., Zhang, B. W., and Zuckerman, D. M. (2010) Steady-state simulations using weighted ensemble path sampling. *J Chem Phys 133*, 014110.

[69] Allen, R. J., Warren, P., and ten Wolde, P. R. (2005) Sampling Rare Switching Events in Biochemical Networks. *Phys Rev Lett 94*, 018104.

[70] Allen, R. J., Frenkel, D., and ten Wolde, P. R. (2006) Simulating rare events in equilibrium or nonequilibrium stochastic systems. *J Chem Phys 124*, 024102.

[71] Dickson, A., and Dinner, A. R. (2010) Enhanced sampling of nonequilibrium steady states. *Annu Rev Phys Chem 61*, 441–459.

[72] Allen, R. J., Frenkel, D., and ten Wolde, P. R. (2006) Forward flux sampling-type schemes for simulating rare events: efficiency analysis. *J Chem Phys 124*, 194111.

[73] Borrero, E. E., and Escobedo, F. A. (2008) Optimizing the sampling and staging for simulations of rare events via forward flux sampling schemes. *J Chem Phys 129*, 024115.

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

[75] Huber, G. A., and Kim, S. (1996) Weighted-Ensemble Brownian Dynamics Simulations for Protein Association Reactions. *Biophys J 70*, 97–110.

[76] Rojnuckarin, A., Kim, S., and Subramaniam, S. (1998) Brownian dynamics simulations of protein folding: access to milliseconds time scale and beyond. *Proc Nat Acad Sci USA 95*, 4288–4292.

[77] Rojnuckarin, A., Livesay, D. R., and Subramaniam, S. (2000) Bimolecular Reaction Simulation Using Weighted Ensemble Brownian Dynamics and the University of Houston Brownian Dynamics Program. *Biophys J 79*, 686–693.

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

[79] Warmflash, A., Bhimalapuram, P., and Dinner, A. R. (2007) Umbrella sampling for nonequilibrium processes. *J Chem Phys 127*, 154112.

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

[81] Zwier, M. C., Kaus, J. W., and Chong, L. T. (2011) 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 7*, 1189–1197.

[82] Zuckerman, D. M., and Woolf, T. B. (2000) Efficient dynamic importance sampling of rare events in one dimension. *Phys Rev E 63*, 1–10.

[83] Zwier, M. C., and Chong, L. T. (2010) Reaching biological timescales with all-atom molecular dynamics simulations. *Curr Opin Pharmacol 10*, 745–752.

[84] Bhatt, D., and Zuckerman, D. M. (2010) Heterogeneous Path Ensembles for Conformational Transitions in Semiatomistic Models of Adenylate Kinase. *J Chem Theory Comput 6*, 3527–3539.

[85] Dang, L. X. (1994) Potential of mean force for the methane–methane pair in water. *J Chem Phys 100*, 9032.

[86] Meng, E. C., and Kollman, P. A. (1996) Molecular Dynamics Studies of the Properties of Water around Simple Organic Solutes. *J Phys Chem 100*, 11460–11470.

[87] Oostenbrink, C., and van Gunsteren, W. F. (2005) Methane clustering in explicit water: effect of urea on hydrophobic interactions. *Phys Chem Chem Phys 7*, 53.

[88] Trzesniak, D., and van Gunsteren, W. F. (2006) Pathway dependence of the efficiency of calculating free energy and entropy of solute – solute association in water. *Chem Phys 330*, 410–416.

[89] Thomas, A. S., and Elcock, A. H. (2007) Molecular dynamics simulations of hydrophobic associations in aqueous salt solutions indicate a connection between water hydrogen bonding and the Hofmeister effect. *J Am Chem Soc 129*, 14887–14898.

[90] Trzesniak, D., Kunz, A.-P. E., and van Gunsteren, W. F. (2007) A Comparison of Methods to Compute the Potential of Mean Force. *Chemphyschem 8*, 162–169.

[91] Belch, A. C., Berkowitz, M. L., and McCammon, J. A. (1986) Solvation structure of a sodium chloride ion pair in water. *J Am Chem Soc 108*, 1755–1761.

[92] Dang, L. X., Rice, J. E., and Kollman, P. A. (1990) The effect of water models on the interaction of the sodium–chloride ion pair in water: Molecular dynamics simulations. *J Chem Phys 93*, 7528.

[93] Guàrdia, E., Rey, R., and Padró, J. A. (1991) Potential of mean force by constrained molecular dynamics: A sodium chloride ion-pair in water. *Chem Phys 155*, 187–195.

[94] Hummer, G., Soumpasis, D., and Neumann, M. (1992) Pair correlations in an NaCl-SPC water model. *Molecular Physics 77*, 769–785.

[95] Pratt, L. R., Hummer, G., and Garcia, A. E. (1994) Ion pair potentials-of-mean-force in water. *Biophysical chemistry 51*, 147–165.

[96] Koneshan, S., and Rasaiah, J. C. (2000) Computer simulation studies of aqueous sodium chloride solutions at 298 K and 683 K. *J Chem Phys 113*, 8125.

[97] Patra, M., and Karttunen, M. (2004) Systematic comparison of force fields for microscopic simulations of NaCl in aqueous solutions: diffusion, free energy of hydration, and structural properties. *J Comput Chem 25*, 678–689.

[98] Baumketner, A. (2009) Removing systematic errors in interionic potentials of mean force computed in molecular simulations using reaction-field-based electrostatics. *J Chem Phys 130*, 104106.

[99] Fennell, C. J., Bizjak, A., Vlachy, V., and Dill, K. A. (2009) Ion pairing in molecular simulations of aqueous alkali halide solutions. *J Phys Chem B 113*, 6782–6791.

[100] Timko, J., Bucher, D., and Kuyucak, S. (2010) Dissociation of NaCl in water from ab initio molecular dynamics simulations. *J Chem Phys 132*, 114510.

[101] Tsuzuki, S., Honda, K., Uchimaru, T., Mikami, M., and Tanabe, K. (2000) The Magnitude of the CH/π Interaction between Benzene and Some Model Hydrocarbons . *J Am Chem Soc 122*, 3746–3753.

[102] Ringer, A. L., Figgs, M. S., Sinnokrot, M. O., and Sherrill, C. D. (2006) Aliphatic C-H / π Interactions : Methane-Benzene , Methane-Phenol , and Methane-Indole Complexes. *J Phys Chem A 110*, 10822–10828.

[103] Dang, L. X., and Kollman, P. A. (1990) Free Energy of Association of the 18-Crown-6:K+ Complex in Water: A Molecular Dynamics Simulation. *J Am Chem Soc 112*, 5716–5720.

[104] Troxler, L., and Wipff, G. (1994) Conformation and Dynamics of 18-Crown-6, Cryptand 222, and Their Cation Complexes in Acetonitrile Studied by Molecular Dynamics Simulations. *J Am Chem Soc 116*, 1468–1480.

[105] Humphrey, W. (1996) VMD: Visual molecular dynamics. *Journal of Molecular Graphics 14*, 33–38.

[106] Efron, B. Y. B., and Tibshirani, R. (1986) Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. *Stat Sci 1*, 54–75.

[107] Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. *Numerical Recipes in C*, 2nd ed.; Cambridge University Press: Cambridge, England, 1992.

[108] Zhang, B. W., Jasnow, D., and Zuckerman, D. M. (2007) Transition-event durations in one-dimensional activated processes. *J Chem Phys 126*, 074504.

[109] Kolmogoroff, A. (1941) Confidence limits for an unknown distribution function. *Ann Math Stat 12*, 461–463.

[110] Kvam, P. H., and Vidakovic, B. *Nonparametric Statistics with Applications to Science and Engineering*; John Wiley & Sons: Hoboken, New Jersey, 2007.

[111] Cambillau, C., Bram, G., Corset, J., Riche, C., and Pascard-Billy, C. (1978) Enolates de l'acetylacetate d'ethyle. Structure et reactivite du sel de potassium en presence d'ether couronne et de cryptand. *Tetrahedron 34*, 2675–2685.

[112] Adelman, S., and Doll, J. (1976) Generalized Langevin equation approach for atom/solid-surface scattering: General formulation for classical scattering off harmonic solids. *J Chem Phys 64*, 2375.

[113] Shirts, M. R., Pitera, J. W., Swope, W. C., and Pande, V. S. (2003) Extremely precise free energy calculations of amino acid side chain analogs: Comparison of common molecular mechanics force fields for proteins. *J Chem Phys 119*, 5740.

[114] Essmann, U., Perera, L., Berkowitz, M. L., Darden, T. A., Lee, H., and Pedersen, L. G. (1995) A smooth particle mesh Ewald method. *J Chem Phys 103*, 8577.

[115] Hess, B., Bekker, H., Berendsen, H. J. C., and Fraaije, J. G. E. M. (1997) LINCS: A linear constraint solver for molecular simulations. *J Comput Chem 18*, 1463–1472.

[116] Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A., and Haak, J. R. (1984) Molecular dynamics with coupling to an external bath. *J Chem Phys 81*, 3684–3690.

[117] Schuler, L., Daura, X., and van Gunsteren, W. F. (2001) An improved GROMOS 96 force field for aliphatic hydrocarbons in the condensed phase. *J Comput Chem 22*, 1205–1218.

[118] Berendsen, H. J. C., Grigera, J. R., and Straatsma, T. P. (1987) The missing term in effective pair potentials. *J Phys Chem 91*, 6269–6271.

[119] Kaminski, G. A., Friesner, R. A., Tirado-Rives, J., and Jorgensen, W. L. (2001) Evaluation and Reparametrization of the OPLS-AA Force Field for Proteins via Comparison with Accurate Quantum Chemical Calculations on Peptides âĂă. *J Phys Chem B 105*, 6474–6487.

[120] Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., and Klein, M. L. (1983) Comparison of simple potential functions for simulating liquid water. *J Chem Phys 79*, 926.

[121] Chandler, D. *Introduction to Modern Statistical Mechanics*; Oxford University Press: New York, New York, 1987.

[122] Lei, H., and Duan, Y. (2007) Improved sampling methods for molecular simulation. *Curr Opin Struct Biol 17*, 187–191.

[123] Huber, T., Torda, A. E., and van Gunsteren, W. F. (1994) Local elevation: a method for improving the searching properties of molecular dynamics simulation. *Journal of Computer-Aided Molecular Design 8*, 695–708.

[124] Laio, A., and Parrinello, M. (2002) Escaping free-energy minima. *Proc Nat Acad Sci USA 99*, 12562–12566.

[125] Schlitter, J., Engels, M., and Kr u ger, P. (1994) Targeted molecular dynamics: a new approach for searching pathways of conformational transitions. *Journal of Molecular Graphics 12*, 84–89.

[126] Izrailev, S., Stepaniants, S., Balsera, M., Oono, Y., and Schulten, K. (1997) Molecular dynamics study of unbinding of the avidin-biotin complex. *Biophys J 72*, 1568–1581.

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

[128] Zheng, L., Chen, M., and Yang, W. (2008) Random walk in orthogonal space to achieve efficient free-energy simulation of complex systems. *Proc Nat Acad Sci USA 105*, 20227–20232.

[129] Májek, P., and Elber, R. (2010) Milestoning without a Reaction Coordinate. *J Chem Theory Comput 6*, 1805–1817.

[130] Gabdoulline, R. R., and Wade, R. C. (2002) Biomolecular diffusional association. *Curr Opin Struct Biol 12*, 204–213.

[131] Wright, P. E., and Dyson, H. J. (1999) Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *Journal of Molecular Biology 293*, 321–331.

[132] Shoemaker, B. A., Portman, J. J., and Wolynes, P. G. (2000) Speeding molecular recognition by using the folding funnel: the fly-casting mechanism. *Proc Nat Acad Sci USA 97*, 8868–8873.

[133] Narayanan, R., Ganesh, O. K., Edison, A. S., and Hagen, S. J. (2008) Kinetics of Folding and Binding of an Intrinsically Disordered Protein: The Inhibitor of Yeast Aspartic Proteinase YPrA. *J Am Chem Soc 130*, 11477–11485.

[134] Sugase, K., Lansing, J. C., Dyson, H. J., and Wright, P. E. (2007) Tailoring Relaxation Dispersion Experiments for Fast-Associating Protein Complexes. *J Am Chem Soc 129*, 13406–13407.

[135] Sugase, K., Dyson, H. J., and Wright, P. E. (2007) Mechanism of coupled folding and binding of an intrinsically disordered protein. *Nature 447*, 1021–1025.

[136] Vamvaca, K., Jelesarov, I., and Hilvert, D. (2008) Kinetics and Thermodynamics of Ligand Binding to a Molten Globular Enzyme and Its Native Counterpart. *Journal of Molecular Biology 382*, 971–977.

[137] Muralidhara, B. K., Rathinakumar, R., and Wittung-Stafshede, P. (2006) Folding of Desulfovibrio desulfuricans flavodoxin is accelerated by cofactor fly-casting. *Archives of biochemistry and biophysics 451*, 51–58.

[138] Lengyel, C. S. E., Willis, L. J., Mann, P., Baker, D., Kortemme, T., Strong, R. K., and McFarland, B. J. (2007) Mutations Designed to Destabilize the Receptor-Bound Conformation Increase MICA-NKG2D Association Rate and Affinity. *Journal of Biological Chemistry 282*, 30658–30666.

[139] Crespin, M. O., Boys, B. L., and Konermann, L. (2005) The reconstitution of unfolded myoglobin with hemin dicyanide is not accelerated by fly-casting. *FEBS Letters 579*, 271–274.

[140] Hoffman, R. M. B., Blumenschein, T. M. A., and Sykes, B. D. (2006) An Interplay between Protein Disorder and Structure Confers the Ca2+ Regulation of Striated Muscle. *Journal of Molecular Biology 361*, 625–633.

[141] Perham, M., Chen, M., Ma, J., and Wittung-Stafshede, P. (2005) Unfolding of Heptameric Co-chaperonin Protein Follows "Fly Casting" Mechanism: Observation of Transient Nonnative Heptamer. *J Am Chem Soc 127*, 16402–16403.

[142] Jemth, P., and Gianni, S. (2007) PDZ Domains: Folding and Binding. *Biochemistry 46*, 8701–8708.

[143] Landfried, D. A., Vuletich, D. A., Pond, M. P., and Lecomte, J. T. J. (2007) Structural and thermodynamic consequences of b heme binding for monomeric apoglobins and other apoproteins. *Gene 398*, 12–28.

[144] Onitsuka, M., Kamikubo, H., Yamazaki, Y., and Kataoka, M. (2008) Mechanism of induced folding: Both folding before binding and binding before folding can be realized in staphylococcal nuclease mutants. *Proteins 72*, 837–847.

[145] Meisner, W. K., and Sosnick, T. R. (2004) Fast folding of a helical protein initiated by the collision of unstructured chains. *Proc. Natl. Acad. Sci. U.S.A. 101*, 13478–13482.

[146] Turjanski, A. G., Gutkind, J. S., Best, R. B., and Hummer, G. (2008) Binding-Induced Folding of a Natively Unstructured Transcription Factor. *PLoS Comput Biol 4*, e1000060.

[147] Levy, Y., Onuchic, J. N., and Wolynes, P. G. (2007) Fly-Casting in ProteinâĹŠDNA Binding: Frustration between Protein Folding and Electrostatics Facilitates Target Recognition. *J Am Chem Soc 129*, 738–739.

[148] Huang, Y., and Liu, Z. (2009) Kinetic advantage of intrinsically disordered proteins in coupled folding-binding process: a critical assessment of the "fly-casting" mechanism. *Journal of Molecular Biology 393*, 1143–1159.

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

[150] Frembgen-Kesner, T., and Elcock, A. H. (2009) Striking Effects of Hydrodynamic Interactions on the Simulated Diffusion and Folding of Proteins. *J Chem Theory Comput 5*, 242–256.

[151] Kussie, P. H., Gorina, S., Marechal, V., Elenbaas, B., Moreau, J., Levine, A. J., and Pavletich, N. P. (1996) Structure of the MDM2 Oncoprotein Bound to the p53 Tumor Suppressor Transactivation Domain. *Science 274*, 948.

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

[153] Takada, S. (1999) Go-ing for the prediction of protein folding mechanisms. *Proc. Natl. Acad. Sci. U.S.A. 96*, 11698–11700.

[154] Ermak, D. L., and McCammon, J. A. (1978) Brownian dynamics with hydrodynamic interactions. *J Chem Phys 69*, 1352.

[155] Elcock, A. H. (2006) Molecular Simulations of Cotranslational Protein Folding: Fragment Stabilities, Folding Cooperativity, and Trapping in the Ribosome. *PLoS Comput Biol 2*, e98.

[156] Lettieri, S., Zwier, M. C., Stringer, C. A., and Suarez, E. (2012) Simultaneous computation of dynamical and equilibrium information using a weighted ensemble of trajectories. *arXiv*

[157] Northrup, S. H., Allison, S. A., and McCammon, J. A. (1984) Brownian dynamics simulation of diffusion-influenced bimolecular reactions. *J Chem Phys 80*, 1517.

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

[159] Frembgen-Kesner, T., and Elcock, A. H. (2010) Absolute Protein-Protein Association Rate Constants from Flexible, Coarse-Grained Brownian Dynamics Simulations: The Role of Intermolecular Hydrodynamic Interactions in Barnase-Barstar Association. *Biophysj 99*, L75–L77.

[160] Kohn, J. E., and Plaxco, K. W. (2005) Engineering a signal transduction mechanism for protein-based biosensors. *Proc. Natl. Acad. Sci. U.S.A. 102*, 10841–10845.

[161] Dror, R. O., Pan, A. C., Arlow, D. H., Borhani, D. W., Maragakis, P., Shan, Y., Xu, H., and Shaw, D. E. (2011) Pathway and mechanism of drug binding to G-protein-coupled receptors. *Proc Nat Acad Sci USA 108*, 13118–13123.

[162] Shan, Y., Kim, E. T., Eastwood, M. P., Dror, R. O., Seeliger, M. A., and Shaw, D. E. (2011) How Does a Drug Molecule Find Its Target Binding Site? *J Am Chem Soc 133*, 9181–9183.

[163] Buch, I., Harvey, M. J., Giorgino, T., Anderson, D. P., and De Fabritiis, G. (2010) High-throughput all-atom molecular dynamics simulations using distributed computing. *J Chem Inf Model 50*, 397–403.

[164] Wang, Y., and Tajkhorshid, E. (2008) Electrostatic funneling of substrate in mitochondrial inner membrane carriers. *Proc Nat Acad Sci USA 105*, 9598–9603.

[165] Giorgino, T., Buch, I., and De Fabritiis, G. (2012) Visualizing the Induced Binding of SH2-Phosphopeptide. *J Chem Theory Comput*

[166] Silva, D.-A., Bowman, G. R., Sosa-Peinado, A., and Huang, X. (2011) A Role for Both Conformational Selection and Induced Fit in Ligand Binding by the LAO Protein. *PLoS Comput Biol 7*, e1002054.

[167] Prinz, J.-H., Wu, H., Sarich, M., Keller, B., Senne, M., Held, M., Chodera, J. D., Schütte, C., and Noé, F. (2011) Markov models of molecular kinetics: Generation and validation. *J Chem Phys 134*, 174105.

[168] Adelman, J. L., Dale, A. L., Zwier, M. C., Bhatt, D., Chong, L. T., Zuckerman, D. M., and Grabe, M. (2011) Simulations of the Alternating Access Mechanism of the Sodium Symporter Mhp1. *Biophys J 101*, 2399–2407.

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

[170] Juraszek, J., Vreede, J., and Bolhuis, P. G. (2011) Transition path sampling of protein conformational changes. *Chem Phys*

[171] Ren, W., Vanden-Eijnden, E., Maragakis, P., and E, W. (2005) Transition pathways in complex systems: Application of the finite-temperature string method to the alanine dipeptide. *J Chem Phys 123*, 134109.

[172] Joerger, A. C., and Fersht, A. R. (2008) Structural biology of the tumor suppressor p53. *Annu Rev Biochem 77*, 557–582.

[173] Toledo, F., and Wahl, G. M. (2006) Regulating the p53 pathway: in vitro hypotheses, in vivo veritas. *Nat Rev Cancer 6*, 909–923.

[174] Massova, I., and Kollman, P. A. (1999) Computational Alanine Scanning To Probe ProteinâĹŠProtein Interactions: A Novel Approach To Evaluate Binding Free Energies. *J Am Chem Soc 121*, 8133–8143.

[175] Zhong, H., and Carlson, H. A. (2005) Computational studies and peptidomimetic design for the human p53-MDM2 complex. *Proteins 58*, 222–234.

[176] Espinoza-Fonseca, L. M., and Trujillo-Ferrara, J. G. (2006) Transient stability of the helical pattern of region F19–L22 of the N-terminal domain of p53: A molecular dynamics simulation study. *Biochem Biophys Res Comm 343*, 110–116.

[177] Xiong, K., Zwier, M. C., Myshakina, N. S., Burger, V. M., Asher, S. A., and Chong, L. T. (2011) Direct Observations of Conformational Distributions of Intrinsically Disordered p53 Peptides Using UV Raman and Explicit Solvent Simulations. *J Phys Chem A 115*, 9520–9527.

[178] Espinoza-Fonseca, L. M., and Trujillo-Ferrara, J. G. (2006) Conformational changes of the p53-binding cleft of MDM2 revealed by molecular dynamics simulations. *Biopolymers 83*, 365–373.

[179] Eyrisch, S., and Helms, V. (2007) Transient Pockets on Protein Surfaces Involved in ProteinâĹŠProtein Interaction. *Journal of Medicinal Chemistry 50*, 3457–3464.

[180] Chen, H.-F., and Luo, R. (2007) Binding induced folding in p53-MDM2 complex. *J Am Chem Soc 129*, 2930–2937.

[181] Dastidar, S. G., Lane, D. P., and Verma, C. S. (2008) Multiple Peptide Conformations Give Rise to Similar Binding Affinities: Molecular Simulations of p53-MDM2. *J Am Chem Soc 130*, 13514–13515.

[182] Mittal, J., Yoo, T. H., Georgiou, G., and Truskett, T. M. (2013) Structural Ensemble of an Intrinsically Disordered Polypeptide. *J Phys Chem B 117*, 118–124.

[183] Dastidar, S. G., Madhumalar, A., Fuentes, G., Lane, D. P., and Verma, C. S. (2009) Forces mediating protein–protein interactions: a computational study of p53 “approaching” MDM2. *Theor Chem Acc 125*, 621–635.

[184] Dastidar, S. G., Lane, D. P., and Verma, C. S. (2009) Modulation of p53 binding to MDM2: computational studies reveal important roles of Tyr100. *BMC Bioinformatics 10 Suppl 1*, S6.

[185] ElSawy, K. M., Verma, C. S., Joseph, T. L., Lane, D. P., Twarock, R., and Caves, L. S. D. (2013) On the interaction mechanisms of a p53 peptide and nutlin with the MDM2 and MDMX proteins: A Brownian dynamics study. *Cell Cycle 12*, 394–404.

[186] Still, W. C., Tempczyk, A., Hawley, R. C., and Hendrickson, T. (1990) Semianalytical treatment of solvation for molecular mechanics and dynamics. *J Am Chem Soc 112*, 6127–6129.

[187] Onufriev, A., Bashford, D., and Case, D. A. (2000) Modification of the Generalized Born Model Suitable for Macromolecules. *J Phys Chem B 104*, 3712–3720.

[188] McCoy, M. A., Gesell, J. J., Senior, M. M., and Wyss, D. F. (2003) Flexible Lid to the p53-Binding Domain of Human Mdm2: Implications for p53 Regulation. *Proc Nat Acad Sci USA 100*, 1645–1648.

[189] Tsai, C. J., Kumar, S., Ma, B., and Nussinov, R. (1999) Folding funnels, binding funnels, and protein function. *Protein Sci 8*, 1181–1190.

[190] Miller, D. W., and Dill, K. A. (2002) Ligand binding to proteins: the binding landscape model. *Protein Sci 6*, 2166–2179.

[191] Zhang, C., Chen, J., and DeLisi, C. (1999) Protein-protein recognition: exploring the energy funnels near the binding sites. *Proteins 34*, 255–267.

[192] Tovchigrechko, A., and Vakser, I. A. (2001) How common is the funnel-like energy landscape in protein-protein interactions? *Protein Sci 10*, 1572–1583.

[193] Selzer, T., and Schreiber, G. (2001) New insights into the mechanism of protein-protein association. *Proteins 45*, 190–198.

[194] Northrup, S. H., and Erickson, H. P. (1992) Kinetics of protein-protein association explained by Brownian dynamics computer simulation. *Proc Nat Acad Sci USA 89*, 3338–3342.

[195] Verkhivker, G. M., Rejto, P. A., Gehlhaar, D. K., and Freer, S. T. (1996) Exploring the energy landscapes of molecular recognition by a genetic algorithm: analysis of the requirements for robust docking of HIV-1 protease and FKBP-12 complexes. *Proteins 25*, 342–353.

[196] Schon, O., Friedler, A., Bycroft, M., Freund, S. M., and Fersht, A. R. (2002) Molecular Mechanism of the Interaction between MDM2 and p53. *Journal of Molecular Biology 323*, 491–501.

[197] Schreiber, G., Haran, G., and Zhou, H.-X. (2009) Fundamental aspects of protein-protein association kinetics. *Chem Rev 109*, 839–860.

[198] Showalter, S. A., Bruschweiler-Li, L., Johnson, E., Zhang, F., and Brüschweiler, R. (2008) Quantitative Lid Dynamics of MDM2 Reveals Differential Ligand Binding Modes of the p53-Binding Cleft. *J Am Chem Soc 130*, 6472–6478.

[199] Setny, P., Baron, R., and al, e. (2013) Solvent fluctuations in hydrophobic cavity–ligand binding kinetics. *Proceedings of the ...*

[200] Zondlo, S. C., Lee, A. E., and Zondlo, N. J. (2006) Determinants of specificity of MDM2 for the activation domains of p53 and p65: proline27 disrupts the MDM2-binding motif of p53. *Biochemistry 45*, 11945–11957.

[201] Li, C., Pazgier, M., Li, C., Yuan, W., Liu, M., Wei, G., Lu, W.-Y., and Lu, W. (2010) Systematic Mutational Analysis of Peptide Inhibition of the p53–MDM2/MDMX Interactions. *Journal of Molecular Biology 398*, 200–213.

[202] Dastidar, S. G., Lane, D. P., and Verma, C. S. (2012) Why is F19Ap53 unable to bind MDM2?: Simulations suggest crack propagation modulates binding. *Cell Cycle 11*, 2239–2247.

[203] Rajamani, D., Thiel, S., Vajda, S., and Camacho, C. J. (2004) Anchor residues in protein-protein interactions. *Proc Nat Acad Sci USA 101*, 11287–11292.

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

[205] Onufriev, A., Bashford, D., and Case, D. A. (2004) Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins 55*, 383–394.

[206] Srinivasan, J., Trevathan, M. W., Beroza, P., and Case, D. A. (1999) Application of a pairwise generalized Born model to proteins and nucleic acids: inclusion of salt effects. *Theor Chem Acc 101*, 426–434.

[207] Mills, R. (1973) Self-diffusion in normal and heavy water in the range 1-45.deg. *J Phys Chem 77*, 685–688.

[208] Gabdoulline, R. R., and Wade, R. C. (1999) On the protein-protein diffusional encounter complex. *J. Mol. Recognit. 12*, 226–234.

[209] Bhatt, D., and Zuckerman, D. M. (2011) Beyond Microscopic Reversibility: Are Observable Nonequilibrium Processes Precisely Reversible? *J Chem Theory Comput 7*, 2520–2527.

[210] Grossfield, A., and Zuckerman, D. (2009) Quantifying uncertainty and sampling quality in biomolecular simulations. *Annual reports in computational chemistry 5*, 23–48.

[211] Shaw, D. E. et al. (2009) Millisecond-scale molecular dynamics simulations on Anton. *Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis - SC '09* 1.

[212] Lindorff-Larsen, K., Piana, S., Dror, R. O., and Shaw, D. E. (2011) How Fast-Folding Proteins Fold. *Science 334*, 517–520.

[213] Clementi, C., Nymeyer, H., and Onuchic, J. N. (2000) 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. *Journal of Molecular Biology 298*, 937–953.

[214] Koga, N., and Takada, S. (2001) Roles of native topology and chain-length scaling in protein folding: A simulation study with a Gō-like model. *Journal of Molecular Biology 313*, 171–180.

[215] Daggett, V., and Fersht, A. (2003) The present view of the mechanism of protein folding. *Nat Rev Mol Cell Biol 4*, 497–502.

[216] Dickson, B. M., Makarov, D. E., and Henkelman, G. (2009) Pitfalls of choosing an order parameter for rare event calculations. *J Chem Phys 131*, 074108.

[217] Hünenberger, P. (2005) Thermostat algorithms for molecular dynamics simulations. *Advanced Computer Simulation* 130–130.

[218] Pronk, S., Pall, S., Schulz, R., Larsson, P., Bjelkmar, P., Apostolov, R., Shirts, M. R., Smith, J. C., Kasson, P. M., van der Spoel, D., Hess, B., and Lindahl, E. (2013) GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics*

[219] (2012) AMBER 12. *University of California, San Francisco*

[220] Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R. D., Kalé, L., and Schulten, K. (2005) Scalable molecular dynamics with NAMD. *J Comput Chem 26*, 1781–1802.