Markov State Modeling of Binding and Conformational Changes of Proteins

publikationsbasierte Dissertation

zur Erlangung des akademischen Grades

doctor rerum naturalium

(Dr. rer. nat.)

in der Wissenschaftsdisziplin Theoretische Physik

eingereicht an der
Mathematisch-Naturwissenschaftlichen Fakultät
der Universität Potsdam

Eingereicht von:

Herrn Fabian Paul, Dipl. Phys. aus Berlin

am 6. September 2017

Hauptbetreuer: PD Dr. Thomas Weikl
Zweitbetreuer: Prof. Dr. Frank Noé
weiterer Gutachter: Prof. Dr. Xuhui Huang
weitere Mitglieder der Prüfungskommission: Prof. Dr. Ralf Metzler, Prof. Dr. Achim Feldmeier, Prof. Dr. Joachim Dzubiella und Prof. Dr. Matthias Holschneider

Tag der Disputation: 16. November 2017

Published online at the
Institutional Repository of the University of Potsdam:
URN urn:nbn:de:kobv:517-opus4-404273
http://nbn-resolving.de/urn:nbn:de:kobv:517-opus4-404273

Ich versichere, dass ich die von mir vorgelegte Arbeit selbstständig verfasst habe, dass ich die verwendeten Quellen, Internet-Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit — einschließlich Tabellen, Abbildungen —, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, in jedem Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.
Ich versichere, dass die Arbeit bisher an keiner anderen Hochschule eingereicht worden ist.

Potsdam, den 6. September 2017

.....
Fabian Paul

The effect of a concept-driven revolution is to explain old things in new ways. The effect of a tool-driven revolution is to discover new things that have to be explained. In almost every branch of science, and especially in biology and astronomy, there has been a preponderance of tool-driven revolutions. We have been more successful in discovering new things than in explaining old ones.

Freeman Dyson in *Imagined Worlds* (1998)

Ohne Mathematik ist diese Moderne gar nicht auszuhalten.

Harald Lesch

I wish to thank PD Dr. Thomas Weikl and Prof. Dr. Frank Noé for giving me the opportunity to work on this interesting topic. I further wish to thank Esam Abualrous, Katja Geiger, Simon Olsson, Guillermo Pérez-Hernández, Nuria Plattner, Sebastian Stolzenberg, Christoph Wehmeyer, Hao Wu, Tim Hempel, Moritz Hoffmann, Jan-Hendrik Prinz, Benjamin Trendelkamp-Schroer, Martin Scherer, Andreas Winkler and Katarzyna Ziółkowska.

Zusammenfassung

Proteine sind für das Leben essentielle Moleküle, die eine Vielzahl von Funktionen in Organismen ausüben. Dazu ändern sie ihre Konformation und binden an andere Moleküle. Jedoch ist das Zusammenspiel zwischen Konformationsänderung und Bindung nicht vollständig verstanden. In dieser Arbeit wird dieses Zusammenspiel mit Molekulardynamik-Simulationen (MD) des Protein-Peptid-Systems Mdm2-PMI und mit der Analyse von Daten aus Relaxationsexperimenten untersucht.
Die zentrale Aufgabe ist, den Bindungsmechanismus aufzudecken, welcher durch die Reihenfolge von (partiellen) Bindungsereignissen und Konformationsänderungsereignissen beschrieben wird, inklusive der Wahrscheinlichkeiten dieser Ereignisse. Im einfachsten Fall lässt sich der Bindungsmechanismus durch ein Zwei-Schritt-Modell beschreiben: erst Bindung, dann Konformationsänderung oder erst Konformationsänderung und dann Bindung. Im allgemeinen Fall sind längere Schrittfolgen mit mehreren Konformationsänderungen und partiellen Bindungsereignissen möglich, ebenso wie parallele Wege, die sich in ihrer Schrittfolge unterscheiden. Die Theorie der Markow-Modelle (MSM) bildet den theoretischen Rahmen, in dem alle diese Fälle modelliert werden können. Dazu werden in dieser Arbeit MSMs aus MD-Daten geschätzt und Ratengleichungsmodelle, die mit MSMs verwandt sind, aus experimentellen Relaxationsdaten abgeleitet.
Die MD-Simulation und Markow-Modellierung des PMI-Mdm2-Systems zeigt, dass PMI und Mdm2 auf verschiedenen Wegen binden können. Ein Hauptergebnis dieser Arbeit ist die durch Markow-Modellierung berechnete Dissoziationsrate von der Größenordnung von einem Ereignis pro Sekunde in Übereinstimmung mit experimentellen Daten. Dissoziations- und Übergangsraten in dieser Größenordnung wurden bisher nur mit Methoden berechnet, die Übergänge beschleunigen, indem mit zeitabhängigen, externen Kräften auf die Bindungspartner eingewirkt wird. Die in dieser Arbeit entwickelte Simulationstechnik dagegen erlaubt die Schätzung von Dissoziationsraten aus der Kombination von Freien-Energie-Rechnungen und direkter MD-Simulation des schnellen Bindungsprozesses. Zwei neue statistische Schätzer, TRAM und TRAMMBAR wurden entwickelt um ein MSM aus dem Gesamtdatensatz aus beiden Simulationstypen zu schätzen.
Zudem wird in dieser Arbeit eine neue Analysetechnik für Zeitreihen aus chemischen Relaxationsexperimenten entwickelt. Sie ermöglicht es einen der beiden oben erwähnten Zwei-Schritt-Mechanismen als den den Daten zugrundeliegenden Mechanismus zu identifizieren. Die neue Methode ist für einen größeren Konzentrationsbereich gültig als frühere Methoden und erlaubt es daher, die Konzentrationen so zu wählen, dass der Mechanismus eindeutig identifiziert werden kann. Sie wurde erfolgreich mit Daten für die Bindung von Recoverin an ein Rhodopsinkinasenpeptid getestet.

Abstract

Proteins are molecules that are essential for life and carry out an enormous number of functions in organisms. To this end, they change their conformation and bind to other molecules. However, the interplay between conformational change and binding is not fully understood. In this work, this interplay is investigated with molecular dynamics (MD) simulations of the protein-peptide system Mdm2-PMI and by analysis of data from relaxation experiments.

The central task it to uncover the binding mechanism, which is described by the sequence of (partial) binding events and conformational change events including their probabilities. In the simplest case, the binding mechanism is described by a two-step model: binding followed by conformational change or conformational change followed by binding. In the general case, longer sequences with multiple conformational changes and partial binding events are possible as well as parallel pathways that differ in their sequences of events. The theory of Markov state models (MSMs) provides the theoretical framework in which all these cases can be modeled. For this purpose, MSMs are estimated in this work from MD data, and rate equation models, which are related to MSMs, are inferred from experimental relaxation data.

The MD simulation and Markov modeling of the PMI-Mdm2 system shows that PMI and Mdm2 can bind via multiple pathways. A main result of this work is a dissociation rate on the order of one event per second, which was calculated using Markov modeling and is in agreement with experiment. So far, dissociation rates and transition rates of this magnitude have only been calculated with methods that speed up transitions by acting with time-dependent, external forces on the binding partners. The simulation technique developed in this work, in contrast, allows the estimation of dissociation rates from the combination of free energy calculation and direct MD simulation of the fast binding process. Two new statistical estimators TRAM and TRAMMBAR are developed to estimate a MSM from the joint data of both simulation types.

In addition, a new analysis technique for time-series data from chemical relaxation experiments is developed in this work. It allows to identify one of the above-mentioned two-step mechanisms as the mechanism that underlays the data. The new method is valid for a broader range of concentrations than previous methods and therefore allows to choose the concentrations such that the mechanism can be uniquely identified. It is successfully tested with data for the binding of recoverin to a rhodopsin kinase peptide.

Contents

1. Introduction

1.1. Proteins

Proteins are molecules that are essential for life by carrying out an enormous number of functions in organisms, such as the catalysis of metabolic processes, energy transfer, gene expression, transport of solutes across biological membranes, cellular communication and information processing, molecular recognition, defense against pathogens and formation of intracellular and extracellular structures. [2, 66] The functions of many proteins can be understood via their structure, that is the tree-dimensional spatial arrangement of the atoms. Proteins consist of one or a few unbranched chains of amino acids. The sequence of amino acids in each chain is unique for every type of protein and is called the *primary structure*. The atoms in the chains interact via different non-bonded attractive and repulsive forces that cause the chains to adopt a three-dimensional structure that is called the *conformation* (of the protein). The conformation of a single chain can be described on two organizational levels: *secondary structure* refers to the three-dimensional structure of amino acids that are nearby in the sequence. The most common elements are α-helices and β-sheets/β-stands. Loops are regions of irregular shape. *Tertiary* structure refers to the global three-dimensional form formed by the polypeptide chain that is stabilized by the interaction of the elements of secondary structure. [27] The disruption of the tertiary structure renders many proteins biologically inactive. [66] That's why it is hypothesized that protein function is determined by protein structure. This hypothesis is called the *structure-function paradigm* and the biologically active tertiary structure is referred to as the *native structure* of a protein. The native structure is thought to be uniquely determined by the primary structure. This hypothesis is called Anfinsen's dogma or the sequence-structure paradigm. [4] In the sequence-structure-function paradigm, both the sequence-structure and structure-function paradigms are assumed, that is protein function follows from protein sequence.
Protein function often involves conformational changes. The most obvious example are perhaps motor proteins that convert chemical energy into directed motion. Other examples are lids or gates in enzymes that close over the chemical substrate to shield it from the interaction with water during the catalytic reaction. In these cases protein function can not be explained by the effect of a single conformation and the structure-function paradigm was therefore extended to the structure-dynamics-function paradigm. [66]

The most dynamic proteins where function can not be understood by the effect of one or a few native conformations are the so-called *intrinsically disordered proteins.* These proteins do not fold to a well-defined structure but exist as an ensemble of widely different conformations. They often have signaling functions and adopt a folded structure when they bind to another molecule. [130]

1.2. Conformational dynamics

To study the conformational changes that are necessary for function, proteins can be investigated *in vitro* which allows much greater control over the experiment compared to the *in vivo* situation. First indirect experimental evidence for conformational changes was found in myoglobin [6]. Later more direct evidence was found with X-ray diffraction and Mössbauer experiments [6, 45, 57, 65] and more recently with Nuclear Magnetic Resonance (NMR) measurements [64, 94, 87, 16, 58, 79, 76] and single-molecule experiments [86, 85, 115, 68]. These experiments show that protein motions and conformational changes take place on a broad range of time scales. Small scale movement like atomic bond vibrations take place on the femtosecond scale, rotation of amino-acid side-chains and loop motions take place on the picosecond to microsecond time scale and larger conformational changes take place on the microsecond any beyond. [58]
The multi-scale conformational dynamics of proteins can be understood as random motion on a high-dimensional energy landscape. [46, 33] In the energy landscape model, every (high-dimensional) point in configuration space is mapped to an energy level. The dynamics on this energy landscape is a random motion that originates from thermal fluctuations.[1] A key assumption is that the energy landscape of proteins is complex: it shows a very high number of minima, separated by barriers of various height. The higher a barrier is, the more attempts it takes the system to overcome it by thermal fluctuation. So protein motion is characterized by a stop-and-go where the system appears to wait in one local minimum and then quickly transitions to another local minimum. The time that the system remains in one minimum is called the *dwell time* and the time it takes for the actual transition between minima is called *transition time*. [26, 137] Due to the complexity of the energy landscape, not all levels of detail can be observed in one experiment. Especially in experiments that probe the slow, often large scale motions, the individual minima of the finer scales are not resolved individually. Under this condition the finer details are described with topographical metaphors like "ruggedness" or "roughness" of the landscape.
In the special case when the energy landscape exhibits a few barriers that exceed

[1] The force field used for molecular dynamics simulations (see section 1.6) is an example of a model for an energy landscape.

in height all other features of the energy landscape, a more simple description of protein dynamics is possible. The conformations can then be grouped into a small number of long-lived *macro-states* (or *metastable states* or just *states*) and the kinetics of conformational change can be described with a more simple model like low-dimensional Master equations, mass-action rate equations or Markov state models with few states (see also section A.1.1). [137] For example it was shown that the folding process of some proteins (that is the process of attaining the native structure) can be well described with only two states: folded and unfolded. [23]

1.3. Binding mechanisms

Proteins typically form complexes with other molecules. Like conformational change, binding is a dynamic process that can involve interconversion between various bound and unbound conformations. Experiments have shown that there exists an interplay of binding and conformational change. X-ray crystallography experiments have shown for many proteins that alternative conformational states, distinct from the ligand-free state, can be stabilized in the ligand-bound form. Moreover NMR experiments show that the conformational state many proteins assume in the ligand-bound (holo) situation can also be assumed in the ligand-free (apo) situation, albeit with lower probability. [40, 10, 17, 59, 76, 68] This observation is interpreted with the population shift model, which states that all states are possible in the unbound and bound situations albeit with different probabilities and that binding changes the probabilities of the conformational states. [80]
It is conceivable that bound and unbound conformations are separated by a single free energy barrier that largely exceeds in height all smaller barriers. This is equivalent to assume that the transition times for binding/unbinding are much smaller than the dwell times in the bound and unbound state. The ligand then appears to "hop" in/out of the binding site. If the bound/unbound states and the conformational states too are separated by high barriers, the kinetics of coupled binding and conformational change can be described by a simple Markov state model (see e. g. figure 1.1). [137]
The four-state model of figure 1.1 shows a clear ordering of events. The conformational change can happen in the unbound state or in the bound state. This gives rise to three different mechanisms: In the *induced fit mechanism* [80] the conformational change (of the bound complex) happens after binding, hence the name that suggest that binding induces the conformational change. In the *conformational selection* mechanism [72], conformational change happens before the binding, so that one of the conformations is selected for complex formation. A third possibility is that both the induced fit pathway and the conformational selection pathway are taken by the molecular system each with a given probability. [48, 138, 55, 50]

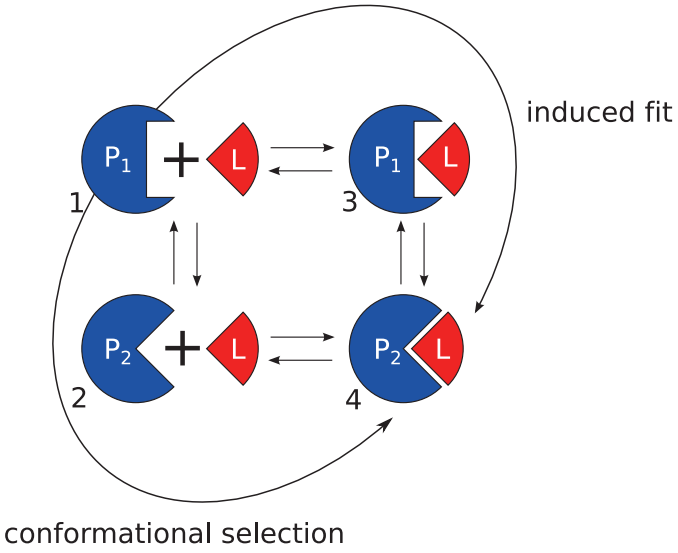

Figure 1.1.: The minimal four-state model with parallel induced fit and conformational selections pathways. [137] Because all state transitions are rare events, it is assumed that the binding/unbinding transition can not coincide with the conformational state transition such that there are no diagonal transitions from state 1 to 4 and from 2 to 3.

Under the conditions of a general, complex energy landscape that is not characterized by a small number of towering barriers, the question about the order of events becomes more involved. One of the subjects studied in this work is coupled binding and folding of a peptide to a protein. The unfolded chain of the peptide can assume a very large set of conformations and it is not clear a priori if (or not) its conformational changes can be modeled with a low number of states. The mechanism in general might not a be sequence of two steps. Also, there might be a large variety of parallel pathways going from the unfolded and unbound state to the folded and bound state, more than just the two pathways of the model figure 1.1. Parallel pathways were observed experimentally in protein binding to small ligands [29] and in protein folding [82, 51].

1.4. Markov state models

In this thesis, coupled binding and coupled conformational change of protein is studied in the theoretical framework of Markov state models (MSMs). In this section, I will briefly introduce MSMs and argue why they are a method of choice. In Markov state models of molecular systems, the long-timescale dynamics is approximated by a Markov chain on a discrete partition of the configuration space.

A MSM consists of a set of microstates and transition probabilities. A *microstate* is a structurally [105] or kinetically [96] related set of conformations. Together all microstates partition the whole conformational space. Traditionally, microstates were chosen to be metastable states. [31, 90] More recent results show that a mixture of metastable and of non-metastable states that are introduced in the transition regions (barriers) gives a more accurate model. [105] The kinetics is modeled in discrete time, that is all quantities are only defined at times that are an integer multiple of the *lag time* τ of the MSM. The stochastic, discrete-time kinetics of a MSM are governed by the transition matrix $\mathbf{T}$. Each element T_{ij} of the transition matrix is the probability that the system is in state i at time $t + \tau$ given that it was is state j at time t. Hence, the future evolution only depends on the current microstate and not on the history of past state visits.
MSMs can be estimated from time series like molecular dynamics trajectories. The estimation procedure consists of three steps: firstly, the microstates are defined with a clustering algorithm and the input time series is converted to a time series of microstate indices. Secondly, transitions between all pairs of microstates are counted and tabulated into a count matrix. And finally, the transition matrix is computed using maximum likelihood estimators.
MSMs have three key advantages over other methods:

1. MSMs can treat complex kinetics that are not well described by few states or reaction coordinates. MSMs of molecular kinetics are routinely constructed with hundreds to a few thousands of microstates and with up to hundred metastable states [91, 114, 20, 102]. Markov modeling integrates well with dimension reduction and clustering techniques [96, 112, 89] that can process high-dimensional data.

2. If they are estimated from time series data, the time series do not have to be in global equilibrium. Because only conditional transition probabilities are estimated, the starting points of the time series does not have to be drawn from the Boltzmann distribution. The time series can be short because they do not need to reach the condition where they sample the Boltzmann distribution. This property is used for distributed computing molecular dynamics where many short trajectories that were initiated from different starting conformations are combined in one MSM.

3. The transition matrix of the MSM can be used to extrapolate to arbitrary long lag times, much longer than the length of the individual trajectories from which the model was estimated, supposed that the transition matrix fulfills the Markov property. Therefore MSMs mitigate the sampling problem by extracting long-time kinetic information from short trajectories.

MSM have the disadvantage that they rely on the assumption of metastability. Because a MSM is built only on a finite number of states, there is the underlying assumption that the fast kinetics that correspond to motion inside a microstate can be neglected. This is equivalent to the assumption of a *spectral gap* or *separation of time scales*. Currently no method is known for testing for the separation of time scales a priori. However the quality of a MSM can be checked a posteriori with the implied time scales test and with the Chapman-Kolmogorov test that have become standard tests accepted by the scientific community. [105] Moreover, because MSMs with a large number of states can be estimated, it is possible to explicitly model a large number of processes that take place on different time scales.

This thesis is organized as follows: Markov state models is the common method that is used throughout the work to describe coupled binding and conformational change. We[2] study these processes by analyzing chemical relaxation experiments and by performing molecular dynamics simulations of the binding process of the 12-mer peptide PMI and the protein Mdm2 as well as of the binding process of the molecule benzamidine to the protein trypsin. In the study of chemical relaxation experiments, we take a phenomenological approach and infer the model that agrees the most with experimentally measured data. In the study of the PMI-Mdm2 system and the trypsin-benzamidine system, we take a reductionist approach by simulating the systems microscopically with molecular dynamics (MD). The MD data is then analyzed with Markov modeling methods. In the following sections 1.5 and 1.6, I will briefly introduce the techniques of chemical relaxation experiments and MD simulation.
MD simulation of coupled binding and conformational change are at the extreme forefront of what can be done with current technology and can only be accomplished with newly developed methods. I will therefore sketch in section 1.7 the basic principles of the new methods that are developed in this work.

1.5. Chemical relaxation

Chemical relaxation experiment are a powerful tool of biochemistry to gain insight into conformational changes and binding mechanisms of macromolecules and to measure transition rates. In a chemical relaxation experiment, the chemical system is prepared in a non-equilibrium initial state and is observed during its relaxation to its equilibrium state. The relaxation of the ensemble to equilibrium is recorded as a time series by monitoring some spectroscopic observable (e. g. fluorescence or absorption) over time. Stopped-flow mixing experiments are a specific type of

[2] In the following I will use "I" when referring to work done by me alone and "we" when referring to joint work done by one or several of my coauthors and me. A detailed listing of my contributions and those of my coauthors is given in appendix C.2.

relaxation experiments where the initial non-equilibrium state is created by mixing two or more solutions of different chemicals. [39]
By analyzing the time series, it is possible to infer the mechanism that the chemical system goes through on its way to equilibrium. The instantaneous amplitude of the spectroscopic signal typically represents the concentration of the bound or the unbound species. However, the instantaneous populations of the individual conformational states are typically not observed. Despite of this partial observation, it is still possible to test for the presence of an induced fit step or a conformational selection step in the mechanism. This is done by comparing the relaxation time series to predictions from models for the induced fit and conformational selection mechanisms. These models are formulated as mass-action rate equations that describe the temporal evolution of all chemical concentrations. [11, 67]
To simplify the mathematical analysis, the rate equations are typically only solved under the so-called *pseudo-first-order conditions* where it is assumed that one of the binding partners is present in large excess over the other binding partner. This mathematical simplification comes at the cost of restricting the experimentalist in the choice of concentrations. Moreover, under pseudo-first-order conditions, induced fit and conformational selection models can produce identical relaxation behavior, which renders the identification of the mechanism impossible for some ligand-protein systems if only chemical relaxation experiments are used. [137, 136, 22]
In this thesis, the model is investigated under a different condition where it is assumed that all concentrations have relaxed and are close to their equilibrium values.

1.6. Molecular dynamics simulation

With molecular dynamics we take a reductionist approach, as we seek to explain structure, mechanism and function of macromolecules from the interaction of their constituents (the atoms) and their interaction with the environment (solvent, temperature and pressure). To predict phenomena on the macromolecular scale that emerge from models formulated at the atomistic scale, computer simulations are an indispensable tool. In particular for biomolecular processes, the approach through atomistic computer simulations has been highly successful and helped to understand transmembrane transport [62, 145, 13, 71], ligand binding and receptor activation [36, 69, 93], and endocytosis [15, 5, 107] among others.
Molecular dynamics simulations are a model for dynamics and thermodynamics. The molecule and its solvent are represented as the Cartesian coordinates $\mathbf{x}$ of their atom centers (or nucleus positions). The dynamics of the system is modeled by a set of differential equations that govern the motion of the atoms, typical choices being Newton's equations of motion, the Langevin equations, or Newton's

equations coupled to "thermostat" and "barostat" models. [47] The potential energy function in these equations is specifically chosen to model atomistic/molecular interactions and is called a "force field". For the simulation of proteins, a variety of well-established force fields exists, which are continuously validated and improved by the scientific community. [77] Thermodynamic properties of the molecular system are modeled by the Boltzmann distribution $p_B(\mathbf{x}, \mathbf{p}) = \exp[\beta F - \beta H(\mathbf{x}, \mathbf{p})]$ where $H(\mathbf{x}, \mathbf{p})$ is the Hamiltonian as defined in the force field, $\mathbf{p}$ are the momenta, $\beta = k_B T$, T is the temperature, k_B is the Boltzmann constant and F is the Helmholtz free energy. The set of all possible conformations $\mathbf{x}$, momenta $\mathbf{p}$ and their probability density $p_B(\mathbf{x}, \mathbf{p})$ for fixed H, β and atom count is called the (canonical) *thermodynamic ensemble*.
In a MD simulation the equations of motion are integrated in time, which results in MD trajectories. Because the initial conditions have random velocities or because the model contains a random force like in the case of the Langevin equations, the trajectories are random. Because of that randomness, MD simulation is typically complemented with statistical analysis to compute averages and probabilities of kinetic and thermodynamic quantities. The easiest way of estimating an equilibrium expectation value is to compute the (empirical) mean from a MD trajectory that is long enough to sample from the Boltzmann distribution. The easiest way to compute kinetic properties is to estimate a MSM or rate model directly from the MD trajectories. These two methods have some drawbacks as will be explained in the next section. Advanced methods, that lead to a reduced error will be introduced in this thesis.

1.7. Methodological developments

Time scales of many important biological processes are still out of reach for unbiased MD simulation. For example, although downhill processes such as protein-ligand association to the bound conformation can be spontaneously sampled [36, 102, 114, 20], the dissociation of stable inhibitory complexes can involve timescales of hours or longer [129], which is beyond the scope of current MD methods. Biased simulations (see section below) are used to speed up events that are rare in the physical ensemble. However, they can not be directly used to compute kinetic properties in the physical ensemble. Approximate methods to reweight kinetic properties from a biased simulation exist, but they rely on a one- or two-dimensional projection of the conformational space [12, 140] and are therefore not suited to study the complicated multi-state kinetics that we expect in the case of coupled folding and binding.
Therefore a larger part of this work is devoted to the development and testing of new enhanced sampling methods. The necessity of that work became evident in

two recent studies about the p53-Mdm2 system, which is a peptide-protein system very similar to the PMI-Mdm2 system studied in this thesis. In none of the studies the authors were able to correctly simulate the unbinding process. [147, 144] In [144], the dissociation rate is over-estimated by about five orders of magnitude. In [147], no complex dissociation is reported.

Before describing the new developments in section 1.7.2 I will first briefly review what is known about free energy calculation and its use for estimation of molecular kinetics in the next section.

1.7.1. Free energy calculation

Stationary expectations for systems with rare events can be calculated with Boltzmann reweighting. For an ergodic system, the stationary (equilibrium) expectation value of an observable can be approximated by computing the empirical mean of the observable from all conformations generated by a sufficiently long MD simulation. In the limit of long simulations, the mean will converge to the true expectation value. If the system has too high free-energy barriers such that the states of the system can not be sampled with available computing resources, a bias energy can be introduced that is added to the physical model Hamiltonian of the system and that reduces or removes the barriers in the system. Popular choices of bias are umbrella sampling simulations [123, 116], multi-temperature simulations [81, 56, 117] and others [54, 78, 73]. Expectation values can still be estimated as (weighted) means from the trajectory data of the MD simulation that was run with the biased Hamiltonian. The effect of the bias energy on the mean is corrected by multiplying every term by a reweighting factor (see section A.2).

In principle, only one biased Hamiltonian is necessary for reweighting approaches. However, in practice, multiple biased Hamiltonians (or multiple temperatures) are defined and used to simulate many biased MD trajectories. [81, 118] The resulting simulation data then have to be combined into one estimate of the observable of interest. Standard algorithms that solve this problem are the weighted histogram analysis method (WHAM) [43, 74] and binless WHAM, also known as multi-state Bennett acceptance ratio (MBAR) [133, 8, 116, 70, 113]. These methods treat their input data as uncorrelated samples of the ensemble distribution and are therefore not suitable for simulation data with long correlation times in some variables, as it is common for unbiased MD simulations and biased simulations with slow unbiased coordinates [108, 63]. This disadvantage can be mitigated by replacing the assumption of uncorrelated input samples by the less strong assumption that the input data was generated by a MSM as will be explained in the next section.

1.7.2. Enhanced estimation of kinetics

Besides reliable estimation of free energies, the goal of this work is to compute kinetic properties of protein-ligand systems. Still, kinetics of reversible systems are inseparably connected to free energies. This can be easily seen for Markov state models: MSMs can not only be used to model trajectories but also to describe the time evolution of state occupancy probabilities (populations). Let $p_i(t)$ be the population of microstate i at time t. Then the population of the microstates at time $t+\tau$ can be approximated with $p_j(t+\tau) = \sum_k T_{ij} p_i(t)$. In the limit of an infinitely long lag time,[3] any initial vector $\mathbf{p}(0)$ will converge to the stationary vector $\boldsymbol{\pi}$. In case of molecular systems $\boldsymbol{\pi}$ can be identified with a coarse-grained version of the Boltzmann distribution, which establishes a direct connection between MSMs and thermodynamic quantities. [111, 124]
MSM theory provides two equations that link kinetics and thermodynamics even closer than that and that are of central importance for enhanced estimation of kinetics:

1. The molecular systems in this work are studied in the absence of external driving forces which implies microscopic reversibility of the dynamics. Under these conditions, the detailed balance relation $\pi_i T_{ij} = \pi_j T_{ji}$ holds for every pair of microstates of the MSM. [111, 124] From this follows that every transition matrix element T_{ij} can be computed from the transition matrix element of the reverse transition T_{ji} and the stationary probabilities π_i and π_j. We make use of this fact to omit sampling the rarest transitions that correspond to the smallest T_{ij}. For this purpose, the reverse transition probability T_{ji} has to estimated. Also π_i needs to be known for all i, e. g. from a previously conducted biased MD simulation / free energy calculation.

2. The transition matrix of the Markov model can be repeatedly applied to any initial probability vector until the stationary vector is found. $\lim_{k\to\infty} T_{ij}(k\tau) = \lim_{k\to\infty} \left(\mathbf{T}^k(\tau)\right)_{ij} = \pi_j$ We estimate $\mathbf{T}(\tau)$ from many short MD trajectories that are shorter than the time required for the system to relax to equilibrium. Thus we do not estimate $\boldsymbol{\pi}$ directly but extrapolate to it with the help of the MSM.

It would be ideal to use these two properties of MSMs in one algorithm to estimate the kinetics and stationary properties for molecular systems with rare events. A sketch of such an ideal algorithm is shown in figure 1.2.

[3] One has to additionally assume irreducibility and positive recurrence of all Markov states. [92] Practically, we work with a reduced set of Markov states defined such that the conditions are always fulfilled.

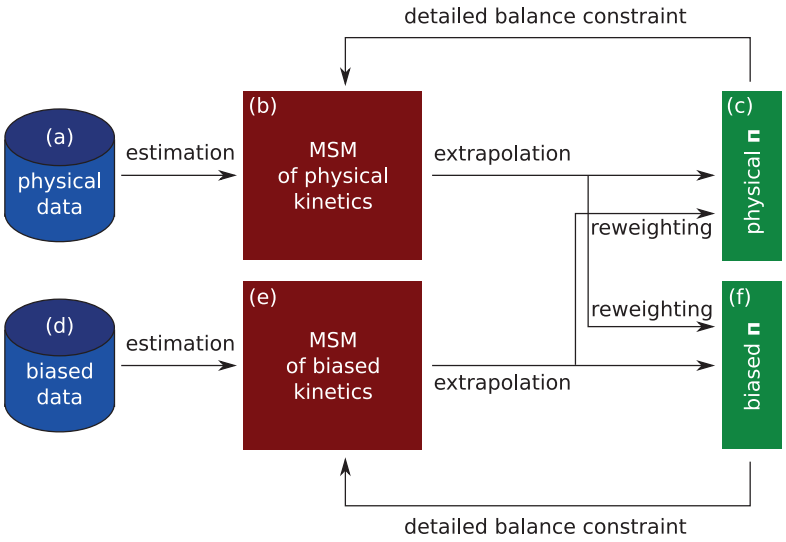

Figure 1.2.: scheme for an algorithm for the enhanced estimation of kinetics that leverages the principles of Boltzmann reweighting, detailed balance, and the Markov property

From a set of short MD simulation trajectories (a in figure 1.2), a MSM (b) is estimated. We use MSM property 2 (extrapolation) to compute the stationary vector of the physical ensemble (c). The physical (unbiased) simulation alone might not have sampled all microstates in both the forward and backward directions, so there might be microstates i whose stationary weight π_i can not be computed from the MSM of the physical ensemble. This missing information is supplied by the stationary vector of the biased ensemble (f) which can be reweighted towards the physical ensemble via Boltzmann reweighting (see appendix A.2). Knowing the full physical stationary vector (c), one can now use MSM property 1 (detailed balance) to compute the missing transitions rates in the transition matrix (b).
Since MSMs are an integral part of this ideal algorithm, we can also permit that the biased simulation data (d) to consist of short out-of-equilibrium trajectories. The biased stationary vector (f) is computed from a MSM of the biased data (e). How to systematically construct such an algorithm is the content of publications [P2] and [P3] and is summarized and discussed in section 2.1. Precursor algorithms that do not exploit all properties of MSMs and Boltzmann reweighting or are restricted to the estimation of stationary probabilities were developed in references [125, 141, 84, 128] by various authors.

2. Summary and Discussion of the content of this thesis

In this thesis, three aspects of macromolecular binding coupled to conformational change are addressed.

1. The study of complicated binding kinetics with MD simulations requires novel methods for rare-event sampling. Existing methods are either suitable to study the kinetics of a) macromolecular systems with a very large number of conformational states that can easily be sampled with MD simulations or b) systems with few but with very stable states. However no method is available for studying systems that have many states, part of them very stable. This is exactly the setting in the binding of flexible ligands to proteins. For this purpose, building on the groundwork done by Trendelkamp-Schroer, Wu, and Mey [125, 141, 84] two new analysis methods named *TRAM* and *TRAMMBAR* are developed and validated with conceptual models.

2. Equipped with these new analysis methods, we study coupled processes of conformational change and binding in two molecular systems: a) the interaction between the protein trypsin and the ligand benzamidine and b) the interaction between the protein fragment $^{25-109}$Mdm2 and the peptide PMI. While the trypsin-benzamidine system is to be seen as a benchmark for the new TRAM method, we present novel results about the dissociation rate and the binding mechanism for the PMI-Mdm2 interaction. For the first time, a dissociation rate on the seconds time-scale is computed from MD simulations without resorting to methods that speed up transitions by externally forcing the system.

3. In the third part, we investigate processes of coupled conformational change and binding by analyzing experimental data from stopped-flow (relaxation) experiments. Typically these experiments are done under pseudo-first-order conditions for the sake of a simple mathematical analysis (see introduction, section 1.5). I generalize the kinetic models employed in the analysis from pseudo-first-order to more general conditions. This allows to infer the binding mechanism for molecular systems for which no such distinction was possible in the past. We correctly identify the mechanism and determine transition rates in the coupled binding and conformational change process between the

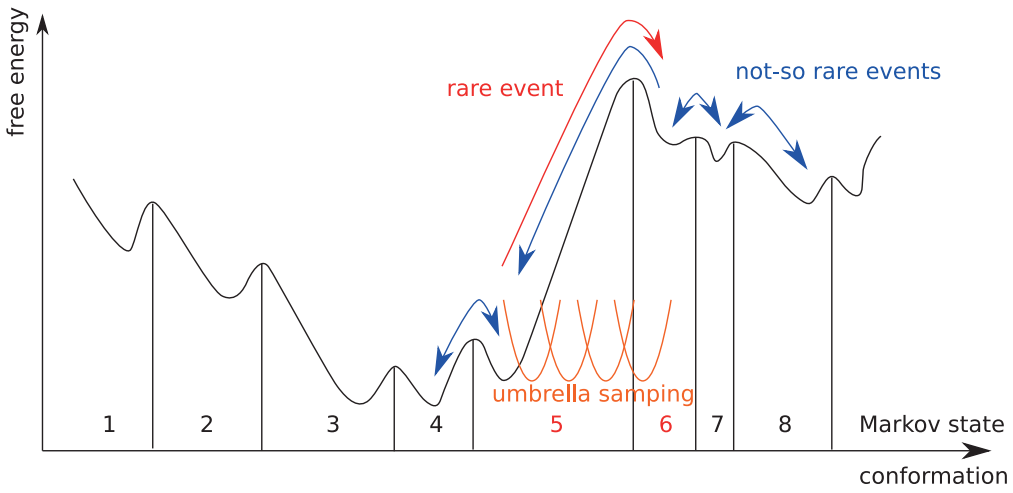

Figure 2.1.: Schematic free-energy landscape in which TRAM and TRAMMBAR can be optimally applied. The exemplary landscape features many small energy barriers that are small enough to be sampled with state-of-the-art brute-force MD simulations and one large hill-slope that is easy to cross in the downhill direction but difficult to cross in the uphill direction. An effective simulation setup for sampling the whole energy landscape would consist in a set of biased MD simulations that sample only the steep hill-slope (orange umbrella potentials) and a set of unbiased MD simulations that were started from many different locations in the free-energy landscape. The biased MD simulations allow to estimate the ratio of stationary weights π_5/π_6 and the unbiased MD simulations allow to estimate the probability of transitioning downhill $T_{6,5}$. The probability of transitioning uphill $T_{5,6}$ can be calculated from the detailed balance condition $\pi_5 T_{5,6} = \pi_6 T_{6,5}$.

protein recoverin and the rhodopsin kinase peptide fused to the B1 domain of immunoglobulin protein G.

2.1. New methods for enhanced estimation of kinetics

The new methods presented in this thesis allow to estimate a MSM that encodes the full multi-state kinetics and thermodynamics of the system. This idea was pioneered in [142, 141, 84] and applied to simple molecular systems and is significantly extended in this thesis.

We want to estimate full kinetics in a situation in which the free-energy landscape exhibits multiple minima separated by barriers of different height (see figure 2.1 as

example). To this end we introduce Markov states, that partition the conformational space. Conventional, unbiased MD simulations that are started from different states are able to explore large parts of the energy landscape. However, some Markov state transitions can only be sampled in one direction with the available computational resources[1].
To infer the probabilities of the missing transitions, the unbiased MD simulations are complemented here by a series of biased MD simulations (see figure 2.1). These biased simulations will be used to estimate the free energy differences between the Markov states, which can be used to infer the missing transition rates via the detailed balance relation. The Hamiltonians in the biased simulations have to be chosen such that the biased simulations sample the transition regions between the Markov states where bidirectional transitions are missing in the unbiased simulations. To implement this idea in a practical algorithm, two tasks had to be solved in this work:

1. A modeling / mathematical task that consists of (i) developing a probabilistic model that describes the simulation data and (ii) deriving a maximum likelihood estimator for the parameters of the model (transition rates and free energies). This model should implement the principles of Markov state modeling, microscopic reversibility, and Boltzmann reweighting as introduced in sections 1.4 and 1.7.

2. A physical / applied task that consists of choosing the Hamiltonians for the biased simulations for the molecular system at hand such that the relevant transitions can be sampled.

In the following, I summarize and discuss results of mathematical modeling (task 1) that lead to the TRAM and TRAMMBAR estimators. Both estimators are tested with conceptual models and their efficiency is compared with other methods. Task 2 is more specific to the applications and will be addressed in sections 2.2 and 2.3.

2.1.1. TRAM

In this section, I assume that the reader is familiar with Markov state modeling, free energy calculation and maximum-likelihood estimators. The non-expert reader is referred to the introductory text in the appendix A.

[1] One could also imagine a situation, where the unbiased simulations are not long enough to sample some state transition *in any* of the forward and backward directions. A conceivable strategy then would be to start simulations from the transition state between the unconnected states. This approach was explored for simple systems by Trendelkamp-Schroer. [124, 126] Identifying the transition state for complicated molecules is a challenge in itself, which can be addressed e. g. with path sampling methods [97].

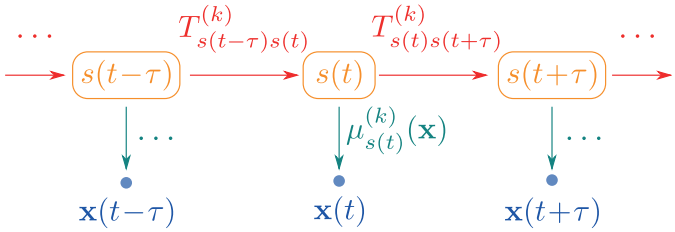

Figure 2.2.: The doubly stochastic model used in TRAM for a single MD trajectory from thermodynamic ensemble k. On the upper level the trajectory is modeled as a discrete-time, discrete-space trajectory that jumps from Markov state to Markov state. On the lower level each conformation $\mathbf{x}(t)$ is modeled as being drawn from a local (conditional) equilibrium distribution $\mu_i^{(k)}(.) = \mu(. \mid S_i, k)$ that belongs to the Markov state S_i which is currently visited by the discrete trajectory.

In this thesis, we formulate a joint probabilistic model for MD trajectories from multiple simulations with different Hamiltonians. This model uses ideas from MSMs (see section A.1.1) and binless reweighting (MBAR, see section A.2.3). The most important element in the model is a set of MSMs, one MSM for every thermodynamic ensemble. We call this set of MSMs a *Multi-Ensemble Markov Model* (MEMM). We call the estimator for the model parameters the *Transition-based Analysis Method* (TRAM).
As detailed in the publication [P2], two mathematical components are needed to formulate the model:

1. Transition matrices $\mathbf{T}^{(k)}$ that model the kinetics in each thermodynamic ensemble k (no conceptional difference is made between biased and unbiased Hamiltonians)

2. Per-microstate reweighting factors $\mu_i^{(k)}(\mathbf{x})$ that reweight a sampled conformation $\mathbf{x}$ towards the Boltzmann distribution of ensemble k restricted to microstate S_i (towards the "local equilibrium distribution" of microstate S_i). $\mu_i^{(k)}(\mathbf{x}) = \mu(\mathbf{x} \mid i, k)$ is defined similarly as the reweighting factor $\mu^{(k)}(\mathbf{x})$ in MBAR (see section A.2.3) with the difference that the probability $\mu_i^{(k)}(\mathbf{x})$ is conditioned on the fact that $\mathbf{x}$ is located in microstate S_i.

Consider an MD trajectory from a simulation with Hamiltonian $H^{(k)}$. The trajectory $\{\mathbf{x}(t)\}_{t=0\ldots T}$ is first converted into a trajectory of microstate indices $\{s(t)\}_{t=0\ldots T}$ (defined such that $\mathbf{x}(t) \in S_{s(t)}$) and then subsampled with a lag time τ to obtain the discrete-space discrete-time sequence $\{s(n\tau)\}_{n=0\ldots N}$. Using this discrete sequence as an auxiliary quantity, $\{\mathbf{x}(n\tau)\}_{n=0\ldots N}$ is modeled with a doubly stochastic process

as shown in figure 2.2. On the upper level of the process, the time evolution is completely modeled as a Markov process between the discrete states visited at different times. On the lower level, the probability of observing the conformation $\mathbf{x}(n\tau)$ given that the Markov state at time $n\tau$ is $s(n\tau)$ is modeled with the reweighting factor $\mu^{(k)}_{s(n\tau)}(\mathbf{x}(n\tau)) = \mu(\mathbf{x}(n\tau) \mid s(n\tau),\, k)$.
This allows to factorize the probability of observing the complete sequences $\{\mathbf{x}(n\tau)\}$ and $\{s(n\tau)\}$ as

$$\mathbb{P}(\{\mathbf{x}(n\tau)\},\, \{s(n\tau)\} \mid k) = \mathbb{P}(\{\mathbf{x}(n\tau)\} \mid \{s(n\tau)\},\, k) \cdot \mathbb{P}(\{s(n\tau)\} \mid k) \tag{2.1}$$

The probability of observing the sequence of Markov states is modeled following the standard approach (A.5) as

$$\mathbb{P}(\{s(n\tau)\} \mid k) = \mathbb{P}(s(0) \mid k) \prod_n T^{(k)}_{s(n\tau-\tau),s(n\tau)} \tag{2.2}$$

where $\mathbb{P}(s(0) \mid k)$ is the probability of starting the trajectory in Markov state with index $s(0)$. The probability of observing the sequence of conformations, given a fixed sequence of Markov states is modeled as

$$\mathbb{P}(\{\mathbf{x}(n\tau)\} \mid \{s(n\tau)\},\, k) = \prod_n \mu(\mathbf{x}(n\tau) \mid s(n\tau),\, k). \tag{2.3}$$

The probability of observing many trajectories, possibly simulated in different thermodynamic ensembles k is simply the product of the probabilities of the individual trajectories (assuming independence of the simulations). This probability of observing all trajectories is used as the likelihood function for the parameters of the MEMM to construct a maximum-likelihood estimator. The likelihood can be written in a more compact form by introducing a count matrix $\mathbf{C}^{(k)}$ for the state transitions in thermodynamic ensemble k like in equation (A.6)

$$\mathcal{L}_{\mathrm{TRAM}} = \prod_k \prod_{i,j} \left(T^{(k)}_{ij}\right)^{C^k_{ij}} \prod_i \prod_{\mathbf{x} \in X^{(k)}_i} \mu^{(k)}_i(\mathbf{x}) \tag{2.4}$$

$X^{(k)}_i$ denotes the set of conformations $\mathbf{x}$ contained in Markov state S_i that have been generated by the simulation run with Hamiltonian $H^{(k)}$.
The principles of Boltzmann reweighting and detailed balance are modeled as constraints on the values of the model parameters and are enforced during the likelihood maximization. To implement Boltzmann reweighting, we couple the per-microstate reweighting factors $\mu^{(k)}_i(\mathbf{x})$ of ensemble k to the global reweighting factors $\mu^{(\mathrm{ref})}(\mathbf{x})$ of an (arbitrarily chosen) reference ensemble. This can be done by first expressing the per-microstate reweighting factors in terms of the global reweighting factors $\mu^{(k)}(\mathbf{x})$ (see equation (A.10)) and a normalization constant

$e^{-f_i^{(k)}} := \mathbb{P}(\mathbf{x} \in S_i,\, k)$ by using the definition of the conditional probability and then relating $\mu^{(k)}(\mathbf{x})$ to $\mu^{(\mathrm{ref})}(\mathbf{x})$ with the help of the bias energy[2]

$$\begin{aligned}\mu(\mathbf{x} \mid S_i, k) &= \frac{\exp\left[-\beta U^{(k)}(\mathbf{x})\right]\chi_i(\mathbf{x})}{\exp[-f_i^{(k)}]} = \\ &= \frac{\mu^{(\mathrm{ref})}(\mathbf{x})\exp\left[-\beta U^{(k)}(\mathbf{x}) + \beta U^{(\mathrm{ref})}(\mathbf{x})\right]\chi_i(\mathbf{x})}{\exp[-f_i^{(k)}]}\end{aligned} \tag{2.5}$$

where χ_i is the indicator function for microstate S_i. This allows to incorporate all bias energies into the estimation of the model parameters, like in MBAR (see section A.2.3). Coupling together the estimates for the reweighting factors $\mu_i^{(k)}(\mathbf{x})$ across different ensembles allows to infer the equilibrium distribution of an ensemble that was not sampled (either not sampled completely or not sampled at all).
To implement detailed balance we couple the stationary probabilities to the transition probabilities:

$$\begin{aligned}e^{-f_i^{(k)}} T_{ij}^{(k)} &= \mathbb{P}(s(t+\tau) = j \mid s(t) = i,\, k)\mathbb{P}(s = i,\, k) = \\ &= \mathbb{P}(s(t+\tau) = i \mid s(t) = j,\, k)\mathbb{P}(s = j,\, k) = e^{-f_j^{(k)}} T_{ji}^{(k)}\end{aligned} \tag{2.6}$$

The algorithm for optimizing the likelihood under these constraints is given in [P2]. We validated TRAM with replica exchange simulations of a small molecule, the alanine dipeptide and with a set of biased and unbiased simulations of diffusion on a two-dimensional energy landscape. For both model systems, the trajectories needed to reach convergence can be shorter for TRAM than for MBAR. This is expected because MBAR requires that all conformations are drawn from the global equilibrium (Boltzmann) distribution which is an assumption that is violated in our test data. TRAM, in contrast, does not assume global equilibration of the data. TRAM relies on the validity of the MSM approximation and uses the MSM to extrapolate towards global equilibrium (see also section 1.7.2) instead of requiring it from the distribution of the input data.
The analysis of the replica exchange simulation data of the alanine dipeptide with TRAM produced results that agree with previous studies but relies on conditions that are typical not fulfilled in replica exchange simulations of larger molecular systems. (See next section for discussion and solution of that problem.)

[2] We assume that the zero of all energies have been shifted such that the partition function of the reference ensemble is one. Alternatively one may interpret all probabilities $e^{-f_i^{(k)}}$ as being implicitly normalized (divided) by the partition function of the reference ensemble.

2.1.2. TRAMMBAR

Coupling biased MD simulations with (Hamiltonian) replica exchange (HREMD) helps in enhancing the exploration of the conformational space and helps to mitigate sampling problems that arise from unfavorable interactions of the bias energy with the underlying physical Hamiltonian (see A.2.4). Therefore, ideally all biased simulations should be carried out with replica exchange. This raises the questions whether HREMD simulations too can be analyzed with TRAM and which physical quantities can be estimated from HREMD data.

When accepted, the replica exchange step interrupts the trajectories simulated with a constant bias potential. Therefore the long-time kinetics of the replica exchange simulations does not correspond to the kinetics of any of the ensembles individually. Since the main interest in this work is to compute the kinetics of the possibly slow conformational changes and unbinding events, it is necessary to find a way to recover the long-time kinetics from HREMD simulations.

The first attempt to estimate kinetics from HREMD data was made in our publication [P2]. In our simulations of the alanine dipeptide, we first split the trajectories into trajectory pieces that end whenever the Hamiltonian changed. We then estimated a MEMM with TRAM from these pieces by treating them as independent trajectories. We found that all transition matrices of the MEMM fulfill the Markov condition (A.4) at a lag time that was shorter than the time interval between exchange attempts. So, one way of using replica exchange data with TRAM is to exchange Hamiltonians infrequently with an interval that is longer than the shortest possible lag-time of the MSM. In general, this is not a useful strategy, for the following reasons: HREMD simulations become more effective, the more frequently exchanges are attempted. [101] In contrast, for the estimation of a MSM for proteins, typically a long lag time between 10 ns and 100 ns is needed [83] which is much longer than the recommended exchange interval for HREMD simulations [101]. Moreover, the length of the HREMD exchange interval is chosen early in the design/preparation stage of a simulation whereas the smallest lag time necessary to estimate the MSM/MEMM can currently only be revealed after the simulation, during the data analysis.

To solve the problem of estimating kinetics from HREMD data, we propose in [P3] to split the computational effort into two types of simulations, called *kinetic* and *equilibrium* simulations. The equilibrium simulations are run simultaneously and coupled with replica exchange. The kinetic simulations are run independently from the equilibrium simulations without exchange of Hamiltonian. Because of the good mixing in the HREMD simulations, one can assume that the equilibrium simulations quickly relax to a state where they sample from the Boltzmann distributions of their respective ensembles. That is, conformations generated with the Hamiltonian $H^{(k)}$ are assumed to be drawn from the corresponding Boltzmann distribution.

The equilibrium simulation data alone then fulfills the necessary assumptions to be analyzed with MBAR[3]. The kinetic simulations, in contrast, are modeled with a Markov model exactly like in TRAM. Because the kinetic simulations do not participate in the exchange of Hamiltonians, relatively long trajectories are available and the choice of lag time is not substantially restricted.
To analyze both data sets simultaneously, we propose a new maximum likelihood estimator called *TRAMMBAR* that combines elements of the TRAM and MBAR estimators. Let $X_{\mathrm{MBAR}}^{(k)}$ denote the equilibrium data from ensemble k i. e. the set of all conformations from the replica exchange simulation that was run with Hamiltonian $H^{(k)}$. Let $X_{\mathrm{MBAR}} = X_{\mathrm{MBAR}}^{(1)} \cup \ldots \cup X_{\mathrm{MBAR}}^{(K)}$ be the set of all conformations from all replica exchange simulations. Let $X_{\mathrm{TRAM}}^{(k)}$ be the conformations sampled by the kinetic simulation that was run with Hamiltonian $H^{(k)}$ and let $X_{\mathrm{TRAM}} = X_{\mathrm{TRAM}}^{(1)} \cup \ldots \cup X_{\mathrm{TRAM}}^{(K)}$. Furthermore let $\mathbf{C}^{(k)}(\tau)$ be the count matrix of the state-to-state transitions in the kinetic simulations conducted with Hamiltonian $H^{(k)}$. Because the equilibrium data and the kinetic data are generated by independent simulations, the TRAMMBAR likelihood takes the form of the product

$$\mathcal{L}_{\mathrm{TRAMMBAR}}(X_{\mathrm{MBAR}},\, X_{\mathrm{TRAM}}, \{\mathbf{C}^{(k)}\}_k) = \mathcal{L}_{\mathrm{TRAM}}(X_{\mathrm{TRAM}}, \{\mathbf{C}^{(k)}\}_k) \cdot \mathcal{L}_{\mathrm{MBAR}}(X_{\mathrm{MBAR}}) \tag{2.7}$$

where $\mathcal{L}_{\mathrm{TRAM}}$ is defined in equation (2.4) and $\mathcal{L}_{\mathrm{MBAR}}$ is defined in equation (A.11). The maximization of the TRAMMBAR likelihood is carried out under the same constraints that are used for the maximization of the TRAM and MBAR likelihoods. For all conformations in X_{TRAM}, microstate-dependent reweighting factors $\mu_i^{(k)}(\mathbf{x})$ are defined just like in TRAM. For all conformations in X_{MBAR} microstate-independent reweighting factors $\mu^{(k)}(\mathbf{x})$ are defined just like for MBAR (see section A.2.3). All reweighting factors can be coupled to the global reweighting factors $\mu^{(\mathrm{ref})}(\mathbf{x})$ of an arbitrarily chosen reference ensemble with the help of the bias energies. Furthermore the detailed balance constraint (2.6) is enforced in all ensembles where kinetic simulation data is available.
The TRAMMBAR algorithm was tested with a 2-D conceptual model of protein-ligand binding in [P3] and applied to the coupled folding and binding in the PMI-Mdm2 system. The test shows that TRAMMBAR needs about 50 times less simulations data than MSMs and than direct sampling for computing the binding free energy and the dwell-time of the bound state. The improvement was expected

[3] MBAR's assumption that simulations in different thermodynamic ensembles are independent is still violated for HREMD data. However practical applications [1, 41] show that MBAR is currently one of the best estimators for analyzing HREMD data. Modeling the dependence between the different replicas would require a much more complicated stochastic model which is beyond the scope of this thesis.

because TRAMMBAR is the only method in this comparison that can integrate the biased simulation data into the estimation. A drastic reduction in simulation effort is possible with TRAMMBAR, especially for the estimation of kinetic quantities which is confirmed in the application to the PMI-Mdm2 system.

2.2. Coupled binding and conformational change in the trypsin-benzamidine system

In [P2], we applied TRAM to the binding between the serine protease trypsin and its inhibitory ligand benzamidine. Benzamidine binds to the Asp-189 side chain of trypsin which is located at the bottom of a deep binding pocket. It was found in [102] that benzamidine binding can involve a conformational change of trypsin. The binding site is accessible via one of two channels, the S1 channel and the S1* channel. Only one of the channels can be present at the same time [102]. Moreover, the side chain of Trp-215, which is located on the protein surface, can flip and close over the exit of the S1 channel. [20, 102]
We were interested in reliably determining the unbinding mechanism and computing the dissociation rate in the trypsin conformation with open S1 channel. Even in that case, the unbinding mechanism is relatively complicated and involves breaking of salt bridges between Asp-189 and benzamidine, going through intermediates with water-mediated interaction between Asp-189 and benzamidine [121, 102], and benzamidine exiting the channel. The conformational flexibility of trypsin is not restrained in our simulations, so it is expected to see conformational changes of trypsin.
In the simulations, at least two slow processes can take place: binding/unbinding and conformational change. We chose to enhance the sampling of the unbinding with biased simulations (umbrella sampling) because dissociation is the slowest process that takes place on the millisecond time-scale. We do not introduce biases to enhance the conformational change (a) because it is probably faster [102] than dissociation, so there is a chance of sampling it spontaneously and (b) because there are multiple possibilities of conformational change. In addition to Trp-215 flipping there is conformational variability in the Asp-189 loop [102] and isomerization of a disulfide bond near the binding pocket among others [P2]. Biasing in multiple dimensions quickly becomes computationally very expensive and is only possible for two or three dimensions [9, 139]. Moreover to define biased Hamiltonians, some initial knowledge about the expected conformational changes is needed, which is not available for all conformational changes before running the simulations.
9.2 μs of umbrella sampling simulations were run using a series of harmonic biasing potentials that restrain benzamidine to different positions along the binding channel. The first umbrella restrains benzamidine to the Aps-198 bound state, the last

umbrella restrains benzamidine to a state that encompasses a mixture of unbound conformations and conformations with benzamidine loosely bound to the surface of trypsin. We additionally included 49.1 μs of unbiased MD simulation data taken from reference [20]. In both the biased and the unbiased simulations, we see closing of Trp-215 over the exit of the channel. This closing is a relatively rare event. Despite the fact that we see a few opening and closing events, we can not assume that the frequencies of observing the open/closed conformations in our simulation data are representative for their true equilibrium probabilities. That is why analysis with TRAM, which does not rely on global equilibration, is necessary in this setting (see section 2.1.1).

We analyze the data with TRAM and find a dissociation rate of $k_{\text{off}} = 1170\,\text{s}^{-1}$ with 95% confidence interval of $[617\,\text{s}^{-1}, 2120\,\text{s}^{-1}]$. The experimental rate is $600\,\text{s}^{-1}$. [52] Along the binding/unbinding pathway that we study, we find that Trp-215 closes over the binding channel either if benzamidine is fully bound or located at the exit of the binding channel (see figure 5 in [P2] and figure 2.3). The closing of Trp-215 while benzamidine is bound could be called an induced fit mechanism. However, the complete binding mechanism is more complicated and involves more conformational changes as was shown in reference [102].

The dissociation rate of benzamidine from trypsin with formed S1 channel was studied by other groups either by using the hyperdynamics method [121], adaptive multilevel splitting [120], or MSMs [20]. The work in [121] is based on the hyperdynamics approximation that requires a careful choice of a bias that does not change the saddle points of the physical energy landscape. The adaptive multilevel splitting method [120] is only efficient if there are no metastable intermediates on the transition from strongly bound to unbound, a problem also found in early versions of the transition interface sampling (TIS) method [131] that was solved by running the TIS method in an adaptive fashion [37].[4] In the simulation study in reference [20], no complete dissociation of trypsin and benzamidine was observed in a single trajectory but only steps in the unbinding direction distributed over multiple trajectories. Under ideal conditions, a MSM *can* be estimated in this situation and the unbinding rate can be estimated from the MSM. However this requires that an accurate reaction coordinate was defined that exclusively describes the binding/unbinding transition. If the reaction coordinate is contaminated with random fluctuations that come from other degrees of freedom, not related to the binding process, this noise could be mistaken for a motion in the unbinding/binding direction and one could gain the impression of seeing partial unbinding while actually no such event took place. Therefore, extra care is necessary in estimating MSMs under these condition. [34, 100] In our work, we took a safer and more robust

[4] private communication on *NAMD Developer's Workshop,* University of Chicago, May 26-27, 2016

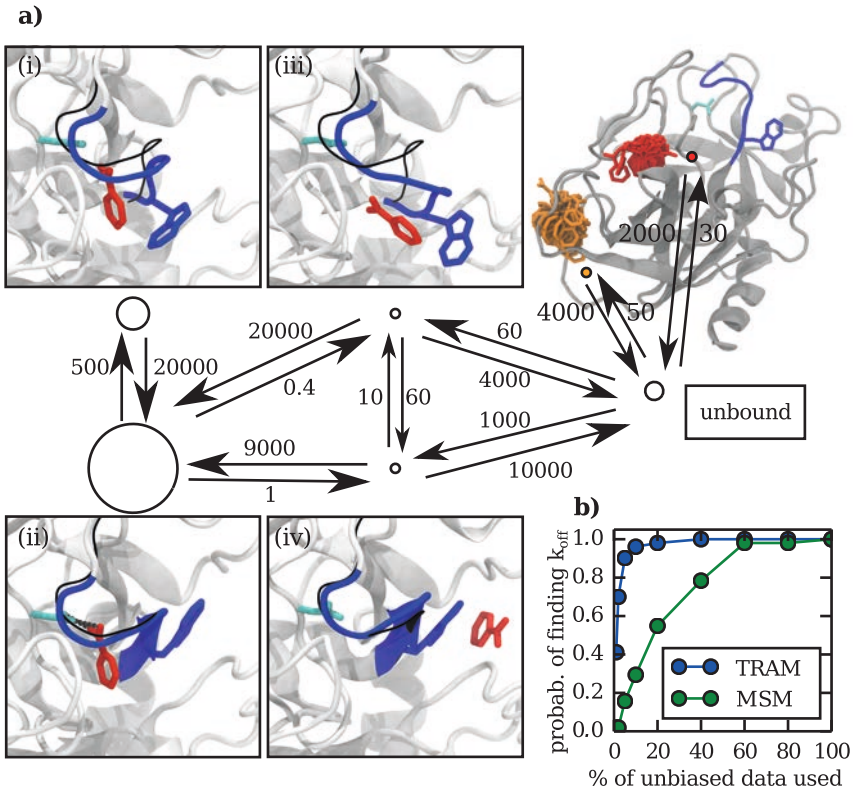

Figure 2.3.: Figure modified from [P2]. a) Coarse-grained kinetic network of the MEMM. The largest transition rates (where at least one direction exceeds 1/ms) between these macrostates, the unbound state and two alternatively bound states are shown as arrows. Units are events per millisecond. b) Efficiency of TRAM in the estimation of unbinding kinetics compared with a MSM built from the same unbiased data. Shown is all all all the probability that k_{off} calculated from a bootstrap sample falls into the interval $[0.5 \log k_{\mathrm{off}}^{\mathrm{all}}, 2 \log k_{\mathrm{off}}^{\mathrm{all}}]$ off where $k_{\mathrm{off}}^{\mathrm{all}}$ is the TRAM estimate calculated using all data.

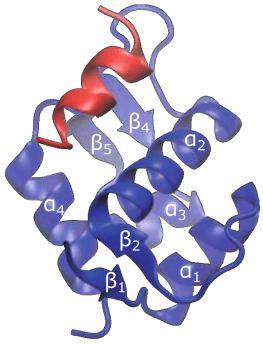

Figure 2.4.: co-crystal structure of the protein fragment $^{25-109}$Mdm2 and the peptide PMI reproduced from the crystal structure 3eqs [95]. Missing residues were remodeled with profix [143].

approach to the problem, because we have trajectories with full binding events and a series of biased simulations that clearly connect the bound and unbound state. The dissociation rate is computed with TRAM which relies on Markov state modeling and on the undisputed physical principles of detailed balance and Boltzmann reweighting.
A important limitation of this work is that we restricted all simulation and analysis to the trypsin conformation with open S1 channel. Work by Plattner *et al.* [102] indicates that this conformation is not the most probable conformation of trypsin in solution. Therefore the experimental dissociation rate of benzamidine and trypsin might be dominated by dissociation from other conformations and comparison of our result for k_{off} with the experimental value has to be taken with care. In future work, additional umbrella sampling simulations of the binding/unbinding process could be conducted with the different conformations that were identified in reference [102]. These biased simulations could be added to simulation data generated in this work and all data could be analyzed together with TRAM.

2.3. Coupled folding and binding in the PMI-Mdm2 system

In [P3] we studied the process of coupled folding of the peptide PMI during binding to the protein fragment $^{25-109}$Mdm2 (from this point called Mdm2). This molecular system is interesting for the following reasons:

1. The co-crystal structure of the Mdm2 fragment with the peptide PMI [95] shows that PMI binds as a helix (figure 2.4). Our MD simulations of PMI in solution without its binding partner show that PMI is at most 40% helical when unbound. [P3] So the binding mechanism must involve folding. The Mdm2 fragment does not show any large conformational changes and PMI is

a small peptide that can fold quickly. Therefore PMI-Mdm2 is a relatively simple example that allows to study coupled binding and folding with MD simulations.

2. Mdm2 is a medically relevant protein [122] and the peptide PMI was engineered to bind to Mdm2 and to block its binding site for the p53 protein. [95] PMI binds very strongly to the Mdm2 fragment with a dissociation constant $K_d = 3.3\,\mathrm{nM}$. [95] It is interesting to understand which PMI conformations are relevant for this strong binding. Additionally the system is a hard test case for the enhanced sampling methods developed in this thesis.

The interaction of the same Mdm2 fragment with the less strongly binding p53 peptide was studied by two other groups with MSMs [88] and with the weighted ensemble method [147]. In both works, only unbiased simulations were used. In none of the works, the dissociation rate could be estimated.
We ran 500 μs of unbiased simulations divided in many short trajectories starting from various bound and unbound conformation and approximately 100 μs of biased simulations that were coupled with HREMD (see section A.2.4). As the bias potential, we selected the so-called boost potential [54]. The joint data set of biased and unbiased simulations was analyzed with TRAMMBAR (see section 2.1.2 and [P3]). We find a dissociation constant $K_{d,\mathrm{sim}} = 0.34\,\mathrm{nM}$ $[0.22\,\mathrm{nM}, 0.44\,\mathrm{nM}]$[5] and a complex residence time of $k_{\mathrm{off,sim}}^{-1} = 883\,\mathrm{ms}$ $[480\,\mathrm{ms}, 1328\,\mathrm{ms}]$. To validate the simulations, we performed binding competition experiments (see supplementary information for [P3]). We found an experimental dissociation constant $K_{d,\mathrm{exp}} = 3.02 \pm 0.31\,\mathrm{nM}$ in agreement with literature data [95] and an experimental residence time of $k_{\mathrm{off,exp}}^{-1} = 26.8\,\mathrm{s}$ $[24.7\,\mathrm{s}, 34.1\,\mathrm{s}]$, which is in good agreement with the simulation results considering expected errors in the simulation force field.
We observe that the bound state is a conformationally very diverse ensemble of structures that bind with different hydrophobic contact surfaces and interconvert on the ten-microsecond timescale. Similar observations of a conformationally diverse hydrophobic encounter complex were made in [7] for the coupled folding and binding of a different molecular system consisting of S-peptide and S-protein.
The bound ensemble of PMI-Mdm2 is dominated by two metastable states. In the most probable state, PMI adopts the same conformation as in the crystal structure with protein data bank identifier 3eqs [95]. In the state with the second largest probability, PMI binds to Mdm2 with a similar contact pattern and an unfolded C-terminus.
The MSM estimated with TRAMMBAR shows that most of the binding/unbinding happens directly without intermediates that have longer life times than 10 μs. The experimental results point in a similar direction. On the time scale of seconds,

[5] All error ranges are given as 95% confidence intervals.

we see no biexponential behavior and no fast initial decay in the relaxation time series (see figure 2f in [P3]), which suggests that there is no intermediate with a life time of similar magnitude as k_{off}^{-1}. The binding/unbinding mechanism features parallel pathways (see figure 3 in [P3]). PMI is completely or predominantly folded in the long-lived on-pathway intermediates. Judging from the secondary structure, this could be called a predominant binding-after-folding mechanism. However the binding sites and binding patterns differ drastically between the intermediates, so a description of the binding mechanism in terms of binding patterns is more adequate for this system than the order of binding and formation of secondary structure (see figures 1 and 3 in [P3]).

Our simulations are good in exploring different conformation of the PMI-Mdm2 complex and their rearrangements. A weak point of our simulations is that they are done with periodic boundary conditions (which are necessary for explicit water models). In our simulations these boundary conditions lead to a small volume of the dissociated state which then leads to fast association ($\approx 10\,\text{ns}$) of the binding partners. This makes it difficult to sample trajectories that follow a binding-after-folding mechanism where PMI first folds in the unbound state and subsequently binds to Mdm2. The fast binding might not leave enough time for unbound PMI to change conformation in our simulations. We correct for the effect of finite simulation volume in the estimation of the residence time (see supplementary information for [P3]). In future work, simulations with a effectively larger box could be attempted by using multi-scale modeling techniques like adaptive resolution molecular-dynamics simulation that work on the atomistic level [103] or by combining MSMs with Brownian dynamics simulations [44, 109, 18].

2.4. Distinguishing induced fit and conformational selection using chemical relaxation rates[6]

Chemical relaxation experiments such as mixing experiments provide information on the binding mechanism. In these experiments, a sequence of time series $\{O_i(t)\}_i$ is measured where each time series $O_i(t)$ is recorded from an experiment started with different initial conditions. In mixing experiments, different initial conditions are created by varying the initial concentrations of the reaction partners. In the experiments considered in this thesis, the initial concentrations of unbound protein $[\text{P}]_0 = [\text{P}](t = 0)$ and ligand $[\text{L}]_0 = [\text{L}](t = 0)$ are varied.

[6] Some passages have been quoted verbatim from [P1]

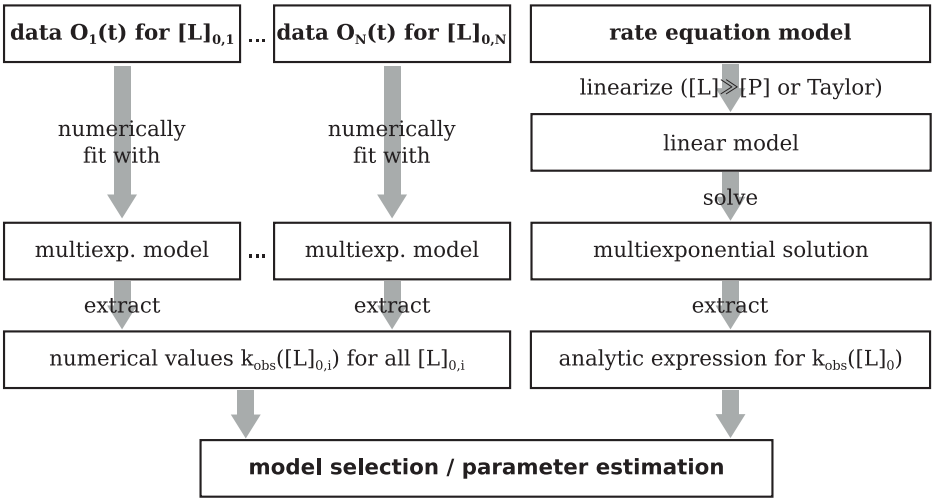

Figure 2.5.: Flow chart for the analysis of relaxation data. For details see main text and [P1].

To test whether the mechanism that underlays the time series data is induced fit or conformational selection (see figure 2.6a and 2.6b), the data is compared to predictions from rate equations that model one of these two mechanisms. Even though a direct comparison between the relaxation time series and the solution of the rate equation models is possible [75, 30], the standard in the field is to map both the experimental observations and the models to the *observed relaxation rate* k_{obs} (see below for definition) and to compare these rates between experiment and model (see figure 2.5). [11, 67] An important task, which is addressed in this thesis, is to develop accurate and general methods to compute k_{obs} from theoretical models and from experimental data.
The observed relaxation rate k_{obs} is defined as the smallest relaxation rate $\min_i(k_i)$ in the set of relaxation rates $\{k_i\}_{i=1,\ldots,n}$ that parameterize the multiexponential model

$$O(t) = \sum_{i=1}^{n} A_i e^{-k_i t} + A_0 \tag{2.8}$$

The constants $\{A_i\}_{i=0,\ldots,n}$ are the relaxation amplitudes. On the experimental side, numerical values for k_{obs} can be found by a numerical fit of the multiexponential model to the data (see figure 2.5, left branch). k_{obs} is often the only relaxation rate that can be determined reliably. [P1] On the theoretical side, analytical expressions for k_{obs} for the induced fit model and the conformational selection model are derived by first linearizing and then solving the rate equation models (see figure 2.5, right branch). The linearization is necessary because the rate equation models involve

binding steps that are second-order reactions and therefore depend on the product of the time-dependent concentrations of unbound proteins and unbound ligands.

The standard method of linearizing the equations is the *pseudo-first-order approximation*, where it is assumed that the total ligand concentration greatly exceeds the total protein concentration, so that the amount of ligand consumed during binding is negligible compared to the total amount of ligand. The concentration of the unbound ligand then can be taken to be constant, and the rate equations only contain terms that are linear in the time-dependent concentration of the protein, which makes them solvable. [P1]

In the more general approach that we propose in this thesis, a linearization of the rate equations is achieved by a Taylor expansion around the equilibrium concentrations of the bound and unbound proteins and ligands. This expansion captures the final relaxation into equilibrium, which is governed by the smallest, dominant relaxation rate k_{obs}, for all concentrations of proteins and ligands, and leads to general results for k_{obs} that include the results from the pseudo-first-order approximation in the limit of large ligand concentrations. [P1]

The standard approach based on the pseudo-first-order assumption has led to mixed success in distinguishing conformational selection and induced fit. [138, 67, 135] Under this assumption, $k_{\text{obs}}([\text{L}]_0)$, regarded as a function of the initial ligand concentration, is monotonic in $[\text{L}]_0$ for both models. For the induced fit model, $k_{\text{obs}}([\text{L}]_0)$ is always an increasing function of $[\text{L}]_0$. For the conformational selection model, $k_{\text{obs}}([\text{L}]_0)$ is either an increasing function or a decreasing function of $[\text{L}]_0$, depending on the numerical values of the rate constants (i. e. on the sign of $k_- - k_e$, see figure 2.6). This monotony can sometimes be used to infer the mechanism. While a decreasing $k_{\text{obs}}([\text{L}]_0)$ points towards the conformational selection mechanism, no statement can be make if $k_{\text{obs}}([\text{L}]_0)$ is an increasing function. [137, 136, 22]

With our more general approach, we find that k_{obs} can exhibit a local minimum. For the induced fit model, the minimum is located at $[\text{L}]_0^{\text{min}} = [\text{P}]_0 - K_d$ (for large enough $[\text{P}_0]$). For the conformational selection model, the minimum is approximately at $[\text{L}]_0^{\text{min}} \approx [\text{P}]_0(k_e + k_-)/(k_e - k_-) - K_d$ (for $k_e > k_-$ and large enough $[\text{P}_0]$). For the induced fit model, $k_{\text{obs}}([\text{L}]_0)$ is an even function symmetric about $[\text{L}]_0^{\text{min}}$. In contrast, $k_{\text{obs}}([\text{L}]_0)$ is in general not symmetric for the conformational selection model. These properties can be used to infer the mechanism even for systems where no distinction between induced fit and conformation selection can be made under pseudo-first-order conditions. We therefore suggest to perform experiments beyond first-order conditions.

To demonstrate the applicability of this suggestion, we tested the approach with synthetic data that we generated by numerically integrating in time the full differential equations of the induced fit model and the conformational selection model and subsequently identifying the correct model and its parameters. In

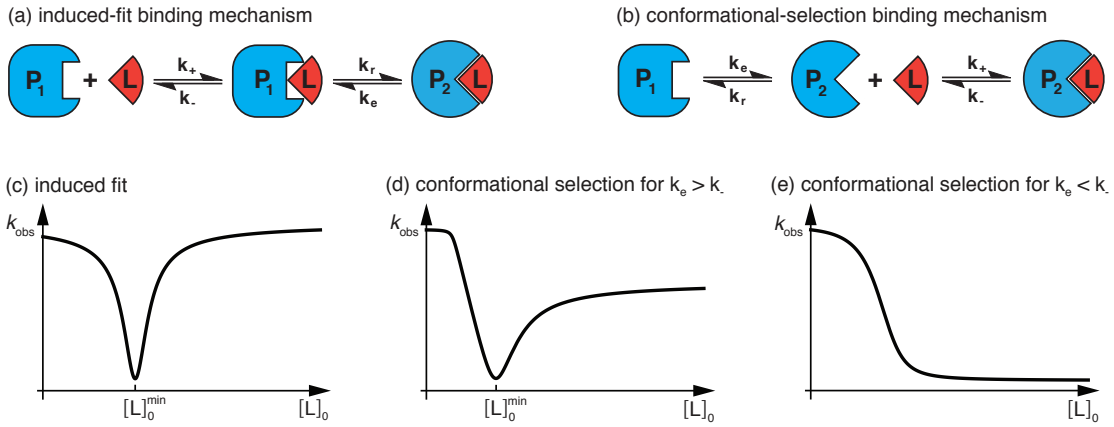

Figure 2.6.: Figure reproduced from [P1]. (a) In induced-fit binding, the change between the conformations P_1 and P_2 of the protein occurs after binding of the ligand L. The intermediate state P_1L relaxes into the bound ground state P_2L with rate k_r , and is excited from the ground state with rate k_e. (b) In conformational-selection binding, the conformational change of the protein occurs prior to ligand binding. The intermediate state P_2 is excited from the unbound ground state P_1 with rate k_e , and relaxes back into the ground state with rate k_r . (c) The dominant, smallest relaxation rate k_{obs} as a function of $[L]_0$. See main text.

addition, we test the approach with experimentally measured data from Chakrabarti *et al.* [21] for the interaction of the protein recoverin with a rhodopsin kinase peptide fused to the B1 domain of immunoglobulin protein G. We confirm the conclusion of reference [21] that the mechanism is conformational selection and obtain estimates of the parameter values k_e and k_- that agree with the result from [21].
An additional benefit of our method is that we find general expressions for k_{obs} that depend on both $[P]_0$ and $[L]_0$ (see e. g. equation (46) in [P1]). Therefore the other pseudo-first-order limit of $[P]_0 \gg [L]_0$ is contained in our equations, and it should be interesting to perform experiments under these conditions. In an even more general experimental approach, the initial concentrations of both species could be varied and a 2-D grid of measured k_{obs} values could be used to fit the models.

3. Conclusions

In this work, we have developed novel methods for the simulation and analysis of macromolecular binding coupled to conformational change and have applied these methods to three molecular systems: trypsin-benzamidine, PMI-Mdm2, and recoverin-rhodopsin kinase peptide. For these systems, binding intermediates were identified and their equilibrium and kinetic properties were computed.

While atomistic simulations are currently the method that gives the most detailed information about a molecular system, the use of MD is impeded by the fact that biological interesting processes often are rare events that are hard to simulate. Complete understanding of macromolecular binding processes in full atomistic resolution would be desirable and can hopefully be achieved in the future by steady but incremental improvements in experimental methods, computer technology, simulation algorithms, and analysis methods. In this work, we developed the TRAM and TRAMMBAR algorithms that allow to estimate Markov state models for systems that exhibit very rare events. These algorithms are particularly efficient in situations where the process to be studied is slow but the reverse process is fast. Unbinding of ligands or peptides from proteins are examples for such processes, and we were able to estimate complex residence times up to hundreds of milliseconds.

Future research could be directed at the problem on how TRAM and TRAMMBAR can be applied to the study of transitions that are hard to sample in both the forward and the backward directions. For very simple molecular systems, progress in this direction has been made [126] by identifying the transitions states and using these states as starting points for the unbiased simulations. However, finding the transitions states in large molecules like proteins is a complicated problem in itself. It may be possible to address this problem with path sampling methods [97] or the string method [38].

Estimating the kinetics of strong binders from simulations might become a useful tool for drug design. Optimizing the residence times of ligand-receptor complexes is increasingly considered a promising strategy in drug design. To ensure contiguous drug effect between subsequent deliveries, the drug's residence time at the receptor should be long enough. [53] Because this effect becomes only relevant for residence times that are comparable to the biological half-life of drugs (hours) [28], it is worth to mention that TRAM and TRAMMBAR can in principle also be used to predict residence times that are much larger than seconds, depending on the efficiency of the biased MD simulations. So, any progress made in method development for biased simulations might directly translate into a corresponding progress in the

estimation of kinetics.
The accurate prediction of kinetic properties with computer simulation relies on an accurate force field. Currently, kinetic properties are often poorly reproduced by MD simulations. This applies to quantitative properties like relaxation time scales [134] and mechanistic properties like the order of events [99]. For protein-protein interaction [98, 14] or natively disordered proteins [106], also free energies are poorly reproduced by MD. In order to attempt a correction of the systematic error of force fields, it is first required that kinetic and equilibrium properties can be estimated with small statistical error. That's where enhanced sampling methods like the ones developed in this work can play an important role.
The simulations and analysis of the PMI-Mdm2 system conducted in this thesis are not exhaustive. The high flexibility of PMI and its unspecific binding to Mdm2 results in a very large number of states that are metastable on a time scale that is comparable to the length of our MD trajectories ($1\,\mu s$). In future research on PMI-Mdm2 interaction or similar peptide-protein interactions, it would be helpful to combine the approach taken in this work with methods for adaptive exploration of phase space and adaptive restarting methods [104, 34, 146] for the reduction of statistical errors.
Finally all the work in this thesis was done in the framework of MSMs and therefore under the implicit assumption that the dynamics that takes place within a microstate is unimportant. A disadvantage of this approach is e. g. that it complicates the identification of transition states. This strong focus on states was mostly taken for technical reasons, because a decomposition of the conformational space into microstates is required to formulate TRAM and TRAMMBAR. Currently many other approaches for estimating kinetic properties are being developed, which describe a mechanism as an ensemble of trajectories and not as a sequence of states (like MSMs). Therefore, these approaches provide a more resolved picture of mechanisms. Among these methods are weighted ensemble [60], path sampling [32, 131, 3, 120], path reweighting [24, 35], and milestoning [42]. Hopefully we will see some convergence in the development of trajectory-based methods and enhanced sampling methods that exploit biased simulations to discover the full picture of macromolecular binding mechanisms.

Bibliography

[1] R. Affentranger, I. Tavernelli, and E. E. Di Iorio. A novel Hamiltonian replica exchange MD protocol to enhance protein conformational space sampling. *J. Chem. Theory Comput.*, 2(2):217–228, 2006.

[2] B. Alberts, A. Johnson, J. Lewis, D. Morgen, M. Raff, K. Roberts, and P. Walter. *Molecular biology of the cell*. Garland Science, Taylor & Francis Group, New York, U.S.A., 2008.

[3] R. J. Allen, C. Valeriani, and P. R. ten Wolde. Forward flux sampling for rare event simulations. *J. Phys.-Condens. Mat.*, 21(46):463102, 2009.

[4] C. B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181(4096):223–230, 1973.

[5] A. Arkhipov, Y. Yin, and K. Schulten. Membrane-bending mechanism of amphiphysin N-BAR domains. *Biophys. J.*, 97:2727–2735, 2009.

[6] R. H. Austin, K. W. Beeson, L. Eisenstein, H. Frauenfelder, and I. C. Gunsalus. Dynamics of ligand binding to myoglobin. *Biochemistry-US*, 14(24):5355–5373, 1975.

[7] A. Bachmann, D. Wildemann, F. Praetorius, G. Fischer, and T. Kiefhaber. Mapping backbone and side-chain interactions in the transition state of a coupled protein folding and binding reaction. *P. Natl. Acad. Sci. USA*, 2011.

[8] C. Bartels. Analyzing biased Monte Carlo and molecular dynamics simulations. *Chem. Phys. Lett.*, 331(5):446–454, 2000.

[9] C. Bartels and M. Karplus. Multidimensional adaptive umbrella sampling: Applications to main chain and side chain peptide conformations. *J. Comput. Chem.*, 18(12):1450–1462, 1997.

[10] H. Beach, R. Cole, M. Gill, and J. Loria. Conservation of μs-ms enzyme motions in the apo- and substrate-mimicked state. *J. Am. Chem. Soc.*, 127:9167–9176, 2005.

[11] C. F. Bernasconi. *Relaxation Kinetics*. Academic Press, 1976.

[12] B. J. Berne, M. Borkovec, and J. E. Straub. Classical and modern methods in reaction rate theory. *J. Phys. Chem.*, 92(13):3711–3725, 1988.

[13] S. Bernèche and B. Roux. Energetics of ion conduction through the K+ channel. *Nature*, 414:73–77, 2001.

[14] R. B. Best, W. Zheng, and J. Mittal. Balanced protein-water interactions improve properties of disordered proteins and non-specific protein association. *J. Chem. Theory Comput.*, 10(11):5113–5124, 2014.

[15] P. D. Blood and G. A. Voth. Direct observation of bin/amphiphysin/rvs (BAR) domain-induced membrane curvature by means of molecular dynamics simulations. *Proc. Natl. Acad. Sci. USA*, 103:15068–15072, 2006.

[16] D. D. Boehr, H. J. Dyson, and P. E. Wright. An NMR perspective on enzyme dynamics. *Chem. Rev.*, 106(8):3055–3079, 2006.

[17] D. D. Boehr, D. McElheny, H. J. Dyson, and P. E. Wright. The dynamic energy landscape of dihydrofolate reductase catalysis. *Science*, 313:1638–1642, 2006.

[18] B. Boras, S. Hirakis, L. Votapka, R. Malmstrom, R. Amaro, and A. McCulloch. Bridging scales through multiscale modeling: a case study on protein kinase A. *Front. Physiol.*, 6:250, 2015.

[19] G. R. Bowman, K. A. Beauchamp, G. Boxer, and V. S. Pande. Progress and challenges in the automated construction of Markov state models for full protein systems. *J. Chem. Phys.*, 131(12):124101, 2009.

[20] I. Buch, T. Giorgino, and G. De Fabritiis. Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations. *P. Natl. Acad. Sci. US*, 2011.

[21] K. S. Chakrabarti, R. V. Agafonov, F. Pontiggia, R. Otten, M. K. Higgins, G. F. X. Schertler, D. D. Oprian, and D. Kern. Conformational selection in a protein-protein interaction revealed by dynamic pathway analysis. *Cell Rep.*, 14(1):32–42, 2016.

[22] P. Chakraborty and E. Di Cera. Induced fit is a special case of conformational selection. *Biochemistry-US*, 56(22):2853–2859, 2017.

[23] H. S. Chan, S. Shimizu, and H. Kaya. Cooperativity principles in protein folding. *Method. Enzymol.*, 380:350–379, 2004. Energetics of Biological Macromolecules, Part E.

[24] J. D. Chodera, W. C. Swope, F. Noé, J.-H. Prinz, M. R. Shirts, and V. S. Pande. Dynamical reweighting: Improved estimates of dynamical properties from simulations at multiple temperatures. *J. Chem. Phys.*, 134(24):244107, 2011.

[25] J. D. Chodera, W. C. Swope, J. W. Pitera, C. Seok, and K. A. Dill. Use of the weighted histogram analysis method for the analysis of simulated and parallel tempering simulations. *J. Chem. Theory Comput.*, 3(1):26–41, 2007.

[26] H. S. Chung and W. A. Eaton. Single-molecule fluorescence probes dynamics of barrier crossing. *Nature*, 502(7473):685–688, 2013.

[27] J. M. B. Clarke, J. L. Tymoczko, and L. Stryer. *Biochemistry*, chapter 3. W. H. Freeman, New York, 2002.

[28] G. Dahl and T. Akerud. Pharmacokinetics and the drug-target residence time concept. *Drug Discov. Today*, 18(15–16):697–707, 2013.

[29] K. G. Daniels, N. K. Tonthat, D. R. McClure, Y. C. Chang, X. Liu, M. A. Schumacher, and et al. Ligand concentration regulates the pathways of coupled protein folding and binding. *J. Am. Chem. Soc.*, 136(3):822–825, 2014.

[30] K. G. Daniels, N. K. Tonthat, D. R. McClure, Y. C. Chang, X. Liu, M. A. Schumacher, and et al. Ligand concentration regulates the pathways of coupled protein folding and binding. *J. Am. Chem. Soc.*, 136(3):822–825, 2014.

[31] B. L. de Groot, X. Daura, A. E. Mark, and H. Grubmüller. Essential dynamics of reversible peptide folding: memory-free conformational dynamics governed by internal hydrogen bonds. *J. Mol. Biol.*, 309(1):299–313, 2001.

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

[33] K. A. Dill and H. S. Chan. From Levinthal to pathways to funnels. *Nat. Struct. Mol. Biol.*, 4(1):10–19, 1997.

[34] S. Doerr and G. De Fabritiis. On-the-fly learning and sampling of ligand binding by high-throughput molecular simulations. *J. Chem. Theory Comput.*, 10(5):2064–2069, 2014.

[35] L. Donati, C. Hartmann, and B. G. Keller. Girsanov reweighting for path ensembles and Markov state models. preprint arXiv:1703.05498, 2017.

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

[37] W.-N. Du and P. G. Bolhuis. Adaptive single replica multiple state transition interface sampling. *J. Chem. Phys.*, 139(4):044105, 2013.

[38] W. E, W. Ren, and E. Vanden-Eijnden. Finite temperature string method for the study of rare events. *J. Phys. Chem. B*, 109(14):6688–6693, 2005.

[39] J. F. Eccleston, J. P. Hutchinson, and H. D. White. *Protein-ligand Interactions, Structure and Spectroscopy: A Practical Approach*, volume 243 of *Practical Approach Series*, chapter 5, pages 201–237. Oxford University Press, 2001.

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

[41] M. Fajer, R. V. Swift, and J. A. McCammon. Using multistate free energy techniques to improve the efficiency of replica exchange accelerated molecular dynamics. *J. Comput. Chem.*, 30(11):1719–1725, 2009.

[42] A. K. Faradjian and R. Elber. Computing time scales from reaction coordinates by milestoning. *J. Chem. Phys.*, 120(23):10880–10889, 2004.

[43] A. M. Ferrenberg and R. H. Swendsen. Optimized Monte Carlo data analysis. *Phys. Rev. Lett.*, 63:1195–1198, 1989.

[44] M. B. Flegg, S. J. Chapman, and R. Erban. The two-regime method for optimizing stochastic reaction-diffusion simulations. *J. Roy. Soc. Interface*, 9(70):859–868, 2012.

[45] H. Frauenfelder and G. Petsko. Structural dynamics of liganded myoglobin. *Biophys. J.*, 32(1):465–483, 1980.

[46] H. Frauenfelder, S. Sligar, and P. Wolynes. The energy landscapes and motions of proteins. *Science*, 254(5038):1598–1603, 1991.

[47] D. Frenkel and B. Smit. *Understanding Molecular Simulation: From Algorithms to Applications*. Computational science series. Elsevier Science, 2001.

[48] C. Frieden. Slow transitions and hysteretic behavior in enzymes. *Ann. Rev. Biochem.*, 48:471–489, 1979.

[49] H. Fukunishi, O. Watanabe, and S. Takada. On the Hamiltonian replica exchange method for efficient sampling of biomolecular systems: Application to protein structure prediction. *J. Chem. Phys.*, 116:9058–9067, 2002.

[50] E. A. Galburt and J. Rammohan. A kinetic signature for parallel pathways: Conformational selection and induced fit. links and disconnects between observed relaxation rates and fractional equilibrium flux under pseudo-first-order conditions. *Biochemistry-US*, 55(50):7014–7022, 2016.

[51] R. A. Goldbeck, Y. G. Thomas, E. Chen, R. M. Esquerra, and D. S. Kliger. Multiple pathways on a protein-folding energy landscape: Kinetic evidence. *Proc. Natl. Acad. Sci. USA*, 96(6):2782–2787, 1999.

[52] F. Guillain and D. Thusius. Use of proflavine as an indicator in temperature-jump studies of the binding of a competitive inhibitor to trypsin. *J. Am. Chem. Soc.*, 92(18):5534–5536, 1970.

[53] D. Guo, L. H. Heitman, and A. P. IJzerman. The role of target binding kinetics in drug discovery. *ChemMedChem*, 10(11):1793–1796, 2015.

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

[55] G. G. Hammes, Y.-C. Chang, and T. G. Oas. Conformational selection or induced fit: a flux description of reaction mechanism. *Proc. Natl. Acad. Sci. USA*, 106(33):13737–13741, 2009.

[56] U. H. Hansmann. Parallel tempering algorithm for conformational studies of biological molecules. *Chem. Phys. Lett.*, 281(1-3):140–150, 1997.

[57] H. Hartmann, F. Parak, W. Steigemann, G. A. Petsko, D. R. Ponzi, and H. Frauenfelder. Conformational substates in a protein: structure and dynamics of metmyoglobin at 80 K. *Proc. Natl. Acad. Sci. USA*, 79(16):4967–4971, 1982.

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

[59] K. A. Henzler-Wildman, V. Thai, M. Lei, M. Ott, M. Wolf-Watz, T. Fenn, E. Pozharski, M. A. Wilson, G. A. Petsko, M. Karplus, C. G. Hübner, and

D. Kern. Intrinsic motions along an enzymatic reaction trajectory. *Nature*, 450(7171):838–844, 2007.

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

[61] A. M. J. Grotendorst, D. Marx, editor. *Quantum simulations of complex many-body systems: From theory to algorithms*, chapter Statistical analysis of simulations: data correlations and error estimation, pages 423–445. NIC Series, 2002.

[62] M. O. Jensen, V. Jogini, D. W. Borhani, A. E. Leffler, R. O. Dror, and D. E. Shaw. Mechanism of voltage gating in potassium channels. *Science*, 336:229–233, 2012.

[63] S. Jo, D. Suh, Z. He, C. Chipot, and B. Roux. Leveraging the information from Markov state models to improve the convergence of umbrella sampling simulations. *J. Phys. Chem. B*, 120(33):8733–8742, 2016.

[64] L. E. Kay, D. A. Torchia, and A. Bax. Backbone dynamics of proteins as studied by nitrogen-15 inverse detected heteronuclear NMR spectroscopy: application to staphylococcal nuclease. *Biochemistry-US*, 28(23):8972–8979, 1989.

[65] H. Keller and P. G. Debrunner. Evidence for conformational and diffusional mean square displacements in frozen aqueous solution of oxymyoglobin. *Phys. Rev. Lett.*, 45:68–71, 1980.

[66] A. Kessel and N. Ben-Tal. *Introduction to Proteins*. CRC Press, Taylor & Francis Group, Boca Raton, FL, U.S.A., 2011.

[67] T. Kiefhaber, A. Bachmann, and K. S. Jensen. Dynamics and mechanisms of coupled protein folding and binding reactions. *Curr. Opin. Struc. Biol.*, 22(1):21–29, 2012.

[68] E. Kim, S. Lee, A. Jeon, J. M. Choi, H.-S. Lee, S. Hohng, and H.-S. Kim. A single-molecule dissection of ligand binding to a protein with intrinsic dynamics. *Nat. Chem. Biol.*, 9(5):313–318, 2013.

[69] K. J. Kohlhoff, D. Shukla, M. Lawrenz, G. R. Bowman, D. E. Konerding, D. Belov, R. B. Altman, and V. S. Pande. Cloud-based simulations on Google Exacycle reveal ligand modulation of GPCR activation pathways. *Nat. Chem.*, 6:15–21, 2014.

[70] A. Kong, P. McCullagh, X.-L. Meng, D. Nicolae, and Z. Tan. A theory of statistical models for Monte Carlo integration. *J. Roy. Stat. Soc. B*, 65(3):585–604, 2003.

[71] D. A. Köpfer, C. Song, T. Gruene, G. M. Sheldrick, U. Zachariae, and B. L. de Groot. Ion permeation in K+ channels occurs by direct Coulomb knock-on. *Science*, 346:352–355, 2014.

[72] D. E. Koshland. Application of a theory of enzyme specificity to protein synthesis. *Proc. Natl. Acad. Sci. USA*, 44(2):98–104, 1958.

[73] M. B. Kubitzki and B. L. de Groot. Molecular dynamics simulations using temperature-enhanced essential dynamics replica exchange. *Biophys. J.*, 92(12):4262–4270, 2007.

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

[75] P. Kuzmič. Chapter 10 - dynafit — a software package for enzymology. In *Method Enzymol.*, volume 467 of *Methods in Enzymology*, pages 247–280. Academic Press, 2009.

[76] O. F. Lange, N.-A. Lakomek, C. Farès, G. F. Schröder, K. F. A. Walter, S. Becker, J. Meiler, H. Grubmüller, C. Griesinger, and B. L. de Groot. Recognition dynamics up to microseconds revealed from an RDC-derived ubiquitin ensemble in solution. *Science*, 320(5882):1471–1475, 2008.

[77] K. Lindorff-Larsen, P. Maragakis, S. Piana, M. P. Eastwood, R. O. Dror, and D. E. Shaw. Systematic validation of protein force fields against experimental data. *PLOS ONE*, 7(2):1–6, 2012.

[78] P. Liu, B. Kim, R. A. Friesner, and B. J. Berne. Replica exchange with solute tempering: A method for sampling biological systems in explicit water. *P. Natl. Acad. Sci. USA*, 102(39):13749–13754, 2005.

[79] J. P. Loria, R. B. Berlow, and E. D. Watt. Characterization of enzyme motions by solution NMR relaxation dispersion. *Acc. Chem. Res.*, 41(2):214–221, 2008.

[80] B. Ma, S. Kumar, C.-J. Tsai, and R. Nussinov. Folding funnels and binding mechanisms. *Protein Eng. Des. Sel.*, 12(9):713–720, 1999.

[81] E. Marinari and G. Parisi. Simulated tempering: A new Monte Carlo scheme. *Euro. Phys. Lett.*, 19(6):451–458, 1992.

[82] A. Matagne, S. E. Radford, and C. M. Dobson. Fast and slow tracks in lysozyme folding: insight into the role of domains in the folding process. *J. Mol. Biol.*, 267(5):1068–1074, 1997.

[83] R. T. McGibbon and V. S. Pande. Variational cross-validation of slow dynamical modes in molecular kinetics. *J. Chem. Phys.*, 142(12):124105, 2015.

[84] A. S. J. S. Mey, H. Wu, and F. Noé. xtram: Estimating equilibrium expectations from time-correlated simulation data at multiple thermodynamic states. *Phys. Rev. X*, 4:041018, 2014.

[85] X. Michalet, S. Weiss, and M. Jäger. Single-molecule fluorescence studies of protein folding and conformational dynamics. *Chem. Rev.*, 106(5):1785–1813, 2006.

[86] W. Min, B. P. English, G. Luo, B. J. Cherayil, S. C. Kou, and X. S. Xie. Fluctuating enzymes: Lessons from single-molecule studies. *Acc. Chem. Res.*, 38(12):923–931, 2005.

[87] A. Mittermaier and L. E. Kay. New tools provide new insights in NMR studies of protein dynamics. *Science*, 312(5771):224–228, 2006.

[88] S. Mukherjee, G. A. Pantelopulos, and V. A. Voelz. Markov models of the apo-MDM2 lid region reveal diffuse yet two-state binding dynamics and receptor poses for computational docking. *Sci. Rep.-UK*, 6:31631, 2016.

[89] F. Noé and C. Clementi. Kinetic distance and kinetic maps from molecular dynamics simulation. *J. Chem. Theory Comput.*, 11(10):5002–5011, 2015.

[90] F. Noé, I. Horenko, C. Schütte, and J. C. Smith. Hierarchical analysis of conformational dynamics in biomolecules: Transition networks of metastable states. *J. Chem. Phys.*, 126(15):155102, 2007.

[91] F. Noé, C. Schütte, E. Vanden-Eijnden, L. Reich, and T. R. Weikl. Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. *P. Natl. Acad. Sci. USA*, 106(45):19011–19016, 2009.

[92] J. R. Norris. *Markov Chains*. Number 2 in Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

[93] R. Nygaard, Y. Zou, R. O. Dror, T. J. Mildorf, D. H. Arlow, A. Manglik, A. C. Pan, C. W. Liu, J. J. Fung, M. P. Bokoch, F. S. Thian, T. S. Kobilka, D. E. Shaw, L. Mueller, R. S. Prosser, and B. K. Kobilka. The dynamic process of β2-adrenergic receptor activation. *Cell*, 152:532–542, 2013.

[94] A. G. Palmer. NMR characterization of the dynamics of biomacromolecules. *Chem. Rev.*, 104(8):3623–3640, 2004.

[95] M. Pazgier, M. Liu, G. Zou, W. Yuan, C. Li, C. Li, J. Li, J. Monbo, D. Zella, S. G. Tarasov, and W. Lu. Structural basis for high-affinity peptide inhibition of p53 interactions with MDM2 and MDMX. *P. Natl. Acad. Sci. USA*, 106(12):4665–4670, 2009.

[96] G. Pérez-Hernández, F. Paul, T. Giorgino, G. De Fabritiis, and F. Noé. Identification of slow molecular order parameters for Markov model construction. *J. Chem. Phys.*, 139(1):015102, 2013.

[97] B. Peters and B. L. Trout. Obtaining reaction coordinates by likelihood maximization. *J. Chem. Phys.*, 125(5):054108, 2006.

[98] D. Petrov and B. Zagrovic. Are current atomistic force fields accurate enough to study proteins in crowded environments? *PLoS Comput. Biol.*, 10(5):1–11, 2014.

[99] S. Piana, K. Lindorff-Larsen, and D. E. Shaw. How robust are protein folding simulations with respect to force field parameterization? *Biophys. J.*, 100:L47–L49, 2011.

[100] N. Plattner, S. Doerr, G. De Fabritiis, and F. Noé. Complete protein-protein association kinetics in atomic detail revealed by molecular dynamics simulations and Markov modelling. *Nat. Chem.*, 2017. advance online publication, DOI: 10.1038/nchem.2785.

[101] N. Plattner, J. D. Doll, P. Dupuis, H. Wang, Y. Liu, and J. E. Gubernatis. An infinite swapping approach to the rare-event sampling problem. *J. Chem. Phys.*, 135:134111, 2011.

[102] N. Plattner and F. Noé. Protein conformational plasticity and complex ligand-binding kinetics explored by atomistic simulations and Markov models. *Nat. Commun.*, 6:7653, 2015.

[103] M. Praprotnik, L. D. Site, and K. Kremer. Adaptive resolution molecular-dynamics simulation: Changing the degrees of freedom on the fly. *J. Chem. Phys.*, 123(22):224106, 2005.

[104] J. Preto and C. Clementi. Fast recovery of free energy landscapes via diffusion-map-directed molecular dynamics. *Phys. Chem. Chem. Phys.*, 16(36):19181–19191, 2014.

[105] J.-H. Prinz, H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J. D. Chodera, C. Schütte, and F. Noé. Markov models of molecular kinetics: Generation and validation. *J. Chem. Phys.*, 134(17):174105, 2011.

[106] S. Rauscher, V. Gapsys, M. J. Gajda, M. Zweckstetter, B. L. de Groot, and H. Grubmüller. Structural ensembles of intrinsically disordered proteins depend strongly on force field: A comparison to experiment. *J. Chem. Theory Comput.*, 11(11):5513–5524, 2015.

[107] T. F. Reubold, K. Faelber, N. Plattner, Y. Posor, K. Branz, U. Curth, J. Schlegel, R. Anand, D. Manstein, F. Noé, V. Haucke, O. Daumke, and S. Eschenburg. Crystal structure of the dynamin tetramer. *Nature*, 525:404–408, 2015.

[108] E. Rosta, H. L. Woodcock, B. R. Brooks, and G. Hummer. Artificial reaction coordinate "tunneling" in free-energy calculations: The catalytic reaction of RNase H. *J. Comput. Chem.*, 30(11):1634–1641, 2009.

[109] J. Schöneberg, A. Ullrich, and F. Noé. Simulation tools for particle-based reaction-diffusion dynamics in continuous space. *BMC Biophys.*, 7(1):11, 2014.

[110] C. Schütte, W. Huisinga, and P. Deuflhard. Transfer operator approach to conformational dynamics in biomolecular systems. Technical report, Konrad-Zuse-Zentrum Berlin, 1999.

[111] C. Schütte and M. Sarich. *Metastability and Markov State Models in Molecular Dynamics*, volume 24 of *Courant Lecture Notes*. American Mathematical Society, 2013.

[112] C. R. Schwantes and V. S. Pande. Improvements in markov state model construction reveal many non-native interactions in the folding of NTL9. *J. Chem. Theory Comput.*, 9(4):2000–2009, 2013.

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

[114] D.-A. Silva, G. R. Bowman, A. Sosa-Peinado, and X. Huang. A role for both conformational selection and induced fit in ligand binding by the lao protein. *PLoS Comput. Biol.*, 7(5):e1002054, 05 2011.

[115] R. D. Smiley and G. G. Hammes. Single molecule studies of enzyme mechanisms. *Chem. Rev.*, 106(8):3080–3094, 2006.

[116] M. Souaille and B. Roux. Extension to the weighted histogram analysis method: combining umbrella sampling with free energy calculations. *Comput. Phys. Commun.*, 135(1):40–57, 2001.

[117] Y. Sugita and Y. Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.*, 314(1-2):141–151, 1999.

[118] R. H. Swendsen and J.-S. Wang. Replica Monte Carlo simulation of spin-glasses. *Phys. Rev. Lett.*, 57:2607–2609, 1986.

[119] Z. Tan, E. Gallicchio, M. Lapelosa, and R. M. Levy. Theory of binless multi-state free energy estimation with applications to protein-ligand binding. *J. Chem. Phys.*, 136(14):144102, 2012.

[120] I. Teo, C. G. Mayne, K. Schulten, and T. Lelièvre. Adaptive multilevel splitting method for molecular dynamics calculation of Benzamidine-Trypsin dissociation time. *J. Chem. Theory Comput.*, 12(6):2983–2989, 2016.

[121] P. Tiwary, V. Limongelli, M. Salvalaglio, and M. Parrinello. Kinetics of protein-ligand unbinding: Predicting pathways, rates, and rate-limiting steps. *Proc. Natl. Acad. Sci. USA*, 112(5):E386–E391, 2015.

[122] F. Toledo and G. M. Wahl. Regulating the p53 pathway: in vitro hypotheses, in vivo veritas. *Nat. Rev. Cancer*, 6(12):909–923, 2006.

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

[124] B. Trendelkamp-Schroer. *Reversible Markov state models*. PhD thesis, Freie Universität Berlin, 2016.

[125] B. Trendelkamp-Schroer and F. Noé. Efficient Bayesian estimation of Markov model transition matrices with given stationary distribution. *J. Chem. Phys.*, 138(16):164113, 2013.

[126] B. Trendelkamp-Schroer and F. Noé. Efficient estimation of rare-event kinetics. *Phys. Rev. X*, 6:011009, 2016.

[127] B. Trendelkamp-Schroer, H. Wu, and F. Noé. Reversible Markov chain estimation using convex-concave programming. arXiv:1603.01640, 2016.

[128] B. Trendelkamp-Schroer, H. Wu, F. Paul, and F. Noé. Estimation and uncertainty of reversible Markov models. *J. Chem. Phys.*, 143(17):174101, 2015.

[129] P. J. Tummino and R. A. Copeland. Residence time of receptor-ligand complexes and its effect on biological function. *Biochemistry*, 47(20):5481–5492, 2008.

[130] V. N. Uversky. Intrinsically disordered proteins from A to Z. *Int. J. Biochem. Cell B.*, 43(8):1090–1103, 2011.

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

[132] N. G. van Kampen. *Stochastic Processes in Physics and Chemistry*, chapter V. The Master Equation, pages 114–117. Elsevier Science B. V., Amsterdam, The Netherlands, 1992.

[133] Y. Vardi. Empirical distributions in selection bias models. *Ann. Stat.*, 13(1):178–203, 1985.

[134] F. Vitalini, A. S. J. S. Mey, F. Noé, and B. G. Keller. Dynamic properties of force fields. *J. Chem. Phys.*, 142(8):084101, 2015.

[135] A. D. Vogt and E. Di Cera. Conformational selection or induced fit? a critical appraisal of the kinetic mechanism. *Biochemistry-US*, 51(30):5894–5902, 2012.

[136] A. D. Vogt and E. Di Cera. Conformational selection is a dominant mechanism of ligand binding. *Biochemistry-US*, 52(34):5723–5729, 2013.

[137] T. R. Weikl and F. Paul. Conformational selection in protein binding and function. *Protein Sci.*, 23(11):1508–1518, 2014.

[138] T. R. Weikl and C. von Deuster. Selected-fit versus induced-fit protein binding: Kinetic differences and mutational analysis. *Proteins*, 75(1):104–110, 2009.

[139] W. Wojtas-Niziurski, Y. Meng, B. Roux, and S. Bernèche. Self-learning adaptive umbrella sampling method for the determination of free energy landscapes in multiple dimensions. *J. Chem. Theory Comput.*, 9(4):1885–1895, 2013.

[140] T. B. Woolf and B. Roux. Conformational flexibility of o-phosphorylcholine and o-phosphorylethanolamine: A molecular dynamics study of solvation effects. *J. Am. Chem. Soc.*, 116(13):5916–5926, 1994.

[141] H. Wu, A. S. J. S. Mey, E. Rosta, and F. Noé. Statistically optimal analysis of state-discretized trajectory data from multiple thermodynamic states. *J. Chem. Phys.*, 141(21):214106, 2014.

[142] H. Wu and F. Noé. Optimal estimation of free energies and stationary densities from multiple biased simulations. *Multiscale Model. Simul.*, 12(1):25–54, 2014.

[143] J. Z. Xiang and B. Honig. JACKAL: a protein structure modeling package. Technical report, Columbia University and Howard Hughes Medical Institute, New York, 2002.

[144] G. Zhou, G. A. Pantelopulos, S. Mukherjee, and V. A. Voelz. Bridging microscopic and macroscopic mechanisms of p53-MDM2 binding using molecular simulations and kinetic network models. *bioRxiv*, 2016.

[145] F. Zhu and G. Hummer. Pore opening and closing of a pentameric ligand-gated ion channel. *Proc. Natl. Acad. Sci. USA*, 107:19814–19819, 2010.

[146] M. I. Zimmerman and G. R. Bowman. FAST conformational searches by balancing exploration/exploitation trade-offs. *J. Chem. Theory Comput.*, 11(12):5747–5757, 2015.

[147] M. C. Zwier, A. J. Pratt, J. L. Adelman, J. W. Kaus, D. M. Zuckerman, and L. T. Chong. Efficient atomistic simulation of pathways and calculation of rate constants for a protein-peptide binding process: Application to the MDM2 protein and an intrinsically disordered p53 peptide. *J. Phys. Chem. Lett.*, 7(17):3440–3445, 2016.

A. Methods appendix

A.1. Markov state models

A.1.1. Definition and metastability

The exposition of Markov state models in this section follows approximately reference [105], see [110] and [124] for a more mathematically rigorous treatment. Capital letters in this section denote vectors in phase space $\mathbf{X} = (\mathbf{x}, \mathbf{p})$, where $\mathbf{x}$ are the Cartesian configurational degrees of freedom and $\mathbf{p}$ are the corresponding momenta.
The stochastic motion of molecules can not only be interpreted as trajectories in phase space but can also seen from the perspective of ensemble probability densities evolving in time. As the equations of motion used for MD are Markovian in the full state space (this means that only the current phase-space coordinates $\mathbf{R}(t)$ are needed to propagate the system in time and no other points $\mathbf{R}(t - \Delta\tau)$ that lie further in the past) one can define the transition probability density

$$p(\mathbf{X}, \mathbf{Y}, \tau)\mathrm{d}y \equiv \mathbb{P}(\mathbf{R}(t + \tau) \in \mathbf{Y} + dY \mid \mathbf{R}(t) = \mathbf{X}) \tag{A.1}$$

This is the probability that a trajectory that was initiated at time t in phase space point $\mathbf{X}$ will be found in an infinitesimal region $\mathrm{d}Y$ around the point $\mathbf{Y}$ at time $t + \tau$. Consider an ensemble of non-interacting copies of the molecular system at time t distributed in phase space according to the probability density $p_t(\mathbf{X})$. At time $t + \tau$ all copies will have evolved according to the transition probability density and will be distributed according to a new probability density $p_{t+\tau}(\mathbf{X})$. From Markovianity and basic identities of probability theory one can derive that

$$p_{t+\tau}(\mathbf{X}) = \int p_t(\mathbf{Y})p(\mathbf{Y}, \mathbf{X}, \tau)\,\mathrm{d}Y \tag{A.2}$$

The integration on the right hand side can be abbreviated with the symbol $\mathcal{P}$ that stands for propagator. The propagator is an integral operator that maps $p_t(\mathbf{X})$ to $p_{t+\tau}(\mathbf{X})$.

$$\mathcal{P}[p_t(\mathbf{Y}); \tau](\mathbf{X}) \equiv \int p_t(\mathbf{Y})p(\mathbf{Y}, \mathbf{X}, \tau)\,\mathrm{d}Y$$

Let $p_B(\mathbf{X})$ be the stationary distribution of the propagator $p_B = \mathcal{P}[p_B]$. In the cases studied in this thesis, the stationary distribution is the Boltzmann distribution. If

the dynamics fulfills the condition of detailed balance[1] then the following equation holds

$$p_B(\mathbf{X})p(\mathbf{X},\mathbf{Y},\tau) = p_B(\mathbf{Y})p(\mathbf{Y},\mathbf{X},\tau)$$

An equivalent description of the dynamics is given by the transfer operator $\mathcal{T}$. Instead of operating on probability densities like the propagator, the transfer operator operates on functions $u_t(\mathbf{X})$ that are related to probability densities by the relation $p_t(\mathbf{X}) = u_t(\mathbf{X})p_B(\mathbf{X})$. It is defined as

$$\mathcal{T}[u_t(\mathbf{Y});\tau](\mathbf{X}) \equiv \frac{1}{p_B(\mathbf{X})}\int u_t(\mathbf{Y})p_B(\mathbf{Y})p(\mathbf{Y},\mathbf{X},\tau)\,\mathrm{d}Y$$

If the molecular system under study shows the phenomenon of metastability, one can discretize the transfer operator to arrive at a numerical method that can be implemented on a computer. Metastability means that after a given lag time the phase space points $\mathbf{X}$ in some set S (i. e. $\mathbf{X} \in S$) are mostly redistributed *within* the set S but the probability of leaving the set is low. It can be shown for systems with metastability, that the transfer operator can be decomposed into a fast and a slow part, based on the eigendecomposition

$$u_{t+\tau}(\mathbf{X}) = \mathcal{T}_{\text{slow}}[u_t;\tau](\mathbf{X}) + \mathcal{T}_{\text{fast}}[u_t;\tau](\mathbf{X}).$$

$\mathcal{T}_{\text{fast}}$ models the fast motions within the metastable states that are usually not interesting (e. g. because they are well below the time resolution of experiments). $\mathcal{T}_{\text{slow}}$ can be expressed using the large eigenvalues λ_i of $\mathcal{T}$ and the corresponding eigenfunctions ψ_i as

$$\mathcal{T}_{\text{slow}}[u_t;\tau](\mathbf{X}) = \sum_{i=1}^{n}\lambda_i(\tau)\psi_i(\mathbf{X})\int \psi_i(\mathbf{Y})p_B(\mathbf{Y})u_t(\mathbf{Y})\,\mathrm{d}Y.$$

The eigenvalues λ_i of $\mathcal{T}_{\text{slow}}$ are related to the autocorrelation times of the system and decay exponentially with the lag time τ. The separation into $\mathcal{T}_{\text{fast}}$ and $\mathcal{T}_{\text{slow}}$ and the eigendecomposition of $\mathcal{T}_{\text{slow}}$ reduce the complexity of the dynamic model by replacing the infinite-dimensional operators by a finite-dimensional representation

[1] Hamilton's equations of motion and the Langevin equations of motion fulfill the condition of *generalized* or *extended detailed balance* which involves reversal of the momenta [132, 111]. That is $p_B(\mathbf{x},\mathbf{p})p(\mathbf{x},\mathbf{p},\mathbf{x}',\mathbf{p}',\tau) = p_B(\mathbf{x},-\mathbf{p})p(\mathbf{x},-\mathbf{p},\mathbf{x}',-\mathbf{p}',\tau)$. As we are eventually interested in the dynamics projected to the configuration space alone (where the slowly-evolving dynamics can be well described), we can integrate the extended detailed balance relation over the momenta to obtain the conventional detailed balance relation in configuration space $\mathbb{P}(\mathbf{r}(t+\tau) \in \mathbf{x}+\mathrm{d}x \text{ and } \mathbf{r}(t) = \mathbf{x}'+\mathrm{d}x') = \mathbb{P}(\mathbf{r}(t+\tau) \in \mathbf{x}'+\mathrm{d}x' \text{ and } \mathbf{r}(t) = \mathbf{x}+\mathrm{d}x)$. See also [124] for a more detailed discussion of this topic.

that contains only the dominant eigenfunctions of $\mathcal{T}_{\text{slow}}$. This provides a method to approximating the dynamics numerically. The problem is even more simplified be the fact that the eigenfunctions ψ_i of $\mathcal{T}_{\text{slow}}$ show a typical structure: they take an approximately constant value on every metastable set and they show a sigmoidal transition in the transition regions between metastable states. This motivates to discretize the phase space into a collection of non-overlapping sets, the so-called *microstates* S_i. Because of the typical structure of the leading eigenfunctions, the microstates can be relatively coarse in the metastable sets (the extreme case would be to resolve every metastable set by a single microstate) and need a finer resolution only in the transition regions. Also the momenta $\mathbf{p}$ are usually not used in the definition of microstates, because their fast relaxation time makes them unsuitable for describing long-lived states. This kind of discretization is referred to in the literature under the names *Galerkin-discretization of the transfer operator* [110] or simply as *Markov state models with many states* [105][2].
The transition matrix is defined as

$$T_{ij}(\tau) \equiv \frac{\int \chi_j(\mathbf{X}) \mathcal{T}[\chi_i](\mathbf{X}) p_B(\mathbf{X}) \,\mathrm{d}X}{\int \chi_j(\mathbf{X}) p_B(\mathbf{X}) \,\mathrm{d}X} = \frac{\int_{S_j} \left(\int_{S_i} p_B(\mathbf{X}) p(\mathbf{X}, \mathbf{Y}; \tau) \,\mathrm{d}X \right) \mathrm{d}Y}{\int_{S_i} p_B(\mathbf{X}) \,\mathrm{d}X} \tag{A.3}$$

where S_i are the microstates and $\chi_i(\mathbf{X})$ is the indicator function of microstate S_i. If the transfer operator admits a decomposition into $\mathcal{T}_{\text{slow}} + \mathcal{T}_{\text{fast}}$ and for a proper choice of microstates, it can be shown that T_{ij} can be interpreted as the conditional probability that the system is in microstate S_j at time $t + \tau$ given that the system was in microstate S_i at the earlier time t. That means that the transition matrix fulfills the Markov property as well

$$T_{ij}(\tau) \approx \mathbb{P}(\mathbf{X}(t + \tau) \in S_j \mid \mathbf{X}(t) \in S_i).$$

If the Markov property holds, the matrix $\mathbf{T}$ can be used to extrapolate to arbitrary long lag times, because the long-time kinetics can be written as

$$\mathbf{T}(k\tau) = (\mathbf{T}(\tau))^k . \tag{A.4}$$

This allows to infer the kinetics on long time scales without needing to simulate them directly. If the propagator fulfills the condition of detailed balance with respect to the stationary distribution $p_B(\mathbf{X})$, the transition matrix fulfills the condition of detailed balance with respect to the stationary vector $\boldsymbol{\pi}$ with $\pi_i = \int_{S_i} p_B(\mathbf{X}) \,\mathrm{d}X$.

[2] as opposed to Markov states models with few states where each metastable state is identical to one microstate

A.1.2. Estimation

The definition of the transition matrix as given by equation (A.3) can not be used directly to compute numerical estimates, because $p(\mathbf{X}, \mathbf{Y}, \tau)$ is not available as an (analytical or numerical) expression that can be directly evaluated on the computer. The high-dimensional integral in equation (A.3) can only be solved using Monte-Carlo methods. Practically, one runs a finite number of MD simulations and records the sequence of microstates that the trajectories visit. From these random realizations, the elements of the transition matrix T_{ij} are estimated. Therefore the transition matrix (and all derived quantities) must be thought of as random variables that are characterized by some probability density function $p(\mathbf{T})$. Often one is not interested in the whole distribution of transition matrices, but only in finding one representative $\mathbf{T}$ that stands for the distribution, especially if the statistical errors of $\mathbf{T}$ are small. A standard procedure for achieving this are *maximum-likelihood methods*. Maximum likelihood methods can be best explained by starting from Bayes' theorem. Bayes' theorem is a general statement about probabilities of two events M and D:

$$\mathbb{P}(M \mid D) = \frac{\mathbb{P}(M)}{\mathbb{P}(D)} \mathbb{P}(D \mid M)$$

The theorem can be used for parameter estimation and model selection if one sets M to be a probabilistic model (in this case the transition matrix) and D (for data) to a set of observations. The quantity $\mathbb{P}(M \mid D)$ is called the *posterior* and according to the Bayesian interpretation of probabilities can be understood as the (subjective) degree of belief in the model, given that one has observed D. Bayes' theorem relates this degree of belief to the *likelihood* $\mathbb{P}(D \mid M)$ which is the probability that the data was generated by the fixed model M and which can be expressed in an analytical form for many probabilistic models. Choosing the "best" model then amounts to maximizing $\mathbb{P}(M \mid D)$ over all possible models. In the optimization, $\mathbb{P}(D)$ can be ignored because it is a data-dependent constant. $\mathbb{P}(M)$ is called the *prior probability* (or just *prior*) and is the degree of belief in a model without having observed any data. The prior can for example be used to exclude models that disagree with prior knowledge or with fundamental physical principles like detailed balance, by setting their probabilities to zero.[3] If no prior knowledge is available, an uniform (also called uninformative) prior is chosen, that assigns equal probabilities to all models. In that case, maximizing the posterior $\mathbb{P}(M \mid D)$ is the same as maximizing the likelihood $\mathbb{P}(D \mid M)$ and the resulting estimator is called the maximum-likelihood estimator of M.
In the case of Markov state models, the model is fully specified by the set of microstates and the transition matrix $\mathbf{T}$. The data are one or many trajectories

[3] See [105, 124] for a more in-depth discussion on the choice of priors for MSM estimation.

(time series) of microstate labels $\Xi = \{s_k\}_{k=0,\dots N}$. The microstate label s_k at time $t = \tau k$ is i if the conformation $\mathbf{x}(t)$ is in microstate S_i. The likelihood $\mathbb{P}(D \mid M)$ for a single trajectory is then given by

$$\begin{aligned}\mathbb{P}(s_0, s_1, \dots s_N \mid \mathbf{T}) &= \prod_{k=1}^{N} \mathbb{P}(s_k \mid s_{k-1})\mathbb{P}(s_0) \\ &= \mathbb{P}(s_0) \prod_{k=1}^{N} T_{s_k s_{k-1}} = \mathbb{P}(s_0) \prod_{i,j}^{n} (T_{ij})^{c_{ij}} \propto \prod_{i,j}^{n} (T_{ij})^{c_{ij}}\end{aligned} \tag{A.5}$$

In the last identity, the count matrix $\mathbf{c}$ was introduced. Each element c_{ij} is the total number of transitions from microstate S_i to microstate S_j observed in a given trajectory. $\mathbb{P}(s_0)$ is the probability of observing the first state in a trajectory. Since it is just a proportionality constant, that doesn't affect the optimal choice of $\mathbf{T}$, it will be ignored in the following. For a collection of many trajectories $\{\Xi_i\}_{i=1,\dots K}$, independence of trajectories and initial states is usually assumed such that the probability of observing all trajectories can be expressed as the product

$$\mathbb{P}(\Xi_1, \Xi_2, \dots, \Xi_K \mid \mathbf{T}) = \prod_{i=1}^{K} \mathbb{P}(\Xi_i \mid \mathbf{T}) \propto \prod_{i,j}^{n} (T_{ij})^{C_{ij}} \equiv \mathcal{L}_{\mathrm{MSM}} \tag{A.6}$$

where $\mathbf{C}$ is the sum of all the count matrices from the individual trajectories. Maximizing $\mathcal{L}_{\mathrm{MSM}}$ under the constraints of non-negative probabilities T_{ij} and normalization $\sum_j T_{ij} = 1$ for all i, results in the well-known estimator

$$\hat{T}_{ij} = \frac{C_{ij}}{\sum_k C_{ik}} \tag{A.7}$$

This is an estimator for a non-reversible MSM, that is in general $\hat{\pi}_i \hat{T}_{ij} \neq \hat{\pi}_j \hat{T}_{ji}$ where $\hat{\boldsymbol{\pi}}$ is the stationary vector of $\hat{\mathbf{T}}$. A maximum-likelihood estimator for a reversible MSM can be formulated analogously. [19] It cannot be expressed in closed form like equation (A.7) but can be implemented on a computer using different variants of numerical optimization algorithms. [19, 105, 127]

A.2. Free-energy calculation

A.2.1. Boltzmann reweighting

For ergodic dynamics the (equilibrium) expectation value of an observable $O(\mathbf{x})$ in the ensemble with equilibrium distribution $p(\mathbf{x}, \mathbf{p})$ can be computed by evaluating an integral over the phase space Ω

$$\langle O \rangle \equiv \int_\Omega O(\mathbf{x}) p(\mathbf{x},\, \mathbf{p})\, \mathrm{d}x\, \mathrm{d}p$$

or by Monte-Carlo integration by computing the mean of a sequence of samples $\{O(t)\}_{t=1}^{N}$ sampled from $p(\mathbf{x},\, \mathbf{p})$

$$\langle O \rangle = \lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} O(t)$$

For molecular systems, the Metropolis Monte-Carlo algorithm and molecular dynamics are popular algorithms for generating the samples $\{O(t)\}_{t=1,\ldots,T}$. Both algorithms employ dynamic systems that evolve according to Markovian dynamics and asymptotically sample from an equilibrium distribution. However they can exhibit metastability, i. e. the Markov chain can get trapped in local minima of the energy landscape and samples $O(t)$ can be highly auto-correlated in time. To reduce the redundancy of the samples and to speed up sampling of interesting regions of phase space, it is possible to introduce a *biased ensemble* with the equilibrium distribution $p^*(\mathbf{x},\, \mathbf{p})$ where the energy landscape is modified such that the metastability is reduced. It is possible to compute the expectation value $\langle O \rangle$ in the original unbiased ensemble by using only the samples $\{O^*(t)\}_{t=1,\ldots,N}$ drawn from $p^*(\mathbf{x},\, \mathbf{p})$. However the samples $O^*(t)$ must be corrected with a factor $p(t)/p^*(t)$. This can be seen from the identity

$$\langle O \rangle = \int_\Omega O(\mathbf{x}) \frac{p(\mathbf{x},\, \mathbf{p})}{p^*(\mathbf{x},\, \mathbf{p})} p^*(\mathbf{x},\, \mathbf{p})\, \mathrm{d}x\mathrm{d}p = \lim_{N^*\to\infty} \frac{1}{N^*} \sum_{t=1}^{N} O^*(t) \frac{p(t)}{p^*(t)}$$

We call the ratio $p(\mathbf{x},\, \mathbf{p})/(N^* p^*(\mathbf{x},\, \mathbf{p}))$ and its generalizations that are introduced in sections A.2.3 and 2.1.1, the *reweighting factor* $\mu(\mathbf{x},\, \mathbf{p})$.
While these general statements hold for any equilibrium distribution, $p(\mathbf{x},\, \mathbf{p})$ in chemistry corresponds to the Boltzmann distribution $\exp[\beta F - \beta H(\mathbf{x},\, \mathbf{p})]$. For these case of a long equilibrium simulation in a single biased ensemble, the reweighting factor can be given in closed form:

$$\mu(t) = \frac{1}{N} \frac{p(t)}{p^*(t)} = \frac{1}{N} \exp\left\{\beta F - \beta^* F^* - \beta H\left[\mathbf{x}(t),\, \mathbf{p}(t)\right] + \beta^* H^*\left[\mathbf{x}(t),\, \mathbf{p}(t)\right]\right\}$$

Here, $\mu(t)$ can be computed from the force field definition $H(\mathbf{x},\, \mathbf{p})$, $H^*(\mathbf{x},\, \mathbf{p})$ up to the constant factor $\exp[\beta F - \beta^* F^*]$.[4] If the temperatures in the biased and the unbiased ensembles are identical $\beta = \beta^*$, the energy difference $H^*(\mathbf{x},\, \mathbf{p}) - H(\mathbf{x},\, \mathbf{p})$

[4] For the simple case with one biased ensemble the missing constant can be found by normalizing a posteriori, that is by using the identity $\langle 1 \rangle = 1$. For the case of multiple ensembles, the normalization constants can be found with the WHAM, MBAR, TRAM or TRAMMBAR estimators among others (see below).

is referred to as the *bias energy* $\Delta E(\mathbf{x}, \mathbf{p})$. In many biased simulation types the kinetic energies $T(\mathbf{p})$ and $T^*(\mathbf{p})$ are identical, so the bias energy is a difference in potential energy $\Delta U(\mathbf{x}) = U^*(\mathbf{x}) - U(\mathbf{x})$. For the case of identical kinetic energy terms and identical temperatures $\beta = \beta^*$, the reweighting factor is independent of the momenta and is only a function $\mu(\mathbf{x})$ of the configurational degrees of freedom $\mathbf{x}$ of the samples. In this thesis, the simulations of the PMI-Mdm2 and trypsin-benzamidine systems fall under this category (identical temperatures and kinetic energy terms). Therefore the discussion is restricted to the case of the purely configuration-dependent reweighting factor $\mu(\mathbf{x})$.

A.2.2. Umbrella sampling

An important question is how to chose the biased equilibrium distribution $p^*(\mathbf{x}, \mathbf{p})$ for the specific molecular system under study. A popular way for constructing this distribution is *umbrella sampling*. [123] One starts by defining a so-called *reaction coordinate* or *order parameter*. An order parameter is a function of the configurational degrees of freedom $\xi(\mathbf{x})$. It is typically chosen based on two sets of configurations A and B, that need to be (approximately) known in advance. The order parameter describes the proximity of every configuration $\mathbf{x}$ to the two sets A and B. The order parameter of configurations in A takes values $\xi \leq \xi_A$ where ξ_A is some threshold and similarly $\xi \geq \xi_B$ for configurations in B. The transition region between A and B is characterized by $\xi_A < \xi < \xi_B$. Typical choices for A and B are the highly populated regions from two different metastable states ("free-energy minima"). Once the order parameter has been defined, the objective is to sample configurations rather uniformly along ξ, in particular in the transition region. A bias potential that flattens the free energy landscape is approximated by a *series* of localized bias potentials $U^{(i)}(\xi)$ centered at different values $\xi^{(i)}$ along the order parameter. Typically a quadratic function of the order parameter is chosen for the bias energy (harmonic spring in order parameter space). Then a series of molecular dynamics simulations are run, each of them with a different bias potential $U^{(i)}(\xi)$.

A.2.3. The weighted histogram analysis method and the multi-state Bennet acceptance ratio

If biased simulation of the same system were conducted with different biased Hamiltonians, like in umbrella sampling or multi-temperature simulations, the question arises how all this data should be combined such that an optimal estimate of expectation values can be obtained. One of the earliest developments that can handle the case of more than two biased Hamiltonians is the weighted histogram analysis method (WHAM) [43, 8]. There the conformational space is discretized into n sets $\{S_i\}_{i=1,\ldots,n}$ and the conformations from the simulations are grouped according

to the set they fall into and the index of the biased Hamiltonian that generated the conformation. The number of conformations in every group is $N_i^{(k)}$ where i stands for the set index and k for the index of the temperature/Hamiltonian. In WHAM a probabilistic model is proposed that describes the probability of observing a given count matrix $N_i^{(k)}$. Let $\pi_i^{(k)}$ be the probability of observing exactly one conformation that was sampled from Hamiltonian k inside set S_i. The probability of observing all counts $N_i^{(k)}$ is then

$$\mathcal{L}_{\mathrm{WHAM}} \equiv \mathbb{P}(\{N_i^{(k)}\}_{i=1,\dots,n}^{k=1,\dots,K} \mid \{\pi_i^{(k)}\}_{i=1,\dots,n}^{k=1,\dots,K}) = \prod_k \prod_i \left(\pi_i^{(k)}\right)^{N_i^{(k)}} \tag{A.8}$$

Taking the product over i requires that all conformations were drawn independently and the product over k requires that the simulations are independent. Especially the independence of the conformations is hard to fulfill for conformations that come from a typically auto-correlated MC or MD time series. [61, 25]
The likelihood $\mathcal{L}_{\mathrm{WHAM}}$ can then be optimized to find the optimal model parameters $\pi_i^{(k)}$. Optimizing only equation (A.8) under the constraint that $\pi_i^{(k)}$ is normalized would result in the maximizer $\hat{\pi}_i^{(k)} = N_i^{(k)} / \sum_j N_j^{(k)}$, from which nothing new could be gained. A more informative estimate can be obtained by incorporating our knowledge of the bias potential into the estimation. This can be done by linking all probabilities $\pi_i^{(k)}$ in the higher ensembles $k > 1$ to the probabilities in the (arbitrarily chosen) reference ensemble $\pi_i^{(1)} = \pi_i^{(\mathrm{ref})}$ by using Boltzmann reweighting

$$\pi_i^{(k)} = \frac{\pi_i^{(\mathrm{ref})} \exp(-\beta U_i^{(k)} + \beta U_i^{(\mathrm{ref})})}{\sum_j \pi_j^{(\mathrm{ref})} \exp(-\beta U_j^{(k)} + \beta U_j^{(\mathrm{ref})})} \tag{A.9}$$

where U_i is some potential energy value that is representative for the whole set S_i. Optimizing the likelihood $\mathcal{L}_{\mathrm{WHAM}}$ under the constraints of equation (A.9) and normalization of $\pi_i^{(k)}$ results in the well-known WHAM equations. [8]
It is a disadvantage of WHAM, that representative energies U_i need to defined for every set. For complicated forms of the bias energy, this is a nontrivial task. However, it was repeatedly realized by various researchers that the discretization step done in WHAM is not necessary. [133, 116, 1, 119] The resulting bin-less estimator was later named *multi-state Bennet acceptance ratio* (MBAR). [113] The main idea is that nowhere in the WHAM equations, the size (or phase-space volume) of the sets needs to be known (very much like the size of microstates in MSM estimation) and all sets can be of different sizes. This can be used to put every sampled conformation $\mathbf{x}_i$ in its own set S_i. Then equation (A.9) becomes an exact identity with $U_i^{(k)} = U^{(k)}(\mathbf{x}_i)$. Mathematically the likelihood (A.8) is barely changed. $N_i^{(k)}$ only takes the values 0 or 1 and usually all factors of the form $(\pi_i^{(k)})^0$ are not written out. [133, 119] In our publication [P2] we introduce the new symbol

$\mu^{(k)}(\mathbf{x})$ that replaces $\pi_i^{(k)}$ and where $\mathbf{x}$ stands for one of the sampled conformations. The $\mu^{(k)}(\mathbf{x})$ of different ensembles k are related *via* Boltzmann reweighting, like the $\pi_i^{(k)}$ in equation (A.9)

$$\mu^{(k)}(\mathbf{x}) := \frac{\mu^{(\mathrm{ref})}(\mathbf{x}) \exp\left[-\beta U^{(k)}(\mathbf{x}) + \beta U^{(\mathrm{ref})}(\mathbf{x})\right]}{\sum_{\mathbf{y}} \mu^{(\mathrm{ref})}(\mathbf{y}) \exp\left[-\beta U^{(k)}(\mathbf{y}) + \beta U^{(\mathrm{ref})}(\mathbf{y})\right]} \tag{A.10}$$

The sum in the denominator runs over all sampled conformations. We introduce the symbol $X^{(k)}$ for the set of all conformations generated with the potential energy $U^{(k)}$ and inverse temperature β. Using these definitions, the likelihood for MBAR is [119]

$$\mathcal{L}_{\mathrm{MBAR}} = \prod_k \prod_{\mathbf{x} \in X^{(k)}} \mu^{(k)}(\mathbf{x}) \tag{A.11}$$

Coming from the perspective of WHAM, $\mu^{(k)}(\mathbf{x})$ can be thought of as the equilibrium weight of a bin that surrounds $\mathbf{x}$. It is not a density and it is not useful to think about $\mu^{(k)}(\mathbf{x})$ as a (scaled) Dirac delta function. However, one can introduce an approximation to the Boltzmann distribution that is defined on the set of all conformations $X = X^{(1)} \cup \ldots \cup X^{(k)}$ sampled in all simulations:

$$p^{(k)}_{B,\mathrm{approx.}}(\mathbf{y}) = \sum_{\mathbf{x} \in X} \delta(\mathbf{x} - \mathbf{y}) \mu^{(k)}(\mathbf{x})$$

Therefore $\mu^{(k)}(\mathbf{x})$ can be interpreted as reweighting factors that reweight every sampled conformation towards the Boltzmann distribution of the thermodynamic ensemble k (ensemble with potential energy $U^{(k)}$). Equilibrium expectation values can then be expressed as

$$\langle O \rangle^{(k)} = \int O(\mathbf{y}) p^{(k)}_{B,\mathrm{approx.}}(\mathbf{y}) \, \mathrm{d}y = \sum_{\mathbf{x} \in X} O(\mathbf{x}) \mu^{(k)}(\mathbf{x})$$

Probabilities of a macroscopic state S (like the unbound state of a protein-ligand system or a metastable state of a MSM) can be calculated by defining an indicator function χ for the state that is 1 if $\mathbf{x} \in S$ and zero otherwise[5]

$$\mathbb{P}(x \in S) = \langle \chi \rangle^{(k)} = \sum_{\mathbf{x} \in X} \chi(\mathbf{x}) \mu^{(k)}(\mathbf{x}) = \sum_{\mathbf{x} \in S} \mu^{(k)}(\mathbf{x})$$

In this thesis, a generalization of MBAR is developed, that relaxes the assumption under which MBAR was derived. In particular it is no longer required that the conformations $\mathbf{x}$ are drawn independently but instead the time correlation between successive frames is modeled explicitly with a MSM (see section 2.1).

[5] For the calculation of macroscopic probabilities, some discrete states have to be introduced. The difference between these discrete states and the bins used in WHAM is, that in WHAM the energy axis always has to be discretized in addition to the macroscopic states that one is interested in.

A.2.4. Replica exchange molecular dynamics simulations

Hamiltonian replica exchange molecular dynamics (HREMD) simulations are a method for accelerating the sampling of the Boltzmann distribution by allowing unphysical transitions between different biased Hamiltonians. [117, 49] HREMD can potentially restore the ergodicity of a set of biased simulations that was broken by a bad choice of bias. It comes at the expense of generating unphysical kinetics that do not correspond to the kinetics of any of the Hamiltonians.

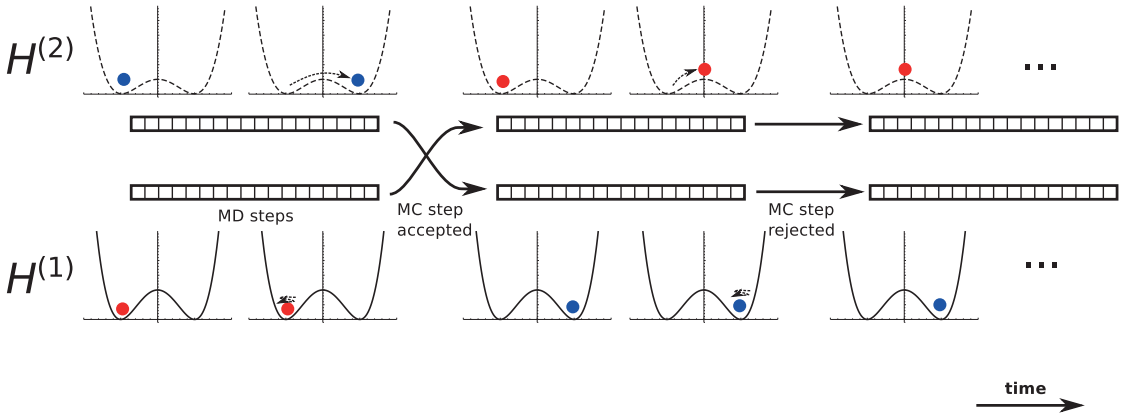

Figure A.1.: Schematic illustration of an HREMD simulation for the double-well potential (unbiased double-well $H^{(1)}$ and biased double-well $H^{(2)}$ with reduced barrier-height). HREMD consists of alternating phases of parallel MD simulation of the all replicas (symbolized by the movie strips) and exchange attempts between replicas (symbolized by the bold arrows) which are accepted or rejected according to the acceptance probability of the Metropolis-Monte-Carlo move.

The general idea of HREMD is to run multiple copies (the replicas) of the molecular system in parallel, each with a different biased Hamiltonian. Every N_{ex} integration steps of the MD simulation, the different Hamiltonians are exchanged among the replicas such that every replica is integrated in time with a new Hamiltonian after the exchange. The exchange of Hamiltonians is a random event that occurs with a probability that is given by the HREMD Metropolis criterion. That criterion guarantees all the conformations generated with Hamiltonian $H^{(i)}$ are distributed according to the Boltzmann distribution of that Hamiltonian for long run times. The algorithm is illustrated in figure A.1: two particles are started in the left well and are propagated with MD. Only with the biased Hamiltonian where the barrier height is reduced, the particle can transition quickly to the right well. The following replica exchange step is accepted and then the right well can be sampled in the simulation with Hamiltonian $H^{(1)}$. The exchange step can also be rejected

when the energy cost of swapping configurations between Hamiltonians is not compatible with the Boltzmann distributions of the Hamiltonians (columns 4, 5 in figure A.1). In general not only two Hamiltonians are used but a whole sequence of Hamiltonians that interpolates between the least biased and the most biased ensemble. These Hamiltonians have to be chosen such that replica exchange steps are frequently accepted. [101]

Because the conformations generated by every Hamiltonian are asymptotically distributed according to the Boltzmann distribution of the respective Hamiltonian, one could in principle define one of the Hamiltonians to be the unbiased, physical Hamiltonian of the system, collect all the conformations that it generated and use them to directly estimate physical equilibrium expectation values. While this might seem like a promising strategy for the system depicted in figure A.1 with two equally deep wells, the situation changes for more asymmetric energy landscapes. If the probability of observing a conformation in the right well in the physical ensemble is sufficiently small (compared to the left well), practically no replica exchange step that would deposit a conformation in the right well, will be accepted. Obeying the physical Boltzmann distribution prevents improbable conformations from being sampled. Therefore it is necessary to use not only conformations generated by the physical Hamiltonian but conformations generated by *all* Hamiltonians, together with an adequate reweighting algorithm to estimate equilibrium expectation values.

Until now the advantage of using HREMD has not been made clear. In particular if all samples from all simulations are used anyway to estimate expectation values, what is the benefit of the exchange step in the HREMD method? The advantage of HREMD is to enhance the exploration of conformational space in situations where an imperfect biasing Hamiltonian is being used. This is illustrated for the case of Umbrella sampling simulations in figure A.2. There the unbinding of a ligand from a protein surface is sampled by imposing a series of Umbrella potentials centered at different values of the protein-ligand separation. The protein surface is rough and shows multiple binding sites which should ideally all be explored by the simulations (e. g. because they contribute both significantly to the free energy of binding). However the simultaneous action of the Umbrella potential $\Delta U^{(1)}$ and steric interaction between the ligand and the protein prevents a transition to binding site (B) if the simulations were initiated with the ligand located in binding site (A). The simulation with the Umbrella potential $\Delta U^{(3)}$ does not allow the exploration of binding site (B) either because the ligand is confined to a too large distance from the protein. In contrast, when the simulations are coupled and exchanges between Umbrella potentials are accepted, the ligand is allowed to freely climb and descend on the ladder of protein-ligand distances and explore both binding sites. Similar problems can appear in many situations whenever there are relevant degrees of freedom orthogonal to the Umbrella sampling order parameter

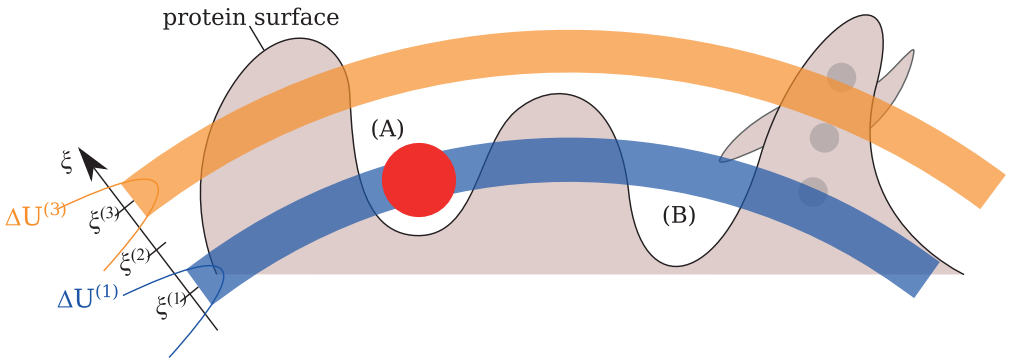

Figure A.2.: Illustration of the necessity of replica exchange of for Umbrella sampling simulations of protein-ligand binding. The red disk represents a ligand molecule that can bind to the surface of a protein. The order parameter ξ describes the distance of the ligand to the center of the protein. The blue (orange) shaded area is the conformational space that can be explored by the ligand if an Umbrella sampling bias potential $\Delta U^{(1)}$ ($\Delta U^{(3)}$) centered at $\xi^{(1)}$ ($\xi^{(3)}$) is active. None of the simulations with fixed Hamiltonian allows the ligand to explore the two binding sites (A) and (B) because the ligand is sterically confided to (A) with active $\Delta U^{(1)}$ and too far from the binding sites with active $\Delta U^{(3)}$. Only Hamiltonian exchange allows full exploration of both binding sites.

ξ. Conformational changes of the protein or the ligand or spatial rearrangement of the ligand are examples for relevant orthogonal degrees of freedom.
In this thesis HREMD is applied to sample the coupled binding and folding process for of the PMI-Mdm2 complex. A new analysis method for HREMD simulation data called TRAMMBAR is developed that combines the benefits of MSMs and the MBAR estimator.

B. Acronyms

DNA	doxyribonucleic acid
GPU	graphics processing unit
HREMD	Hamiltonian replica exchange molecular dynamics
MBAR	multi-state Bennet acceptance ratio
MC	Monte-Carlo
MD	molecular dynamics
MDM2	mouse double minute 2 homolog (protein name)
MFPT	mean first passage time
MSM	Markov state model
NMR	nuclear magnetic resonance
PMF	potential of mean force
PMI	potent MDM2 inhibitor (peptide name)
QSSA	quasi-steady state approximation
RMSD	root-mean-square deviation
TAD	transactivation domain
TICA	time-structure-based independent component analysis
TPT	transition path theory
TRAM	transition-based analysis method
TRAMMBAR	transition-based analysis method with multi-state Bennet acceptance ratio
WHAM	weighted histogram analysis method

C. Publications

C.1. Full list of publications

First-author and shared first-author publications

These publications constitute this cumulative dissertation.

[P1] Paul, Weikl
How to distinguish conformational selection and induced fit based on chemical relaxation rates
PLOS Comput. Biol. **12** e1005067 (2016)
doi: 10.1371/journal.pcbi.1005067

[P2] Wu, Paul, Wehmeyer, Noé
Multiensemble Markov models of molecular thermodynamics and kinetics
Proc. Natl. Acad. Sci. USA. **113** E3221 (2016)
doi: 10.1073/pnas.1525092113

[P3] Paul, Wehmeyer, Abualrous, Wu, Crabtree, Schöneberg, Clarke, Freund, Weikl, Noé
Protein-peptide association kinetics beyond the seconds timescale from atomistic simulations
Nat. Commun. **8** 1095 (2017)
doi: 10.1038/s41467-017-01163-6

Other publications

[P4] Pérez-Hernández, Paul, Giorgino, De Fabritiis, Noé
Identification of slow molecular order parameters for Markov model construction
J. Chem. Phys. **139**, 015102 (2013)

[P5] Weikl, Paul
Conformational selection in protein binding and function
Protein Science **23**, 1508 (2014)

[P6] Scherer, Trendelkamp-Schroer, Paul, Pérez-Hernández, Hoffmann, Plattner, Wehmeyer, Prinz, Noé
PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models
J. Chem. Theory Comput. **11** 5525 (2015)

[P7] Trendelkamp-Schroer, Wu, Paul, Noé
Estimation and uncertainty of reversible Markov models
J. Chem. Phys. **143** 174101 (2015)

[P8] Pinamonti, Zhao, Condon, Paul, Noé, Turner, Bussi **Predicting the kinetics of RNA oligonucleotides using Markov state models** *J. Chem. Theory Comput.* **13** 926 (2017)

[P9] Wu, Nüske, Paul, Klus, Koltai, Noé
Variational Koopman models: slow collective variables and molecular kinetics from short off-equilibrium simulations
J. Chem. Phys., **146** 154104 (2017)

[P10] Olsson, Wu, Paul, Clementi, Noé
Combining experimental and simulation data of molecular processes via augmented Markov models
Proc. Natl. Acad. Sci. USA. **114** 8265 (2017)

C.2. Author contributions

Clarification of the contributions of Fabian Paul to [P1] "How to Distinguish Conformational Selection and Induced Fit Based on Chemical Relaxation Rates"

Thomas Weikl proposed the project: to study the dependence of the dominant relaxation rate of the initial ligand concentration $k_{\mathrm{obs}}([\mathrm{L}]_0)$ for the "induced fit" model and the "conformational selection" model. He proposed to search for curves $k_{\mathrm{obs}}([\mathrm{L}]_0)$ that show a local minimum.

Fabian Paul proposed to study the rate equation in the limit of equilibrium concentration. He derived all mathematical equations and inequalities in this publication.

Fabian Paul chose all methods for data analysis: the choice of using a global optimizer for fitting multi-exponential models, the choice of the AIC and the Bayes factor for model comparison

Together, Thomas Weikl and Fabian Paul selected the numerical examples in figures 2 and 3.

Fabian Paul proposed to analyze the data from [Chakrabarti *et al*, *Cell Reports.* **14**, 32 (2016)].

Together, Thomas Weikl and Fabian Paul analyzed the numerical examples and the data from Chakrabarti *et al.*

Together Thomas Weikl and Fabian Paul proposed to use the symmetry of $k_{\mathrm{obs}}([\mathrm{L}]_0)$ as a criterion to distinguish "induced fit" and the "conformational selection" mechanism.

Thomas Weikl proposed to use the mathematical models of $k_{\mathrm{obs}}([\mathrm{L}]_0)$ for estimating transition rates of the conformational selection model and the induced fit model.

Approved: _____

PD Dr. Thomas Weikl

Clarification of the individual contributions of Fabian Paul and Hao Wu in the shared first-author publication [P2] "Multiensemble Markov models of molecular thermodynamics and kinetics"

Frank Noé proposed to use bin-less reweighting together with MSMs to estimate stationary equilibrium observables.

Fabian Paul proposed to combine the MBAR (bin-less) reweighting method with the maximum likelihood estimation of Markov state models for a given stationary vector to compute the transition probabilities of rare events. Together with Frank Noé, he derived equations [**13**], [**16**] and [**18**] of the publication for the more simple case of MSM estimation with a given, fixed stationary vector. In sum, he proposed the following *suboptimal* iterative equations for the estimation of a Multiensemble Markov model:

$$v_i^{k,\mathrm{new}} := v_i^k \sum_j \frac{c_{ij}^k + c_{ji}^k}{\exp\left[f_j^k - f_i^k\right] v_j^k + v_i^k}$$

$$f_i^{k,\mathrm{new}} := -\ln \sum_{x \in X_i} \frac{\exp\left[-b^k(x)\right]}{\sum_l \frac{\sum_i N_j^l \exp\left[-b^l(x)\right]}{\sum_j \exp\left[-f_j^l\right]}}$$

This algorithm doesn't use state transitions in the estimation of the thermodynamic quantities f_i^k.

Hao Wu lied out the mathematical framework to derive an estimator for Multiensemble Markov models that uses all data in an optimal way. He formulated the likelihood functions (equations [**7**], [**8**], [**9**]) and derived the dual Lagrangian function of the constrained optimization problem [**9**]-[**12**]. He derived the optimality conditions and the iterative algorithm.

Hao Wu conducted the simulation and analysis of the Alanine dipeptide system and the two-dimensional toy model.

Fabian Paul conducted the umbrella sampling simulation and analysis of the Trypsin-Benzamidine protein-ligand system.

Hao Wu showed the asymptotic correctness of the TRAM estimator.

Together Fabian Paul and Hao Wu showed that dTRAM is a special case of TRAM and derived the acceleration method given by equations [**36**]-[**38**].

Fabian Paul gave an interpretation to the mathematical expression **[15]** for the effective counts.

Christoph Wehmeyer, Hao Wu and Fabian Paul wrote computer implementations of TRAM.

Hao Wu, Fabian Paul, and Frank Noé wrote the paper.

Approved: _____
PD Dr. Thomas Weikl

This publication has not been used as part of any other dissertation except for the present dissertation "Markov State Modeling of Binding and Conformational Changes of Proteins" by Fabian Paul.

Approved: _____
Fabian Paul

Approved: _____
PD Dr. Thomas Weikl

Clarification of the contributions of Fabian Paul to [P3] "Protein-peptide association kinetics beyond the seconds timescale from atomistic simulations"

Thomas Weikl proposed to study coupled folding and binding in the PMI-$^{25-109}$Mdm2 system.

Frank Noé proposed to address this problem with a mixture of unbiased MD simulations and HREMD simulations and to analyze the data with MSMs.

Fabian Paul developed the TRAMMBAR algorithm with the help of Hao Wu. Fabian Paul conducted all simulations and analyses of the PMI-$^{25-109}$Mdm2 system. He implemented all simulation and major analysis software: implementation of the biasing potential in the MD software, the TRAMMBAR algorithm, the methods from [F. Zeller, M. Zacharias, *J. Comput. Chem.* **35**, 2256 (2014)] for the computation of differences in binding free energy upon alanine mutation and methods for rate matrix estimation. He proposed to estimate probabilities of metastable states upon alanine mutation. He proposed to estimate dissociation rates by estimating a continuous time MSM (rate matrix). Together with Frank Noé, Christoph Wehmeyer and Esam T. Abualrous, Fabian Paul wrote the publication.

Johannes Schöneberg provided an expression vector for the $^{25-109}$Mdm2 protein fragment.

Christoph Wehmeyer performed all modeling, simulations, analysis and necessary computer programming for the conceptual model for ligand binding.

Esam T. Abualrous and Michael D. Crabtree conducted all lab experiments.

Hao Wu corrected an early version of the TRAMMBAR algorithm that was proposed by Fabian Paul. He re-derived TRAMMBAR based on the likelihood formulation given in the publication.

Approved: _____

PD Dr. Thomas Weikl

PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

How to Distinguish Conformational Selection and Induced Fit Based on Chemical Relaxation Rates

Fabian Paul[1,2]*, **Thomas R. Weikl**[1]*

1 Max Planck Institute of Colloids and Interfaces, Department of Theory and Bio-Systems, Potsdam, Germany, **2** Free University Berlin, Department of Mathematics and Computer Science, Berlin, Germany

* fabian.paul@mpikg.mpg.de (FP); thomas.weikl@mpikg.mpg.de (TRW)

doi: 10.1371/journal.pcbi.1005067

Abstract

Protein binding often involves conformational changes. Important questions are whether a conformational change occurs prior to a binding event ('conformational selection') or after a binding event ('induced fit'), and how conformational transition rates can be obtained from experiments. In this article, we present general results for the chemical relaxation rates of conformational-selection and induced-fit binding processes that hold for all concentrations of proteins and ligands and, thus, go beyond the standard pseudo-first-order approximation of large ligand concentration. These results allow to distinguish conformational-selection from induced-fit processes—also in cases in which such a distinction is not possible under pseudo-first-order conditions—and to extract conformational transition rates of proteins from chemical relaxation data.

Author Summary

The function of proteins is affected by their conformational dynamics, i.e. by transitions between lower-energy ground-state conformations and higher-energy excited-state conformations of the proteins. Advanced NMR and single-molecule experiments indicate that higher-energy conformations in the unbound state of proteins can be similar to ground-state conformations in the bound state, and vice versa. These experiments illustrate that the conformational change of a protein during binding may occur before a binding event, rather than being induced by this binding event. However, determining the temporal order of conformational transitions and binding events typically requires additional information from chemical relaxation experiments that probe the relaxation kinetics of a mixture of proteins and ligands into binding equilibrium. These chemical relaxation experiments are usually performed and analysed at ligand concentrations that are much larger than the protein concentrations. At such high ligand concentrations, the temporal order of conformational transitions and binding events can only be inferred in special cases. In this article, we present general equations that describe the dominant chemical relaxation kinetics for all protein and ligand concentrations. Our general equations allow

OPEN ACCESS

Citation: Paul F, Weikl TR (2016) How to Distinguish Conformational Selection and Induced Fit Based on Chemical Relaxation Rates. PLoS Comput Biol 12(9): e1005067. doi:10.1371/journal.pcbi.1005067

Editor: Nikolay V. Dokholyan, University of North Carolina at Chapel Hill, UNITED STATES

Received: March 18, 2016

Accepted: July 14, 2016

Published: September 16, 2016

Data Availability Statement: All relevant data are within the paper.

Funding: This work was supported by Deusche Forschungsgemeinschaft http://www.dfg.de/en/ collective grant SFB 1114 http://gepris.dfg.de/gepris/projekt/235221301?language=en (to FP TRW). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

to clearly infer from relaxation data whether a conformational transition occurs prior to a binding event, or after the binding event.

Introduction

Protein function often involves conformational changes during the binding to ligand molecules [1]. Advanced NMR experiments [2–7] and single-molecule spectroscopy [8–10] indicate that these conformational changes can occur without ligand, or with bound ligand and thus point to an intrinsic conformational dynamics of the proteins. An important question is how the conformational dynamics is coupled to the binding events. Two mechanisms for this coupling are 'conformational selection' [11] and 'induced fit' [12] (see Fig 1(a) and 1(b)). In conformational-selection binding, a conformational change occurs *prior to* the binding of a ligand molecule, as a conformational excitation from the unbound-ground state conformation of the protein. In this mechanism, the ligand seems to 'select' and stabilize a higher-energy conformation for binding. In induced-fit binding, the conformational change occurs *after* ligand binding and is a conformational relaxation into the bound ground-state conformation that is apparently 'induced' by the ligand. These two mechanisms are in particular plausible for small ligand molecules that can quickly 'hop' in and out of the protein binding pocket, i.e. that can enter and exit this binding pocket within transition times that are significantly smaller than the residence or dwell times of the proteins in the different conformations [13].

A central problem is to identify protein binding mechanisms based on experimental data [13–24]. Advanced NMR experiments and single-molecule spectroscopy can reveal higher-energy conformations that are necessary for conformational-selection or induced-fit binding,

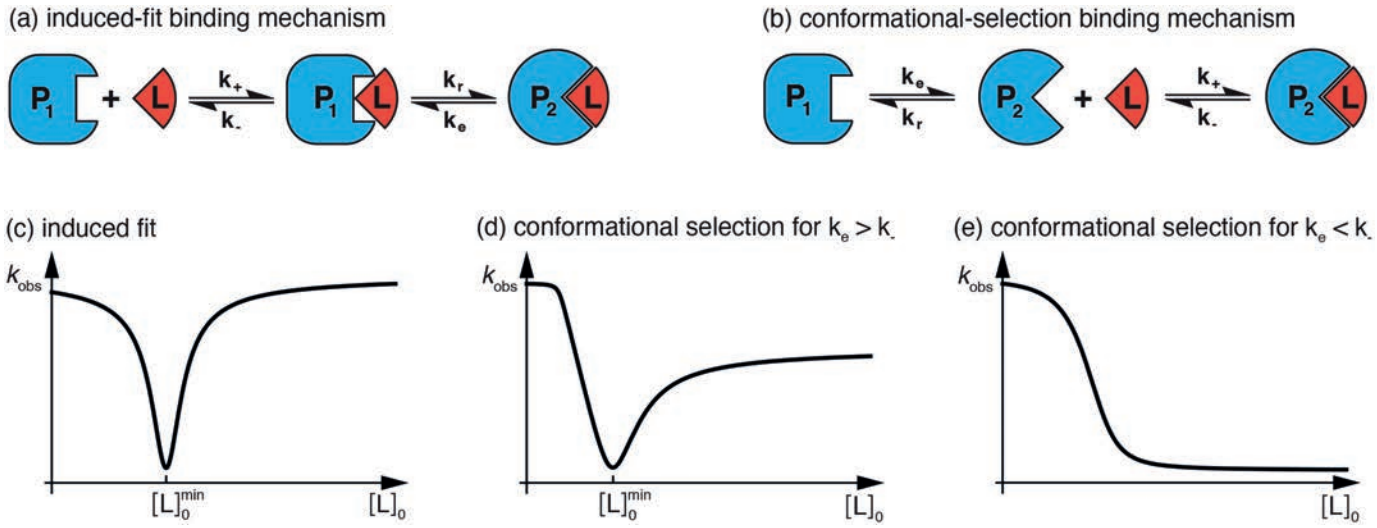

Fig 1. Characteristic chemical relaxation of induced-fit and conformational-selection binding. (a) In induced-fit binding, the change between the conformations P_1 and P_2 of the protein occurs after binding of the ligand L. The intermediate state P_1L relaxes into the bound ground state P_2L with rate k_r, and is excited from the ground state with rate k_e. (b) In conformational-selection binding, the conformational change of the protein occurs prior to ligand binding. The intermediate state P_2 is excited from the unbound ground state P_1 with rate k_e, and relaxes back into the ground state with rate k_r. (c) The dominant, smallest relaxation rate k_{obs} of induced-fit binding is minimal at the total ligand concentration $[L]_0^{min} = [P]_0 - K_d$ where $[P]_0$ is the total protein concentration and K_d the overall dissociation constant. As a function of $[L]_0$, the dominant rate k_{obs} is symmetric with respect to this minimum. (d) The dominant, smallest relaxation rate k_{obs} of conformational-selection binding has a characteristic minimum as a function of $[L]_0$ for $k_e > k_-$, but is not symmetric with respect to this minimum. (e) The dominant rate k_{obs} of conformational-selection binding decreases monotonically with $[L]_0$ for $k_e < k_-$.

doi:10.1371/journal.pcbi.1005067.g001

but do not directly indicate the binding mechanism because such higher-energy conformations may exist both in the bound and unbound state of the protein [4, 8]. In principle, both conformational-selection or induced-fit binding then are possible. Standard mixing or temperature-jump experiments that probe the chemical relaxation into the binding equilibrium can provide additional information that allows to identify the binding mechanism [22, 25–28]. Of particular interest is the dominant, slowest relaxation rate $k_{\rm obs}$ observed in the experiments, and how this rate depends on the total ligand concentration $[L]_0$ [22, 25, 28]. The chemical relaxation experiments are often performed and analysed under pseudo-first-oder conditions, i.e. at ligand concentrations that greatly exceed the protein concentrations [22, 25, 29–36]. In the case of induced-fit binding, the dominant relaxation rate $k_{\rm obs}$ *increases* monotonically with the ligand concentration $[L]_0$ under pseudo-first-oder conditions. In the case of conformational-selection binding, $k_{\rm obs}$ *decreases* monotonically with increasing $[L]_0$ for conformational excitation rates $k_e < k_-$, and *increases* monotonically with $[L]_0$ for $k_e > k_-$ where k_- is the unbinding rate of the ligand from the bound ground-state conformation of the protein (see Fig 1(b)). A decrease of the dominant relaxation rate $k_{\rm obs}$ with increasing ligand concentration $[L]_0$ thus indicates conformational-selection binding [25]. However, an increase of $k_{\rm obs}$ with $[L]_0$ under pseudo-first-order conditions is possible both for induced-fit binding and conformational-selection binding and does not uniquely point towards a binding mechanism [22].

In this article, we present general analytical results for the dominant relaxation rate $k_{\rm obs}$ of induced-fit binding and conformational-selection binding processes that hold for all ligand and protein concentrations. Our general results are based on an expansion of the rate equations for these binding processes around the equilibrium concentrations of ligands and proteins, and include the pseudo-first-order results in the limit of large ligand concentrations. In the case of induced-fit binding, we find that $k_{\rm obs}$ exhibits a minimum at the total ligand concentration $[L]_0^{\rm min} = [P]_0 - K_d$ for total protein concentrations $[P]_0$ that are larger than the overall dissociation constant K_d of the binding process. As a characteristic feature, the function $k_{\rm obs}([L]_0)$ for induced-fit binding is symmetric with respect to this minimum. At sufficiently large protein concentrations $[P]_0$, the function $k_{\rm obs}([L]_0)$ tends to identical values for small ligand concentrations $[L]_0 \ll [P]_0$ and for large ligand concentrations $[L]_0 \gg [P]_0$ because of its symmetry (see Fig 1(c)). In the case of conformational-selection binding, we find that $k_{\rm obs}$ exhibits a minimum for conformational excitation rates $k_e > k_-$ and sufficiently large protein concentrations $[P]_0$ (see Fig 1(d)). The location $[L]_0^{\rm min}$ of this minimum depends on $[P]_0$, K_d, and the rates k_e and k_- (see Eq (10) below). In contrast to induced-fit binding, the function $k_{\rm obs}([L]_0)$ for conformational-selection binding is not symmetric with respect to this minimum. At sufficiently large protein concentrations $[P]_0$, the function $k_{\rm obs}([L]_0)$ attains values for small ligand concentrations $[L]_0 \ll [P]_0$ that can greatly exceed the values for large ligand concentrations $[L]_0 \gg [P]_0$ (see Fig 1(d)). For excitation rates $k_e < k_-$ of conformational-selection binding processes, the dominant relaxation rate $k_{\rm obs}$ decreases monotonically with increasing ligand concentration $[L]_0$ (see Fig 1(e)). Our general results for the dominant relaxation rate $k_{\rm obs}$ of induced-fit and conformational-selection binding processes allow to clearly distinguish between these two binding mechanisms for sufficiently large protein concentrations $[P]_0$ (see Figs 2 and 3 below for numerical examples).

Results

Solving the rate equations of the induced-fit and conformational-selection binding models shown in Fig 1(a) and 1(b) is complicated by the fact that the binding steps in these models are second-order reactions that depend on the product of the time-dependent concentrations of unbound proteins and unbound ligands. In the standard pseudo-first-order approximation, the rate equations are simplified by assuming that the total ligand concentration greatly

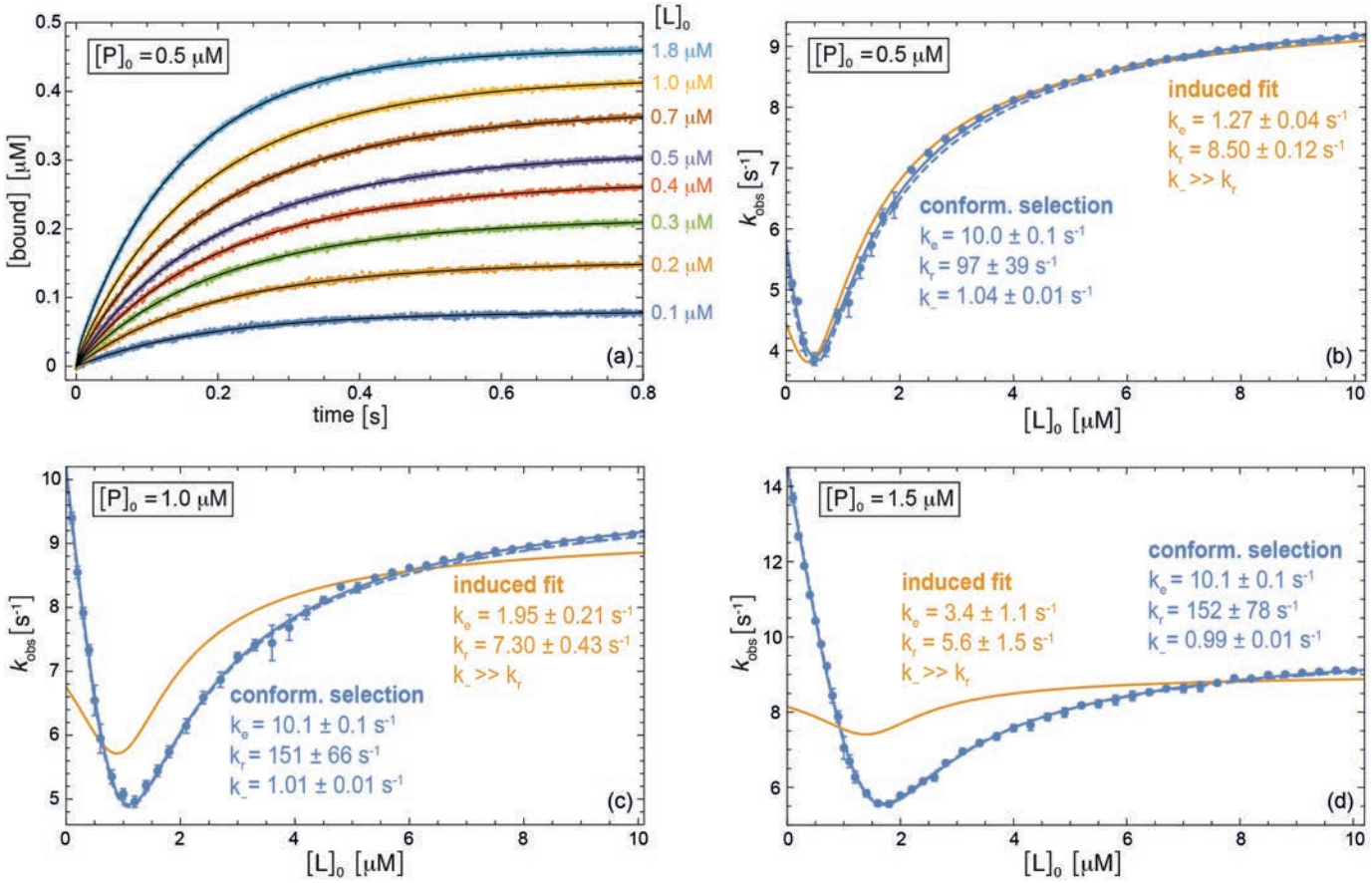

Fig 2. Numerical example for conformational-selection binding with the rate constants $k_e = 10\ s^{-1}$, $k_r = 100\ s^{-1}$, $k_+ = 100\ \mu M^{-1}s^{-1}$, and $k_- = 1\ s^{-1}$. (a) Relaxation data for the bound complex obtained by numerical integration of the rate equations and subsequent addition of Gaussian noise with amplitude 0.002 μM at the total protein concentration $[P]_0 = 0.5$ μM and exemplary total ligand concentrations $[L]_0$. The black lines represent multi-exponential fits of the data points. (b) to (d) Comparison of k_{obs} values obtained from multi-exponential fits of numerical relaxation data (points) to our theoretical results for k_{obs} (lines) at the three different total protein concentrations $[P]_0 = 0.5$ μM, 1.0 μM, and 1.5 μM and total ligand concentrations $[L]_0$ between 0.1 μM and 10 μM. The full lines represent fits of Eq (6) for conformational-selection binding (blue) and of Eq (1) for induced-fit binding (orange), with fit parameter values specified in the figure. In these fits, the dissociation constant $K_d = 0.11$ μM is assumed to be known from equilibrium data. The dashed blue lines are obtained from Eq (6) for the 'true' rate constants of the numerical example.

doi:10.1371/journal.pcbi.1005067.g002

exceeds the total protein concentration, so that the amount of ligand consumed during binding is negligible compared to the total amount of ligand. The concentration of the unbound ligand then can be taken to be constant, and the rate equations only contain terms that are linear in the time-dependent concentration of the protein, which makes them solvable. In our more general approach, a linearization of the rate equations is achieved by expanding around the equilibrium concentrations of the bound and unbound proteins and ligands (see Methods). This expansion captures the final relaxation into equilibrium, which is governed by the smallest, dominant relaxation rate k_{obs}, for all concentrations of proteins and ligands, and leads to general results for k_{obs} that include the results from the pseudo-first-order approximation in the limit of large ligand concentrations.

Dominant relaxation rate of induced-fit binding

Expanding the rate equations of the induced-fit binding mechanism shown in Fig 1(a) around the equilibrium concentrations of proteins and ligands leads to the dominant, smallest

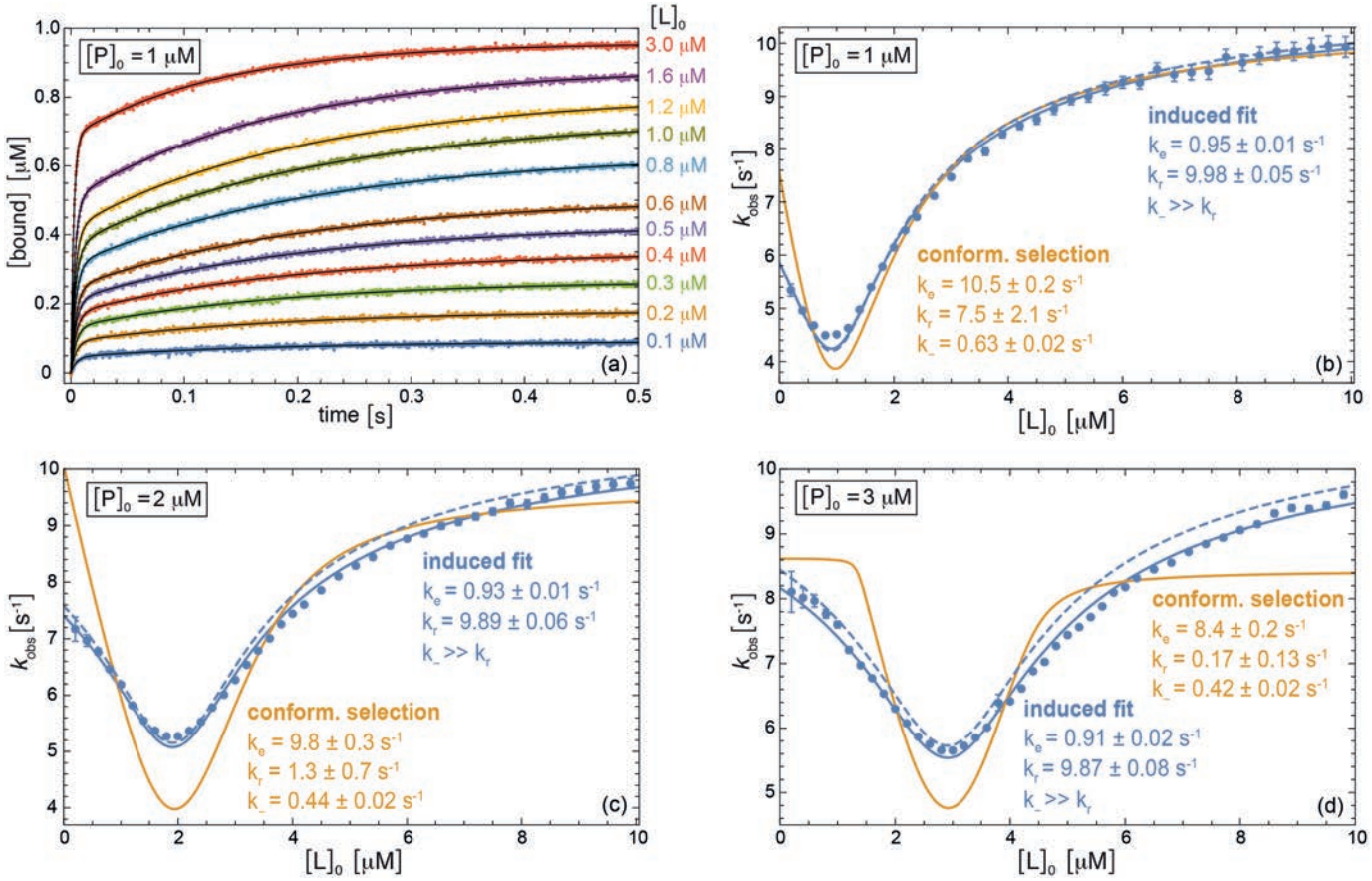

Fig 3. Numerical example for induced-fit binding with the rate constants $k_+ = 100\ \mu M^{-1}s^{-1}$, $k_- = 100\ s^{-1}$, $k_e = 1\ s^{-1}$, and $k_r = 10\ s^{-1}$. (a) Relaxation data for the bound complex obtained by numerical integration of the rate equations and subsequent addition of Gaussian noise with amplitude 0.004 μM at the total protein concentration $[P]_0 = 1$ μM and exemplary total ligand concentrations $[L]_0$. The black lines represent multi-exponential fits of the data points. (b) to (d) Comparison of k_{obs} values obtained from multi-exponential fits of numerical relaxation data (points) to our theoretical results for k_{obs} (lines) at the three different total protein concentrations $[P]_0 = 1$ μM, 2 μM, and 3 μM and total ligand concentrations $[L]_0$ between 0.1 μM and 10 μM. The full lines represent fits of Eq (1) for induced-fit binding (blue) and of Eq (6) for conformational-selection binding (orange), with fit parameter values specified in the figure. In these fits, the dissociation constant $K_d = 1/11$ μM is assumed to be known from equilibrium data. The dashed blue lines are obtained from Eq (1) for the 'true' rate constants of the numerical example.

doi:10.1371/journal.pcbi.1005067.g003

relaxation rate (see Methods)

$$k_{\text{obs}} = k_e + k_r + \frac{1}{2}\gamma - \frac{1}{2}\sqrt{\gamma^2 + 4k_-k_r} \tag{1}$$

with

$$\gamma = -k_e - k_r + k_- + k_+(\delta - K_d) \tag{2}$$

$$\delta = \sqrt{\left([\mathrm{L}]_0 - [\mathrm{P}]_0 + K_d\right)^2 + 4[\mathrm{P}]_0 K_d} \tag{3}$$

and with the overall dissociation constant

$$K_d = \frac{k_-k_e}{k_+(k_e + k_r)} \tag{4}$$

of induced-fit binding. This general result for k_{obs} holds for all total ligand concentrations $[L]_0$ and protein concentrations $[P]_0$. In the limit of large ligand concentrations $[L]_0 \gg [P]_0$, we obtain $\delta \simeq [L]_0 + K_d$ and $\gamma \simeq -k_e - k_r + k_- + k_+[L]_0$ from Eqs (2) and (3), which agrees with results derived in pseudo-first-order approximation [21, 22].

As a function of the total ligand concentration $[L]_0$, the dominant relaxation rate k_{obs} exhibits a minimum at

$$[L]_0^{min} = [P]_0 - K_d \tag{5}$$

for total protein concentrations $[P]_0 > K_d$. The function $k_{obs}([L]_0)$ is symmetric with respect to $[L]_0^{min}$ (see (Fig 1(c)). This symmetry and the location $[L]_0^{min}$ of the minimum result from the fact that k_{obs} depends on $[L]_0$ only via the term δ, which is minimal at $[L]_0^{min}$ and symmetric with respect to $[L]_0^{min}$. The dominant relaxation rate k_{obs} is minimal when δ is minimal. For large ligand concentrations $[L]_0$, k_{obs} tends towards the maximum value $k_e + k_r$ as in pseudo-first-order approximation. The location $[L]_0^{min}$ of the minimum and the symmetry of the function $k_{obs}([L]_0)$ with respect to this minimum are properties that the induced-fit binding model appears to 'inherit' from the elementary binding model P + L $\rightleftharpoons$ PL (see Eq (46) in Methods section). However, the function $k_{obs}([L]_0)$ of the elementary binding model is V-shaped and does not tend to a constant maximum value for large ligand concentrations $[L]_0$.

Dominant relaxation rate of conformational-selection binding

For the conformational-selection binding mechanism shown in Fig 1(b), an expansion of the rate equations around the equilibrium concentrations of proteins and ligands leads to the dominant, smallest relaxation rate (see Methods)

$$k_{obs} = k_e + \frac{1}{2}\alpha - \frac{1}{2}\sqrt{\alpha^2 + \beta} \tag{6}$$

with

$$\alpha = k_r - k_e + \frac{k_-\left((2k_e + k_r)\delta + k_r\left([L]_0 - [P]_0 - K_d\right)\right)}{2k_eK_d} \tag{7}$$

$$\beta = 2k_r\left(2k_e - k_- - \frac{k_-\left(\delta - [L]_0 + [P]_0\right)}{K_d}\right) \tag{8}$$

and δ as in Eq (3), and with the overall dissociation constant

$$K_d = \frac{k_-(k_e + k_r)}{k_+k_e} \tag{9}$$

of conformational-selection binding. This general result for k_{obs} holds for all total ligand concentrations $[L]_0$ and protein concentrations $[P]_0$. In the limit of large ligand concentrations $[L]_0 \gg [P]_0$, we obtain $\alpha \simeq -k_e + k_r + k_- + k_+[L]_0$ and $\beta \simeq 4k_r(k_e - k_-)$ from Eqs (3), (7) and (8), in agreement with results derived in pseudo-first-order approximation [21, 22].

For conformational-selection binding, the shape of the function $k_{obs}([L]_0)$ depends on the values of the conformational excitation rate k_e and the unbinding rate k_- (see Fig 1(d) and 1(e)). For $k_e < k_-$, the dominant relaxation rate k_{obs} decreases monotonically with increasing total ligand concentration $[L]_0$. For $k_e > k_-$, the dominant relaxation rate k_{obs} exhibits a minimum as a function of $[L]_0$ at sufficiently large total protein concentrations $[P]_0$. The minimum

is located at (see Methods)

$$[\mathrm{L}]_0^{\min} \simeq \frac{k_e + k_-}{k_e - k_-}[\mathrm{P}]_0 - K_d \tag{10}$$

if the conformational relaxation rate k_r is much larger than the excitation rate k_e, which typically holds for the conformational exchange between ground-state and excited-state conformations of proteins. In contrast to induced-fit binding, the function $k_{\mathrm{obs}}([\mathrm{L}]_0)$ is not symmetric with respect to this minimum. For large ligand concentrations, the limiting value of the dominant relaxation rate is $k_{\mathrm{obs}}(\infty) = k_e$ as in pseudo-first-order approximation. For vanishing ligand concentrations $[\mathrm{L}]_0 \to 0$, the limiting value is $k_{\mathrm{obs}}(0) = k_e + k_r$ for total protein concentrations $[\mathrm{P}]_0 > K_d(k_e + k_r - k_-)/k_-$ and $k_{\mathrm{obs}}(0) = k_-([\mathrm{P}]_0 + K_d)/K_d$ for $[\mathrm{P}]_0 < K_d(k_e + k_r - k_-)/k_-$.

Distinguishing induced fit and conformational selection

The general results for the dominant relaxation rate k_{obs} presented in the previous sections allow to clearly distinguish induced-fit from conformational-selection binding processes. In Fig 2, we consider a conformational-selection binding process with the rate constants $k_e = 10\ \mathrm{s}^{-1}$, $k_r = 100\ \mathrm{s}^{-1}$, $k_+ = 100\ \mu\mathrm{M}^{-1}\mathrm{s}^{-1}$, and $k_- = 1\ \mathrm{s}^{-1}$ as a numerical example. The data points in Fig 2(a) represent relaxation curves for the bound complex that have been generated by numerical integration of the rate equations and subsequent addition of Gaussian noise to mimic measurement errors. The black lines in Fig 2(a) are multi-exponential fits of the data points. The number of exponentials in these fits has been determined with the Akaike information criterion (AIC), which is a standard criterion for the trade-off between quality of fit and number of fit parameters, and ranges from 2 to 4. The data points in Fig 2(b) to 2(d) represent the dominant relaxation rates k_{obs} that are obtained from multi-exponential fits of relaxation curves for different total ligand concentrations $[\mathrm{L}]_0$ and total protein concentrations $[\mathrm{P}]_0$. The dominant relaxation rate k_{obs} here is identified as the smallest relaxation rate of a multi-exponential fit. The full blue lines in Fig 2(b) to 2(d) result from fitting our general result Eq (6) for conformational-selection binding to the k_{obs} data points. The full orange lines represent fits of our general result Eq (1) for induced-fit binding. For all fits, we assume that the dissociation constant $K_d = 0.11\ \mu\mathrm{M}$ is known from equilibrium data, and use k_e, k_r, and k_- as fit parameters. Finally, the blue dashed lines in Fig 2(b) to 2(d) are the k_{obs} curves obtained from Eq (6) for the 'true' rate constants of the conformational-selection binding process given above. These dashed lines agree with the data points, which indicates that the k_{obs} values from multi-exponential fits as in Fig 2(a) are identical to the values obtained from Eq (6) within the statistical errors of the numerical example.

The fits in Fig 2(b) to 2(d) clearly identify conformational selection as the correct binding mechanism in this example. The blue fit curves for conformational selection agree with the data points within statistical errors, while the orange fit curves for induced fit deviate from the data. For conformational-selection binding, the fit values of the conformational transition rates k_e and k_r and of the unbinding rate k_- specified in the figure agree with the correct values $k_e = 10\ \mathrm{s}^{-1}$, $k_r = 100\ \mathrm{s}^{-1}$, and $k_- = 1\ \mathrm{s}^{-1}$ of the numerical example within statistical errors.

In Fig 3, we consider an induced-fit binding process with rate constants $k_+ = 100\ \mu\mathrm{M}^{-1}\mathrm{s}^{-1}$, $k_- = 100\ \mathrm{s}^{-1}$, $k_e = 1\ \mathrm{s}^{-1}$, and $k_r = 10\ \mathrm{s}^{-1}$ as a second numerical example. The k_{obs} data points in Fig 3(b) to 3(d) are again obtained from multi-exponential fits of relaxation curves that have been generated by numerical integration of the rate equations and subsequent addition of Gaussian noise (see Fig 3(a)). The fits in Fig 3(b) to 3(d) clearly identify induced-fit binding as the correct mechanism in this example. The full blue curves that represent fits of Eq (1) for induced-fit binding are in overall agreement with the k_{obs} points, while the orange fit curves of

Eq (6) for conformational-selection binding deviate from the data. The fit values of the conformational transition rates k_e and k_r for the induced-fit binding model are in good agreement with the correct values $k_e = 1\ s^{-1}$, and $k_r = 10\ s^{-1}$ of the example. The dashed blue curves in Fig 3(b) to 3(d), which are obtained from Eq (1) for the 'true' rate constants of the induced-fit binding process, are in overall agreement with the data points. Slight deviations result from the fact that the amplitude of the slow relaxation mode with rate k_{obs} is rather small compared to the amplitude of the fast modes (see Fig 3(a)), which can lead to numerical inaccuracies.

In both numerical examples of Figs 2 and 3, the correct binding mechanism cannot be identified under pseudo-first-order conditions because k_{obs} is monotonically increasing with $[L]_0$ for ligand concentrations that greatly exceed the protein concentration $[P]_0$ [22].

Analysis of chemical relaxation rates for recoverin binding

Chakrabarti et al. [28] have recently investigated the conformational dynamics and binding kinetics of the protein recoverin with chemical relaxation and advanced NMR experiments. Recoverin exhibits a conformational change during binding of its ligand, which is a rhodopsin kinase peptide fused to the B1 domain of immunoglobulin protein G in the experiments of Chakrabarti et al. [28]. The data points in Fig 4 represent the dominant relaxation rates k_{obs} obtained by Chakrabarti et al. from relaxation experiments at the temperatures 30°C and 10°C for a recoverin concentration of 10μM. The lines in Fig 4 result from fitting our general results Eqs (1) and (6) for the dominant relaxation rate k_{obs} of induced-fit and conformational-selection binding processes. In these fits, we have used the values $K_d = 1.0 \pm 0.2$ μM and $K_d = 1.8 \pm 0.2$ μM obtained by Chakrabarti et al. from isothermal titration calometry experiments at 30°C and 10°C, which reduces the parameters to k_e, k_r, and k_-. The fits of our general result Eq (6) for conformational-selection binding are rather insensitive to the relaxation rate k_r, which is illustrated in Fig 4 by nearly identical fits for $k_r = 100\ s^{-1}$ and $k_r = 1000\ s^{-1}$

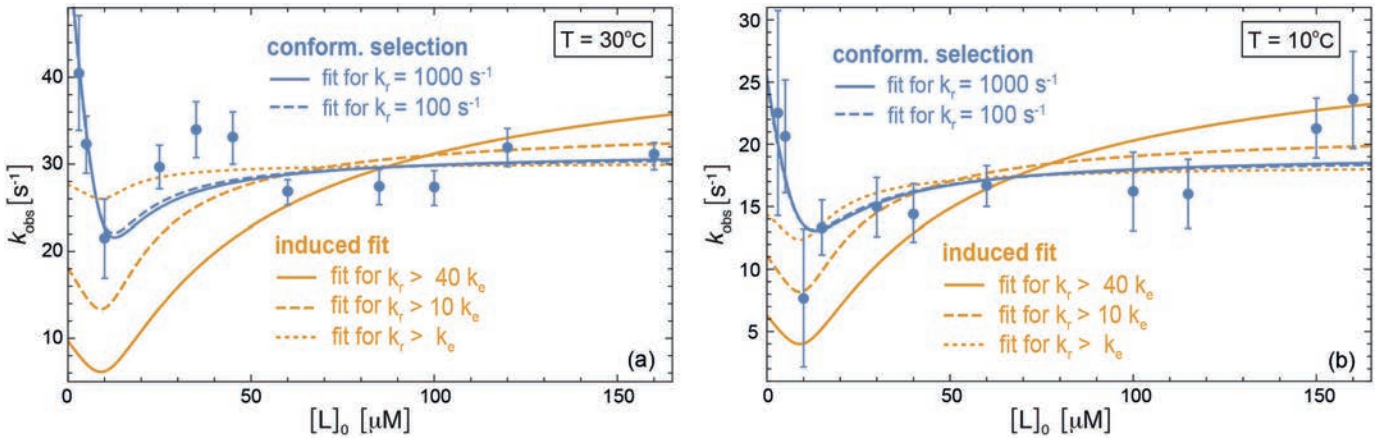

Fig 4. Analysis of experimentally determined relaxation rates k_{obs} for the binding of recoverin to a rhodopsin kinase peptide ligand. The data points represent results of Chakrabarti et al. [28] obtained from chemical relaxation experiments at the temperatures 30°C and 10°C for a recoverin concentration of 10 μM. The blue lines result from fits of Eq (6) for conformational-selection binding with the values $k_r = 1000\ s^{-1}$ (full) and $k_r = 100\ s^{-1}$ (dashed) of the conformational relaxation rate. At 30°C, the parameter values obtained from fitting are $k_e = 31.5 \pm 0.8\ s^{-1}$ and $k_- = 5.1 \pm 0.4\ s^{-1}$ for $k_r = 1000\ s^{-1}$, and $k_e = 31.1 \pm 0.8\ s^{-1}$ and $k_- = 5.0 \pm 0.4\ s^{-1}$ for $k_r = 100\ s^{-1}$. At 10°C, the fit parameter values are $k_e = 19.3 \pm 1.4\ s^{-1}$ and $k_- = 3.9 \pm 0.7\ s^{-1}$ for $k_r = 1000\ s^{-1}$, and $k_e = 19.0 \pm 1.3\ s^{-1}$ and $k_- = 3.8 \pm 0.7\ s^{-1}$ for $k_r = 100 s^{-1}$. The yellow lines represent fits of Eq (1) for induced-fit binding with constraints on the conformational excitation and relaxation rates k_e and k_r. At 30°C, the obtained fit values for the conformational exchange rates are $k_e = k_r = 15 \pm 10\ s^{-1}$ for the constraint $k_r > k_e$, $k_e = 3.1 \pm 1.9\ s^{-1}$ and $k_r = 31 \pm 4 s^{-1}$ for the constraint $k_r > 10k_e$, and $k_e = 1.1 \pm 0.8\ s^{-1}$ and $k_r = 44 \pm 8\ s^{-1}$ for $k_r > 40k_e$. At 10°C, the fit values are $k_e = 4.5 \pm 4.0\ s^{-1}$ and $k_r = 14 \pm 10\ s^{-1}$ for the constraint $k_r > k_e$, $k_e = 1.9 \pm 1.5\ s^{-1}$ and $k_r = 19 \pm 5\ s^{-1}$ for $k_r > 10k_e$, and $k_e = 0.7 \pm 0.5\ s^{-1}$ and $k_r = 28 \pm 11\ s^{-1}$ for $k_r > 40k_e$. In all fits of Eq (1) for induced-fit binding, we obtain $k_- \gg k_r$, i.e. the fit values of the unbinding rate k_- are much larger than the conformational relaxation rate k_r and cannot be specified.

doi:10.1371/journal.pcbi.1005067.g004

(see dashed and full blue lines). Our fit values for the conformational excitation rate k_e specified in the figure caption agree with the values $k_e = 33 \pm 5\ s^{-1}$ and $k_e = 23 \pm 5\ s^{-1}$ obtained by Chakrabarti et al. from advanced NMR experiments at 30°C and 10°C, respectively. From these experiments, Chakrabarti et al. obtain the values $k_r = 990 \pm 100\ s^{-1}$ and $k_r = 920 \pm 200\ s^{-1}$ at 30°C and 10°C, which cannot be deduced from our fits of the k_{obs} data because these fits are insensitive to k_r. The NMR experiments indicate that the higher-energy conformation of unbound recoverin resembles the ground-state conformation of bound recoverin [28] as required for the conformational-selection binding mechanism illustrated in Fig 1(b), and that the excited-state conformation of unbound recoverin has the equilibrium occupancy $P_e = k_e/(k_r + k_e) = 3.2\% \pm 0.5\%$ at 30°C and $P_e = 2.4\% \pm 0.7\%$ at 10°C, relative to the ground-state conformation.

Fits of our general result Eq (1) for the dominant relaxation rate k_{obs} of induced-fit binding with unconstrained parameters k_e, k_r, and k_- lead to fit values for the conformational exchange rates k_e and k_r with $k_e \gg k_r$. For such values of k_e and k_r, the conformation 1 of the induced-fit binding model illustrated in Fig 1(a) is the ground-state conformation both for the unbound state and the bound state of recoverin, which contradicts the experimental observation that recoverin changes its conformation during binding [28]. Distinct ground-state conformations for the unbound and bound state of recoverin can be enforced by constraining k_r to values larger than k_e. The yellow lines in Fig 4 result from fits with the constraints $k_r > k_e$, $k_r > 10k_e$, and $k_r > 40k_e$. These constraints correspond to equilibrium occupancies P_e of the excited-state conformation of bound recoverin with $P_e < 50\%$, $P_e < 9.1\%$, and $P_e < 2.4\%$, respectively. The fits of Eq (1) for induced-fit binding with the constraints $k_r > 10k_e$ and $k_r > 40k_e$ deviate rather strongly from the two data points with the smallest ligand concentrations $[L]_0 = 3\ \mu M$ and $5\ \mu M$, in contrast to fits of Eq (6) for conformational-selection binding (blue lines). A Bayesian model comparison of conformational-selection binding and induced-fit binding based on Eqs (1) and (6) leads to Bayes factors of $9.8 \cdot 10^{13}$ and $1.5 \cdot 10^{23}$ at 30°C for the constraints $k_r > 10k_e$ and $k_r > 40k_e$, and to Bayes factors of $4.2 \cdot 10^3$ and $9.6 \cdot 10^9$ at 10°C for $k_r > 10k_e$, and $k_r > 40k_e$, respectively (see Methods for details). These Bayes factors indicate that the k_{obs} data of Fig 4 strongly point towards conformational-selection binding. Bayes factors larger than 10^2 are generally considered to be decisive [37]. For the bound recoverin complex, Chakrabarti et al. did not observe an excited-state conformation in NMR experiments, which limits the excited-state occupancy P_e to undetectable values smaller than 1% for a conformational exchange that is fast compared to the NMR timescale as in the case of unbound recoverin. The analysis of the experimental data for the dominant relaxation rate k_{obs} of recoverin binding based on our general results Eqs (1) and (6) thus indicates a conformational-selection binding mechanism, in agreement with a numerical analysis of Chakrabarti et al. [28]. In this numerical analysis, Chakrabarti et al. include the chemical relaxation data for recoverin binding, additional relaxation data from dilution experiments, the values for the conformational exchange rates k_e and k_r obtained from NMR experiments, and the K_d values deduced from isothermal titration calometry [28]. In contrast, our analysis of the k_{obs} data in Fig 4 from the chemical relaxation experiments of recoverin binding only includes the K_d values from isothermal titration calometry as additional input.

Discussion

We have shown here that the dominant rate k_{obs} of chemical relaxation experiments with total protein and ligand concentrations of comparable magnitude conveys information on the binding mechanism and conformational transition rates of proteins. For sufficiently large protein concentrations $[P]_0$, the function $k_{obs}([L]_0)$ obtained from such experiments has characteristic

features that are clearly distinct for induced-fit binding and conformational-selection binding. The function $k_{obs}([L]_0)$ of induced-fit binding exhibits a characteristic symmetry around a minimum and tends to identical values for small and large ligand concentrations $[L]_0$ as in Fig 1(c) if the protein concentration $[P]_0$, which determines the location of the minimum, is sufficiently large. In contrast, the function $k_{obs}([L]_0)$ of conformational-selection binding is either monotonically decreasing for $k_e < k_-$, or asymmetric around a minimum for $k_e > k_-$. In both cases, $k_{obs}([L]_0)$ tends for small ligand concentrations $[L]_0$ to values that exceed the values for large ligand concentrations as in Fig 1(d) and 1(e) if the protein concentration $[P]_0$ is sufficiently large.

Our general results for the dominant rate k_{obs} of chemical relaxation experiments thus provide a transparent route to distinguish induced-fit binding from conformational-selection binding based on the shape of the function $k_{obs}([L]_0)$, and to infer conformational transition rates from fitting. Alternatively, these binding mechanisms can be identified from a numerical analysis of time-dependent relaxation curves [26–28], based on steric effects that may prohibit ligand entry and exit in the bound ground-state conformation of the protein and, thus, rule out conformational-selection binding [15], from a comparison of conformational excitation rates to overall, effective binding and unbinding rates [4, 13], or from the effect of distal mutations that mainly affect the conformational exchange, but not the binding kinetics in different protein conformations [13, 16, 21, 38]. Of particular interest is how such mutations change the overall binding and unbinding rates. If both conformational-selection and induced-fit binding are viable, increasing the ligand concentration may shift the binding mechanism from conformational selection to induced fit [16, 18, 26, 39, 40].

Methods

Near-equilibrium relaxation of induced-fit binding

The induced-fit binding model of Fig 1(a) leads to the four rate equations

$$\frac{\mathrm{d}}{\mathrm{d}t}[P_1] = -k_+[P_1][L] + k_-[P_1L] \tag{11}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}[L] = -k_+[P_1][L] + k_-[P_1L] \tag{12}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}[P_1L] = k_+[P_1][L] - k_-[P_1L] + k_e[P_2L] - k_r[P_1L] \tag{13}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}[P_2L] = k_r[P_1L] - k_e[P_2L] \tag{14}$$

that describe the time-dependent evolution of the concentration $[P_1]$ of the unbound protein, the concentration $[L]$ of the unbound ligand, and the concentrations $[P_1L]$ and $[P_2L]$ of the bound complexes. These four rate equations are redundant because the total concentrations $[P]_0$ and $[L]_0$ of proteins and ligands are conserved:

$$[P_1L] + [P_2L] + [P_1] = [P]_0 \tag{15}$$

$$[L] + [P_1L] + [P_2L] = [L]_0 \tag{16}$$

With Eqs (15) and (16), the concentrations $[P_1]$ and $[P_1L]$ can be expressed in terms of $[L]$ and

$[P_2L]$, which results in the two non-redundant rate equations

$$\frac{d}{dt}[L] = -k_+([L] - [L]_0 + [P]_0)[L] + k_-([L]_0 - [L] - [P_2L]) \tag{17}$$

$$\frac{d}{dt}[P_2L] = k_r([L]_0 - [L] - [P_2L]) - k_e[P_2L] \tag{18}$$

These rate equations can be written in the vectorial form

$$\frac{d}{dt}\mathbf{c} = \mathbf{F}(\mathbf{c}) \tag{19}$$

with

$$\mathbf{c}(t) \equiv \begin{pmatrix} [L](t) \\ [P_2L](t) \end{pmatrix} \tag{20}$$

The two components of the vector $\mathbf{F}(\mathbf{c})$ in Eq (19) are the right-hand sides of the Eqs (17) and (18). The rate equations describe the temporal evolution of the concentrations $[L]$ and $[P_2L]$ towards equilibrium, and are nonlinear because of the quadratic term in $[L]$ on the right-hand side of Eq (17).

To obtain linearized versions of the rate equations that describe the slow processes corresponding to the final relaxation into equilibrium, we expand the vector $\mathbf{F}(\mathbf{c})$ in Eq (19) around the equilibrium concentrations $\mathbf{c}_{eq}$:

$$\mathbf{F}(\mathbf{c}) = \mathbf{F}(\mathbf{c}_{eq} + \Delta\mathbf{c}) \simeq \mathbf{F}(\mathbf{c}_{eq}) + J(\mathbf{c}_{eq})\Delta\mathbf{c} = J(\mathbf{c}_{eq})\Delta\mathbf{c} \tag{21}$$

Here, J is the Jacobian matrix of $\mathbf{F}$ with elements $J_{ij} = \partial F_i/\partial c_j$. The right-hand side of Eq (21) follows from $\mathbf{F}(\mathbf{c}_{eq}) = 0$. Inserting the expansion (21) into Eq (19) and making use of $\frac{d}{dt}\mathbf{c} = \frac{d}{dt}(\mathbf{c}_{eq} + \Delta\mathbf{c}) = \frac{d}{dt}\Delta\mathbf{c}$ leads to the linearized rate equations

$$\frac{d}{dt}\Delta\mathbf{c} = J(\mathbf{c}_{eq})\Delta\mathbf{c} \tag{22}$$

with

$$J(\mathbf{c}_{eq}) = \begin{pmatrix} k_+\left([L]_0 - 2[L]_{eq} - [P]_0\right) - k_- & -k_- \\ -k_r & -k_e - k_r \end{pmatrix} \tag{23}$$

and the equilibrium concentration of the unbound ligand

$$[L]_{eq} = \frac{1}{2}\left([L]_0 - [P]_0 - K_d + \sqrt{\left([L]_0 - [P]_0 + K_d\right)^2 + 4[P]_0K_d}\right) \tag{24}$$

The overall dissociation constant K_d of the induced-fit binding process is given in Eq (4). The relaxation rates of the linearized rate Eq (22) are the two eigenvalues of the matrix $-J(\mathbf{c}_{eq})$. These eigenvalues are k_{obs} given in Eq (1) and

$$k_2 = k_e + k_r + \frac{1}{2}\gamma + \frac{1}{2}\sqrt{\gamma^2 + 4k_-k_r} \tag{25}$$

with γ and δ given in Eqs (2) and (3). The relaxation rate k_{obs} is smaller than k_2 and, thus, dominates the final relaxation into equilibrium.

Near-equilibrium relaxation of conformational-selection binding

The four rate equations of the conformational-selection binding model of Fig 1(b) are

$$\frac{\mathrm{d}}{\mathrm{d}t}[\mathrm{P}_1] = -k_e[\mathrm{P}_1] + k_r[\mathrm{P}_2] \tag{26}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}[\mathrm{P}_2] = k_e[\mathrm{P}_1] - k_r[\mathrm{P}_2] + k_-[\mathrm{P}_2\mathrm{L}] - k_+[\mathrm{P}_2][\mathrm{L}] \tag{27}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}[\mathrm{L}] = k_-[\mathrm{P}_2\mathrm{L}] - k_+[\mathrm{P}_2][\mathrm{L}] \tag{28}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}[\mathrm{P}_2\mathrm{L}] = -k_-[\mathrm{P}_2\mathrm{L}] + k_+[\mathrm{P}_2][\mathrm{L}] \tag{29}$$

The total concentrations $[\mathrm{L}]_0$ and $[\mathrm{P}]_0$ of the ligands and proteins are conserved:

$$[\mathrm{L}] + [\mathrm{P}_2\mathrm{L}] = [\mathrm{L}]_0 \tag{30}$$

$$[\mathrm{P}_1] + [\mathrm{P}_2] + [\mathrm{P}_2\mathrm{L}] = [\mathrm{P}]_0 \tag{31}$$

With these equations, the concentrations $[\mathrm{P}_1]$ and $[\mathrm{P}_2\mathrm{L}]$ can be expressed in terms of $[\mathrm{L}]$ and $[\mathrm{P}_2]$, which leads to the two rate equations

$$\frac{\mathrm{d}}{\mathrm{d}t}[\mathrm{P}_2] = k_e\big([\mathrm{P}]_0 - [\mathrm{P}_2]\big) - (k_r + k_+[\mathrm{L}])[\mathrm{P}_2] + (k_- - k_e)([\mathrm{L}]_0 - [\mathrm{L}]) \tag{32}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}[\mathrm{L}] = k_-([\mathrm{L}]_0 - [\mathrm{L}]) - k_+[\mathrm{P}_2][\mathrm{L}] \tag{33}$$

These rate equations can be written in the vectorial form of Eq (19) with

$$\mathbf{c}(t) \equiv \begin{pmatrix} [\mathrm{P}_2](t) \\ [\mathrm{L}](t) \end{pmatrix} \tag{34}$$

and with a vector $\mathbf{F}(\mathbf{c})$ that contains the right-hand sides of the Eqs (32) and (33) as components. An expansion of the vector $\mathbf{F}(\mathbf{c})$ around the equilibrium concentrations $\mathbf{c}_{\mathrm{eq}}$ leads to Eq (22) with the Jacobian matrix

$$J(\mathbf{c}_{\mathrm{eq}}) = -\begin{pmatrix} k_r + k_e + k_+[\mathrm{L}]_{\mathrm{eq}} & -k_e + k_- + k_+[\mathrm{P}_2]_{\mathrm{eq}} \\ k_+[\mathrm{L}]_{\mathrm{eq}} & k_- + k_+[\mathrm{P}_2]_{\mathrm{eq}} \end{pmatrix} \tag{35}$$

and the equilibrium concentrations

$$[\mathrm{P}_2]_{\mathrm{eq}} = \frac{1}{2K_d}\frac{k_-}{k_+}\left([\mathrm{P}]_0 - [\mathrm{L}]_0 - K_d + \sqrt{([\mathrm{P}]_0 - [\mathrm{L}]_0 - K_d)^2 + 4K_d[\mathrm{P}]_0}\right) \tag{36}$$

$$[\mathrm{L}]_{\mathrm{eq}} = \frac{1}{2}\left([\mathrm{L}]_0 - [\mathrm{P}]_0 - K_d + \sqrt{([\mathrm{P}]_0 - [\mathrm{L}]_0 - K_d)^2 + 4K_d[\mathrm{P}]_0}\right) \tag{37}$$

The overall dissociation constant K_d of the conformational-selection binding process is given in Eq (9). The relaxation rates of the linearized rate equations are the two eigenvalues of the

matrix $-J(\mathbf{c}_{eq})$. These eigenvalues are k_{obs} given in Eq (6) and

$$k_2 = k_e + \frac{1}{2}\alpha + \frac{1}{2}\sqrt{\alpha^2 + \beta} \tag{38}$$

with α and β given in Eqs (7) and (8). The relaxation rate k_{obs} is smaller than k_2 and therefore dominates the final relaxation into equilibrium.

To derive Eq (10) for the location of the minimum of k_{obs} as a function of the total ligand concentration $[L]_0$, we now consider the near-equilibrium relaxation of the conformational-selection model in quasi-steady-state approximation (qssa), which assumes that the concentration of the intermediate $[P_2]$ does not change in time. The left-hand side of Eq (32) then is equal to zero, and the two Eqs (32) and (33) reduce to the single equation

$$\frac{\mathrm{d}}{\mathrm{d}t}[L] = -k_e k_- \frac{([L] + K_d)([L] - [L]_0) + [L][P]_0}{k_-[L] + k_e K_d} = f([L]) \tag{39}$$

An expansion of the function $f([L])$ around the equilibrium concentration $[L]_{eq}$ leads to the linear equation $\mathrm{d}[L]/\mathrm{d}t \simeq -k_{obs}^{(qssa)}([L] - [L]_{eq})$ with

$$k_{obs}^{(qssa)} = -\left.\frac{\mathrm{d}f([L])}{\mathrm{d}[L]}\right|_{[L]=[L]_{eq}} = \frac{k_- k_e \delta}{k_e K_d + k_-[L]_{eq}} \tag{40}$$

and δ and $[L]_{eq}$ given in Eqs (3) and (37). The derivative of $k_{obs}^{(qssa)}$ is zero at $[L]_0 = [L]_0^{min}$ with $[L]_0^{min}$ given in Eq (10). In general, the quasi-steady-state result $k_{obs}^{(qssa)}$ is a good approximation of k_{obs} if the rates for the transitions out of the intermediate state P_2 of conformational-selection binding are much larger than the rates for the transitions to P_2. A numerical analysis shows that the location $[L]_0^{min}$ of the minimum of $k_{obs}^{(qssa)}([L])$ is in good agreement with the location of the minimum of $k_{obs}([L])$ for conformational transitions rates with $k_r \gg k_e$.

Multi-exponential relaxation

In the numerical examples illustrated in Figs 2 and 3, chemical relaxation curves for conformational-selection and induced-fit binding are fitted with a multi-exponential model. Such multi-exponential models are an adequate description for the time evolution of concentrations in first-order chemical reactions. However, the binding steps of the induced-fit and conformational-selection models of Fig 1(a) and 1(b) are of second order. To justify that multi-exponential models can also be used to approximate the chemical relaxation of second-order reactions, we consider here the elementary binding model

$$P + L \underset{k_-}{\overset{k_+[P][L]}{\rightleftharpoons}} PL \tag{41}$$

of a protein P and ligand L. For the initial condition $[PL](0) = 0$, the rate equation of the elementary binding model can be written as

$$\frac{\mathrm{d}}{\mathrm{d}t}[PL] = k_+([P]_0 - [PL])([L]_0 - [PL]) - k_-[PL] \tag{42}$$

and has the analytical solution [38]

$$[PL](t) = -\frac{\lambda_1(\mathrm{e}^{(\lambda_1 - \lambda_2)t} - 1)}{k_+(\mathrm{e}^{(\lambda_1 - \lambda_2)t} - \lambda_1/\lambda_2)} \tag{43}$$

with

$$\lambda_{1,2} = -\frac{1}{2}k_+\left([\mathrm{P}]_0 + [\mathrm{L}]_0 + K_d \pm \sqrt{([\mathrm{P}]_0 + [\mathrm{L}]_0 + K_d)^2 - 4[\mathrm{P}]_0[\mathrm{L}]_0}\right) \tag{44}$$

where $K_d = k_-/k_+$ is the dissociation constant of the elementary binding model.

We first show that $\lambda_2 - \lambda_1$ is identical to the dominant relaxation rate k_{obs} obtained from a linear expansion around equilibrium. An expansion of the right-hand side of Eq (42) around the equilibrium concentration

$$[\mathrm{PL}]_{\mathrm{eq}} = \frac{1}{2}\left([\mathrm{P}]_0 + [\mathrm{L}]_0 + K_d - \sqrt{([\mathrm{L}]_0 - [P]_0 + K_d)^2 + 4K_d[P]_0}\right) \tag{45}$$

leads to the linear equation $\mathrm{d}[\mathrm{PL}]/\mathrm{d}t \simeq -k_{\mathrm{obs}}([\mathrm{PL}] - [\mathrm{PL}]_{\mathrm{eq}})$ with

$$k_{\mathrm{obs}} = k_+\sqrt{([\mathrm{L}]_0 - [P]_0 + K_d)^2 + 4K_d[P]_0} \tag{46}$$

This dominant relaxation rate k_{obs} is identical to $\lambda_2 - \lambda_1$. As a function of $[\mathrm{L}]_0$, the dominant rate k_{obs} of the elementary binding model exhibits a minimum at $[\mathrm{L}]_0^{\mathrm{min}} = [P]_0 - K_d$ and is symmetric with respect to this minimum.

We next use the limit of the geometric series $\sum_{n=0}^{\infty} q^n = 1/(1-q)$ with $q = e^{-k_{obs}\,t}\,\lambda_2/\lambda_1$ to rewrite Eq (43) as

$$[\mathrm{PL}](t) \propto \lambda_2 + (\lambda_2 - \lambda_1)\sum_{n=1}^{\infty}\frac{\mathrm{e}^{-nk_{\mathrm{obs}}t}}{(\lambda_1/\lambda_2)^n} \tag{47}$$

which shows that the chemical relaxation of the elementary binding model can be described as an infinite sum of exponential functions. The exponents of these functions are integer multiples of k_{obs}, which is reminiscent of the higher harmonics in oscillatory phenomena. The prefactors $(\lambda_2/\lambda_1)^n$ in Eq (47) decay exponentially with the order n of the harmonic because of $\lambda_2/\lambda_1 < 1$. The infinite sum of Eq (47) therefore can be truncated in practical situations. Under pseudo-first-order conditions, Eq (47) reduces to a single-exponential relaxation.

In analogy to the elementary binding model, we propose that the time evolution of the concentrations in the induced-fit and conformational-selection models can be represented as a sum of exponentials where the exponents are integer combinations $-ik_{\mathrm{obs}} - jk_2$ with $i, j = 0, 1, 2, 3, \ldots$ of the relaxation rates k_{obs} and k_2 obtained from a linear expansion around the equilibrium concentrations. Under pseudo-first-order conditions, the chemical relaxation reduces to a double-exponential relaxation [16, 21, 22].

In the numerical examples of Figs 2 and 3, the chemical relaxation of the bound complexes is fitted with a multi-exponential model

$$[\mathrm{bound}](t) = A_0 + \sum_{n=1}^{N} A_n e^{-k_n t} \tag{48}$$

with $k_n > 0$ for all n. We have used the routine NonlinearModelFit of the software Mathematica [41] with the differential evolution algorithm [42], which was repeatedly run with different values of its F parameter ranging from 0.1 to 1 for a given number of exponentials N. Among different runs, we have selected fit results based on the residual sum of squares, after discarding fits with singular results in which two rates k_n coincide within 95% confidence intervals, or in which one or more rates k_n are identical to 0 within 95% confidence intervals. We have then determined the number of exponentials N based on the small-sample-size corrected version of Akaike's information criterion (AIC) [43].

Bayes factors

The Bayes factor K is as measure for how plausible one model is relatively to an alternative model, given experimental data [44]. The Bayes factor for the plausibility of the conformational-selection binding model relative to induced-fit binding model is

$$K = \frac{\int p(\text{data} \mid \text{conformational} - \text{selection binding}, \theta) p(\theta) \mathrm{d}\theta}{\int p(\text{data} \mid \text{induced} - \text{fit binding}, \theta) p(\theta) \mathrm{d}\theta} \tag{49}$$

Here, $p(\text{data} \mid M, \theta)$ is the probability that the data were produced by the model M with given parameters θ, where M either stands for conformational-selection binding or induced-fit binding, and $p(\theta)$ is the prior distribution on the parameter values, which encodes any prior knowledge that we have about the parameters. The integrals of Eq (49) are taken over all parameter values and result in the probability $p(\text{data} \mid M)$ that the data were produced by the model, regardless of specific parameter values. The data here consist of the slowest relaxation rates $k_{\text{obs}}^{(i)}$ with $i = 1, 2, \ldots, N$ obtained from multi-exponential fits of the N time series with ligand concentrations $[L]_0^{(i)}$, and the errors σ_i of these rates. Following standard approaches [44], the probability that the data were generated by the model M with parameters $\theta = (k_e, k_r, k_-, K_d, [P]_0)$ is

$$p(\text{data} \mid M, \theta) \propto \prod_{i=1}^{N} \exp\left[-\frac{\left(k_{\text{obs}}^{(i)} - k_{\text{obs}}^{M}\left(\theta, [L]_0^{(i)}\right)\right)^2}{2\sigma_i^2} \right] \tag{50}$$

for $k_r > nk_e$, and 0 otherwise. The inequality $k_r > nk_e$ reflects constraints on the conformational relaxation rate k_r and excitation rate k_e of the models (see section "Analysis of chemical relaxation rates for recoverin binding"). Eq (50) implies that the errors $k_{\text{obs}}^{(i)} - k_{\text{obs}}^{M}(\theta, [L]_0^{(i)})$ are independently and normally distributed random variables with standard deviations σ_i. Depending on the model M, we either use Eqs (1) or (6) to determine $k_{\text{obs}}^{M}(\theta, [L]_0^{(i)})$. For simplicity, K_d and $[P]_0$ are kept fixed at the experimentally measured values. We choose a prior $p(\theta)$ that is uniform in the logarithm of the rates k_e, k_r, k_-. Taking the logarithm of the rates is not crucial, as a uniform prior on the rates gives similar results in the analysis of recoverin binding and, thus, leads to the same conclusions. The prior $p(\theta)$ here can be chosen to be uniform because it is identical for both the induced-fit and conformational-selection binding models due to the equivalent parameters of the models [45].

Acknowledgments

F. P. would like to thank Christof Schütte for insightful discussions.

Author Contributions

Analyzed the data: FP TRW.

Wrote the paper: FP TRW.

Conceived and designed the theory: FP TRW.

Performed the numerical analysis: FP TRW.

References

1. Gerstein M, Krebs W. A database of macromolecular motions. Nucleic Acids Res. 1998; 26:4280–4290. doi: 10.1093/nar/26.18.4280 PMID: 9722650

2. Eisenmesser EZ, Millet O, Labeikovsky W, Korzhnev DM, Wolf-Watz M, Bosco DA, et al. Intrinsic dynamics of an enzyme underlies catalysis. Nature. 2005; 438:117–121. doi: 10.1038/nature04105 PMID: 16267559
3. Beach H, Cole R, Gill M, Loria J. Conservation of μs-ms enzyme motions in the apo- and substrate-mimicked state. J Am Chem Soc. 2005; 127:9167–9176. doi: 10.1021/ja0514949 PMID: 15969595
4. Boehr DD, McElheny D, Dyson HJ, Wright PE. The dynamic energy landscape of dihydrofolate reductase catalysis. Science. 2006; 313:1638–1642. doi: 10.1126/science.1130258 PMID: 16973882
5. Henzler-Wildman KA, Thai V, Lei M, Ott M, Wolf-Watz M, Fenn T, et al. Intrinsic motions along an enzymatic reaction trajectory. Nature. 2007; 450:838–844. doi: 10.1038/nature06410 PMID: 18026086
6. Tang C, Schwieters CD, Clore GM. Open-to-closed transition in apo maltose-binding protein observed by paramagnetic NMR. Nature. 2007; 449:1078–1082. doi: 10.1038/nature06232 PMID: 17960247
7. Lange OF, Lakomek NA, Fares C, Schröder GF, Walter KFA, Becker S, et al. Recognition dynamics up to microseconds revealed from an RDC-derived ubiquitin ensemble in solution. Science. 2008; 320:1471–1475. doi: 10.1126/science.1157092 PMID: 18556554
8. Kim E, Lee S, Jeon A, Choi JM, Lee HS, Hohng S, et al. A single-molecule dissection of ligand binding to a protein with intrinsic dynamics. Nat Chem Biol. 2013; 9:313–318. doi: 10.1038/nchembio.1213 PMID: 23502425
9. Munro JB, Gorman J, Ma X, Zhou Z, Arthos J, Burton DR, et al. Conformational dynamics of single HIV-1 envelope trimers on the surface of native virions. Science. 2014; 346:759–763. doi: 10.1126/science.1254426 PMID: 25298114
10. Ghoneim M, Spies M. Direct correlation of DNA binding and single protein domain motion via dual illumination fluorescence microscopy. Nano Lett. 2014; 14:5920–5931. doi: 10.1021/nl502890g PMID: 25204359
11. Ma B, Kumar S, Tsai CJ, Nussinov R. Folding funnels and binding mechanisms. Protein Eng. 1999; 12:713–720. doi: 10.1093/protein/12.9.713 PMID: 10506280
12. Koshland DE. Application of a theory of enzyme specificity to protein synthesis. Proc Natl Acad Sci USA. 1958; 44:98–104. doi: 10.1073/pnas.44.2.98 PMID: 16590179
13. Weikl TR, Paul F. Conformational selection in protein binding and function. Protein Sci. 2014; 23:1508–1518. doi: 10.1002/pro.2539 PMID: 25155241
14. Bosshard HR. Molecular recognition by induced fit: How fit is the concept? News Physiol Sci. 2001; 16:171–1733. PMID: 11479367
15. Sullivan SM, Holyoak T. Enzymes with lid-gated active sites must operate by an induced fit mechanism instead of conformational selection. Proc Natl Acad Sci USA. 2008; 105:13829–13834. doi: 10.1073/pnas.0805364105 PMID: 18772387
16. Weikl TR, von Deuster C. Selected-fit versus induced-fit protein binding: kinetic differences and mutational analysis. Proteins. 2009; 75:104–110. doi: 10.1002/prot.22223 PMID: 18798570
17. Boehr DD, Nussinov R, Wright PE. The role of dynamic conformational ensembles in biomolecular recognition. Nat Chem Biol. 2009; 5:789–796. doi: 10.1038/nchembio.232 PMID: 19841628
18. Hammes GG, Chang YC, Oas TG. Conformational selection or induced fit: a flux description of reaction mechanism. Proc Natl Acad Sci USA. 2009; 106:13737–13741. doi: 10.1073/pnas.0907195106 PMID: 19666553
19. Wlodarski T, Zagrovic B. Conformational selection and induced fit mechanism underlie specificity in noncovalent interactions with ubiquitin. Proc Natl Acad Sci USA. 2009; 106:19346–19351. doi: 10.1073/pnas.0906966106 PMID: 19887638
20. Changeux JP, Edelstein S. Conformational selection or induced fit? 50 years of debate resolved. F1000 Biol Rep. 2011; 3:19. doi: 10.3410/B3-19 PMID: 21941598
21. Weikl TR, Boehr DD. Conformational selection and induced changes along the catalytic cycle of *Escherichia coli* dihydrofolate reductase. Proteins. 2012; 80:2369–2383. doi: 10.1002/prot.24123 PMID: 22641560
22. Vogt AD, Di Cera E. Conformational selection or induced fit? A critical appraisal of the kinetic mechanism. Biochemistry. 2012; 51:5894–5902. doi: 10.1021/bi3006913 PMID: 22775458
23. Kiefhaber T, Bachmann A, Jensen KS. Dynamics and mechanisms of coupled protein folding and binding reactions. Curr Opion Struct Biol. 2012; 22:21–29. doi: 10.1016/j.sbi.2011.09.010
24. Vogt AD, Pozzi N, Chen Z, Di Cera E. Essential role of conformational selection in ligand binding. Biophys Chem. 2014; 186:13–21. doi: 10.1016/j.bpc.2013.09.003 PMID: 24113284
25. Pozzi N, Vogt AD, Gohara DW, Di Cera E. Conformational selection in trypsin-like proteases. Curr Opin Struct Biol. 2012; 22:421–431. doi: 10.1016/j.sbi.2012.05.006 PMID: 22664096

26. Daniels KG, Tonthat NK, McClure DR, Chang YC, Liu X, Schumacher MA, et al. Ligand concentration regulates the pathways of coupled protein folding and binding. J Am Chem Soc. 2014; 136:822–825. doi: 10.1021/ja4086726 PMID: 24364358

27. Daniels KG, Suo Y, Oas TG. Conformational kinetics reveals affinities of protein conformational states. Proc Natl Acad Sci USA. 2015; 112:9352–9357. doi: 10.1073/pnas.1502084112 PMID: 26162682

28. Chakrabarti KS, Agafonov RV, Pontiggia F, Otten R, Higgins MK, Schertler GFX, et al. Conformational selection in a protein-protein interaction revealed by dynamic pathway analysis. Cell Reports. 2016; 14:32–42. doi: 10.1016/j.celrep.2015.12.010 PMID: 26725117

29. James LC, Roversi P, Tawfik DS. Antibody multispecificity mediated by conformational diversity. Science. 2003; 299:1362–1367. doi: 10.1126/science.1079731 PMID: 12610298

30. Heredia VV, Thomson J, Nettleton D, Sun S. Glucose-induced conformational changes in glucokinase mediate allosteric regulation: transient kinetic analysis. Biochemistry. 2006; 45:7553–7562. doi: 10.1021/bi060253q PMID: 16768451

31. Kim YB, Kalinowski SS, Marcinkeviciene J. A pre-steady state analysis of ligand binding to human glucokinase: Evidence for a preexisting equilibrium. Biochemistry. 2007; 46:1423–1431. doi: 10.1021/bi0617308 PMID: 17260972

32. Tummino PJ, Copeland RA. Residence time of receptor-ligand complexes and its effect on biological function. Biochemistry. 2008; 47:5481–5492. doi: 10.1021/bi8002023 PMID: 18412369

33. Antoine M, Boutin JA, Ferry G. Binding kinetics of glucose and allosteric activators to human glucokinase reveal multiple conformational states. Biochemistry. 2009; 48:5466–5482. doi: 10.1021/bi900374c PMID: 19459610

34. Vogt AD, Di Cera E. Conformational selection is a dominant mechanism of ligand binding. Biochemistry. 2013; 52:5723–5729. doi: 10.1021/bi400929b PMID: 23947609

35. Gianni S, Dogan J, Jemth P. Distinguishing induced fit from conformational selection. Biophys Chem. 2014; 189:33–39. doi: 10.1016/j.bpc.2014.03.003 PMID: 24747333

36. Vogt AD, Chakraborty P, Di Cera E. Kinetic dissection of the pre-existing conformational equilibrium in the trypsin fold. J Biol Chem. 2015; 290:22435–22445. doi: 10.1074/jbc.M115.675538 PMID: 26216877

37. Jarosz AF, Wiley J. What are the odds? A practical guide to computing and reporting Bayes factors. J Problem Solving. 2014; 7:2. doi: 10.7771/1932-6246.1167

38. Peuker S, Cukkemane A, Held M, Noé F, Kaupp UB, Seifert R. Kinetics of ligand-receptor interaction reveals an induced-fit mode of binding in a cyclic nucleotide-activated protein. Biophys J. 2013; 104:63–74. doi: 10.1016/j.bpj.2012.11.3816 PMID: 23332059

39. Greives N, Zhou HX. Both protein dynamics and ligand concentration can shift the binding mechanism between conformational selection and induced fit. Proc Natl Acad Sci USA. 2014; 111:10197–10202. doi: 10.1073/pnas.1407545111 PMID: 24982141

40. Suddala KC, Wang J, Hou Q, Walter NG. Mg^{2+} shifts ligand-mediated folding of a riboswitch from induced-fit to conformational selection. J Am Chem Soc. 2015; 137:14075–14083. doi: 10.1021/jacs.5b09740 PMID: 26471732

41. Mathematica, Version 10.3. Wolfram Research, Inc., Champaign, Illinois; 2015.

42. Storn R, Price K. Differential evolution—a simple and efficient heuristic for global optimization over continuous spaces. J Global Optim. 1997; 11:341–359. doi: 10.1023/A:1008202821328

43. Cavanaugh JE. Unifying the derivations of the Akaike and corrected Akaike information criteria. Stat Probabil Lett. 1997; 31:201–208. doi: 10.1016/S0167-7152(96)00128-9

44. Jaynes ET. Probability theory: the logic of science. Cambridge University Press; 2003.

45. Strachan RW, van Dijk HK. Improper priors with well-defined Bayes factors. Liverpool, L69 7ZA, United Kingdom: Department of Economics and Accounting, University of Liverpool; 2005. EI 2004–18.

Multiensemble Markov models of molecular thermodynamics and kinetics

Hao Wu*, Fabian Paul*, Christoph Wehmeyer, Frank Noé
*) *equal contribution*
Freie Universität Berlin, Department of Mathematics, Computer Science and Bioinformatics, Arnimallee 6, 14195 Berlin

(Dated: Edited and approved April 22, 2016 (received for review December 21, 2015))

This is an Author's Accepted Manuscript of an article published the Proceedings of the National Academy of Sciences, volume 113, number 23 in 2016, pages E3221-E3230, available online at: http://www.pnas.org/content/113/23/E3221.full (doi: 10.1073/pnas.1525092113).

ABSTRACT

We introduce the general transition-based reweighting analysis method (TRAM), a statistically optimal approach to integrate both unbiased and biased molecular dynamics simulations, such as umbrella sampling or replica exchange. TRAM estimates a multiensemble Markov model (MEMM) with full thermodynamic and kinetic information at all ensembles. The approach combines the benefits of Markov state models—clustering of high-dimensional spaces and modeling of complex many-state systems—with those of the multistate Bennett acceptance ratio of exploiting biased or high-temperature ensembles to accelerate rare-event sampling. TRAM does not depend on any rate model in addition to the widely used Markov state model approximation, but uses only fundamental relations such as detailed balance and binless reweighting of configurations between ensembles. Previous methods, including the multistate Bennett acceptance ratio, discrete TRAM, and Markov state models are special cases and can be derived from the TRAM equations. TRAM is demonstrated by efficiently computing MEMMs in cases where other estimators break down, including the full thermodynamics and rare-event kinetics from high-dimensional simulation data of an all-atom protein–ligand binding model.

SIGNIFICANCE

Molecular dynamics simulations can provide mechanistic understanding of biomolecular processes. However, direct simulation of slow transitions such as protein conformational transitions or protein–ligand dissociation are unfeasible with commonly available computational resources. Two typical strategies are (*i*) conducting large ensembles of short simulations and estimating the long-term kinetics with a Markov state model, and (*ii*) speeding up rare events by bias potentials or higher temperatures and estimating the unbiased thermodynamics with reweighting estimators. In this work, we introduce the transition-based reweighting analysis method (TRAM), a statistically optimal approach that combines the best of both worlds and estimates a multiensemble Markov model (MEMM) with full thermodynamic and kinetic information at all simulated ensembles.

Computer simulations have become important tools in the investigation of biomolecular processes, including transmembrane transport [1–4], ligand reception and receptor activation [5–7], and endocytosis [8–10]. Unbiased atomistic molecular dynamics (MD) simulations have recently reached the ability to extensively sample biomolecular processes on timescales up to milliseconds, including protein folding [11], conformational changes [5, 12], and protein–ligand association and dissociation [13–15]. In addition to breakthroughs in computer hardware [16], simulation software [17–19], and distributed computing [20, 21], a key technology to reconcile swarms of individually short simulations to long-time kinetics are kinetic models, such as Markov state models (MSMs) [22–28]. A key advantage of Markov state modeling over many other approaches is that it integrates well with dimension reduction and clustering techniques [29–31] that can process high-dimensional data, and can thus treat complex kinetics that are not well described by few states or reaction co-ordinates [13, 32–34]. A limitation of MSMs is that they rely on the rare events being reversibly sampled in the underlying MD simulation data.

However, important biological processes are still out of reach for unbiased MD simulation. For example, although downhill processes such as protein–ligand association to the bound pose can now be spontaneously sampled [6, 13, 15, 34], the dissociation of stable inhibitory complexes can involve timescales of hours or longer [35]. Enhanced sampling methods such as umbrella sampling (US) [36, 37], parallel or simulated tempering [38–40], metadynamics [41], and others [42–44] use simulations at different ensembles in which bias potentials or higher temperatures are used to speed up events that are rare in the physical ensemble [45, 46]. Reweighting methods such as the weighted histogram analysis method (WHAM) [37, 47, 48] and binless WHAM, also known as multistate Bennett acceptance ratio (MBAR) [49–51], can combine multiple-ensemble simulation data to estimates of the unbiased thermodynamics (free energies or probabilities). These methods treat their input data as uncorrelated samples of the ensemble distribution and are therefore not suitable for simulation data with long correlation times in some variables, as it is common for unbiased MD simulations and biased simulations with slow unbiased coordinates [32, 52].

To overcome individual limitations of MSMs and enhanced sampling techniques, we propose to integrate simulation data from multiple ensembles in a multi-ensemble Markov model (MEMM) (Fig. 1), in such a way as to (*i*) work with high-dimensional data and coarse state-space discretizations, (*ii*) use unbiased MD simulations from nonequilibrium starting points but avoid rate models beyond MSMs, and (*iii*) optimally combine data to full thermo- dynamics and kinetics at all simulated ensembles.

Here, we develop the generic transition-based reweighting analysis method (TRAM), an estimation method for MEMMs that combines the above features as follows. (*i*) Statistical weights of sampled configurations are reweighted between ensembles in a binless manner, a key property for working in high-dimensional spaces. (*ii*) Conditional transition statistics are used in an MSM-based likelihood, and thus simulations only need to be in local but not in global equilibrium. TRAM only relies on the MSM approximation and detailed balance relations to predict rare-event kinetics [53] and avoids the use of additional rate models. (*iii*) TRAM provides a maximum-likelihood MEMM with full thermodynamic and kinetic information at all ensembles. In summary, TRAM goes significantly beyond previously proposed transition-based reweighting methods [54–57] and other methods to estimate thermodynamics and kinetics from multiensemble data [53, 58–61], which offer some but not all of the above properties (for more detailed discussion, see below). TRAM is a formal generalization to WHAM, MBAR, reversible MSMs, and discrete TRAM that can all be derived from TRAM.

We apply TRAM on two benchmark systems and an all-atom model of the trypsin protein with the benzamidine inhibitor. We illustrate that TRAM can estimate the thermodynamics at ensembles more accurately and with less simulation data than previous estimation methods, and that additionally unbiased models of the kinetics can be built. We demonstrate that our MEMM approach offers a systematic treatment of the common problem of slow unbiased coordinates in US simulations and provides efficient estimates of rare-event kinetics, such as protein–ligand dissociation.

TRAM

Basics.

Let us consider a molecular system in a reference ensemble with configuration x in a configuration space and dimensionless potential function $u(x)$. $u(x)$ has units of thermal energy $k_B T = \beta^{-1}$, where T is the temperature. $u(x)$ is a sum of terms, including $\beta U(x)$ with the potential energy function U, and pressure–volume or chemical potential terms, depending on the ensembles under consideration [49]. The system has an equilibrium distribution as follows:

$$\mu(x) = \mathrm{e}^{f-u(x)}, \tag{1}$$

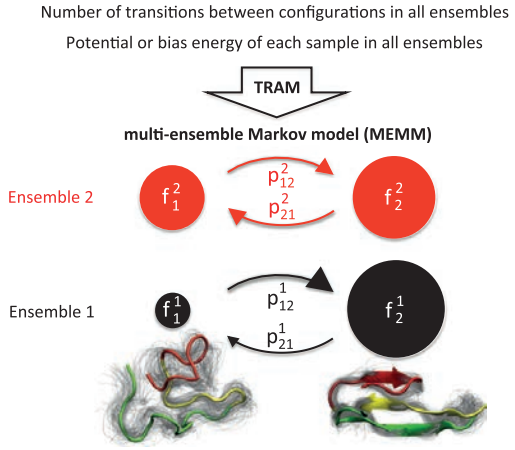

Figure 1. Multiensemble Markov models (MEMMs) contain thermodynamic (e.g., free energies) and kinetic (e.g., transition probabilities) information among all configuration states (subscript index) and ensembles (superscript index). TRAM estimates MEMMs by combining transitions observed between configuration states, and the statistical weights/reduced energies of samples in all ensembles (reweighting).

where the free energy f is the negative logarithm of the partition function and has the role to normalize $\mu(x)$.

Suppose we are given simulations from K different ensembles (indexed by the superscript k), which may comprise an arbitrary combination of the unbiased ensemble, simulations with biased energy functions, or different temperatures. We can formally relate any ensemble with dimensionless potential $u^k(x)$ to the reference ensemble by introducing a bias potential $b^k(x)$ such that $u^k(x) = u(x) + b^k(x)$. The corresponding equilibrium distribution $\mu^k(x)$ of the kth ensemble can be expressed as follows:

$$\mu^k(x) = \mathrm{e}^{f^k - b^k(x)} \mu(x) \tag{2}$$

where the relative free energy f^k of ensemble k is chosen such that $\mu^k(x)$ is normalized. Consider the following examples to see how $b^k(x)$ must be chosen to model commonly used enhanced sampling methods:

i) In US, the potential energy function of each simulation is $U(x) + B^k(x)$, where $B^k(x)$ is the kth umbrella potential. The bias potential is as follows:

$$b^k(x) = \beta B^k(x). \tag{3}$$

ii) Replica exchange or parallel tempering simulations are performed at different temperatures $T^1, \ldots, T^K$, the bias of the kth temperature with respect to the reference ensemble (e.g., the lowest temperature) is as follows:

$$b^k(x) = U(x)\left(\beta^k - \beta\right). \tag{4}$$

MSMs for Molecular Kinetics.

An MSM at ensemble k consists of a partition of the molecular configuration space into m discrete and nonoverlapping configuration states $S_1, \ldots, S_m$ and the conditional transition probabilities $p_{ij}^k(\tau)$ that a system that is in state S_i at time t will be found in state S_j at time $t + \tau$.

We first define the local free energy f_i^k of configuration state S_i in ensemble k. The exponential of f_i^k is proportional to the statistical weight of this state:

$$\mathrm{e}^{-f_i^k} = \mathrm{e}^{-f^k} \int_{S_i} \mu^k(x)\,\mathrm{d}x \tag{5}$$

where the integral evaluates to the equilibrium probability of the system to be in state S_i when simulated in ensemble k.

For given simulation data from ensemble k that contains c_{ij}^k transitions from state S_i at time t and to state S_j at time $t + \tau$, the likelihood of an MSM with transition matrix $\mathbf{P}^k = \left[p_{ij}^k\right]$ is as follows:

$$L_{\mathrm{MSM}}^k = \prod_{i=1}^{m} \prod_{j=1}^{m} \left(p_{ij}^k\right)^{c_{ij}^k}. \tag{6}$$

When simulations are conducted at thermal equilibrium (i.e., without adding or removing energy to the system) in ensemble k, equilibrium and transition probabilities are related by the detailed balance equations $\mathrm{e}^{-f_i^k} p_{ij}^k = \mathrm{e}^{-f_j^k} p_{ji}^k$, and the Markov model is said to be reversible. With detailed balance constraints, the maximum likelihood of Eq. 6 has no closed-form solution but can be iteratively solved [28, 62, 63].

Local Equilibrium Model.

If simulations sample from multiple ensembles, a central problem is to infer the equilibrium distribution $\mu(x)$ at a reference ensemble, given the simulation data at all ensembles. The principle behind such inference is that we can reweight the equilibrium probability of a sample x between different ensembles by means of Eq. 2.

A widely used estimator is the binless WHAM method, [50, 51], also called MBAR [49], which provides an optimal estimate of $\mu(x)$ under the assumption that at each ensemble k, the samples x are drawn independently from their global equilibrium distribution $\mu^k(x)$. MBAR can be derived by maximizing a likelihood that is simply given by the product of $\mu^k(x)$ over all samples x and all ensembles k [49–51].

However, we do not want to depend on the global equilibrium assumption. Hence we define the local equilibrium distribution for each configuration state S_i:

$$\mu_i^k(x) = \begin{cases} \mathrm{e}^{f_i^k - f^k}\mu^k(x) & x \in S_i \\ 0 & \text{else.} \end{cases} \tag{7}$$

We assume that simulations are sampling from these local equilibrium distributions, but they do not need to be in equilibrium between configuration states, which is key for invoking the MSM framework. We obtain the following likelihood of generating the simulation data for a given sequence of discrete states:

$$L_{\mathrm{LEQ}}^k = \prod_{i=1}^{m} \prod_{x \in X_i^k} \mu_i^k(x) \tag{8}$$

where X_i^k denotes the set of all samples generated from the kth ensemble and in configuration state S_i. As $\mu_i^k(x)$ can be related to $\mu(x)$ via Eqs. 2 and 7, the local equilibrium model is key to reweight samples between different ensembles.

TRAM likelihood.

We develop the TRAM estimator. The TRAM likelihood combines the MSM likelihood (6) and local equilibrium likelihood (8). Inserting Eqs. 2 and 7, we obtain the following:

$$L_{\mathrm{TRAM}} = \prod_{k=1}^{K} \underbrace{\left(\prod_{i,j} \left(p_{ij}^k\right)^{c_{ij}^k} \right)}_{L_{\mathrm{MSM}}^k} \underbrace{\left(\prod_{i=1}^{m} \prod_{x \in X_i^k} \mu(x)\, \mathrm{e}^{f_i^k - b^k(x)} \right)}_{L_{\mathrm{LEQ}}^k}. \tag{9}$$

This likelihood expresses the probability that a given set of trajectories sampling from different ensembles has visited a particular sequence of discrete states (L_{MSM}^k) and has sampled the local configurations inside these discrete states (L_{LEQ}^k). The structure of the TRAM likelihood is similar to that of a hidden Markov model [64].

The trajectory statistics include the bias potentials $b^k(x)$ that are defined by the simulation protocol [e.g., US or replica exchange molecular dynamics (REMD)], and the number of observed transitions c_{ij}^k. The unknown variables in the TRAM likelihood are the point densities $\mu(x)$, the local free energies f_i^k, and the transition probabilities p_{ij}^k. The TRAM problem is to maximize the likelihood (9) in the variable space subject to the following constraints:

$$\mathrm{e}^{-f_i^k} p_{ij}^k = \mathrm{e}^{-f_j^k} p_{ji}^k, \qquad \text{for all } i,j,k \tag{10}$$

$$\sum_j p_{ij}^k = 1, \qquad \text{for all } i,k \tag{11}$$

$$\sum_{x \in X} \mu(x) = 1. \tag{12}$$

where (11) and (12) are simple normalization constraints, and $\mu(x)$ is considered as a discrete distribution on the set of all samples, X. The detailed balance condition (10) couples the dynamical (MSM) part to the local equilibrium part. Unfortunately, the detailed balance constraints make the above problem very hard to solve.

Maximum-Likelihood Solution.

The TRAM problem contains $\left(m^2K + |X|\right)$ unknowns, $(mK+1)$ linear equality constraints (normalization), and $Km(m-1)/2$ nonlinear equality constraints (detailed balance), so finding the optimal solution by directly using gradient- or Newton-type methods is difficult even for systems with only few configuration states or ensembles. Fortunately, we can transform the TRAM problem into a more tractable system of nonlinear algebraic equations and solve the resulting system by an iterative algorithm. By using the Lagrange duality theory (Appendix), it can be proved that the maximum of the TRAM likelihood satisfies the following equations:

$$\sum_j \frac{c_{ij}^k + c_{ji}^k}{\exp\left[f_j^k - f_i^k\right] v_j^k + v_i^k} = 1, \quad \text{for all } i,k \tag{13}$$

$$\sum_{x \in X_i} \frac{\exp\left(f_i^k - b^k(x)\right)}{\sum_l R_i^l \exp\left[f_i^l - b^l(x)\right]} = 1, \quad \text{for all } i,k \tag{14}$$

where v_i^k are Lagrange multipliers that can be interpreted as counts (for infinite statistics $v_i^k = \sum_j c_{ij}^k$, see Appendix). X_i is the set of all samples in configuration state S_i, no matter from which ensemble. The factor R_i^l is given by the following:

$$R_i^k = \sum_j \frac{\left(c_{ij}^k + c_{ji}^k\right) v_j^k}{v_j^k + \exp\left[f_i^k - f_j^k\right] v_i^k} + N_i^k - \sum_j c_{ji}^k \tag{15}$$

where N_i^k is the number of samples in X_i^k. R_i^k are effective state counts (see below). When (9-12) are fulfilled and in the limit of infinite statistics, $R_i^k = N_i^k$.

In contrast to Eqs. 9-12, the formulation in (13) and (14) only contains the $2mK$ unknowns v_i^k and $f_i^k$and does not involve $\mathbf{P}^k$ and $\mu(x)$ explicitly. Given the solution of Eqs. 13 and 14, we can compute all MSM transition matrices by the following:

$$p_{ij}^k = \frac{c_{ij}^k + c_{ji}^k}{\exp\left[f_j^k - f_i^k\right] v_j^k + v_i^k} \tag{16}$$

and the unbiased statistical weights of all samples by the following:

$$\mu(x) = \frac{1}{\sum_k R_{i(x)}^k \exp\left[f_{i(x)}^k - b^k(x)\right]} \tag{17}$$

where we have defined $i(x)$ such that $i(x) = j$ when $x \in X_j$. The TRAM estimator defined by Eqs. 13 and 14 is statistically optimal, asymptotically correct, i.e., converges to the correct results of f_i^k, p_{ij}^k, and $\mu(x)$ as the length or number of simulation trajectories increases, and the most general multiensemble Markov model estimator (see below and Appendix).

Eqs. 13 and 14 are reminiscent of other estimators: Eq. 13 arises when optimizing an MSM transition matrix with given stationary weights $\exp(-f_i^k)$ [63]. Eq. 14 has the same form as the self-consistent MBAR equation [49] for the ensemble free energies f_i^k of a single configuration state S_i, but instead of the number of samples in that state N_i^k, the modified counts R_i^k are used (detailed interpretation in Appendix). The TRAM equations can therefore be thought of expressing two optimization problems simultaneously: (i) at each ensemble k, the optimization of the MSMs for given free energies, f_i^k for all configurations S_i. (ii) At each configuration S_i, the optimization of the free energies, f_i^k for all ensembles.

Optimization Algorithm.

The TRAM equations 13 and 14 are coupled and can only be solved numerically. Here, we transform them into a simple fixed-point problem, in which the following equations need to be iterated until convergence:

$$v_i^{k,\mathrm{new}} := v_i^k \sum_j \frac{c_{ij}^k + c_{ji}^k}{\exp\left[f_j^k - f_i^k\right] v_j^k + v_i^k} \tag{18}$$

$$f_i^{k,\mathrm{new}} := -\ln \sum_{x \in X_i} \frac{\exp\left[-b^k(x)\right]}{\sum_l R_i^l \exp\left[f_i^l - b^l(x)\right]} \tag{19}$$

More implementation details of this algorithm, including initialization, termination, and convergence acceleration are given in Appendix. Note that, instead of a fixed-point iteration, we could attempt a Newton-based [65, 66] or stochastic optimization method [67, 68].

Thermodynamics and Kinetics from TRAM.

Thermodynamics. The correct calculation of stationary (thermodynamic) properties does not rely on Markovianity, but only requires the unbiased estimation of free energies f_i^k or the stationary density $\mu^k(x)$ at the chosen lag time τ. However, it is required that the simulations are in local equilibrium within the configuration states, and violations of local equilibrium can be compensated by using longer lag times τ. The robustness of TRAM estimates should therefore be tested as a function of lag time (see results, Fig. 3).

Kinetics. Asymptotic correctness of all p_{ij}^k at the selected lag time τ does not imply that powers of the matrix $\mathbf{P}^k$ are a good prediction of the transition probabilities at longer lag times. Whether the multiensemble Markov model is able to predict long-term kinetics depends on the quality of the discretization and on τ being sufficiently large, as usual for MSMs [28, 56] (see results, Fig. 3D). Note that this behavior does not change if a rate matrix is used instead of a transition matrix.

Generality of TRAM.

TRAM is a generalization of discrete TRAM, binless WHAM/MBAR, binned WHAM, and reversible MSMs (Fig. 2). These specialized estimators can be derived from TRAM by adding the specific assumptions made by them. MBAR can be derived from TRAM by assuming that samples are drawn from the global equilibrium distribution of each ensemble. Discrete TRAM can be derived by assuming that the bias energies are piecewise constant and pointwise reweighting can be replaced by histogram reweighting. WHAM is derived using a combination of both assumptions. Finally, if we have only a single ensemble, the TRAM solution is identical to the reversible MSM estimator (derivations in Appendix).

TRAM has the Markovianity assumption in its likelihood model, but otherwise only uses fundamental relations such as detailed balance and pointwise reweighting. It is therefore the most general MSM-based estimator for simulation data from multiple ensembles. Other transition-based reweighting methods are related as follows: trajectory reweighting techniques [58–60] are applicable without any state space discretization, but assume the trajectory starting points to emerge from a global equilibrium distribution. The dynamic weighted histogram analysis method (DHAM) [57] uses a kinetic reweighting scheme that can predict kinetics at ensembles not simulated from, but is based on a rate model, uses histogram binning and does not optimize with respect to detailed balance. An advantage in not enforcing the detailed balance constraint is that DHAM is a

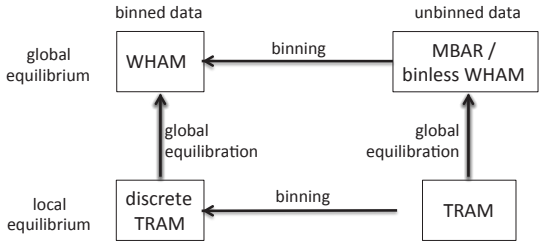

Figure 2. Relationship of different statistically optimal reweighting estimators. TRAM is the most general method considered here, it can be specialized to MBAR, discrete TRAM, and WHAM by adding the assumption of global equilibrium or performing a binning of sampled configurations.

single-shot estimator, while TRAM and dTRAM are estimators that need to be iterated to solution. xTRAM [55] is a bin-less TRAM method, but in contrast to the present method not statistically optimal for finite data (Applications).

APPLICATIONS

Multitemperature Replica-Exchange.

We compare the performance of MBAR and TRAM for REMD simulations of solvated alanine dipeptide using 33 exponentially spaced temperatures in the range of 300–600 K (Fig. 3A) (see ref. [55] for simulation protocol). In multitemperature simulations, the bias potentials between ensembles depend on the potential energy (Eq. 4). To analyze such data with histogram-based methods, one would have to bin the potential energy axis in addition to the coordinate(s) of interest [69]. For many-body systems such as solvated macromolecules, the large range of potential energies sampled and the required resolution to approximate the bias energies (Eq. 4) disables binned estimators such as WHAM, discrete TRAM, and DHAM, and instead require binless methods such as MBAR and TRAM.

The configuration space of alanine dipeptide is partitioned into 20 discrete states using k-means clustering in the space of the coordinates $\{\cos\phi, \sin\phi, \cos\psi, \sin\psi\}$ (Fig. 3A). The equilibrium probabilities on the sets I–IV are compared between estimators. Within statistical error, both MBAR and TRAM converge to the same values, whereas TRAM converges significantly faster (Fig. 3B). TRAM outperforms MBAR because TRAM relies only on local rather than global equilibria. As a result, TRAM does not suffer from the fact that initial structures are not sampled from a global equilibrium distribution, and the REMD simulation must first relax to sample from global equilibrium. Even after this relaxation phase, TRAM uses the data more effectively because a smaller number of simulation steps is required to generate an uncorrelated sample from local equilibrium compared with global equilibrium.

Next, we test the robustness of TRAM estimates as a function of the lag time (Fig. 3 C and D). It is seen that the stationary probabilities, and thus the results in Fig. 3B, are independent of the lag time, demonstrating that the Markov property is not required to get correct estimates of the stationary properties. Unbiased estimates of equilibrium properties only require that the simulations are in local equilibrium. For REMD simulations with a good state space discretization, this is fulfilled even at short lag times τ. In contrast, unbiased estimates of the kinetic properties require sufficiently long lag times for the Markov property to be valid. The estimated relaxation timescale at temperature 366 K is constant above $\tau = 10$ ps (Fig. 3D), which was used for all TRAM estimates in Fig. 3. Only trajectory segments in which no temperature swap was executed for 10 ps or longer were used for this estimate.

Fig. 3 E and F show thermodynamics and kinetics obtained from the multiensemble Markov model as a function of the temperature. The probabilities of metastable states become more similar with increasing temperatures, but the temperature dependence is very weak, indicating that entropy differences play a minor role (Fig. 3E). The mean first passage times (inverse transition rates) from I/II to III/IV and back decrease strongly with temperature (Fig. 3F). The decrease is exponential (Arrhenius-like) up to 450 K, but shows a weaker temperature dependence for higher temperatures, indicating that the kinetics are limited by diffusion rather than barrier crossing in this range.

Biased Simulations with Slow Orthogonal Degrees of Freedom.

Simulations in which sampling is enhanced along predefined reaction coordinates (e.g., using bias potentials) are often hampered by unforeseen rare transitions in other coordinates [52]. For illustration, we use a pathological 2D toy potential with three wells (Fig. 4A). US simulations are conducted using only the x coordinate as bias coordinate (details in Appendix).

As the potential wells I and III cannot be separated on the x axis, it takes a long time to converge to the global equilibrium when the simulations are confined to values of $x < 15$. Especially simulations with the second umbrella potential centered at $x = 8.33$ exhibit rare-event transitions along the y axis (Fig. 4B), characterized by

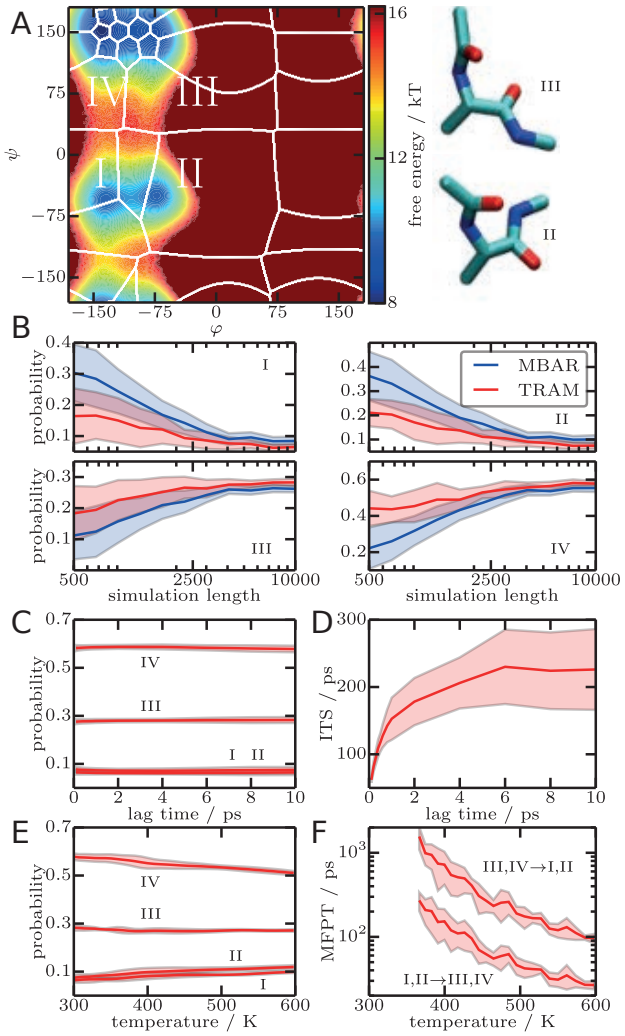

Figure 3. MBAR versus TRAM using REMD simulations of alanine dipeptide; mean and SD over seven independent simulations are shown. (A) Histogram-based free-energy landscape in the backbone torsions ϕ, ψ at a temperature of 300 K. Thin lines: borders of 20 clusters used to partition the space $\{\cos\phi, \sin\phi, \cos\psi, \sin\psi\}$ (hence the nonlinear boundaries). Thick lines: borders of four metastable sets analyzed in B. (B) Convergence of equilibrium probability estimates of sets I–IV at 300 K, as a function of simulation length. (C–F) TRAM estimates. (C) Equilibrium probabilities as a function of the lag time τ. (D) Slowest relaxation timescale at 366 K as a function of the lag time. (E) Equilibrium probabilities of sets I–IV as a function of temperature. (F) Mean first passage times at temperatures where transitions were found.

an autocorrelation time of 500 steps (Fig. 4C) [70]. In contrast, the largest value of the autocorrelation time along the x axis is only about 22 steps.

Rare events in nonenhanced coordinates are a common problem in enhanced sampling simulations and cause major problems in their analysis. For estimation methods relying on global equilibrium sampling such as MBAR, the statistically correlated samples should be discarded before running the estimator [49]. In umbrella simulation 2, this would result in retaining only one effective sample for each 500 samples in the simulation, resulting in the loss of almost all data and requiring very long simulation times.

TRAM does not require global equilibrium sampling and can therefore use simulation data much more efficiently. We discretize the configuration space Ω into 20 states as shown in Fig. 4A, and then use TRAM with lag time $\tau = 1$ to estimate the unbiased equilibrium distribution from the US data. Fig. 4D summarizes estimation errors of TRAM for different lengths of each simulation trajectory and compares them with those of MBAR and the previously described xTRAM method [55]. Here, the error is evaluated as the Kullback–Leibler divergence between the estimated probability distributions of the three macrostates I, II, and III and the true reference. In contrast to MBAR, TRAM can effectively overcome the influence of the nonequilibrium distribution of the data through Markov state modeling and achieve accurate estimates even in the case of trajectory length smaller than autocorrelation times $\tau_{\mathrm{ess}}(y)$ of some biased simulations. Furthermore, TRAM also significantly outperforms xTRAM, which is a consistent estimator under the MSM assumption but not statistically optimal for finite data.

Protein–Ligand Binding and Kinetics.

Finally, we demonstrate that TRAM can help to resolve the problem of rare events in orthogonal degrees of freedom and provides efficient estimates of rare-event kinetics in all-atom, explicit-solvent simulations of the serine protease trypsin and its inhibitor benzamidine (see ref. [34] for detailed setup). This illustrates the usefulness of the estimator in high-dimensional spaces where binning of all relevant coordinates is not an option.

We first analyze pure US simulations with 150 umbrella windows used to sample the position of benzamidine between the bound pose and a prebinding site (Fig. 5A, structures *i*–*iv*; details in Appendix).

To detect rare events in the unbiased coordinates, time-lagged independent component analysis (TICA) [29, 30] was used with the Cartesian coordinates of

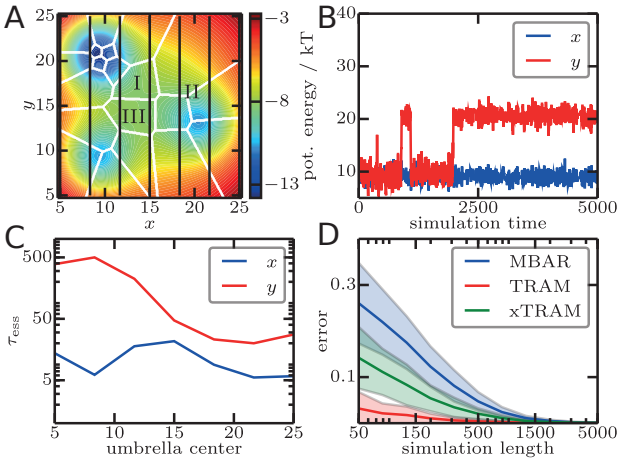

Figure 4. Comparison of MBAR, xTRAM, and TRAM for estimating the equilibrium distribution of a three-well potential from US data. (A) Potential function $u(x, y)$, where thin white lines represent the borders of 20 discrete states, thick white lines represent the borders of the three potential wells, and the dashed black lines indicate the umbrella centers. (B) A simulation trajectory with the bias potential centered at $x = 8.33$. (C) Autocorrelation times $\tau_{\mathrm{ess}}(x)$ and $\tau_{\mathrm{ess}}(y)$ with respect to x axis and y axis for different umbrella centers. (D) Average estimation errors and their SDs of MBAR, xTRAM, and TRAM for different simulation trajectory lengths over 30 independent realizations of US.

residues around the binding site. The first independent component (IC) is strongly correlated with the US coordinate. From the remaining ICs, two had timescales implied by the TICA eigenvalues larger than the trajectory length, indicating undesirable metastable transitions orthogonal to the umbrella coordinate. The second IC corresponds to closing of the binding pocket by the Trp 215 side chain (Fig. 5D, structures *i* and *iii*). The third IC corresponds to an isomerization of the disulfide bond between Cys 191 and Cys 220. An analysis using MBAR or WHAM is thus unfeasible or inefficient, as the global equilibrium assumption is strongly violated.

One strategy to deal with this very common problem is to restrain coordinates orthogonal to the umbrella coordinate, to avoid undesirable degrees of freedom from switching [45]. Although this approach is useful for computing energy differences between end states, it may change or restrain the transition mechanism and artificially increase free-energy barriers along the pathway. With TICA and TRAM, we now have the possibility to allow these orthogonal dynamics to happen, and to treat these events explicitly.

The space spanned by the US coordinate and the second IC was discretized into 100 Voronoi cells with the k-means algorithm (Fig. 5A). This number of states is far smaller than the number of bins that would be required with a binned estimator such as WHAM or discrete TRAM. A count matrix c_{ij}^k was estimated for every umbrella at a lag time of 11 ns, and the largest strongly connected component $\mathcal{S}$ of the summed count matrix $c_{ij} = \sum_k c_{ij}^k$ was determined. The initial set was strongly disconnected, and we therefore adaptively started new umbrella simulations in nine rounds, to improve the connectivity (Appendix). In the complete dataset, some clusters are still disconnected (red clusters in Fig. 5A).

In particular, these disconnected states include structures in which the binding site is occluded by a tryptophan side chain, while benzamidine is still inside, and structures in which the binding site attempts to close during the exit pathway. TRAM is applied on the connected subset of states (white clusters in Fig. 5A). The TRAM results show that the Trp-occluded conformation is a local minimum in the free-energy landscape (Fig. 5C). This is confirmed by refs. [13] and [61] where the Trp-occluded conformation is shown to be a metastable conformation of the protein. In contrast, this local minimum is not found by MBAR, and several disconnected minima are spuriously estimated (boxes in Fig. 5B).

To analyze the full high-dimensional binding mechanism and estimate unbiased kinetics, we must go beyond US simulations. We therefore used TRAM to combine the US data with up to 49.1 μs of unbiased MD data (details in Appendix). The unbiased trajectories started in the unbound state, such that many binding events are present. Individual steps of dissociation events are found in some trajectories, but no complete dissociation event is found in any single trajectory. By combining the free-energy information inherent in the biased trajectories with the binding kinetics from the unbiased trajectories, the full unbinding kinetics can be estimated with TRAM. TRAM gives the estimate $k_{\mathrm{off}}^{\mathrm{TRAM}} = 1170\,\mathrm{s}^{-1}$, with 95% confidence intervals of $[617\,\mathrm{s}^{-1}, 2120\,\mathrm{s}^{-1}]$. For comparison, the MSM estimated from the unbiased simulation data only, using the same state definition and lag time as for TRAM, provides an estimate of $k_{\mathrm{off}}^{\mathrm{MSM}} = 1863\,\mathrm{s}^{-1}$ with a larger uncertainty of $[876\,\mathrm{s}^{-1}, 4816\,\mathrm{s}^{-1}]$ [all errors estimated using bootstrap, experimental dissociation rate 600 s^{-1} [71]].

To assess the data efficiency of TRAM and the MSM, we varied the amount of unbiased MD data that was used for the estimation. With TRAM, only 5–10% of the unbiased MD data are needed compared with an MSM to reliably estimate k_{off} (Fig. 5E).

Fig. 5D shows a kinetic network of the binding/dissociation events at the unbiased ensemble of the multiensemble Markov model. The kinetic data include

association to several secondary binding sites; two of them are shown in Fig. 5D. At the lag time that we chose (30 ns), the prebound states *iii* and *iv* are not metastable and indirect transitions where these states are skipped during binding/unbinding appear in the transition matrix (not shown in the figure for clarity).

CONCLUSIONS

We have derived the TRAM for estimating MEMMs from simulation data comprising arbitrary combinations of unbiased MD, biased enhanced sampling simulations such as US, or multitemperature simulations such as REMD. TRAM does not require binning of the bias energies and is therefore suitable for the analysis of multitemperature simulations and of high-dimensional state spaces. TRAM is a Markov modeling method as it only requires local equilibrium and uses conditional transition statistics to estimate the MEMM—i.e., it can use short trajectories whose starting points were not sampled from global equilibrium. Even when just being used for estimating thermodynamics, e.g., the equilibrium distribution at the unbiased ensemble, or temperature-dependent free energies, TRAM is superior to global equilibrium-based estimators such as WHAM or MBAR.

In an application to US simulations of protein–ligand binding, we have used MSM concepts of finding slow coordinates and detecting a connected set of states to define a meaningful subspace for computing a ligand dissociation pathway and its free-energy profile. We have also sketched an approach to identify sampling bottlenecks and extend simulations in nonconverged umbrella windows to adaptively improve the convergence of the umbrella simulation, in line with other adaptive approaches [52, 72]. In this example, combining US with replica exchange simulations may also have improved the sampling [73, 74].

We demonstrated that TRAM can be used to compute an unbiased estimate of protein–ligand dissociation kinetics on the order of a millisecond by using only a few microseconds of simulation data. Beyond the simple two-state rate, TRAM is ideally suited to estimate the full multistate kinetics that was found in refs. [46] and [13] with rate models or much more simulation data. TRAM significantly expands the power of the MSM framework by allowing to integrate the full power of enhanced sampling simulations. The TRAM estimator is included in PyEMMA [75] as of version 2.2. Tutorials can be found under pyemma.org.

MATERIALS AND METHODS

Three-Well Potential Setup.

The potential shown in Fig. 4A is defined by a sum of four Gaussians $u(x,y) = -\sum_{i=1}^{4} a_i g_{h_1,h_2,\sigma_1,\sigma_2}(x,y)$ with parameters (8,15,15,10,10), (4.8,9,9,2.5,2.5), (8,9,21,2.5,2.5), and (4,21,13,2.5,2.5), and $g_{h_1,h_2,\sigma_1,\sigma_2}(x,y) = \exp\left[-(x-h_1)^2/2\sigma_1^2 - (y-h_2)^2/2\sigma_1^2\right]$ on a square $[5,25]\times[5,25]$ and ∞ outside. US simulations are conducted using bias potentials $b^k(x,y) = (x-\bar{x}^k)^2/5$ for $k = 1, ..., 7$ with umbrella centers $\{\bar{x}^k\}$ positioned at $\bar{x}^k = (10k+5)/3$. We generate 20 independent simulation trajectories $\{(x_t^k, y_t^k)\}$ for each biased potential using the Metropolis sampling algorithm, where $x_0^k = \bar{x}^k$, y_0^k is randomly drawn from $[5,25]$ and the candidate sample follows the uniform distribution on $[x_t^k - 3, x_t^k + 3]\times[y_t^k - 3, y_t^k + 3]$ for a given (x_t^k, y_t^k). The autocorrelation time of $\{(x_t^k, y_t^k)\}$ in x is given by $\tau_{\mathrm{ess}} = 1 + 2\sum_{s=1}^{\infty}\rho_s(x)$, and likewise in y [70], where $\rho_s(h)$ denotes the autocorrelation at lag s of h. We compute τ_{ess} from a long trajectory with 10^6 steps as described in ref. [76].

Trypsin–Benzamidine Setup.

US. The US coordinate is defined as the distance from the center of mass of all backbone atoms of Asp 189 and Pro 161 to the center of mass of the benzamidine ring atoms. The setup consists of 150 harmonic umbrellas with uniform force constant of $100\,\mathrm{kcal\,mol^{-1}\,Å^{-2}}$ and umbrella centers positioned along the US coordinate according to $x_{\mathrm{rest},i} = 10.5\,\mathrm{Å} + i\cdot 0.05\,\mathrm{Å}$ for $i = 0\ldots149$. For each umbrella, multiple independent runs were generated all starting from the same initial conditions that come from an unbiased binding trajectory. In total, 459 trajectories each having a length of 20 ns were generated adaptively in nine rounds of restarts. After an initial exploratory round, eight additional rounds were started to increase the overlap $o_{k,k+1} = \sum_{i\in\mathcal{S}}\min(N_i^k, N_i^{k+1})$ between ensembles. Restarts were done in ensembles k, $k+1$ where $o_{k,k+1} < 100$. For the analysis, TICA [29] was used at a lag time 5.5 ns on the Cartesian coordinates of all heavy atoms within a 15 Å radius around Asp 189 in the Protein Data Bank structure.

Molecular dynamics. A total of 491 unbiased MD simulation trajectories of length 100 ns each (data from ref. [34]) were discretized by selecting the nearest neighbor heavy-atom contacts between benzamidine and all trypsin residues as input features [75]. The features were transformed by TICA (lag time, 5 ns) to a kinetic map

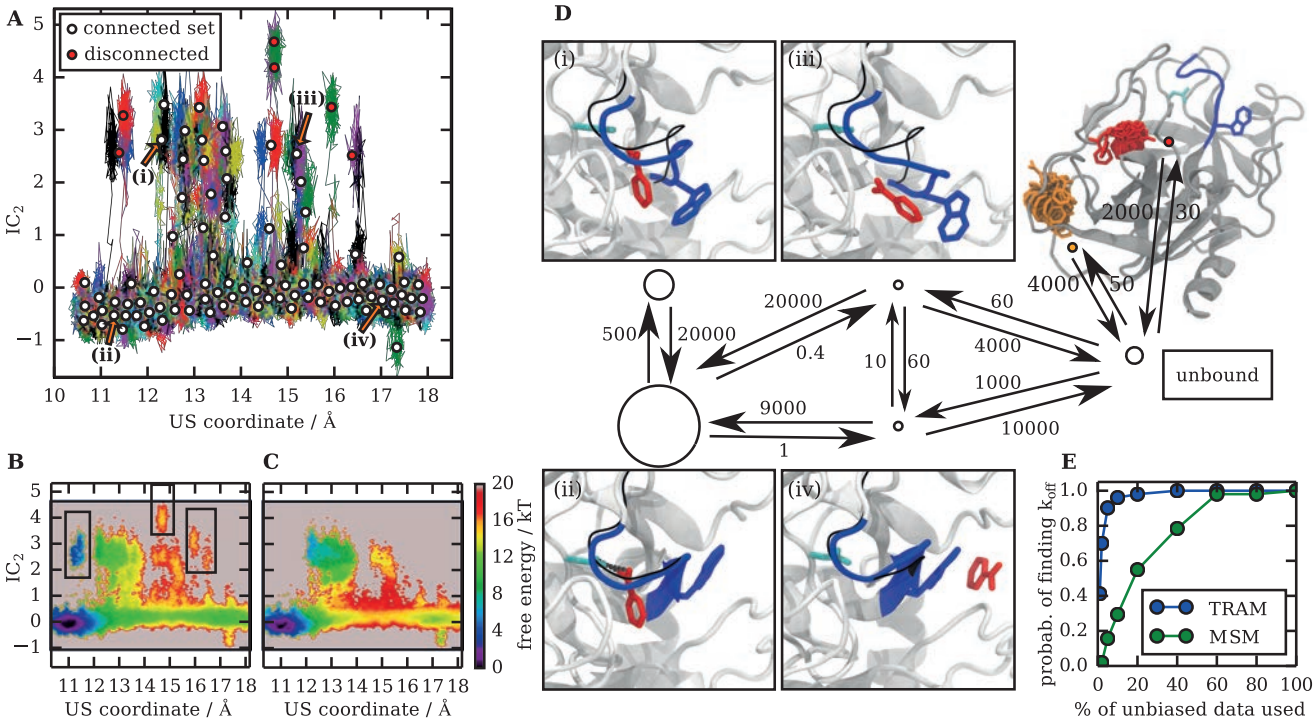

Figure 5. Thermodynamics and kinetics of all-atom protein–ligand binding model for trypsin–benzamidine. (A–C) US simulations. (D and E) MEMM using both unbiased and US simulations. (A) Trajectories projected on the space of the umbrella sampling (US) coordinate and the second independent component (IC). The US coordinate describes a transition from benzamidine bound to Asp-189 to benzamidine located outside the binding pocket on the surface of trypsin. The second IC corresponds to concerted opening of loop (Trp-215-Gln-221) and flipping of Trp-215. The Voronoi centers of the Markov states are shows as disks. Markov states that are irreversibly connected to the data set are shown as red disks and are excluded from the MEMM. (B) Potential of mean force (PMF) in the same coordinate space computed with MBAR; (C) PMF computed with TRAM. Besides a higher barrier along the US coordinate, the TRAM-PMF gives the Trp-occluded conformation a lower free energy compared with the MBAR result. (D) Coarse-grained kinetic network of the MEMM. Structures (*i*, *ii*, *iii*, and *iv*) are found in the four quadrants of A. The largest transition rates (where at least one direction exceeds 1/ms) between these macrostates, the unbound state and two alternatively bound states are shown as arrows. Units are events per millisecond. (E) Efficiency of TRAM in the estimation of unbinding kinetics compared with an MSM built from the same unbiased data. Shown is the probability that $\log k_{\mathrm{off}}$ calculated from a bootstrap sample falls into the interval $[0.5 \log k_{\mathrm{off}}^{\mathrm{all}}, 2 \log k_{\mathrm{off}}^{\mathrm{all}}]$ where $k_{\mathrm{off}}^{\mathrm{all}}$ is the TRAM estimate calculated using all data.

preserving 95% of the kinetic variance [31], resulting in a 31-dimensional transformed space. Discretization of all data in the joint space of 31 ICs and the two coordinates shown in Fig. 5A was done with the k-means algorithm, using $k = 500$. Microstates were grouped into seven macrostates. Four macrostates correspond to the quadrants of Fig. 5A, splitting microstates near the binding site at the US coordinate 14.5 Å and the TICA coordinate $\mathrm{IC}_2 = 1$. Nearness to the binding site is defined by an US coordinate < 18.2 Å and being inside a binding funnel, defined by $\cos\gamma \leq 0.74$, where γ measures the angle between the vectors connecting centers of mass of benzamidine with Pro 161 and Trp 215 with Pro 161. The remaining microstates were grouped in three macrostates: the unbound state and two alternatively bound states where benzamidine binds to secondary binding sites of trypsin, found with PCCA++ [77].

Kinetics. Using TRAM, MEMMs were estimated combining the US data and the unbiased MD data. The MEMM lag times were chosen as 30 ns for the unbiased data and as 10 ns for the US data (chosen from the interval where k_{off} appears to be independent of both lag times). A transition matrix for the unbiased ensemble was computed according to Eq. 16. k_{off} was computed as the reciprocal of the mean-first-passage time from

the bound macrostates (US coordinate < 14.5 Å) to the unbound state.

Bootstrapping: Errors bars for the different estimates were obtained from a bootstrap. Every sample of the bootstrap was generated by first partitioning the trajectories by ensemble, and then independently for every partition drawing whole trajectories and finally merging the trajectories.

APPENDIX

Solution of the TRAM Problem.

Ignoring constants, the constrained optimization problem of the TRAM log likelihood (9) can be written as follows:

$$\begin{aligned} \min_{\{\mu(x)\},\{\mathbf{P}^k\}} \quad & -\sum_{k,i,j} c_{ij}^k \ln p_{ij}^k - \sum_{k,i} N_i^k f_i^k - \sum_{x\in X} \ln \mu(x) \\ s.t. \quad & \mathrm{e}^{-f_i^k} p_{ij}^k = \mathrm{e}^{-f_j^k} p_{ji}^k \qquad \text{for all } i,j,k \\ & \sum_j p_{ij}^k = 1 \qquad \text{for all } i,j \end{aligned} \tag{20}$$

with

$$f_i^k = -\ln \sum_{x\in X_i} \mu(x)\, \mathrm{e}^{-b^k(x)}. \tag{21}$$

We omit the normalization constraint (17), because the normalization of $\mu(x)$ does not affect the optimality of the solution of Eq. 20, and we can thus normalize $\mu(x)$ a posteriori.

Using the Lagrange duality lemma of discrete TRAM problem [56], it can be shown that Eq. 20 is equivalent to the following unconstrained min-max optimization problem:

$$\begin{aligned} &\min_{\{\mu(x)\}} \max_{\{\mathbf{v}^k\}} L_{\mathrm{dual}} = \sum_{k,i,j} c_{ij}^k \ln\left(\mathrm{e}^{-f_i^k} v_j^k + \mathrm{e}^{-f_j^k} v_i^k\right) \\ &- \sum_{k,i}\left(N_i^k - \sum_j c_{ji}^k\right) f_i^k - \sum_{i,k} v_i^k - \sum_{x\in X} \ln \mu(x) \end{aligned} \tag{22}$$

where $\mathbf{v}^k = [v_i^k]$ are Lagrange multipliers. Equivalence means that the optimal solution of Eq. 20 can be obtained from that of Eq. 22 by using Eq. 16.

We now consider solving Eq. 22. Because L_{dual} is a concave function of $\{\mathbf{v}^k\}$ and a convex function of $\{\ln \mu(x)\}$, the optimal solution of Eq. 22 can be characterized as a saddle point with $\partial L_{\mathrm{dual}}/\partial v_i^k = 0$ and $\partial L_{\mathrm{dual}}/\partial \mu(x) = 0$ for all i, k and x (see section 10.3.4 in ref. [78]). Because

$$\frac{\partial L_{\mathrm{dual}}}{\partial v_i^k} = \sum_j \frac{c_{ij}^k + c_{ji}^k}{\exp\left(f_j^k - f_i^k\right) v_j^k + v_i^k} - 1 \tag{23}$$

$$\frac{\partial L_{\mathrm{dual}}}{\partial \mu(x)} = \sum_k R_{i(x)}^k \mathrm{e}^{f_{i(x)}^k - b^k(x)} - \mu(x)^{-1}, \text{ for } x \in S_i \tag{24}$$

where R_i^k is defined in Eq. 15, we can conclude that the optimal solution of Eq. 22 should satisfy Eq. 13, and Eq. 17 holds. Substituting Eq. 17 into Eq. 21, we can get the optimality condition (14).

Asymptotic Correctness of TRAM.

We use $\bar{b}$ to denote the exact value of an unknown variable b without any statistical error, and denote by $c_i^k = \sum_j c_{ij}^k$ the sum of row i in count matrix $\mathbf{C}^k = [c_{ij}^k]$, and by N_i the number of samples in S_i (c_i^k is different from N_i^k for finite statistics).

Now we show that the TRAM estimates of local partition functions, transition matrices, and reference distribution converge to the correct ones under the condition that the size of simulation data (either length or number of simulation trajectories) tends to infinity and the ratio N_i^k/N_i tends to a constant w_i^k for any i,k under the assumption that the local equilibrium within each configuration S_i is achieved in simulations. In this limit, the transition counts become the following:

$$c_{ij}^k = c_i^k \bar{p}_{ij}^k \tag{25}$$

Substituting Eq. 25 into Eqs. 16, 15 and 13, and replacing f_i^k, v_i^k with $\bar{f}_i^k, c_i^k$, we have $p_{ij}^k = \bar{p}_{ij}^k$, $R_i^k = N_i^k$, and

$$\sum_j \frac{c_{ij}^k + c_{ji}^k}{\exp\left[\bar{f}_j^k - \bar{f}_i^k\right] c_j^k + c_i^k} = \sum_j \bar{p}_{ij}^k = 1 \tag{26}$$

Define the average equilibrium distribution within S_i over multiple ensembles as $\bar{\mu}_i'(x) = \sum_k w_i^k \bar{\mu}_i^k(x)$. According to the law of large numbers, we can get

$$\frac{1}{N_i} \sum_{x\in X\cap S_i} a(x) = \int a(x)\, \bar{\mu}_i'(x)\, \mathrm{d}x \tag{27}$$

in the statistical limit for an arbitrary function $a(x)$. According to the above equation, we have the following:

$$\begin{aligned}
&\sum_{x\in X\cap S_i}\frac{\exp\left[\bar{f}_i^k-b^k\left(x\right)\right]}{\sum_l N_i^l\exp\left[\bar{f}_i^l-b^l\left(x\right)\right]}\\
&=\frac{1}{N_i}\sum_{x\in X\cap S_i}\frac{\exp\left[\bar{f}_i^k-b^k\left(x\right)\right]}{\sum_l w_i^l\exp\left[\bar{f}_i^l-b^l\left(x\right)\right]}\\
&=\int\frac{\exp\left[\bar{f}_i^k-b^k\left(x\right)\right]}{\sum_l w_i^l\exp\left[\bar{f}_i^l-b^l\left(x\right)\right]}\left(\sum_l w_i^l\bar{\mu}_i^l\left(x\right)\right)\mathrm{d}x\\
&=\int_{S_i}\exp\left[\bar{f}_i^k-b^k\left(x\right)\right]\bar{\mu}\left(x\right)\mathrm{d}x=1
\end{aligned}\tag{28}$$

From the above, we can conclude that in the statistical limit, the TRAM iterative algorithm converges to $v_i^k=c_i^k$ and $f_i^k=\bar{f}_i^k$, and the estimates of p_{ij}^k given by Eq. 16 are also equal to $\bar{p}_{ij}^k$ in the limit. Moreover, the corresponding estimated reference distribution is as follows:

$$\mu\left(x\right)=\frac{1}{\sum_k N_{i(x)}^k\exp\left[\bar{f}_{i(x)}^k-b^k\left(x\right)\right]},\quad\text{for } x\in S_i\tag{29}$$

and it satisfies that

$$\begin{aligned}
\mathbb{E}_\mu\left[a\left(x\right)\right]&=\sum_i\sum_{x\in X\cap S_i}\frac{a\left(x\right)}{\sum_l N_i^l\exp\left[\bar{f}_i^l-b^l\left(x\right)\right]}\\
&=\sum_i\frac{1}{N_i}\sum_{x\in X\cap S_i}\frac{a\left(x\right)}{\sum_l w_i^l\exp\left[\bar{f}_i^l-b^l\left(x\right)\right]}\\
&=\sum_i\int_{S_i}a\left(x\right)\bar{\mu}\left(x\right)\mathrm{d}x=\mathbb{E}_{\bar{\mu}}\left[a\left(x\right)\right]
\end{aligned}\tag{30}$$

for any function $a\left(x\right)$ of the system configuration. So the discrete distribution $\mu\left(x\right)$ given by the TRAM algorithm is also a consistent estimate of the reference distribution $\bar{\mu}\left(x\right)$.

Proofs That TRAM Is a Generalization of Discrete TRAM, WHAM, MSMs, and MBAR.

MBAR/binless WHAM. Suppose that all simulations are in global equilibrium and there is only one configuration state S_1 for the whole configuration space, i.e., $S_1=\Omega$. Then we can rewrite the TRAM equations 13 and 14 by dropping all of the subscripts as $v^k=c^k$ and

$$\sum_{x\in X}\frac{\exp\left[f^k-b^k\left(x\right)\right]}{\sum_l N^l\exp\left[f^l-b^l\left(x\right)\right]}=1\tag{31}$$

Eq. 31 is exactly the MBAR estimation equation for free energies f^k [49, 51, 66, 79].

Discrete (histogram-based) TRAM. Discrete (histogram-based) TRAM [56] can be expressed in the TRAM nomenclature by using bias potentials $b^k\left(x\right)$ that are step functions with

$$\mathrm{e}^{-b^k\left(x\right)}\equiv\gamma_i^k,\ \text{for } x\in S_i\tag{32}$$

Then, $\mu\left(x\right)$ in Eq. 17 takes a constant value $\mu_i=\left(\sum_k R_i^k\exp\left[f_i^k\right]\gamma_i^k\right)^{-1}$ on S_i, yielding the following estimate of the stationary probability of S_i in the unbiased ensemble:

$$\pi_i=N_i\mu_i\tag{33}$$

Substituting Eqs. 32, 33 and 17 into the TRAM equation (14), we can obtain $\exp\left[-f_i^k\right]=\gamma_i^k\pi_i$ and rewrite Eq. 33 as follows:

$$\begin{aligned}
&\frac{N_i}{\pi_i}=\mu_i^{-1}=\sum_{k,j}\frac{\left(c_{ij}^k+c_{ji}^k\right)v_j^k\gamma_i^k}{\gamma_i^k\pi_i v_j^k+\gamma_j^k\pi_j v_i^k}+\frac{N_i}{\pi_i}-\frac{\sum_{k,j}c_{ji}^k}{\pi_i}\\
&\Rightarrow\frac{\sum_{k,j}c_{ji}^k}{\pi_i}=\sum_{k,j}\frac{\left(c_{ij}^k+c_{ji}^k\right)v_j^k\gamma_i^k}{\gamma_i^k\pi_i v_j^k+\gamma_j^k\pi_j v_i^k}
\end{aligned}\tag{34}$$

Eqs. 13 and 34 are identical to the self-consistent equations of discrete TRAM [56] with bias factors γ_i^k, which means that discrete TRAM is a special case of TRAM and applies if the bias energies can be discretized without error.

WHAM. From the discrete TRAM equations (13) and (34), we can further derive the WHAM equations under the assumption that the global equilibrium is achieved with $p_{ij}^k=\pi_j^k\propto\gamma_j^k\pi_j$ and $c_{ij}^k=\pi_j^k\sum_{j'}c_{ij'}^k$ [56].

MSMs. If simulations are only performed at one ensemble, the self-consistent equations for reversible maximum-likelihood estimation of MSMs are a special case of discrete TRAM [56].

Interpretation of the Effective Counts. R_i^k

Supposing that we apply MBAR only to the samples in a given configuration state S_i, the estimates of the local free energies $\{f_i^k\}$ are given by the following:

$$\sum_{x\in X_i}\frac{\exp\left(f_i^k-b^k\left(x\right)\right)}{\sum_l N_i^l\exp\left(f_i^l-b^l\left(x\right)\right)}=1\tag{35}$$

This equation has the same form as the TRAM Eq. (14) except that R_i^k is replaced by N_i^k. Thus, we can interpret R_i^k as counts. By using Eqs. 15 and 16, we find $R_i^k=N_i^k-\sum_j c_{ji}^k+\sum_j v_j^k p_{ji}^k$. $N_i^k-\sum_j c_{ji}^k$ is the number of visits to S_i in the initial frames of all trajectories. $\sum_j v_j^k p_{ji}^k$ can be interpreted as the corrected number of

incoming transitions to S_i: First, $\sum_j v_j^k p_{ji}^k$ converges to $\sum_j c_{ji}^k$ in the limit of infinite statistics. Second, the term $N_i^k - \sum_j c_{ji}^k$ in R_i^k accounts for the first visit to S_i (which cannot be computed from the MSM alone). What remains to be included into R_i^k is the effective number of visits to S_i after the first state transition has happened. A suitable candidate would be the number of incoming transitions to S_i. What distinguishes $\sum_j v_j^k p_{ji}^k$ from $\sum_j c_{ji}^k$ is that the transition matrix is used in the computation of the former. Moreover, although $\sum_j c_{ji}^k$ and $\sum_j c_{ij}^k$ in principle are two independent variables, the quantities $\sum_j v_j^k p_{ji}^k$ and v_i^k (which can be interpreted as the corrected number of outgoing transitions) are linked by the equation $\sum_j v_j^k p_{ji}^k + v_i^k = \sum_j c_{ji}^k + \sum_j c_{ij}^k$ (which can be derived from Eq. 16). So both $\sum_j v_j^k p_{ji}^k$ and v_i^k are counts that are corrected by the Markov model, which itself fulfills detailed balance.

Implementation Notes.

In applications of this paper, we initialize the TRAM iteration with $v_i^k := 1$ and $f_i^k := 1$ as the convergence of TRAM does not seem to depend on the choice of initial point. We terminate the TRAM algorithm when the maximum change in normalized free energies $\max_{i,k} \left| f_i^k - f_i^{k,\mathrm{new}} \right| < tol$ with tol being a small number (e.g., 10^{-10}). Considering that the TRAM equations are invariant with respect to a global shift $f_i^k \to \alpha + f_i^k$, we perform the normalization after every iteration such that $\sum_i \exp\left[-f_i^k\right] = 1$ for the first ensemble $k = 1$ to avoid an uncontrolled drift of the f_i^k.

The bias factors $\exp\left[-b^k(x)\right]$ can easily exceed the maximum range of double-precision floating point numbers, so we perform most calculations in log-space to avoid the numerical overflow or underflow. For all summations of the form $\log \sum_i \exp[a_i]$, we use the log-sum-exp formula $\log \sum_i \exp[a_i] = \hat{a} + \log \sum_i \exp[a_i - \hat{a}]$, where $\hat{a} = \max_i(a_i)$.

In addition, according to our experience, the convergence of the TRAM algorithm can be significantly sped up by adding an extra update step to each iteration that shifts local free energies f_i^k by δ_i as follows:

$$f_i^{k,\mathrm{new}} = f_i^k + \delta_i \tag{36}$$

with

$$\delta_i = \ln \sum_{k,j} \frac{\left(c_{ij}^k + c_{ji}^k\right) v_j^k}{v_j^k + \exp\left[f_i^k - f_j^k\right] v_i^k} - \ln \sum_{k,j} c_{ji}^k \tag{37}$$

Note that we can obtain from Eqs. 15 and 14 that

$$\begin{aligned}
&\sum_{k,j} \frac{\left(c_{ij}^k + c_{ji}^k\right) v_j^k}{v_j^k + \exp\left[f_i^k - f_j^k\right] v_i^k} - \sum_{k,j} c_{ji}^k \\
&\quad = \sum_k R_i^k \cdot \sum_{x \in X \cap S_i} \frac{\exp\left[f_i^k - b^k(x)\right]}{\sum_l R_i^l \exp\left[f_i^l - b^l(x)\right]} - \sum_k N_i^k \\
&\quad = \sum_{x \in X \cap S_i} 1 - \sum_k N_i^k = 0
\end{aligned} \tag{38}$$

Hence $\delta_i = 0$ is a necessary condition for the TRAM equations, and the update step (36) does not influence the optimality of the limit of the algorithm.

Acknowledgments

We are grateful to A. S .J. S. Mey for sharing the Alanine dipeptide simulations and to P. G. Bolhuis, J. D. Chodera, C. Clementi, G. De Fabritiis, G. Hummer, A. Laio, J.-H. Prinz, E. Rosta, B. Roux, B. Trendelkamp-Schroer, and G. A. Voth for enlightening discussions.

[1] Jensen MO, et al. (2012) Mechanism of voltage gating in potassium channels. *Science* 336:229–233.

[2] Zhu F, Hummer G (2010) Pore opening and closing of a pentameric ligand-gated ion channel. *Proc. Natl. Acad. Sci. USA* 107:19814–19819.

[3] Bernèche S, Roux B (2001) Energetics of ion conduction through the K+ channel. *Nature* 414:73–77.

[4] Köpfer DA, et al. (2014) Ion permeation in K+ channels occurs by direct Coulomb knock-on. *Science* 346:352–355.

[5] Kohlhoff KJ, et al. (2014) Cloud-based simulations on Google exacycle reveal ligand modulation of GPCR activation pathways. *Nat. Chem.* 6:15–21.

[6] Dror RO, et al. (2011) Pathway and mechanism of drug binding to g-protein-coupled receptors. *Proc. Natl. Acad. Sci. USA* 108:13118–13123.

[7] Nygaard R, et al. (2013) The dynamic process of β2-adrenergic receptor activation. *Cell* 152:532–542.

[8] Blood PD, Voth GA (2006) Direct observation of bin/amphiphysin/rvs (bar) domain-induced membrane curvature by means of molecular dynamics simulations. *Proc. Natl. Acad. Sci. USA* 103:15068–15072.

[9] Arkhipov A, Yin Y, Schulten K (2009) Membrane-bending mechanism of amphiphysin n-bar domains. *Biophys. J.* 97:2727–2735.

[10] Reubold TF, et al. (2015) Crystal structure of the dynamin tetramer. *Nature* 525:404–408.

[11] Voelz VA, Bowman GR, Beauchamp KA, Pande VS (2010) Molecular Simulation of ab Initio Protein Folding for a Millisecond Folder NTL9. *J. Am. Chem. Soc.* 132:1526–1528.

[12] Silva DA, et al. (2014) Millisecond dynamics of rna polymerase ii translocation at atomic resolution. *Proc. Natl. Acad. Sci. USA* 111:7665–7670.

[13] Plattner N, Noé F (2015) Protein conformational plasticity and complex ligand binding kinetics explored by atomistic simulations and markov models. *Nature Commun.* 6:7653.

[14] Sadiq SK, Noé F, De Fabritiis G (2012) Kinetic characterization of the critical step in HIV-1 protease maturation. *Proc. Natl. Acad. Sci. USA* 109:20449–20454.

[15] Silva DA, Bowman GR, Sosa-Peinado A, Huang X (2011) A role for both conformational selection and induced fit in ligand binding by the lao protein. *PLoS Comput. Biol.* 7:e1002054.

[16] Shaw DE, et al. (2010) Atomic-Level Characterization of the Structural Dynamics of Proteins. *Science* 330:341–346.

[17] Salomon-Ferrer R, Goetz AW, Poole D, Grand SL, Walker RC (2013) Routine microsecond molecular dynamics simulations with amber - part ii: Particle mesh ewald. *J. Chem. Theory Comput. 9* 9:3878–3888.

[18] Harvey M, Giupponi G, Fabritiis GD (2009) Acemd: Accelerated molecular dynamics simulations in the microseconds timescale. *J. Chem. Theory Comput.* 5:1632–1639.

[19] 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.

[20] Shirts M, Pande VS (2000) Screen savers of the world unite! *Science* 290:1903–1904.

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

[22] Schütte C, Fischer A, Huisinga W, Deuflhard P (1999) A Direct Approach to Conformational Dynamics based on Hybrid Monte Carlo. *J. Comput. Phys.* 151:146–168.

[23] Swope WC, Pitera JW, Suits F (2004) Describing protein folding kinetics by molecular dynamics simulations: 1. Theory. *J. Phys. Chem. B* 108:6571–6581.

[24] Singhal N, Pande VS (2005) Error analysis and efficient sampling in Markovian state models for molecular dynamics. *J. Chem. Phys.* 123:204909.

[25] Sriraman S, Kevrekidis IG, Hummer G (2005) Coarse Master Equation from Bayesian Analysis of Replica Molecular Dynamics Simulations. *J. Phys. Chem. B* 109:6479–6484.

[26] Noé F, Horenko I, Schütte C, Smith JC (2007) Hierarchical Analysis of Conformational Dynamics in Biomolecules: Transition Networks of Metastable States. *J. Chem. Phys.* 126:155102.

[27] Chodera JD, et al. (2007) Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics. *J. Chem. Phys.* 126:155101.

[28] Prinz JH, et al. (2011) Markov models of molecular kinetics: Generation and validation. *J. Chem. Phys.* 134:174105.

[29] Perez-Hernandez G, Paul F, Giorgino T, D Fabritiis G, Noé F (2013) Identification of slow molecular order parameters for markov model construction. *J. Chem. Phys.* 139:015102.

[30] Schwantes CR, Pande VS (2013) Improvements in markov state model construction reveal many non-native interactions in the folding of ntl9. *J. Chem. Theory Comput.* 9:2000–2009.

[31] Noé F, Clementi C (2015) Kinetic distance and kinetic maps from molecular dynamics simulation. *J. Chem. Theory Comput.* 22:5002–5011.

[32] Noé F, Schütte C, Vanden-Eijnden E, Reich L, Weikl TR (2009) Constructing the full ensemble of folding pathways from short off-equilibrium simulations. *Proc. Natl. Acad. Sci. USA* 106:19011–19016.

[33] Bowman GR, Voelz VA, Pande VS (2011) Atomistic Folding Simulations of the Five-Helix Bundle Protein Lambda 6-85. *J. Am. Chem. Soc.* 133:664–667.

[34] Buch I, Giorgino T, De Fabritiis G (2011) Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations. *Proc. Natl. Acad. Sci. USA* 108:10184–10189.

[35] Tummino PJ, Copeland RA (2008) Residence time of receptor-ligand complexes and its effect on biological function. *Biochemistry* 47:5481–5492.

[36] Torrie GM, Valleau JP (1977) Nonphysical Sampling Distributions in Monte Carlo Free-Energy Estimation: Umbrella Sampling. *J. Comp. Phys.* 23:187–199.

[37] Souaille M, Roux B (2001) Extension to the weighted histogram analysis method: combining umbrella sampling with free energy calculations. *Comput. Phys. Commun.* 135:40–57.

[38] Marinari E, Parisi G (1992) Simulated tempering: A new monte carlo scheme. *Euro. Phys. Lett.* 19:451–458.

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

[40] Sugita Y, Okamoto Y (1999) Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* 314:141–151.

[41] Laio A, Parrinello M (2002) Escaping free energy minima. *Proc. Natl. Acad. Sci. USA* 99:12562–12566.

[42] Grubmüller H (1995) Predicting slow structural transitions in macromolecular systems: conformational flooding. *Phys. Rev. E* 52:2893.

[43] Hénin J, Chipot C (2004) Overcoming free energy barriers using unconstrained molecular dynamics simulations. *J. Chem. Phys.* 121:2904–2914.

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

[45] Gumbart JC, Roux B, Chipot C (2013) Efficient determination of protein–protein standard binding free energies from first principles. *J. Chem. Theory Comput.* 9:3789–3798.

[46] Tiwary P, Limongelli V, Salvalaglio M, Parrinello M (2014) Kinetics of protein–ligand unbinding: Predicting pathways, rates, and rate-limiting steps. *Proc. Natl. Acad. Sci. USA* 112:E386–E391.

[47] Ferrenberg AM, Swendsen RH (1989) Optimized Monte Carlo data analysis. *Phys. Rev. Lett.* 63:1195–1198.

[48] Kumar S, Rosenberg JM, Bouzida D, Swendsen RH, Kollman PA (1992) The weighted histogram analysis method for free-energy calculations on biomolecules. i. the method. *J. Comput. Chem.* 13:1011–1021.

[49] Shirts MR, Chodera JD (2008) Statistically optimal analysis of samples from multiple equilibrium states. *J. Chem. Phys.* 129:124105.

[50] Kong A, McCullagh P, Meng XL, Nicolae D, Tan Z (2003) A theory of statistical models for Monte Carlo integration. *J. Roy. Stat. Soc. B Met.* 65:585–604.

[51] Bartels C (2000) Analyzing biased Monte Carlo and molecular dynamics simulations. *Chem. Phys Lett.* 331:446–454.

[52] Rosta E, Woodcock HL, Brooks BR, Hummer G (2009) Artificial reaction coordinate "tunneling" in free-energy calculations: The catalytic reaction of RNase H. *J. Comput. Chem.* 30:1634–1641.

[53] Trendelkamp-Schroer B, Noé F (2016) Efficient estimation of rare-event kinetics. *Phys. Rev. X* 6:011009.

[54] Wu H, Noé F (2014) Optimal estimation of free energies and stationary densities from multiple biased simulations. *SIAM Multiscale Model. Simul.* 12:25–54.

[55] Mey ASJS, Wu H, Noé F (2014) xTRAM: Estimating equilibrium expectations from time-correlated simulation data at multiple thermodynamic states. *Phys. Rev. X* 4:041018.

[56] Wu H, Mey ASJS, Rosta E, Noé F (2014) Statistically optimal analysis of state-discretized trajectory data from multiple thermodynamic states. *J. Chem. Phys.* 141:214106.

[57] Rosta E, Hummer G (2015) Free energies from dynamic weighted histogram analysis using unbiased Markov state model. *J. Chem. Theory Comput.* 11:276–285.

[58] Minh DDL, Chodera JD (2009) Optimal estimators and asymptotic variances for nonequilibrium path-ensemble averages. *J. Chem. Phys.* 131:134110.

[59] Chodera JD, Swope WC, Noé F, Prinz JH, Pande VS (2011) Dynamical reweighting: Improved estimates of dynamical properties from simulations at multiple temperatures. *J. Phys. Chem.* 134:244107.

[60] Prinz JH, et al. (2011) Optimal use of data in parallel tempering simulations for the construction of discrete-state Markov models of biomolecular dynamics. *J. Chem. Phys.* 134:244108.

[61] Tiwary P, Parrinello M (2013) From metadynamics to dynamics. *Phys. Rev. Lett.* 111:230602.

[62] Bowman GR, Beauchamp KA, Boxer G, Pande VS (2009) Progress and challenges in the automated construction of Markov state models for full protein systems. *J. Chem. Phys.* 131:124101.

[63] Trendelkamp-Schroer B, Wu H, Paul F, Noé F (2015) Estimation and uncertainty of reversible Markov models. *J. Chem. Phys.* 143:174101.

[64] Rabiner LR (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* 77:257–286.

[65] Zhu F, Hummer G (2012) Convergence and error estimation in free energy calculations using the weighted histogram analysis method. *J. Comput. Chem.* 33:453–465.

[66] Tan Z, Gallicchio E, Lapelosa M, Levy RM (2012) Theory of binless multi-state free energy estimation with applications to protein-ligand binding. *J. Chem. Phys.* 136:144102.

[67] Zhang BW, Xia J, Tan Z, Levy RM (2015) A stochastic solution to the unbinned WHAM equations. *J. Phys. Chem. Lett.* 6:3834–3840.

[68] Tan Z (2015) Optimally adjusted mixture sampling and locally weighted histogram analysis. *J. Comp. Graph. Stat. (published online).*

[69] Gallicchio E, Andrec M, Felts AK, Levy RM (2005) Temperature weighted histogram analysis method, replica exchange, and transition paths. *J. Phys. Chem. B* 109:6722–6731.

[70] Kass RE, Carlin BP, Gelman A, Neal RM (1998) Markov chain Monte Carlo in practice: a roundtable discussion. *Am. Stat.* 52:93–100.

[71] Guillain F, Thusius D (1970) Use of proflavine as an indicator in temperature-jump studies of the binding of a competitive inhibitor to trypsin. *J. Am. Chem. Soc.* 92:5534–5536.

[72] Wojtas-Niziurski W, Meng Y, Roux B, Bernèche S (2013) Self-learning adaptive umbrella sampling method for the determination of free energy landscapes in multiple dimensions. *J. Chem. Theory Comput.* 9:1885–1895.

[73] Chipot C, Pohorille A (2007) *Free energy calculations* (Springer, Berlin).

[74] Frenkel D, Smit B (2001) *Understanding Molecular Simulation: From Algorithms to Applications*, Computational Science Series (Elsevier Science, Burlington, MA).

[75] Scherer MK, et al. (2015) PyEMMA 2: A software package for estimation, validation and analysis of markov models. *J. Chem. Theory Comput.* 11:5525–5542.

[76] Geyer CJ (1992) Practical Markov chain Monte Carlo. *Stat. Sci.* 7:473–483.

[77] Röblitz S, Weber M (2013) Fuzzy spectral clustering by PCCA+: application to Markov state models and data classification. *Adv. Data Anal. Classif.* 7:147–179.

[78] Boyd S, Vandenberghe L (2004) *Convex optimization* (Cambridge University Press).

[79] Vardi Y (1985) Empirical distributions in selection bias models. *Ann. Stat.* 13:178–203.

nature COMMUNICATIONS

ARTICLE

DOI: 10.1038/s41467-017-01163-6 OPEN

Protein-peptide association kinetics beyond the seconds timescale from atomistic simulations

Fabian Paul[1,2], Christoph Wehmeyer[1], Esam T. Abualrous[1], Hao Wu[1], Michael D. Crabtree[3], Johannes Schöneberg[1], Jane Clarke[3], Christian Freund[4], Thomas R. Weikl[2] & Frank Noé[1]

Understanding and control of structures and rates involved in protein ligand binding are essential for drug design. Unfortunately, atomistic molecular dynamics (MD) simulations cannot directly sample the excessively long residence and rearrangement times of tightly binding complexes. Here we exploit the recently developed multi-ensemble Markov model framework to compute full protein-peptide kinetics of the oncoprotein fragment $^{25-109}$Mdm2 and the nano-molar inhibitor peptide PMI. Using this system, we report, for the first time, direct estimates of kinetics beyond the seconds timescale using simulations of an all-atom MD model, with high accuracy and precision. These results only require explicit simulations on the sub-milliseconds timescale and are tested against existing mutagenesis data and our own experimental measurements of the dissociation and association rates. The full kinetic model reveals an overall downhill but rugged binding funnel with multiple pathways. The overall strong binding arises from a variety of conformations with different hydrophobic contact surfaces that interconvert on the milliseconds timescale.

[1] Department of Molecular and Cell Biology and California Institute for Quantitative Biosciences, University of California, Berkeley, CA 94720, USA. [2] Max Planck Institute of Colloids and Interfaces, Department of Theory and Bio-Systems, 14476 Potsdam, Germany. [3] Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, UK. [4] Institute of Chemistry and Biochemistry, Freie Universität Berlin, Thielallee 63, 14195 Berlin, Germany. Christoph Wehmeyer and Esam T. Abualrous contributed equally to this work. Correspondence and requests for materials should be addressed to F.N. (email: frank.noe@fu-berlin.de)

In the past, drug design has primarily focused on finding inhibitors with maximal binding affinity to the target. Recently, there has been a growing interest in optimizing target-drug kinetics[1, 2]. A direct strategy to exploit kinetics is the maximization of the drug's residence time at the receptor in order to ensure contiguous drug effect between subsequent deliveries[3, 4]. Protein–ligand kinetics may involve more than two kinetically relevant states, either due to different ligand binding poses, different protein conformations or their coupling[5–10]. While this multi-state nature is not always apparent in ensemble kinetic experiments[11], accounting for it may help during multiple stages of the drug design process[12, 13]. On the molecular scale, targeting receptor binding pockets that open transiently can lead to allosteric inhibitors[14, 15]. On the pharmacokinetic scale, a complete assessment of protein–drug kinetics can provide more accurate models and offer additional freedom to optimize the drug delivery strategy[2, 16]. Multi-state kinetics are especially relevant in multivalent binders, which are characterized by highly non-exponential kinetics and nonlinear amplification of the binding strength through multiple parallel binding interfaces[17, 18].

Simultaneous study of molecular structure and kinetics at high resolution is possible with fully flexible all-atom molecular dynamics (MD) simulation in explicit solvent. However, such simulations are limited to lengths of few microseconds on publicly available hardware. Few milliseconds can be reached on specialized hardware[19] or in aggregate times using distributed computing[20–23]. These simulation times are short compared to residence times of most high-affinity binders.

Calculating unbiased long-term kinetics for all-atom MD models is one of the hardest problems in molecular simulation, as it depends upon the solution of three difficult tasks simultaneously: (A) the ability to explore initially unknown states and conformational changes, (B) the repeated sampling of the slowest transitions, (C) the computation of unbiased transition rates from such simulation data. Fortunately, tools have been established that each excel at one or two of these tasks, and that can be combined to a powerful framework.

Path sampling and milestoning-based methods[24–27] enhance the probability of transition pathways between a priori known end-states and can be extended to compute transition rates (tasks B, C), but offer only limited help in exploring the state space. In contrast, unbiased MD simulations, especially high-throughput MD simulations[28, 29] can explore the state space without hindrance from constraints (task A). When analyzed with kinetic models, such as Markov state models (MSMs)[30–33], the unbiased long-term kinetics can be approximated[34, 35], without required initial knowledge of relevant states, coordinates or a timescale separation (task C). However, this approach relies on having sampled the rare-event transitions in the data. While MSMs help with parallelizing this problem and rare events can be sampled, in particular when adaptive sampling strategies are combined with high-throughput simulation[23], the sampling of very rare events such as protein-inhibitor dissociation can still be very inefficient. In practice, this difficulty may result in not properly connected models and underestimated or imprecisely estimated residence times. While MSM analyses have the advantage of being able to detect these problems with carefully conducted Markovianity tests[36] and by computing binding free energies as a function of the MSM lag time[37, 38], the typical solution involves running more simulations, which is unpractical when computational resources are limited. Enhanced sampling methods such as umbrella sampling, flooding, metadynamics, or replica exchange[39–42] are specialized in rare-event sampling (task B), and some of them can significantly help to explore states with low populations (task A), however they rely on a priori knowledge of good collective coordinates. Kinetic quantities cannot be directly computed from such data and the data analysis relies on the applicability of macroscopic rate theories[43]. This has been mitigated by recent progress in hyper-dynamics which allows to predict transition rates between long-lived states when good collective coordinates are known[44–48].

In order to combine the advantages of enhanced sampling methods and MSMs, we recently developed the concept of multi-ensemble Markov models (MEMMs)[49]. MEMMs rely on the idea of combining unbiased simulations of fast events (such as rapid binding) with efficient sampling of the rare events in biased ensembles (such as biased unbinding) within a reweighting framework that can extract full and unbiased kinetics. Several MEMM estimators have been developed[50–52], including the statistically optimal transition-based reweighting analysis method (TRAM), which exploits detailed balance to extract unbiased kinetics of the slow steps from equilibrium properties harvested at biased ensembles[49, 53]. The recently introduced binless TRAM version can compute complex multi-state kinetics without requiring pre-defined collective variables[49], which allows kinetics in very high-dimensional and complex examples to be studied.

Here we show how enhanced MD simulation techniques can be combined to compute unbiased multi-state kinetics of the onco-protein fragment $^{25–109}$Mdm2 with the nano-molar peptide inhibitor PMI in all-atom resolution. MEMMs are the key technology for this achievement, and allow us to obtain the residence time that is beyond the seconds timescale with high accuracy and precision, from sub-millisecond simulations. Multiple intermediates and mis-bound modes are found, the equilibrium folding–binding pathways are computed. The simulations are tested against previous mutagenesis experiments and binding–unbinding kinetics experiments conducted here.

Results

Direct MD simulation of protein–ligand complex Mdm2–PMI. Mdm2 is a major therapeutic target that antagonizes the tumor suppressor p53 by ubiquitinating it or by binding the N-terminal trans-activation domain (TAD) of p53. In certain cancers, Mdm2 is over-expressed leading to excessive inactivation of p53[54]. Therefore the Mdm2–p53 interaction is a primary target for inhibitor design[55–57]. The 12-amino-acid peptide PMI (p53–Mdm2/MdmX inhibitor) is one of the strongest known Mdm2 binders, with a dissociation constant of $K_d = 3.3$ nM[57]. In the co-crystal structure of PMI with the protein fragment $^{25–109}$Mdm2, PMI binds as a helix[57] while our MD simulations of PMI without its binding partner suggest that PMI is at most 40% helical in isolation. Thus the binding mechanism must involve PMI folding. The binding of PMI to the Mdm2 protein fragment is a particular challenging system for MD not only because of the high affinity but also because of the abundance of metastable states that act as traps on achievable simulation lengths of microseconds. In Zwier et al.[58], 120 μs of implicit solvent simulations of the same Mdm2 fragment were conducted with a different p53-peptide and only 10% of the simulations reached the crystallographic binding pose.

We conducted 500 μs of unbiased atomistic MD simulations of the protein fragment $^{25–109}$Mdm2 and the PMI peptide from different initial structures, especially dissociated states. A preliminary analysis showed that these trajectories contain five complete binding events from dissociated to crystal-like states, several tens of partial binding events via intermediates. A variety of intermediates and trap states were found (Fig. 1). However, not a single clear dissociation event was observed, and a MSM constructed from the unbiased MD data contained many disconnected states.

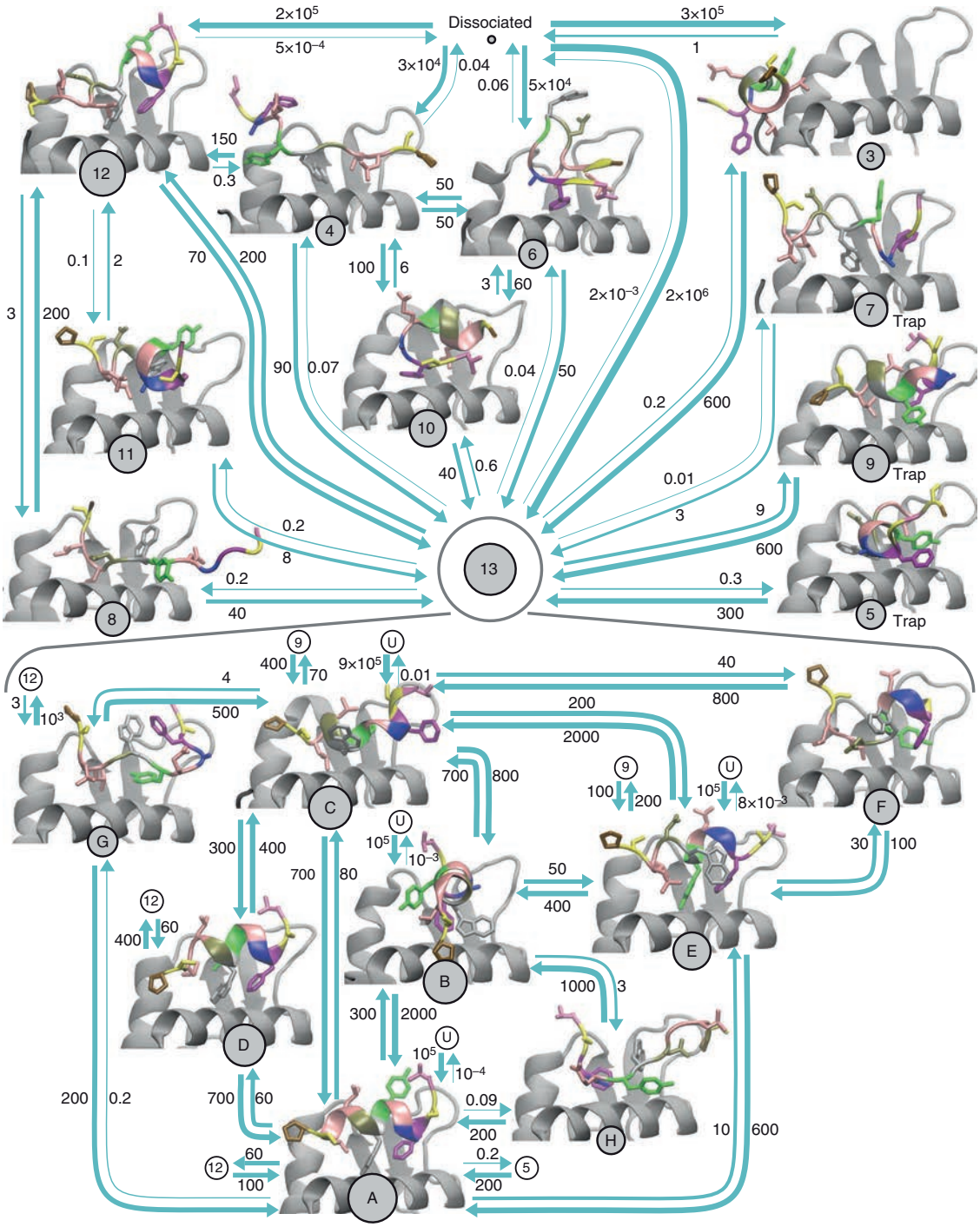

Fig. 1 Metastable states and transition rates for the binding of PMI to Mdm2. The PMI peptide is colored according to (TSFAEYWNLLSP). States are represented by discs with areas proportional to the natural logarithm of the equilibrium probability. Arrows indicate transitions with rate constants of at least 1 ms^{-1} in either direction. Numbers quantify transition rate constants in $ms^{-1}\,M^{-1}$ for association events and in ms^{-1} for all other transitions. The definition of the states is hierarchical: between top-level states 0 and 13, transitions happen on timescales of 10 μs or slower. States in the lower part of the figure are sub-states of top-level state 13. There, PMI transitions between different states in the main binding pocket of Mdm2 on timescales of microseconds or slower (only states with large probabilities are shown)

Biased simulations predict the binding affinity. Consequently, we added biased simulations with the aim of reversibly sampling bound, unbound and intermediate states. MEMMs can in principle be built using any biased sampling protocol, including umbrella sampling[39] or metadynamics[41]. Here, six independent Hamiltonian replica-exchange simulations were conducted, each about 1 μs long and with 14 replicas. The first Hamiltonian is unbiased while the other Hamiltonians have gradually reduced protein–ligand interaction strengths (see "Methods"). In contrast to unbiased MD, these simulations do not provide direct kinetic information, but sampled efficiently different binding sites and binding modes. After discarding the initial equilibration phase of 50 ns (Supplementary Note 3.3) these data still contained six full binding and 26 full

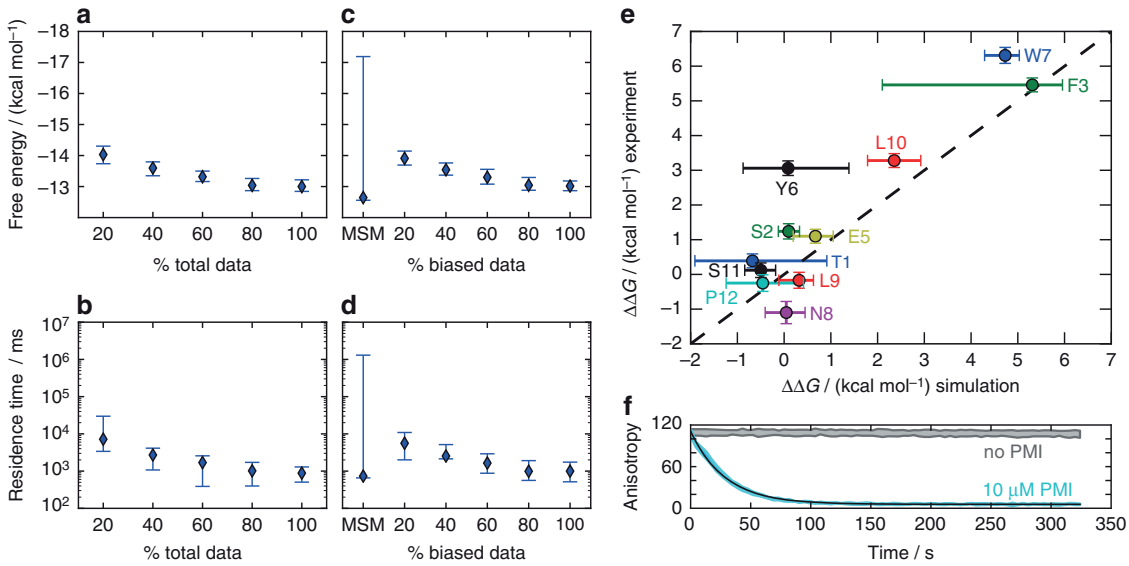

Fig. 2 Computational predictions and experimental validations of binding affinities and kinetics of PMI-Mdm2. **a**, **b** Maximum likelihood estimates of the binding free energy and residence time, respectively, as a function of the amount of data used. Diamonds mark maximum likelihood estimates, error bars indicate 95% confidence intervals. *x*% of total data means that *x*% of all biased data and *x*% of all unbiased data were used. **c**, **d** Estimates of binding free energy and residence time as a function of data composition. The fraction of biased simulation data is varied between 0 and 100% of all biased data while keeping the sum of the amount of biased and the amount of unbiased data constant at 502 μs. **e** Validation of the simulation model: binding free energy differences (ΔΔ*G*) of PMI-Mdm2 upon alanine mutations of the PMI peptide, compared between the present simulations and experiments[59]. Only biased simulation data was used and analyzed with MBAR[62]. Error bars mark standard deviations (simulation error computed using bootstrap, see Supplementary Note 3.3). **f** Cyan: average and standard deviations of anisotropy time traces from three repeats of a binding competition experiment, Mdm2 (10 nM), pre-incubated with labeled PMI (10 nM), was mixed with unlabeled PMI (10 μM). A mono-exponential function (black) was fitted to the average time trace. Gray: control without unlabeled PMI

dissociation events, as well as many transitions between intermediates (Fig. 1).

To test the enhanced sampling simulation, we determined the dissociation constant between Mdm2 and PMI experimentally using fluorescence anisotropy (see Supplementary Note 4.2 and Supplementary Fig. 11), obtaining $K_d^{\mathrm{exp}} = 3.02 \pm 0.31$ nM, in agreement with previous data[59]. Computationally, $K_d^{\mathrm{sim}} = 0.34$ nM (95% confidence interval: [0.22 nM, 0.44 nM]) was determined by applying the PyEMMA implementation[60] of the MBAR estimator[61, 62] on the replica-exchange data (Supplementary Note 3.6). The difference between the computational and the experimental value corresponds to 1.3 kcal mol^{-1}, which is in the expected range of force field inaccuracies[63, 64]. As a more comprehensive test, previously measured changes in binding free energies (ΔΔ*G*) upon mutation of PMI residues to alanine were predicted using perturbation theory[65, 66]. We find good agreement of the ΔΔ*G* values between simulation and experiment[59] within statistical uncertainties, in particular for the amino acids that are important for binding: Phe3, Trp7, and Leu10 (Fig. 2e and Supplementary Note 3.2).

Multi-ensemble Markov models reveal slow unbinding kinetics. We developed an extension of the recent TRAM estimator[49] called TRAMMBAR for combining unbiased MD simulations with replica-exchange simulations (see "Methods"). While TRAM requires all simulations to be longer than its lag time (often on the order of tens to hundreds of nanoseconds), this is not the case for replica-exchange simulations with rapid exchanges. TRAMMBAR can employ such replica-exchange data, by assuming global equilibrium for that part of the simulation, which is justified when statistical tests indicate short correlation times[62]. The present replica-exchange data has a correlation time of 40 ns, compared to simulation lengths of about 1 μs (Supplementary Note 3.3). Using TRAMMBAR, all unbiased and biased simulation data were combined to a MEMM with 1056 states at a lag time of 150 ns, and its self-consistency was validated using standard tests[35] (Supplementary Note 3.4 and Supplementary Figs. 5 and 6). The kinetics of the unbiased ensemble was then analyzed.

The association rate is predicted to be 3.3×10^9 M^{-1} s^{-1} (see "Methods", Supplementary Note 3.7) which is faster than the association of similar p53-peptides to the full-length N-terminal domain of Mdm2 (on the order of 10^7 M^{-1} s^{-1})[67] and still faster than the association of the $^{17-29}$p53 peptide to the $^{25-109}$Mdm2 fragment ($k_{\mathrm{on}} = 7 \times 10^7$ M^{-1} s^{-1})[58]. The majority of association trajectories enter basin 13 that contains the crystallographic complex and is correctly predicted as the most populous state (Fig. 1).

Computing the residence time of the complex from the transition matrix may lead to a systematic overestimate, because the dissociated state lifetime is shorter than the lag time used to estimate the transition matrix. To avoid this bias, we estimated rate matrices. Rate matrix estimation is not unique and we considered the maximum likelihood approach of Kalbfleisch and Lawless[68] which gives an estimate of the residence time of 0.88 s (95% confidence interval [0.48 s, 1.33 s], see Fig. 2b, d), and the least-squares approach of Crommelin and Vanden-Eijnden[69], which gives an estimate of 8 s (confidence interval [1.5 s, 40 s]). To test the predicted values from the simulations, we decided to measure the binding kinetics of PMI to Mdm2 experimentally. We performed a binding competition experiment with a fluorescence anisotropy readout to measure the PMI dissociation rate and stopped-flow kinetics experiments to measure the association rate (see Supplementary Notes 4.1–4.3, Fig. 2f and Supplementary Fig. 12). We measured a residence time of

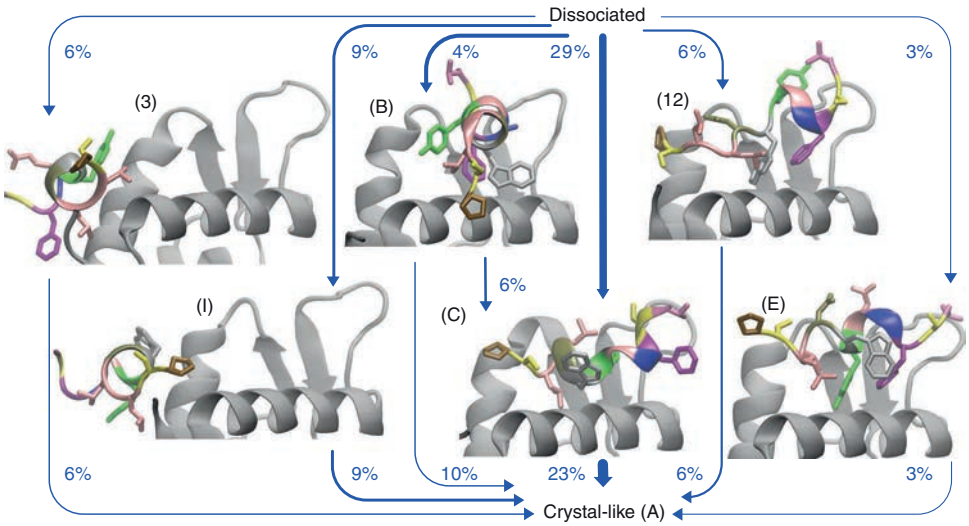

Fig. 3 Binding mechanism comprised by the 60% most probable pathways. Structures of metastable (on-pathway) intermediates are shown, labels are as in Fig. 1. Arrows indicate the direction and relative magnitude of the reactive flux from the dissociated state to the crystal-like bound state. PMI residues that form PMI-Mdm2 contacts with at least a probability of 0.5 in a given macro-state are shown as sticks

26.8 s (confidence interval [24.7 s, 34.1 s]) and an association rate constant of $5.27 \times 10^8\,M^{-1}\,s^{-1}$ (confidence interval $[5.17, 5.37] \times 10^8\,M^{-1}\,s^{-1}$). Interestingly, our simulation-based predictions and the experimental estimate for the residence time all lie in the range of seconds to tens of seconds, which is a good agreement considering expected errors in the simulation force field[63, 64] and influence of the measurement by the fluorescence label. About 50% of the simulation data, i.e., a total of 300 µs of mixed unbiased and biased data, are sufficient to get estimates that are statistically indistinguishable from the estimates using all data (Fig. 2a, b and Supplementary Fig. 9a).

To assess the importance of the biased simulations for the computation of the binding free energy and the residence time, we varied the fraction of biased data used for the estimation (Fig. 2c, d and Supplementary Fig. 9b). Both quantities converge within statistical uncertainty if at least 50% of the biased data is included in the estimation (i.e., 450 µs unbiased data and a total of 50 µs biased data in all replicas). If no biased simulation data is used and a conventional MSM is estimated (using 500 µs unbiased data) the errors increase by a magnitude that makes the estimate practically useless. Note that it is not easy to determine whether a MSM is truly connected, and it is possible that this large error actually indicates that the dissociation pathway has not been sampled in the unbiased simulations alone.

Analysis of the full kinetic network. To obtain an overview of structure-kinetics relationships, we analyzed the MEMM kinetics between the dissociated state (protein-peptide distances larger than 1 nm) and 14 metastable states that interchange on the timescale of 10 µs or slower (Fig. 1 upper half). At this resolution, the binding is overall downhill with fast direct association rates on the order of $10^9\,M^{-1}\,s^{-1}$ into the native basin 13 that dominate the experimentally measurable on-rate. Association can also occur to non-native intermediates (3, 4, 6, 12) with smaller rates of 10^7 to $10^8\,M^{-1}\,s^{-1}$ (Fig. 1).

In the most populous state 13, PMI is folded and anchored, with a high probability, to the binding pocket with its hydrophobic residues Phe3 and Trp7. In the second-most populous state 12, PMI has the folded crystallographic N-terminal conformation, but the C-terminus is unfolded and forms a different contact pattern: while Leu9 forms multiple contacts with Mdm2 helix 2 (Supplementary Fig. 4), Leu10 has no contact to Leu54, Val93, and Ile99. Ser11 forms a contact with Tyr100 and Pro12 forms contacts with Arg97, His96, and Try100 of Mdm2 (Supplementary Table 1).

To examine the importance of different PMI side chains for the observed binding modes, we computed the change in binding free energy upon mutation $\Delta\Delta G$ but with the free energy of the associated state replaced by the free energy of macro-state S_i (see Supplementary Note 3.2 and Supplementary Table 2). We observe that Phe3 and Trp7 are most important for stable binding. The role of the other side chains depends on the binding mode. For example, Thr1, Tyr6, Leu9, and Pro12 stabilize state 12 but not state 13. Alanine scanning experiments (Fig. 2e) have revealed that the Tyr6Ala mutant shows a similar $\Delta\Delta G$ to that of the Leu10Ala mutant even though the crystal structure shows no binding of Tyr6 to the inside of the hydrophobic cleft of Mdm2[59]. Our results thus suggest that the higher K_d of the Tyr6Ala PMI mutant is not due to a destabilization of the crystal-like state, but may rather be explained by the destabilization of alternative bound states.

Other binding modes that involve more flexible PMI configurations do not strongly contribute to the binding affinity, but are relevant for the association process by "catching" PMI and funneling it into state 13. In the non-native states, PMI binds in different locations (3), in different orientations (5, 10, 11), or in unfolded conformations that dissociate relatively easily, but otherwise fold during the binding transition (4, 6, 8). The slowest transitions occur between states 12 and 11 and between states 13 and 7 that happen on milliseconds to hundreds of milliseconds. Non-native states that do not significantly contribute to binding pathways are briefly denoted as "trap". Trap states 5, 7, and 9 are predominantly reachable from state 13. Additional traps with lifetimes larger than 10 µs but not significant population were found, in which PMI binds far away from the binding site (structures not shown).

To resolve the dynamics inside the main binding pocket in greater detail, we split state 13 into the sub-states A–H with kinetics on time scales of a single microsecond or slower (Fig. 1 lower half and Supplementary Note 3.5). Sub-state A is structurally well-defined and contains the crystal structure (pdb code 3eqs), the crystallographically unresolved Pro12 forms contacts with Mdm2 Tyr100. Many of the sub-states (B, C, E, I)

are intermediates in the binding process (Fig. 3). In the crystal-like state (A) the Tyr6 side chain of PMI is not buried in the binding cleft. However in many non-native states, Tyr6 can either bind to the inner cleft together with Trp7 and Phe3 (D, B, H) or take the role of Trp7/Phe3 by anchoring PMI to the cleft (C, G, F and 5). Tyr6 can even take the place of Trp7 in a helically bound conformation that is similar to the crystallographic mode (E).

With the simulations conducted here, we find that state A has a stationary probability of 72%. Together states A and 12 have a joint stationary probability of 86%. Thus a large fraction of the strong affinity between PMI and Mdm2 is due to the two distinct but individually well-defined conformations 12 and A that interconvert directly on the timescale of 10 µs.

It is possible that the number of discovered non-natively bound structures, and their combined equilibrium probability, would continue to grow if the simulations would be extended. However, almost all metastable states found here are already visited in the first 60% of our simulation data (Supplementary Fig. 7) and the estimate for the binding free energy is converged (Fig. 2a). These indicators suggest that the non-natively bound structures with significant probabilities have been found.

Binding mechanism. To investigate the binding mechanism, we computed the reactive flux using transition path theory[36, 70] from the dissociated state to the crystal-like bound state (Supplementary Note 3.5). There are multiple parallel pathways and the metastable states can be grouped into on-pathway intermediates and off-pathway trap states—see Fig. 3 for an illustration of the major 60% of binding pathways. The most populous pathway (29%) goes through a partially folded state (C) that is anchored by Leu10 and Tyr6 to the binding cleft, while Phe3 and Trp7 form contacts with the outer surface of helix 2 of Mdm2. 15% of the reactive binding flux goes through states where PMI binds to the terminal region of the Mdm2 fragment that is located at the end of the binding cleft. A similar pathway was found for the p53-peptide in Ref. [71]. The terminally bound states form a conformational ensemble with various unfolded (not shown) and folded (3, I) PMI conformations. Among the terminally bound states, the macro-states that carry most reactive flux exhibit folded PMI. The folded conformations differ in the (hydrophobic) interface that they form with Mdm2 (3, I). Nine percent of the flux go through states 12 and E where PMI is almost in the crystal-like fold but the binding pattern is non-native. Inspection of the MD trajectories shows that during the fast transition from state E to the crystal-like state, the Tyr6 side chain leaves the binding cleft first and is then replaced by the Trp7 side chain all while Phe3 remains anchored to the cleft. In the transition between state 12 and the crystal-like state, the flexible C-terminus of PMI is rearranged such that Leu10 takes the place of Leu9 at the binding interface.

Discussion

Multi-ensemble Markov models can be used to probe full multi-state kinetics of strong binders by combining conventional MD simulations of the binding process with biased MD simulations that spontaneously sample bound and unbound states. While standard analyses of enhanced sampling simulations do not readily provide kinetic information, MEMM estimators provide direct estimates of the kinetics without invoking macroscopic rate models. Using the nano-molar complex PMI–$^{25-109}$Mdm2 as an example, we obtained robust estimates of residence times that exceed the total amount of simulation data by three to four orders of magnitude and the individual simulation lengths by six to seven orders of magnitude.

Importantly, the inclusion of relatively little biased data enables us to sample rare events such as the protein-inhibitor dissociation steps, and drastically reduces the statistical error of rates and binding free energies compared to a MSM of purely unbiased MD data. In particular, we have demonstrated that MEMMs can effectively mitigate the problem of trajectories getting trapped in long-lived states. While direct estimation of MSMs requires that the visited states are reversibly connected—a condition that is difficult to test in high-dimensional systems—MEMMs only require irreversible visits to metastable states if those states were sampled reversibly in a biased simulation. On the other hand, in contrast to standard analysis methods such as WHAM or MBAR, MEMM estimators such as TRAM or TRAMMBAR do not require the full simulation data to be sampled from global equilibrium, thus greatly alleviating the sampling problem.

The binding/unbinding mechanism of PMI and Mdm2 was elucidated in full atomistic detail. While the binding is overall funnel-like, the detailed kinetics are quite complex. Rebinding can occur via multiple non-native intermediates on multi-milliseconds timescales. Another slow process is the interconversion of the crystallographic PMI–Mdm2 state with a newly identified state in which the C-terminus of PMI is unraveled and forms a new interaction pattern with Mdm2. Both states contribute significantly to the PMI–Mdm2 binding affinity and will inhibit binding to p53. The identification of such conformations gives us additional flexibility in optimizing the inhibitor.

Some minor trap states were found that do not significantly contribute to the binding affinity, but have lifetimes on the order of microseconds. Although such states may be overrepresented by current atomistic force fields[63], their existence implies that even for fast binders, around 100 µs of unbiased MD simulation are needed in order to characterize the association kinetics with statistical confidence.

The current study is a proof of principle—making optimal choice of starting structures and amount of data in unbiased vs. biased simulations depends on the molecular system, and a logical next step would be to make these choices iteratively within an adaptive sampling framework[28, 37, 72]. The present simulation approach makes progress towards the routine computation of residence times, and the identification of non-native or allosteric binding sites for protein-inhibitor systems. Because the approach does not require a priori knowledge of order parameters and structures, it can potentially be fully automated. With ever increasing computing power, this approach may become part of a high-throughput framework to compute protein–drug kinetics that may serve both pharmacological applications and the improvement of force fields towards the more accurate prediction of kinetics[73, 74].

MEMMs combine methods of free energy calculation and MSM estimation. Therefore any progress made in the development of protocols for free energy calculation might directly translate into a corresponding progress in the estimation of kinetics. We are confident that the seconds timescale is not the limit and that timescales comparable to the biological half-life of drugs (hours)[2], or the excessively long lifetimes of multivalent binders[17, 18] are, in principle, accessible. Equilibrium kinetic models of protein binding kinetics harvested with MEMMs can be embedded into particle-based reaction-diffusion simulations in order to probe the kinetics emerging from non-equilibrium conditions and the behavior of entire cellular signaling pathways[75].

Methods

MDM2-PMI simulation setup. MD simulations were conducted with the Amber99SB-ILDN force field[76] and TIP3P water model[77] in the canonical

ensemble at temperature $T = 300$ K. To generate a starting structure, we used the heavy atom positions from the protein data bank (PDB) file 3eqs[57] and moved the peptide out of the binding pocket. Missing residues of the PDB structure (PMI Pro12 and Mdm2 Glu25) were modeled in standard conformations. Hydrogen atoms were added with AmberTools[78], the complex was solvated in a cubic box of edge length 7.62 nm with 13,698 water molecules, and five Cl^- and one Na^+ counter ions were added. The two histidine residues of Mdm2 were protonated at the $\epsilon 2$ site. Simulations were performed with the ACEMD computer code[79] using the Langevin integrator using a damping constant $\gamma = 0.1$ ps^{-1}, constraints on the bonds that involve hydrogen atoms, and with heavy hydrogen atoms (four times the natural mass) to allow for an integration time step of 4 fs. Electrostatics were computed using Particle Mesh Ewald using a real-space cutoff of 0.9 nm.

Hamiltonian replica-exchange simulations. Since the relevant conformational states were a priori unknown, we avoided choosing structure-based collective variables but instead employed a so-called boost potential that was developed in the context of accelerated MD[80] and works by reducing the depth of the minima in the potential energy landscape. As the interaction of Mdm2 with PMI and other peptides is mostly hydrophobic[57, 81], the boost potential was applied to the Lennard-Jones interactions between the two chains and not the electrostatic interactions (see Supplementary Note 3.1 for simulation details). Six independent simulations starting from the crystallographic pose of about 1 μs length each were carried out with replica exchange[82] between 14 ensembles that interpolate between unbiased and strongly boosted potentials (Supplementary Note 3.1). The simulation took approximately 42×10^3 GPU hours.

Unbiased MD simulations. Short MD simulations of a total of 20 μs were used to explore the conformational space. From these simulations and all replicas of the replica-exchange simulations, starting conformations were uniformly sampled, generating various bound and unbound structures. In total, 502.597 μs of unbiased MD simulations were run. The initial structures were resolvated, energy minimized with 100 steps of conjugate-gradient descent, temperature equilibrated for 100 ps with harmonic constraints on protein and peptide atoms, followed by a 1 ns pressure equilibration with the Berendsen barostat. Finally the box size was set to the fixed cube with 7.62 nm edge length and an additional equilibration run of 1 ns was performed with active harmonic constraints. The production run generated 481 trajectories with varying lengths (between 945 and 1211 ns per trajectory). The simulation took approximately 115×10^3 GPU hours.

TRAMMBAR is a new estimator for MEMMs. Replica-exchange MD between different bias potentials can be extremely effective in exploring complex molecular state spaces[82]. Here we develop an extension of the bin-less TRAM method[49] to compute MEMMs in order to facilitate the integration of replica-exchange MD with unbiased MD. TRAM's ability to estimate unbiased kinetics relies on counting transitions between states within each simulation ensemble, but the contiguous simulation times between ensemble changes in a replica-exchange scheme are usually too short for that. We address the problem by splitting the data into two sets: (a) data from replica-exchange simulations for which we assume that it samples the equilibrium distributions of the respective ensembles and is thus analyzed with the MBAR framework[61, 62]; (b) data from unbiased MD simulations that are not long enough to sample the equilibrium distribution of the respective ensemble and are analyzed with bin-less TRAM. These two parts need to be coupled, and we call the resulting hybrid analysis method TRAMMBAR. Following Ref. [49], we denote the set of equilibrium samples (a) from ensemble k by X^k_{MBAR} and the set of time-correlated samples (b) from ensemble k by X^k_{TRAM}. We approximate the reference equilibrium distribution as a point-wise distribution on all data by maximizing the likelihood

$$L_{\mathrm{TRAMMBAR}} = L_{\mathrm{TRAM}} \cdot L_{\mathrm{MBAR}} \quad (1)$$

where L_{TRAM} is defined as in Ref. [49]

$$L_{\mathrm{TRAM}} = \prod_{k,i,j} \left(p^k_{ij}\right)^{c^k_{ij}} \prod_{\mathbf{x}\in X^k_{\mathrm{TRAM}}\cap S_i} \mu(\mathbf{x}) e^{f^k - b^k(\mathbf{x})} \quad (2)$$

and L_{MBAR} is the standard MBAR likelihood[61]

$$L_{\mathrm{MBAR}} = \prod_k \prod_{\mathbf{x}\in X^k_{\mathrm{MBAR}}} \mu(\mathbf{x}) e^{f^k - b^k(\mathbf{x})} \quad (3)$$

Here, $b^k(\mathbf{x})$ denotes the known unit-less bias energy of configuration $\mathbf{x}$ evaluated in the kth ensemble that can be obtained from the MD software. c^k_{ij} are the observed transition counts from the time-correlated data X^k_{TRAM}, and $e^{-f^k} := \sum_i e^{-f^k_i}$ are the ensemble free energies. The likelihood is optimized by varying: the unbiased configuration weights $\mu(\mathbf{x})$, the joint equilibrium probabilities $e^{-f^k_i}$ to be in Markov state S_i and ensemble k, and the transition probabilities p^k_{ij}, from which the kinetics at every ensemble can be computed. The TRAMMBAR algorithm is equivalent to the MBAR algorithm if X_{TRAM} is empty, and equivalent to the TRAM algorithm with empty X_{MBAR}. The algorithm for maximizing the above likelihood is described in Supplementary Notes 1.1 and 1.2.

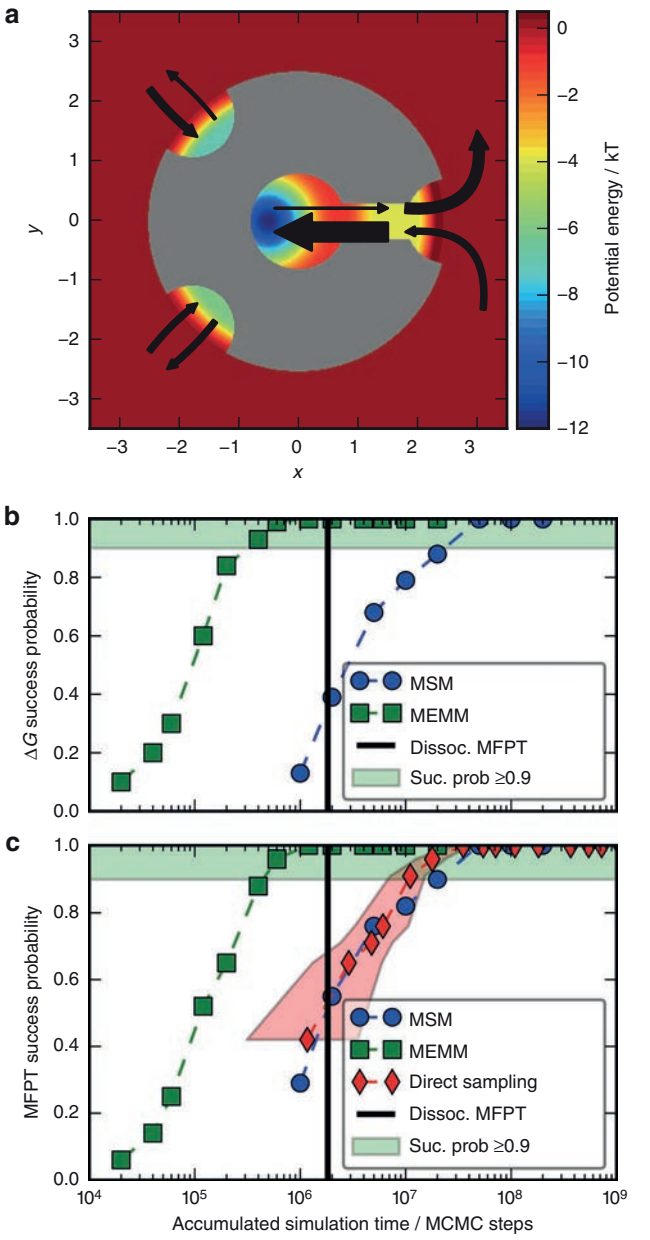

Fig. 4 Illustration of computing rare-event kinetics with TRAMMBAR using a model for protein ligand binding. **a** Potential energy surface and transition rates between five states (bound, pre-bound, two mis-bound states, dissociated). Arrow thickness is proportional to rate. **b** Probability of computing the binding free energy ΔG within $1k_BT$ accuracy of the exact value for a given amount of simulation data using MEMMs (TRAMMBAR estimator) or MSMs. The vertical bar indicates the mean-first-passage time (MFPT) for dissociation. **c** Probability of computing the dissociation rate within factor $\frac{1}{2}$ to 2 accuracy of the exact value

In order to illustrate our approach for computing rare-event kinetics for strong binders, consider the two-dimensional potential energy landscape in Fig. 4a. The gray shape represents a protein to which a small molecule ligand can bind. The protein has two shallow minima representing non-native binding sites on the surface, and a deep energy minimum representing an internal binding pocket at the end of a channel.

The mean dissociation time for this system is about 1.8×10^6 Monte Carlo steps (vertical bar in Fig. 4b, c). To approximate this time with direct simulation, multiple trajectories with lengths of at least 10^7 steps need to be launched from the bound state (Supplementary Fig. 2). Using MSMs, still a total of 10^7 steps of simulation time in shorter trajectories is needed to obtain accurate estimates of

both thermodynamics and kinetics (Fig. 4b, c). Using a MEMM with the TRAMMBAR estimator, we can get accurate estimates for both the binding free energy and the dissociation time with a total of only 5×10^5 steps that include short unbiased binding simulations and biased simulations on a flattened potential (Fig. 4b, c, Supplementary Note 2 and Supplementary Figs. 1 and 3). In contrast to most other enhanced sampling methods, a MEMM allows the computation of unbiased kinetics despite the fact that biases are used in the simulation. Moreover, MEMMs provide not only selected macroscopic rates, but full kinetics such as the whole set of transition rates shown in Fig. 4a.

Multi-ensemble Markov model for Mdm2–PMI. A MEMM was build from the MD and replica-exchange trajectories. To define MEMM states, we first chose the following set of features: all 1086 nearest-neighbor heavy atom distances between PMI residues and PMI residues (a) or Mdm2 residues (b) and the sine and cosine of the χ_1 dihedral angle of Mdm2 Tyr100 (c), which is a known "gate-keeper" residue for ligand association[83]. The time-lagged independent component analysis (TICA) algorithm[84] with a lag time of 10 ns was used to obtain 20 independent components containing the slow kinetics. To these, trajectories of the minimal distance between PMI and Mdm2 were added to facilitate a clear definition of the fully dissociated state. The resulting feature trajectories were clustered with k-means ($k = 1000$). In total, 56 microstates discretizing the dissociated state were defined based on the minimal heavy atom distance between PMI and Mdm2 and added to the set of the 1000 k-means clusters. The dissociated states had to be defined explicitly because of the low metastability of the dissociated state in the simulation box which prevents that the TICA algorithm finds a dimension that describes the full association/dissociation process of the binding partners (see Supplementary Fig. 10 for the influence of the definition of the dissociated states on the estimates of the binding free energy and of the residence time). Transition counts were computed for TRAMMBAR and for the MSM. For TRAMMBAR the initial 50 ns of the replica-exchange trajectories were discarded and the rest was subsampled, taking only one frame every 0.1 ns. We picked a lag time of 150 ns for TRAMMBAR based on the convergence of the implied time scales and mean-first-passage-times (Supplementary Notes 3.4 and 3.7, Supplementary Figs. 5 and 8). All analyses were done using PyEMMA[60] and MDTraj[85].

Experimental binding kinetics. The association and dissociation rate measurements were performed in stopped-flow and competition fluorescence anisotropy experiments. For the association measurements, FITC-PMI and Mdm2 were rapidly mixed using an SX20 stopped-flow spectrometer (Applied Photophysics). The temperature was maintained at 25 °C, and an excitation wavelength of 493 nm, in conjunction with a 515 nm long-pass filter was utilized. For the dissociation measurements, 10 nM FITC-PMI peptide was incubated with Mdm2, then excess of the unlabeled PMI (10 μM) was added and the dissociation was followed with a Multilabel 384-well plate reader (Tecan, Infinite M1000 PRO) with excitation at 494 nm and emission at 517 nm (see Supplementary Notes 4.1–4.3 and Supplementary Figs. 11 and 12 for details).

Code availability. TRAMMBAR has been implemented in the PyEMMA software. PyEMMA is available free of charge at http://pyemma.org.

Data availability. The molecular dynamics data that support the findings of this study are available in the Edmond Open Access Data Repository with the identifier doi:10.17617/3.x.[86]. All relevant data is available from the authors upon request.

Received: 5 May 2017 Accepted: 22 August 2017
Published online: 23 October 2017

References

1. Copeland, R. A. The drug-target residence time model: a 10-year retrospective. *Nat. Rev. Drug Discov.* **15**, 87–95 (2016).
2. Dahl, G. & Akerud, T. Pharmacokinetics and the drug-target residence time concept. *Drug. Discov. Today* **18**, 697–707 (2013).
3. Copeland, R. A., Pompliano, D. L. & Meek, T. D. Drug-target residence time and its implications for lead optimization. *Nat. Rev. Drug Discov.* **5**, 730–739 (2006).
4. Tummino, P. J. & Copeland, R. A. Residence time of receptor-ligand complexes and its effect on biological function. *Biochemistry* **47**, 5481–5492 (2008).
5. Plattner, N. & Noé, F. Protein conformational plasticity and complex ligand binding kinetics explored by atomistic simulations and Markov models. *Nat. Commun.* **6**, 7653 (2015).
6. Boehr, D. D., Nussinov, R. & Wright, P. E. The role of dynamic conformational ensembles in biomolecular recognition. *Nat. Chem. Biol.* **5**, 789–796 (2009).
7. Silva, D.-A., Bowman, G. R., Sosa-Peinado, A. & Huang, X. A role for both conformational selection and induced fit in ligand binding by the LAO protein. *PLoS Comput. Biol.* **7**, e1002054 (2011).
8. Henzler-Wildman, K. & Kern, D. Dynamic personalities of proteins. *Nature* **450**, 964–972 (2007).
9. Kiefhaber, T., Bachmann, A. & Jensen, K. S. Dynamics and mechanisms of coupled protein folding and binding reactions. *Curr. Opin. Struct. Biol.* **22**, 21–29 (2012).
10. Weikl, T. R. & Paul, F. Conformational selection in protein binding and function. *Protein Sci.* **23**, 1508–1518 (2014).
11. Noé, F. et al. Dynamical fingerprints for probing individual relaxation processes in biomolecular dynamics with simulations and kinetic experiments. *Proc. Natl Acad. Sci. USA* **108**, 4822–4827 (2011).
12. Clancy, C. E., Zhu, Z. I. & Rudy, Y. Pharmacogenetics and anti-arrhythmic drug therapy: a theoretical investigation. *Am. J. Physiol. Heart C* **292**, H66–H75 (2007).
13. Bennett, K. A. et al. Pharmacology and structure of isolated conformations of the adenosine A2A receptor define ligand efficacy. *Mol. Pharmacol.* **83**, 949–958 (2013).
14. Bowman, G. R. & Geissler, P. L. Equilibrium fluctuations of a single folded protein reveal a multitude of potential cryptic allosteric sites. *Proc. Natl Acad. Sci. USA* **109**, 11681–11686 (2012).
15. Bowman, G. R., Bolin, E. R., Harta, K. M., Maguire, B. & Marqusee, S. Discovery of multiple hidden allosteric sites by combining Markov state models and experiments. *Proc. Natl Acad. Sci. USA* **112**, 2734–2739 (2015).
16. Copeland, R. A. The dynamics of drug-target interactions: drug-target residence time and its impact on efficacy and safety. *Expert Opin. Drug Discov.* **5**, 305–310 (2010).
17. Fasting, C. et al. Multivalency as a chemical organization and action principle. *Angew. Chem. Int. Ed.* **51**, 10472–10498 (2012).
18. Rao, J., Lahiri, J., Weis, R. M. & Whitesides, G. M. Design, synthesis, and characterization of a high-affinity trivalent system derived from vancomycin and l-Lys-d-Ala-d-Ala. *J. Am. Chem. Soc.* **122**, 2698–2710 (2000).
19. Shaw, D. E. et al. Atomic-level characterization of the structural dynamics of proteins. *Science* **330**, 341–346 (2010).
20. Bowman, G. R., Voelz, V. A. & Pande, V. S. Atomistic folding simulations of the five-helix bundle protein λ_{6-85}. *J. Am. Chem. Soc.* **133**, 664–667 (2011).
21. Beauchamp, K. A., Ensign, D. L., Das, R. & Pande, V. S. Quantitative comparison of villin headpiece subdomain simulations and triplet–triplet energy transfer experiments. *Proc. Natl Acad. Sci. USA* **108**, 12734–12739 (2011).
22. Stanley, N., Esteban-Martin, S. & Fabritiis, G. D. Kinetic modulation of a disordered protein domain by phosphorylation. *Nat. Commun.* **5**, 5272 (2014).
23. Plattner, N., Doerr, S., De Fabritiis, G. & Noé, F. Complete protein-protein association kinetics in atomic detail revealed by molecular dynamics simulations and Markov modelling. *Nat. Chem.* http://dx.doi.org/10.1038/nchem.2785 (2017).
24. Dellago, C., Bolhuis, P. G., Csajka, F. S. & Chandler, D. Transition path sampling and the calculation of rate constants. *J. Chem. Phys.* **108**, 1964–1977 (1998).
25. Faradjian, A. K. & Elber, R. Computing time scales from reaction coordinates by milestoning. *J. Chem. Phys.* **120**, 10880–10889 (2004).
26. Du, W.-N., Marino, K. A. & Bolhuis, P. G. Multiple state transition interface sampling of alanine dipeptide in explicit solvent. *J. Chem. Phys.* **135**, 145102 (2011).
27. Votapka, L. W., Jagger, B. R., Heyneman, A. L. & Amaro, R. E. SEEKR: simulation enabled estimation of kinetic rates, a computational tool to estimate molecular kinetics and its application to trypsin–benzamidine binding. *J. Phys. Chem. B* **121**, 3597–3606 (2017).
28. Preto, J. & Clementi, C. Fast recovery of free energy landscapes via diffusion-map-directed molecular dynamics. *Phys. Chem. Chem. Phys.* **16**, 19181–19191 (2014).
29. Doerr, S., Harvey, M. J., Noé, F. & Fabritiis, G. D. HTMD: high-throughput molecular dynamics for molecular discovery. *J. Chem. Theory Comput.* **12**, 1845–1852 (2016).
30. Swope, W. C., Pitera, J. W. & Suits, F. Describing protein folding kinetics by molecular dynamics simulations: 1. Theory. *J. Phys. Chem. B* **108**, 6571–6581 (2004).
31. Noé, F., Horenko, I., Schütte, C. & Smith, J. C. Hierarchical analysis of conformational dynamics in biomolecules: transition networks of metastable states. *J. Chem. Phys.* **126**, 155102 (2007).
32. Buchete, N.-V. & Hummer, G. Coarse master equations for peptide folding dynamics. *J. Phys. Chem. B* **112**, 6057–6069 (2008).
33. Bowman, G. R., Pande, V. S. & Noé, F. (eds) *Advances in Experimental Medicine and Biology*, Vol. 797 (Springer, 2014).
34. Sarich, M., Noé, F. & Schütte, C. On the approximation quality of Markov state models. *Multiscale Model. Simul.* **8**, 1154–1177 (2010).
35. Prinz, J.-H. et al. Markov models of molecular kinetics: generation and validation. *J. Chem. Phys.* **134**, 174105 (2011).
36. Noé, F., Schütte, C., Vanden-Eijnden, E., Reich, L. & Weikl, T. R. Constructing the full ensemble of folding pathways from short off-

equilibrium simulations. *Proc. Natl Acad. Sci. USA* **106**, 19011–19016 (2009).
37. Doerr, S. & Fabritiis, G. D. On-the-fly learning and sampling of ligand binding by high-throughput molecular simulations. *J. Chem. Theory Comput.* **10**, 2064–2069 (2014).
38. Wu, H. et al. Variational Koopman models: slow collective variables and molecular kinetics from short off-equilibrium simulations. *J. Chem. Phys.* **146**, 154104 (2017).
39. Torrie, G. M. & Valleau, J. P. Nonphysical sampling distributions in Monte Carlo free-energy estimation: umbrella sampling. *J. Comp. Phys.* **23**, 187–199 (1977).
40. Grubmüller, H. Predicting slow structural transitions in macromolecular systems: conformational flooding. *Phys. Rev. E* **52**, 2893–2906 (1995).
41. Laio, A. & Parrinello, M. Escaping free energy minima. *Proc. Natl Acad. Sci. USA* **99**, 12562–12566 (2002).
42. Sugita, Y. & Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **314**, 141–151 (1999).
43. Berne, B. J., Borkovec, M. & Straub, J. E. Classical and modern methods in reaction rate theory. *J. Phys. Chem.* **92**, 3711–3725 (1988).
44. Voter, A. F. Hyperdynamics: accelerated molecular dynamics of infrequent events. *Phys. Rev. Lett.* **78**, 3908–3911 (1997).
45. Tiwary, P. & Parrinello, M. From metadynamics to dynamics. *Phys. Rev. Lett.* **111**, 230602 (2013).
46. Tiwary, P., Limongelli, V., Salvalaglio, M. & Parrinello, M. Kinetics of protein-ligand unbinding: predicting pathways, rates, and rate-limiting steps. *Proc. Natl Acad. Sci. USA* **112**, E386–E391 (2015).
47. Casasnovas, R., Limongelli, V., Tiwary, P., Carloni, P. & Parrinello, M. Unbinding kinetics of a p38 MAP kinase type II inhibitor from metadynamics simulations. *J. Am. Chem. Soc.* **139**, 4780–4788 (2017).
48. Tiwary, P., Mondal, J. & Berne, B. J. How and when does an anticancer drug leave its binding site? *Sci. Adv.* **3**, e1700014 (2017).
49. Wu, H., Paul, F., Wehmeyer, C. & Noé, F. Multiensemble Markov models of molecular thermodynamics and kinetics. *Proc. Natl Acad. Sci. USA* **113**, E3221–E3230 (2016).
50. Mey, A. S. J. S., Wu, H. & Noé, F. xTRAM: estimating equilibrium expectations from time-correlated simulation data at multiple thermodynamic states. *Phys. Rev. X* **4**, 041018 (2014).
51. Rosta, E. & Hummer, G. Free energies from dynamic weighted histogram analysis using unbiased Markov state model. *J. Chem. Theory Comput.* **11**, 276–285 (2015).
52. Stelzl, L. S. & Hummer, G. Kinetics from replica exchange molecular dynamics simulations. *J. Chem. Theory Comput.* **13**, 3927–3935 (2017).
53. Wu, H., Mey, A. S. J. S., Rosta, E. & Noé, F. Statistically optimal analysis of state-discretized trajectory data from multiple thermodynamic states. *J. Chem. Phys.* **141**, 214106 (2014).
54. Toledo, F. & Wahl, G. M. Regulating the p53 pathway: in vitro hypotheses, in vivo veritas. *Nat. Rev. Cancer* **6**, 909–923 (2006).
55. Vassilev, L. T. et al. In vivo activation of the p53 pathway by small-molecule antagonists of MDM2. *Science* **303**, 844–848 (2004).
56. Shangary, S. et al. Temporal activation of p53 by a specific MDM2 inhibitor is selectively toxic to tumors and leads to complete tumor growth inhibition. *Proc. Natl Acad. Sci. USA* **105**, 3933–3938 (2008).
57. Pazgier, M. et al. Structural basis for high-affinity peptide inhibition of p53 interactions with MDM2 and MDMX. *Proc. Natl Acad. Sci. USA* **106**, 4665–4670 (2009).
58. Zwier, M. C. et al. Efficient atomistic simulation of pathways and calculation of rate constants for a protein-peptide binding process: application to the MDM2 protein and an intrinsically disordered p53 peptide. *J. Phys. Chem. Lett.* **7**, 3440–3445 (2016).
59. Li, C. et al. Systematic mutational analysis of peptide inhibition of the p53-MDM2/MDMX interactions. *J. Mol. Biol.* **398**, 200–213 (2010).
60. Scherer, M. K. et al. PyEMMA 2: a software package for estimation, validation and analysis of Markov models. *J. Chem. Theory Comput.* **11**, 5525–5542 (2015).
61. Vardi, Y. Empirical distributions in selection bias models. *Ann. Stat.* **13**, 178–203 (1985).
62. Shirts, M. R. & Chodera, J. D. Statistically optimal analysis of samples from multiple equilibrium states. *J. Chem. Phys.* **129**, 124105 (2008).
63. Best, R. B., Zheng, W. & Mittal, J. Balanced protein-water interactions improve properties of disordered proteins and non-specific protein association. *J. Chem. Theory Comput.* **10**, 5113–5124 (2014).
64. Rauscher, S. et al. Structural ensembles of intrinsically disordered proteins depend strongly on force field: A comparison to experiment. *J. Chem. Theory Comput.* **11**, 5513–5524 (2015).
65. Matysiak, S. & Clementi, C. Optimal combination of theory and experiment for the characterization of the protein folding landscape of S6: how far can a minimalist model go? *J. Mol. Biol.* **343**, 235–248 (2004).
66. Zeller, F. & Zacharias, M. Efficient calculation of relative binding free energies by umbrella sampling perturbation. *J. Comput. Chem.* **35**, 2256–2262 (2014).
67. Schon, O., Friedler, A., Bycroft, M., Freund, S. M. V. & Fersht, A. R. Molecular mechanism of the interaction between MDM2 and p53. *J. Mol. Biol.* **323**, 491–501 (2002).
68. Kalbfleisch, J. D. & Lawless, J. F. The analysis of panel data under a Markov assumption. *J. Am. Stat. Assoc.* **80**, 863–871 (1985).
69. Crommelin, D. & Vanden-Eijnden, E. Data-based inference of generators for Markov jump processes using convex optimization. *Multiscale Model. Simul.* **7**, 1751–1778 (2009).
70. Metzner, P., Schütte, C. & Vanden-Eijnden, E. Transition path theory for Markov jump processes. *Multiscale Model. Simul.* **7**, 1192–1219 (2009).
71. ElSawy, K. M., Lane, D. P., Verma, C. S. & Caves, L. S. D. Recognition dynamics of p53 and MDM2: implications for peptide design. *J. Phys. Chem. B* **120**, 320–328 (2016).
72. Zimmerman, M. I. & Bowman, G. R. FAST conformational searches by balancing exploration/exploitation trade-offs. *J. Chem. Theory Comput.* **11**, 5747–5757 (2015).
73. Vitalini, F., Mey, A. S. J. S., Noé, F. & Keller, B. G. Dynamic properties of force fields. *J. Chem. Phys.* **142**, 084101 (2015).
74. Piana, S., Lindorff-Larsen, K. & Shaw, D. E. How robust are protein folding simulations with respect to force field parameterization? *Biophys. J.* **100**, L47–L49 (2011).
75. Schöneberg, J., Ullrich, A. & Noé, F. Simulation tools for particle-based reaction-diffusion dynamics in continuous space. *BMC Biophys.* **7**, 11 (2014).
76. Lindorff-Larsen, K. et al. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins* **78**, 1950–1958 (2010).
77. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926–935 (1983).
78. Case, D. *et al. AMBER 2015. Tech. Rep*. (University of California, San Francisco, 2015).
79. Harvey, M. J., Giupponi, G. & Fabritiis, G. D. ACEMD: accelerating biomolecular dynamics in the microsecond time scale. *J. Chem. Theory Comput.* **5**, 1632–1639 (2009).
80. Hamelberg, D., Mongan, J. & McCammon, J. A. Accelerated molecular dynamics: A promising and efficient simulation method for biomolecules. *J. Chem. Phys.* **120**, 11919–11929 (2004).
81. Kussie, P. H. et al. Structure of the MDM2 oncoprotein bound to the p53 tumor suppressor transactivation domain. *Science* **274**, 948–953 (1996).
82. Fukunishi, H., Watanabe, O. & Takada, S. On the Hamiltonian replica exchange method for efficient sampling of biomolecular systems: application to protein structure prediction. *J. Chem. Phys.* **116**, 9058–9067 (2002).
83. Dastidar, S. G., Lane, D. P. & Verma, C. S. Modulation of p53 binding to MDM2: computational studies reveal important roles of Tyr100. *BMC Bioinform.* **10**, 1–11 (2009).
84. Pérez-Hernández, G., Paul, F., Giorgino, T., De Fabritiis, G. & Noé, F. Identification of slow molecular order parameters for Markov model construction. *J. Chem. Phys.* **139**, 015102 (2013).
85. McGibbon, R. T. et al. MDTraj: a modern open library for the analysis of molecular dynamics trajectories. *Biophys. J.* **109**, 1528–1532 (2015).
86. Paul, F. Molecular dynamics trajectories for the 25–109-Mdm2-PMI system. *Edmond Open Access Data Repository* doi:10.17617/3.x (2017).

Acknowledgements

We thank Matt Harvey (Acellera Ltd.) and Markus Rampp and Ingeborg Weidl (Max Planck Computing and Data Facility) for technical support and the Max Planck Society for usage of the Hydra supercomputer. We thank the group of Prof. Tad Holak in the Jagiellonian University, Poland for assistance with the Mdm2 purification and refolding protocols. We thank Prof. Sebastian Springer in Jacobs University Bremen for the usage of plate reader (Tecan, Infinite M1000 PRO). We thank Dr. S.L. Shammas, who kindly provided the script used to analyze the association kinetics data. We are grateful to Kresten Lindorff-Larsen, Vincent A. Voelz, John D. Chodera as well as all members of the Computational Molecular Biology group for insightful discussions. Funding is acknowledged by European Commission (ERC StG "pcCells" to F.N.), Deutsche Forschungsgemeinschaft (SFB 1114/C3, SFB 740/D7, and TRR 186/A12 to F.N. and SFB 1114/A4 to F.N. and T.R.W.). J.C. is a Wellcome Trust Senior Research Fellow (WT 095195MA). J.S. is a Marie Skłodowska-Curie Internationally outgoing fellow. M.D.C. is supported by a Biotechnology and Biological Sciences Research Council (BBSRC) studentship.

Author contributions

F.P., C.W., E.T.A., T.R.W. and F.N. have designed research. F.P. and H.W. have developed methods. F.P. and C.W. have developed/implemented software. F.P. and C.W. have conducted simulations. F.P. and C.W. have analyzed simulations. E.T.A., C.F. and J.C.

have designed experiments, E.T.A., M.D.C. and J.S. have conducted experiments. F.P., C.W., E.T.A., and F.N. have written the paper.

Additional information

Supplementary Information accompanies this paper at doi:10.1038/s41467-017-01163-6.

Competing interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.