UNIVERSITY OF CALIFORNIA SAN DIEGO

Multiscale Simulation Approaches for Predicting Protein-Ligand Binding Kinetics

A dissertation submitted in partial satisfaction of the requirements for the degree

Doctor of Philosophy

in

Chemistry

by

Benjamin Robert Jagger

Committee in Charge:

Professor Rommie E. Amaro, Chair
Professor J. Andrew McCammon, Co-Chair
Professor Michael Galperin
Professor Michael Gilson
Professor Simpson Joseph
Professor Andrew McCulloch

2020

The Dissertation of Benjamin Robert Jagger is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____
Co-Chair

_____
Chair

University of California San Diego

2020

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

To my advisors, Rommie Amaro and Andy McCammon: Thank you for sharing your passion and commitment to science with me. I am immensely grateful for you support and mentorship.

To my committee, Michael Galperin, Michael Gilson, Simpson Joseph, and Andrew McCulloch: Thank you for your thoughtful advice and careful guidance throughout my study.

To my collaborators and colleagues: Thank you for many insightful discussions and your brilliant and creative ideas that have pushed our projects forward. It was a pleasure working with each of you. Christopher Lee, thank you for you mentorship, friendship, and for always pushing me to achieve more. You have impacted my career in so many ways, for which I am so thankful. Lane Votapka, thank you for your support as I began my graduate career. Andy Stokley and Anupam Ojha, it has been a pleasure working with both of you and seeing your excitement and passion for science. Thank you for your contributions and dedication to our projects.

To the Amaro and McCammon lab members: I am forever grateful for the positive and caring environment created by each of you.

To Patti Craft and Teri Simas: You do an amazing job supporting us both scientifically and personally. Thank you for all you have done and continue to do.

To my friends: You bring such joy into my life. It has been a pleasure getting to know all of you. Thank you for all of our wonderful memories.

To my fencing family: Thank you for the support, friendship, and happiness your provided outside of my graduate studies.

To Sarah Kochanek: Thank you for you unwavering support, for having confidence in me even when I don't have it in myself, and for helping me to grow and achieve my best both professionally and personally.

Finally, to my family: Mom and Dad, thank you for giving me the amazing foundation and skills to become the person I am today. Your constant love, support, and belief in me is so special and I am so grateful. To my larger, extended family: you have each played an important role in shaping me and helping me to grow, thank you.

Chapter 1, in full, is a reprint of the material as it appears in: "Jagger, B. R. †; Kochanek, S. E. †; Haldar, S.; Amaro, R. E.; Mulholland, A. J. Multiscale Simulation Approaches to Modeling Drug–Protein Binding. *Curr. Opin. Struct. Biol.* 2020, *61*, 213–221." The dissertation author was a primary coinvestigator and author of this work.

Chapter 2 , in full, is a reprint of the material as it appears in: "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* 2017, *121* (15), 3597–3606." The dissertation author was a primary coinvestigator and author of this work.

Chapter 3 , in full, is a reprint of the material as it appears in: "Jagger, B. R.; Lee, C. T.; Amaro, R. E. Quantitative Ranking of Ligand Binding Kinetics with a Multiscale Milestoning Simulation Approach. *J. Phys. Chem. Lett.* 2018, *9* (17), 4941–4948." The dissertation author was a primary investigator and author of this work.

Chapter 4, in full, has been submitted for publication and is presented as it may appear in: "Jagger, B. R.; Ojha, A. A.; Amaro, R. E. Predicting Ligand Binding Kinetics Using a Markovian

Milestoning with Voronoi Tessellation Multiscale Approach. *J. Chem. Theory Comput. Submitted.*

The dissertation author was a primary investigator and author of this work.

VITA

2011 - 2015	Undergraduate Research Fellow Wheeler Lab, Duquesne University
2015	Bachelor of Science in Chemistry, Minor in Mathematics Duquesne University
2015 - 2020	Graduate Research Fellow Amaro and McCammon Labs, University of California San Diego
2020	Doctor of Philosophy in Chemistry University of California San Diego

PUBLICATIONS

Jagger, B. R.; Koval, A. M.; Wheeler, R. A. Distinguishing Protonation States of Histidine Ligands to the Oxidized Rieske Iron-Sulfur Cluster through 15 N Vibrational Frequency Shifts. *ChemPhysChem* **2016**, *17* (2), 216–220. https://doi.org/10.1002/cphc.201500838.

Koval, A. M.; **Jagger, B. R**.; Wheeler, R. A. Distinguishing the Protonation State of the Histidine Ligand to the Oxidized Iron–Sulfur Cluster from the MitoNEET Family of Proteins. *ChemPhysChem* **2017**, *18* (1). https://doi.org/10.1002/cphc.201600957.

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* **2017**, *121* (15), 3597–3606. https://doi.org/10.1021/acs.jpcb.6b09388.

Jagger, B. R.; Lee, C. T.; Amaro, R. E. Quantitative Ranking of Ligand Binding Kinetics with a Multiscale Milestoning Simulation Approach. *J. Phys. Chem. Lett.* **2018**, *9* (17), 4941–4948. https://doi.org/10.1021/acs.jpclett.8b02047.

Jagger, B. R.[†]; Kochanek, S. E.[†]; Haldar, S.; Amaro, R. E.; Mulholland, A. J. Multiscale Simulation Approaches to Modeling Drug–Protein Binding. *Curr. Opin. Struct. Biol.* **2020**, *61*, 213–221. https://doi.org/10.1016/j.sbi.2020.01.014.

Ahn, S; **Jagger, B. R.**; Amaro, R. E. Ranking of Ligand Binding Kinetics Using a Weighted Ensemble Approach and Comparison with a Multiscale Milestoning Approach. *Submitted*

Jagger, B. R.; Ojha, A. A.; Amaro, R. E. Predicting Ligand Binding Kinetics Using a Markovian Milestoning with Voronoi Tessellation Multiscale Approach. *J. Chem. Theory Comput. Submitted*

ABSTRACT OF THE DISSERTATION

Multiscale Simulation Approaches for Predicting Protein-Ligand Binding Kinetics

by

Benjamin Robert Jagger

Doctor of Philosophy in Chemistry

University of California San Diego, 2020

Professor Rommie E. Amaro, Chair
Professor J. Andrew McCammon, Co-Chair

A detailed understanding of the interaction between a drug candidate molecule and its target is essential for the development, optimization, and efficacy prediction of a drug. Kinetic parameters such as the association rate and residence time of a molecule have been shown to better correlate with *in vivo* efficacy than more commonly used thermodynamic parameters. Efficient and accurate computational predictions of these quantities are therefore of great interest for their potential to inform and improve the development of novel pharmaceuticals. In this dissertation, I present the development and application of a multiscale molecular simulation approach which

combines molecular dynamics and Brownian dynamics simulations with the theory of milestoning to efficiently calculate protein-ligand binding and unbinding rates. I begin with an overview of many of the existing multiscale simulation approaches for studying drug-protein binding. Then I present the methodology we have developed, Simulation Enabled Estimation of Kinetic Rates (SEEKR), and demonstrate its effectiveness for predicting the association and dissociation rates of the inhibitor, benzamidine, to the trypsin protein; a common model system. I then present the effectiveness of our multiscale milestoning approach for rank-ordering a series of chemically diverse ligands to the model system β-cyclodextrin. This study includes a direct comparison of both efficiency and accuracy to long timescale molecular dynamics simulations and also outlines best practices for the use of our approach and the assessment of sampling convergence. Finally, I present the implementation of a new milestoning algorithm, Markovian Milestoning with Voronoi Tesselations, in our multiscale methodology to significantly decrease the simulation cost of kinetics calculations, improve the assessment of sampling convergence, and provide a framework for the future development of additional capabilities with the SEEKR method. This study also includes the development and deployment of our toolkit along with documentation and tutorials to facilitate its use and continued improvement by the scientific community.

Chapter 1

Multiscale Simulation Approaches to Modeling Drug-Protein Binding

1.1 Abstract

Simulations can provide detailed insight into the molecular processes involved in drug action, such as protein-ligand binding, and can therefore be a valuable tool for drug design and development. Processes with a large range of length and timescales may be involved, and understanding these different scales typically requires different types of simulation methodology. Ideally, simulations should be able to connect across scales, to analyze and predict how changes at one scale can influence another. Multiscale simulation methods, which combine different levels of treatment, are an emerging frontier with great potential in this area. Here we review multiscale frameworks of various types, and selected applications to biomolecular systems with a focus on drug-ligand binding.

1.2 Introduction

Protein-ligand interactions are integral to coordinating the complex functions of cellular activity. Such interactions include the binding of signaling molecules, enzyme substrates, toxins,

regulating factors, or other proteins to the protein of interest. Of particular interest for pharmaceutical development is the binding of drug molecules that mimic, inhibit, or modulate native protein-ligand interactions for therapeutic effect. Molecular simulations are increasingly involved in drug discovery pipelines in understanding protein-drug binding interactions, and also have the potential to reduce cost and time of drug discovery associated with synthesizing and experimentally testing many compounds.[1] Simulations can be used as screens during the hit identification phase[2], provide insight for lead optimization[3], and aid in analyzing drug resistance.[4] A particular focus is understanding and predicting drug binding and kinetics:[5,6] increasingly it is clear that the biological activity of many drugs depends on the rates of association or dissociation from their targets, rather than their binding affinity.[5–7] Developments in computer architecture, such as GPUs, and the promise of exascale computing power, are transforming the range and scope of biomolecular simulations.[8] Simulations can reveal molecular mechanisms and analyze them in a level of detail and dynamic resolution beyond the reach of experiment.

Simulations face conflicting challenges in this arena, with a tension between the need to address long timescales and large spatial scales for some relevant processes (large-scale conformational changes, macromolecular association, and beyond to changes in organelles and cells) and the requirement for accurate description of molecular interactions and reactions for reliable predictions.[9] Methods exist to simulate biomolecules at different length scales, ranging from quantum mechanical electronic structure calculations to atomistic, coarse-grained, mesoscale and continuum models, each with well-established domains of applicability, and can provide useful predictions of biologically relevant properties when applied with simulation techniques to sample underlying structure, organization and dynamics.[10] Furthermore, many enhanced sampling simulation techniques exist to more efficiently access the wide range of timescales relevant for

drug action. No one methodology is capable of completely and accurately describing the broad range of time- and length scales of the protein-drug binding process and the molecular changes that result from binding of a drug to its target. Therefore, there is great potential impact from the combination of different types of methods, e.g. for predicting the higher-level effects of changes at the molecular level by connecting across scales and understanding the mechanisms responsible. There is further promise in leveraging and connecting to the increasing wealth of experimental, genetic, and other biological data and ultimately informing future experiments to push forward drug development campaigns.

Here, we review emerging multiscale methods for studying protein-ligand binding relevant to drug design and the development of small molecule therapeutics. Multiscale techniques bridge spatial and/or temporal scales, coupling together two or more different types of modelling approaches, with varying degrees of 'tightness' of coupling (Figure 1.1). Two or more different levels of representation may coexist within a single simulation, a switch between independent levels can be triggered when a critical configuration or milestone is reached, or sampling at a lower, computationally cheaper level can enhance and/or be informed by calculations at a higher level, with information being directly passed between different types of simulation. We also include in this review a brief description of enhanced sampling techniques, primarily focused on the atomistic scale, as these methodologies are essential for studying the longer temporal scales on which drug binding occurs and offer immense potential through combination with other multiscale techniques designed to access multiple length scales. The simulation methods we highlight are ordered by increasing scale (time and length) which, generally, is inversely correlated with computational cost.

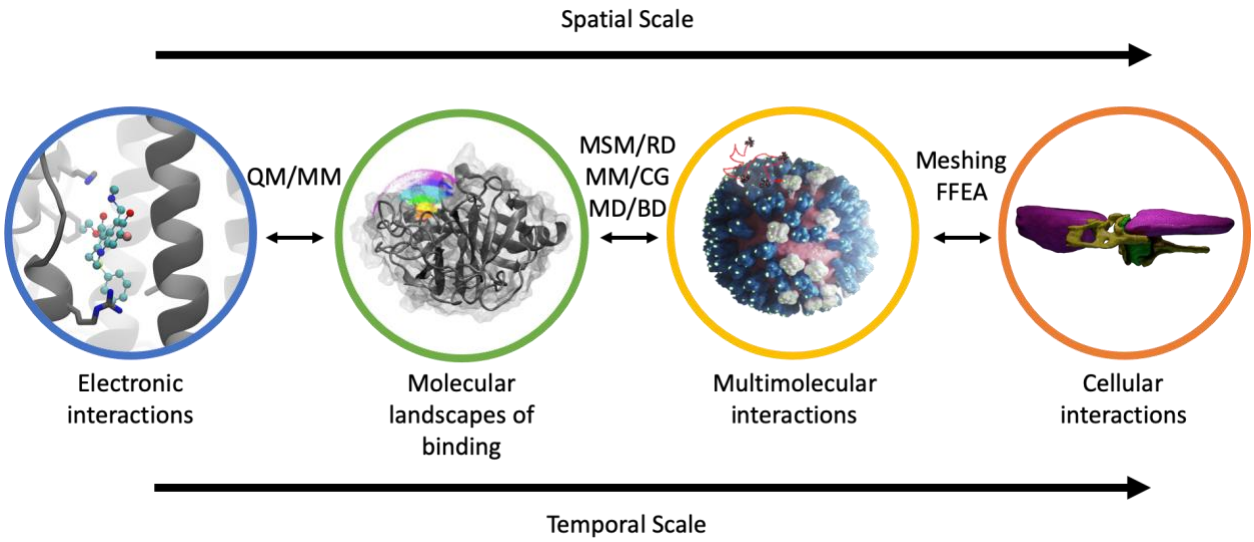


Figure 1.1. Depiction of the range of scales relevant for drug binding and action, from atomic to cellular scales. Multiscale simulation techniques bridge two or more of these scales by combining different molecular representations or simulation modalities with varying degrees of coupling. Through multiscale approaches, one can potentially obtain information and make predictions at larger scale methods without losing the detail associated with smaller scale techniques.

1.3 Combining QM methods with atomistic and coarse-grained representations

Multiscale approaches can provide a route to combine the accuracy and detail of high-level methods with the ability to model large systems and perform significant sampling. Quantum mechanics/molecular mechanics (QM/MM) methods are a paradigm of multiscale molecular modelling. They combine an electronic structure description of a small region with a simpler empirical MM (usually atomistic) representation of the surroundings (e.g. protein, solvent). The QM treatment can provide an accurate description of the electrostatics and polarization of the high-level region, and model chemical reactivity, e.g. to study covalent reactivity *in situ*.[4] Inclusion of environmental effects may generally be crucial for reliable prediction of reactivity of drug-like molecules, for which ligand-only prediction approaches may fail.[11] QM/MM methods have important relevant applications beyond chemical reactivity in investigations of drug binding. As a ligand binds to a protein, its environment changes significantly, e.g. from fully solvated to buried within the protein. This may cause significant changes in the polarization of the ligand preferably

when a charged residue is close to ligand, affecting its interactions, but the effects of such changes are not included in invariant charge MM models. The effects of electronic polarization changes on binding kinetics can be investigated by combining QM/MM free energy calculations with enhanced sampling simulations of binding (Figure 1.2). An example is a study of the anticancer drug imatinib binding to c-Src kinase,[12] which combined metadynamics simulation of binding with QM/MM free energy corrections at critical points along the (un)binding pathway. In this approach, the free energy change for changing from a MM to a QM treatment of the ligand is calculated, for the bound complex, for the transition state (TS), and for the unbound ligand in solution, by replica exchange Monte Carlo simulations using a Metropolis-Hastings-Warshel algorithm (Figure 2).[13] The results show that there is a significant difference between these environments in the free energy for the change from a MM to a QM treatment of the ligand. This indicates that that polarization (and therefore interactions) of the ligand are significantly different in different environments, and changes as the drug binds to the protein. Inclusion of electronic polarization in this way has the effect of increasing the off rate, bringing it closer to the experimental value.

QM/MM methods allow chemical reactions to be investigated in proteins and are now widely applied in modelling enzyme-catalyzed reactions, and increasingly in other relevant areas such as covalent inhibition and prediction of drug metabolism.[14] In modelling drug metabolism, they have been combined with coarse-grained and atomistic molecular dynamics (MD) simulations:[9] coarse-grained MD is used to generate a model of the membrane-bound enzyme investigate association with a drug (e.g. warfarin) in the membrane; coarse-grained models are converted into atomistic MD simulations to investigate drug binding within the enzyme; and these atomistic simulations are used to generate QM/MM models to investigate reaction in the active site . The simulations showed important effects of the membrane, e.g. on the channels controlling

access to the active site and gating residues. This multiscale coarse-grained-atomistic-QM/MM protocol is applicable to other membrane-bound enzymes. QM/MM methods with impressive scalability are also being developed.[15] QM/MM methods can also be combined with coarse-grained representations, in triple resolution models directly containing QM, MM and CG regions, to model large systems.[16] It should always be remembered that approximate QM methods suffer from limitations (e.g. some density functionals may fail to model some types of reactivity correctly);[9–11] these can be overcome by multiscale methods that embed a high-level ab initio QM treatment within a larger region treated by density functional theory, effectively partitioning the QM region within a QM/MM framework. This embedding approach removes uncertainty due to the functional and allows calculation of reaction barriers with high accuracy.[17]

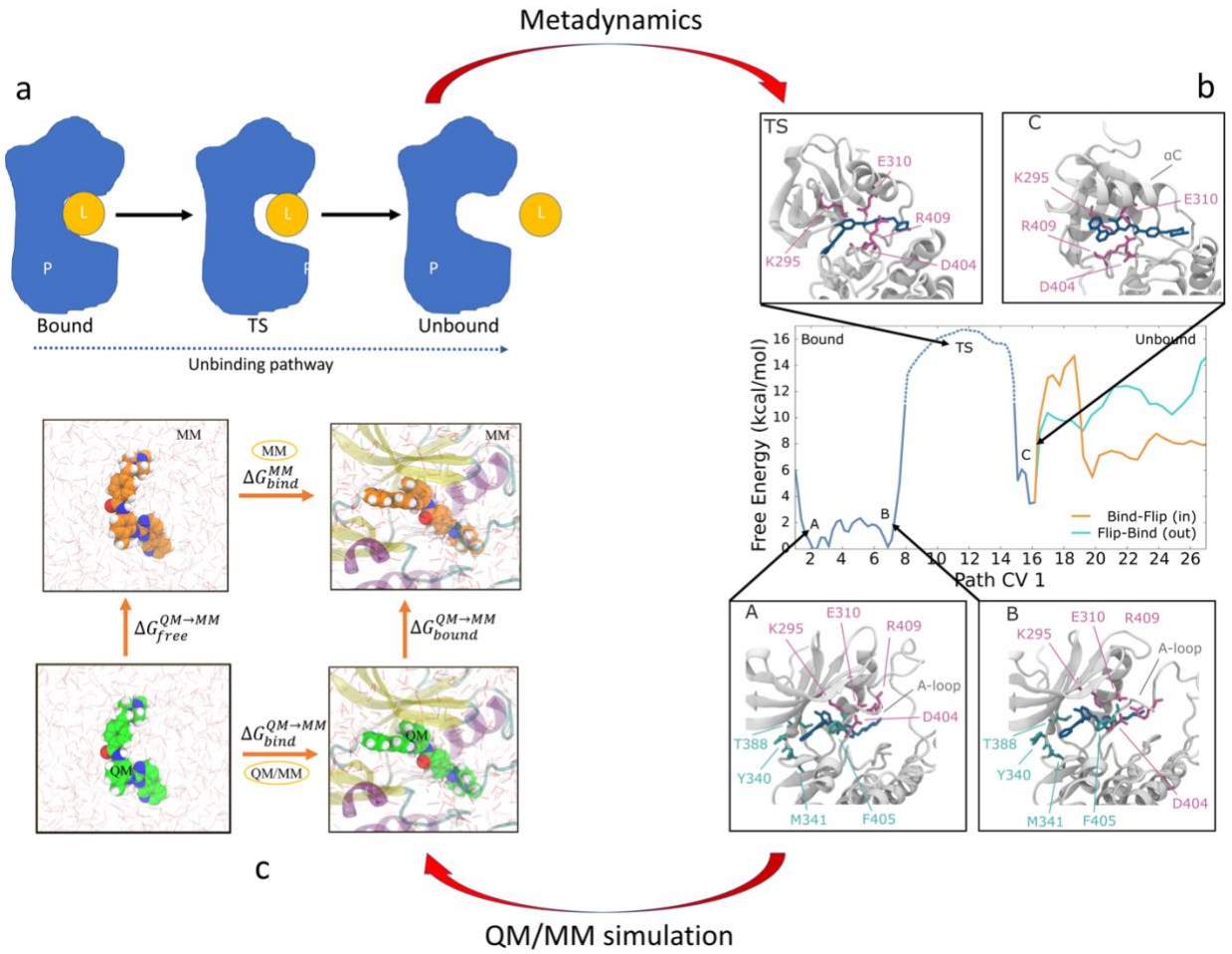


Figure 1.2 A practical multiscale simulation approach to modeling drug-protein binding kinetics combing atomistic metadynamics simulations and QM/MM free energy calculations.
a.) A schematic representation of the unbinding pathway of the ligand from its protein target showing the most important states along the (un)binding pathway: the bound protein-ligand complex, transition state (TS) for binding and the unbound state (with the ligand free in solution). **b.)** Shows the one-dimensional free energy profile for binding of the cancer drug imatinib binding to c-Src kinase: this was calculated using parallel tempering metadynamics with path collective variables. **c.)** Representative structures of the imatinib-src bound complex (A), TS and encounter complex (C). **d.)** QM/MM corrections to the profile are calculated from the free energy of changing from a MM to a QM representation of the ligand for each state. The free energy cycle shows this for the bound and unbound states. $\Delta G_{free}^{QM \to MM}$ and $G_{bound}^{QM \to MM}$ are the QM/MM correction free energy for the bound and unbound states, respectively. The thermodynamic cycle shown provides a QM/MM estimate of the binding affinity, by correcting the binding free energy calculated at the MM level.

1.4 Accessing drug-relevant timescales

Enhanced sampling simulations (usually based on atomistic MD) are increasingly used to study and predict drug binding kinetics.[1,5–7,10,12] There is potential to apply enhanced sampling methods in multiscale frameworks to extend their scope.

Atomistic MD simulations can provide good descriptions of biomolecular interactions, and with free energy approaches can analyze determinants of binding affinity. However, even with supercomputer resources, timescales are limited to the nanosecond-microsecond regime, so it is in general difficult to simulate multiple binding events unless a biasing or enhanced sampling method is applied to accelerate the process or to focus sampling on a desired region of phase space. Many such methods exist and can be applied with different potential functions (atomistic, coarse-grained, QM/MM, etc). An example of an enhanced sampling simulations approach is the calculation of residence times τ ($\tau = 1/k_{off}$) using τ-random acceleration molecular dynamics (τ-RAMD) for a diverse set of inhibitors of an important cancer target, the human N-terminal domain of heat shock protein 90α (N-HSP90). The τ-RAMD method relies on generating a random force which allows exit of the ligands within a short simulation time. τ-RAMD gives an excellent correlation between computed residence time (τ_{comp}) and measured τ_{expt} values for 78% of the compounds.[18] Other bias-based MD approaches have also been successful for predicting drug-target residence time. [19,20]

Metadynamics (MetaD) simulations[21] of various types are being increasingly widely used in drug discovery, e.g. for prediction of binding kinetics, exploration of ligand binding or unbinding pathways,[13,22,23] or analysis of conformational behavior e.g. relating to drug resistance,[4] and can be applied in multiscale frameworks and/or with multiscale potentials. MetaD sampling techniques rely on choosing appropriate collective variables (CV) to describe the slow degrees of freedom of interest. It is a nontrivial task to choose or identify effective and representative CVs to

describe a process effectively, often involving trial and error. Several recent developments address this issue.[24] Bernetti *et al.* showed that with a Path Collective Variables (PCVs) description coupled with metadynamics can be combined with Markov state models (MSM).[25] They applied this integrated method to model the binding of alprenolol to the β2-adrenergic receptor, providing as estimate of the binding free energy and identifying the minimum free energy path for formation of the protein-ligand complex. McCarty and Parrinello combined well-tempered metadynamics and time-lagged independent component analysis to obtain efficient collective variables.[26] Brotzakis *et al.* combined this variational approach to conformational dynamics with funnel metadynamics to calculate the absolute protein-ligand binding free energy and study energetic and structural details of benzamidine binding to trypsin at relatively low computational cost.[27]

Alternative approaches to simulating molecular association and other slow processes, such as MSM[28] and milestoning[29,30] rely on statistical reconstruction of simulation data from many short, independent simulations. MSMs can describe small-molecule binding kinetics in good agreement with experiment[31] and have also been used to elucidate the effects of protein dynamics for drug binding.[32] The transition-based reweighting analysis method allows integration of unbiased and biased MD simulations, estimating a multiensemble Markov model, combining the advantages of MSMs and rare-event sampling simulations.[33] Experimental data can also be integrated into augmented Markov models[34] There are also multiple examples using milestoning with MD simulations to study ligand binding.[35,36] Weighted ensemble approaches also show great promise.[37]

Such approaches are powerful, but the most important limitation is generally the high computational costs of sampling of atomistic MD simulations of protein-drug binding. With sufficient sampling, the accuracy of predictions may be limited by the potential functions used:[38] typical atomistic MM forcefields do not allow for changes in electronic polarization and are limited

in their description of electrostatics and dispersion effects, so may not describe binding interactions correctly.[39] Coupling to simulations with more physically correct models (e.g. QM electronic structure methods) in multiscale frameworks potentially allows for such limitations to be tested and corrected (Figure 1.2).[12,13] Connecting higher level simulations with more approximate models potentially allows on-the-fly (and/or machine learned) refinement of the approximate models, which will help provide improved potential functions for specific systems. Enhanced sampling methods such as metadynamics can also be used to generate reactive conformations for subsequent QM/MM modelling of reactivity in investigations of targeted covalent inhibitors.[4]

1.5 Combining molecular level descriptions

Drug-protein binding typically occurs in two main steps: first, non-specific and long-range interactions, such as electrostatics, drive initial association to the binding site region, subsequently, more specific short-range interactions (covalent bonds, hydrogen bonds, salt bridges, etc.) within the binding site determine the final binding pose (where specificity of interactions is key). This immediately suggests that the whole binding and recognition process can be simulated by a combination of different methods to describe these two aspects, and indeed this has inspired multiscale approaches. From Smoluchowski theory, the maximum diffusion-limited rate coefficient of two molecules approximated as spheres of similar size in aqueous solution is on the order of 10^9-10^{10} $M^{-1}s^{-1}$.[40] However, as a result of molecule size, molecular and hydrodynamic interactions, crowding, geometric constraints of binding sites, gating effects, etc. the observed rate coefficients of ligand binding occur over a much broader range, 10^3 to 10^{10} $M^{-1}s^{-1}$.[40–42] Brownian dynamics (BD) simulations solve the Langevin equation in the overdamped limit, and often achieve decreased computational cost by neglecting internal degrees of freedom and describing solvent implicitly with a dielectric constant and viscosity term.[43] As such, trajectory-based BD

simulations are well-suited for studying the long-range interactions that dominate ligand association, particularly the electrostatic steering involved in forming the initial ligand-protein encounter complex. BD simulations were combined with all-atom MD simulations in one of the earliest multiscale approaches to protein-ligand binding.[44] Chang *et al.* presented a multiscale approach to model binding pathways of ligands to HIV-1 protease that involved initial coarse-grained BD simulations followed by all-atom MD simulations initiated from snapshots of the BD trajectories.[45] The low cost of the coarse-grained BD simulations allows for extensive sampling of multiple association pathways. The pathway data generated from these simulations served as the starting point for multiple follow-up studies, providing detailed descriptions of the drug binding process.[46,47] Another approach to model ligand binding combines a molecular mechanics description of the ligand and water molecules with a coarse-grained Gō model description of the protein.[48] This approach was applied to multiple G-protein coupled receptors and successfully predicted important residues for ligand binding, as well as binding poses.[49]

Milestoning theory can also be leveraged to combine MD and BD simulations; balancing atomistic detail and accuracy with efficiency. In particular, Simulation Enabled Estimation of Kinetic Rates (SEEKR) is a multiscale milestoning simulation technique that directly combines MD and BD simulations to calculate k_{on}, k_{off}, and the free energy of protein-ligand binding, with a focus on small molecule drugs.[50–52] Atomistic MD simulations are used in regions close to the binding site, where molecular flexibility and atomic-level detail are essential. Rigid-body BD simulations are used in the regions farther away from the binding site, where molecular flexibility is less important (Figure 1.3). The use of BD results in dramatic computational savings compared to purely MD simulations, as millions of individual millisecond-to-second trajectories can be generated overnight on a standard desktop computer, in contrast to MD simulations which require

multiple weeks and supercomputers, GPUs or other specialized hardware to produce a single trajectory on the order of 1 microsecond. SEEKR further reduces the compute time required for calculations via an enhancement in sampling of rare events due to statistical bootstrapping, and it is 'embarrassingly' parallel, as each independent milestone can be simulated concurrently. SEEKR has been shown to effectively rank-order ligands by both k_{off} and binding free energy for the biosynthetic receptor, β-cyclodextrin.[52] SEEKR has also been employed to calculate k_{on}, k_{off}, and the binding free energy for the well-studied model protein system trypsin with the noncovalent binder, benzamidine.[51] In a similar study, Zeller et al. combined MD and BD simulations to obtain association rates and pathways for two clinically relevant inhibitors of the influenza neuraminidase enzyme.[53] Using an MD/BD multiscale approach resulted in a reduction in computational cost while still retaining detailed information about the association pathway, such as intermediate states, that can be useful to inform inhibitor design. The binding and unbinding rates calculated with these methods can be used as parameters for larger-scale phenomenological and diffusion-based models, adding a further multiscale dimension.

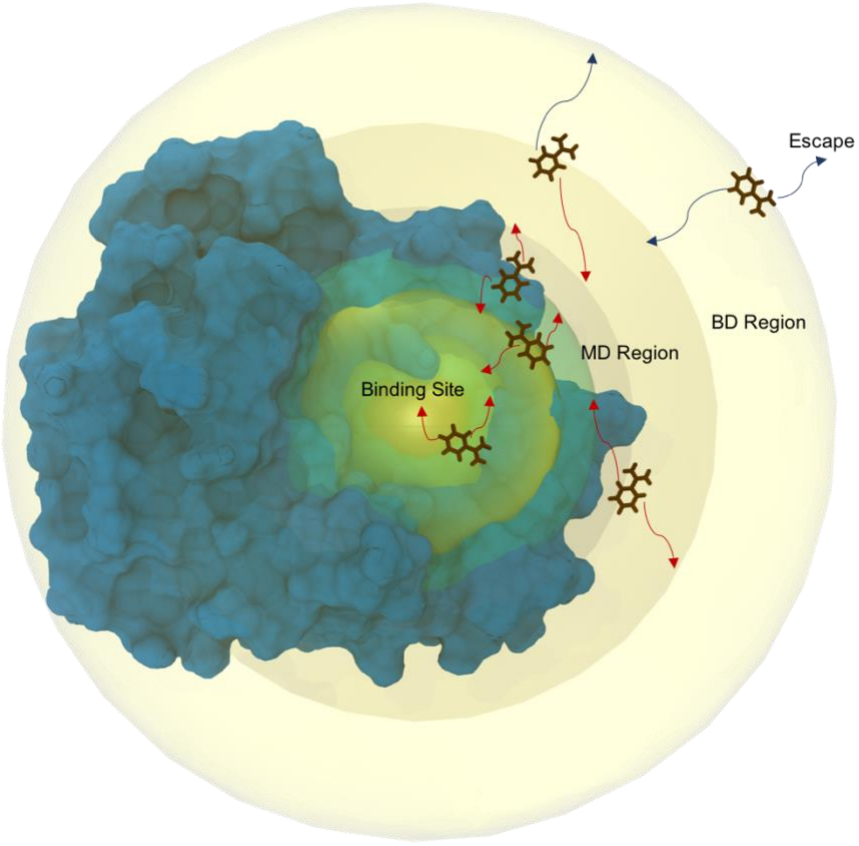


Figure 1.3 SEEKR is designed for calculations of ligand receptor binding and unbinding kinetics in a multiscale framework using molecular dynamics and Brownian dynamics simulations.
Regions closest to the binding site are simulated with atomistic MD and the regions furthest away is simulated using rigid body BD. Ligands are placed on each spherical milestone and only simulated until an adjacent milestone is touched. Arrows in red represent MD trajectories and blue arrows represent BD trajectories. Statistics from each of the independent simulations are combined to estimate association and dissociation rates, as well as binding affinity.

1.6 Linking molecular detail to sub-cellular and beyond

The methods discussed so far can model biomolecular (drug and protein) systems; however, the size and timescales that these methods can address prevent them from being able to directly address the complexities (e.g., off-target effects, cooperativity, metabolism, etc.) of drug binding in subcellular, cellular, and larger environments. Such increased system complexity typically necessitates a reduction in the level of detail with which each component can be described in a model. This can be accomplished using large-scale particle- or continuum-based methods.

Multiscale techniques capable of leveraging data obtained from more detailed (e.g., atomistic, coarse-grained, etc.) approaches and incorporating it such large-scale methods show great promise. One such approach is particle-based reaction-diffusion (RD) simulations. By combining MD-based MSMs with RD simulations in a technique called MSM/RD, processes such as drug-protein binding can be modeled at large time and length scales, while conserving atomistic details.[28] More specifically, MSM/RD can be used to model intracellular dynamics and describe diffusion, association, and dissociation on the cellular scale. Dibak *et al*. demonstrate the utility of MSM/RD approaches for biomolecular systems by application to carbon monoxide diffusion into the heme cavity of myoglobin.[28] Extensions of this methodology that incorporate more complex cellular environments have the potential to become a powerful tool for studying off-target effects of drug molecules. The MSM/RD framework is also highly generalizable, with the potential to be incorporated into many of the existing powerful RD tools.[54–58]

At larger length and time scales, continuum approaches that utilize partial differential equations to model diffusing substrates can be employed. Here, recent developments in computational geometric meshing algorithms and software have facilitated the incorporation of high-resolution reconstructions of experimentally realistic cellular and organelle geometries for biophysical simulation.[59] Such meshing software allows the seamless connection of particle-based (RD, BD, MD, QM methods) with continuum approaches that utilize numerical techniques such as finite element methods in realistic (not highly idealized) geometries. This methodology has recently been used to model calcium dynamics in realistic geometries of a dendritic spine[59] and a cardiac calcium release unit,[60] demonstrating a cohesive workflow for mesh generation and refinement. Using a different approach, cardiac thin filament activation has been modeled using a combination of coarse-grained molecular scale BD simulations and filament level stochastic

Langevin dynamics to investigate muscle contraction.[61] BD simulations of the association of individual tropomyosin and actin molecules were used to generate an interaction energy landscape that was then coupled to mathematical models for sarcomere-level activation dynamics. Finally, linkage to mesoscale models for receptor dynamics is also evolving as a viable strategy,[62,63] with much potential.

1.7 Conclusions and Outlook

Multiscale biomolecular simulation methods are emerging and developing rapidly,[64] promising increasing insight and impact in drug development. Multiscale simulation methods connect across two or more scales to investigate, for example, how changes at one level drive or are affected by changes at another and, by doing so, will bring a new depth of mechanistic understanding and unprecedented level of predictive power to drug discovery. Drug action is intrinsically multiscale, and understanding it requires an understanding of how molecular-level changes lead to macroscopic changes in biological systems. Small molecule binding to a receptor of interest leads to changes at many levels; modelling its effects and how binding is affected by the cellular milieu requires tools able to integrate biological and biochemical phenomena across a range of scales. While we have primarily focused here on methods for understanding such binding and unbinding processes at the molecular level, new developments in meshing seem poised to contribute to understanding how drug molecules exert their effects at higher levels of biological organization, in realistic, experimentally determined system structures.

The challenges involved are many and varied, reflecting the complexity of biological systems and the dynamics and fluctuating interactions of drug targets. As the examples reviewed briefly here show, significant progress has been made in integrating different types of simulation methods to link across diverse time- and length- scales. They have provided insight into factors

determining drug association rates and residence times, and the causes of drug resistance. Together with more expansive studies carried out on larger datasets, continued improvements to force fields with better connections to quantum mechanical methods, and treatment of complex biological environments, the scope and power of multiscale simulation will certainly increase. To access cellular time- and length scales, much remains to be done at the intersection of particle and continuum approaches. Ongoing challenges include making the transition between representations more seamless and routine, and the simulation of meshes that deform and reshape or remodel due to biological forces. Detailed comparison with experiment is essential in developing and testing such methods, which in turn will inform experimental design and analysis, and data engineering. Furthermore, additional potential will be realized through the incorporation of experimental and genetic data in fully integrative biological simulation methods.

1.8 Acknowledgements

S.H. and A.J.M thank EPSRC for funding [grant nos. EP/M015378/1, EP/M022609/1], and the Advanced Computing Research Centre (ACRC), University of Bristol - *http://www.bris.ac.uk/acrc/* for computer time. B.R.J., S.E.K., and R.E.A. acknowledge support from the National Biomedical Computation Resource (NBCR) NIH P41- GM103426. B.R.J. and S.E.K. also acknowledge support from the NIH Molecular Biophysics Training Program (T32-GM008326).

Chapter 1, in full, is a reprint of the material as it appears in: "Jagger, B. R. [†]; Kochanek, S. E. [†]; Haldar, S.; Amaro, R. E.; Mulholland, A. J. Multiscale Simulation Approaches to Modeling Drug–Protein Binding. *Curr. Opin. Struct. Biol.* 2020, *61*, 213–221." The dissertation author was a primary coinvestigator and author of this work.

Chapter 2

SEEKR: Simulation Enabled Estimation of Kinetic Rates, A Computational Tool to Estimate Molecular Kinetics and Its Application to Trypsin–Benzamidine Binding

2.1 Abstract

We present the Simulation Enabled Estimation of Kinetic Rates (SEEKR) package, a suite of open-source scripts and tools designed to enable researchers to perform multi-scale computation of the kinetics of molecular binding, unbinding, and transport using a combination of molecular dynamics, Brownian dynamics, and milestoning theory. To demonstrate its utility, we compute the k_{on}, $k_{off,}$ and ΔG_{bind} for the protein trypsin with its noncovalent binder, benzamidine, and examine the kinetics and other results generated in the context of the new software, and compare our findings to previous studies performed on the same system. We compute a k_{on} estimate of $2.1\pm0.3\bullet10^{7}$ $M^{-1}s^{-1}$, a k_{off} estimate of 83 ± 14 s^{-1}, and a ΔG_{bind} of -7.4 ± 0.2 kcal•mol^{-1}, all of which compare closely to the experimentally measured values of $2.9\bullet10^{7}$ $M^{-1}s^{-1}$, 600 ± 300 s^{-1}, and -6.7 kcal•mol^{-1}, respectively.

2.2 Introduction

Elucidating the kinetics and thermodynamics of binding and unbinding processes between a biomolecule and a substrate remains an important challenge in the field of molecular biophysics. Countless processes within the cell involve the association of a biomolecule with a metabolite, signaling molecule, toxin, drug, or another biomolecule.[65,66] Many of these interactions have important kinetic considerations: for instance, the speed of reactions or the residence time of an intermolecular encounter.[66]

Significant effort has been expended to accurately estimate the thermodynamics of binding using a variety of methods, particularly in the field of drug discovery, where the identification of a tight binder is an integral step towards obtaining a potential drug molecule that would accomplish a desired medical result.[22,35,67–69] While the thermodynamics of binding, encapsulated in the quantity of the free energy ΔG_{bind} of receptor-ligand complex formation, is an important factor in the binding process, a comprehensive understanding of the binding process requires consideration of binding kinetics and reaction rates.

Many theoretical approaches and simulation methods have been used to estimate both the thermodynamics and kinetics of binding. For instance, specialized machinery and long molecular dynamics (MD) simulations can be used in a 'brute force' approach, although it is relatively costly compared to other methods.[70–73] Markov models[74–80] can also be used to investigate the kinetics of binding,[31,44,81] as can milestoning.[35] Additional clever methodologies can be used to speed the computation using MD.[22,69,82–84] Brownian dynamics (BD) can also be used to approach the problem of binding kinetics,[68,85–88] as can Smoluchowski equation solvers.[89]

Our past work[50,90,91] has focused on using a multi-scale combination of MD and BD, unified through the theoretical framework of milestoning. In our previous study, we presented a

hybrid MD/BD/milestoning methodology to conduct our investigations into the kinetics of binding between superoxide dismutase and its natural substrate, the superoxide anion, and between troponin C and its natural substrate, the calcium ion.[50] Here, we make available a software package, SEEKR, that implements this method with significant improvements in automation, usability, and analysis. We demonstrate the utility of SEEKR by applying it to estimate the k_{on}, the k_{off}, and ΔG_{bind} between the serine protease trypsin and its ligand, benzamidine. In addition to the SEEKR software to perform milestoning calculations on any receptor-ligand system, we also make available a user guide, tutorial, and workflow to allow users to repeat our simulations and analysis for the trypsin-benzamidine system, and compute kinetics and thermodynamics for additional receptor-ligand systems.

2.3 Theory

The rationale and methodology behind our usage of milestoning to estimate kinetics using both MD and BD has been described recently in detail[50] for multiple applications. Our implementation to the trypsin-benzamidine receptor-ligand system in this study was adapted with few changes, the majority consisting of improvements in software efficiency.

In the case of bimolecular association, the kinetics of binding and unbinding can be represented respectively by two quantities, k_{on} and k_{off}, which are frequently depicted according to the following equation:

$$\boldsymbol{A + B \underset{k_{off}}{\overset{k_{on}}{\rightleftarrows}} AB} \qquad (2.1)$$

Which is shorthand for specifying that the values k_{on} and k_{off} function as parameters within the following differential equations:

$$\frac{d[AB]}{dt} = k_{on}[A][B] - k_{off}[AB] \quad (2.2)$$

$$\frac{d[A]}{dt} = k_{off}[AB] - k_{on}[A][B] \quad (2.3)$$

$$\frac{d[B]}{dt} = k_{off}[AB] - k_{on}[A][B] \quad (2.4)$$

Where *[A]*, *[B]*, and *[AB]* represent the concentrations of chemical species *A*, *B*, and their complex *AB*. The k_{on} and k_{off} relate to the dissociation constant K_D, and by extension, a free energy of association ΔG_{bind}:[69]

$$\frac{k_{off}}{k_{on}} = K_D = K_{\ominus} e^{\Delta G_{bind}/RT} \quad (2.5)$$

Where R is the gas constant, T is temperature, and $K_{\ominus}$ is a factor equal to one, in units of concentration.

The theory of milestoning has been formulated to compute kinetic and thermodynamic details of a process if the states of that process are represented as carefully chosen surfaces in phase space. These surfaces are known as "milestones".[92] In this study, the milestones are represented as concentric spherical shells (Figure 2.1) that encapsulate the binding site of the receptor. These spherical milestones are used for the computation of k_{on}, k_{off}, and ΔG_{bind}. Milestoning theory allows us to approach the problem of kinetics by utilizing a multi-scale strategy. We use highly-detailed, but computationally expensive MD simulations to observe transitions between milestones closer to the binding site so that molecular flexibility will be a component of the transitions between milestones. We then use BD for the larger and more widely-spaced milestones far from the binding site, where fast sampling of long trajectories is required and rigid body dynamics and implicit solvent are adequate[44,87,93] to model transition times and probabilities. In this way, we take

advantage of fully flexible MD where molecular flexibility is required, and also take advantage of the computation efficiency of BD where molecular flexibility is less important. Milestoning is the theory that combines the MD and BD components, by allowing statistics to be obtained in each regime independently, and then unifying the statistics through a rigorous theory that is agnostic to the method that was used to obtain them. Since the statistics of each milestone are obtained independently from the others, and since milestoning theory is a robust framework that can utilize information obtained by either Brownian or Newtonian dynamics,[94] we can choose whichever simulation method is most appropriate and convenient for that milestone.

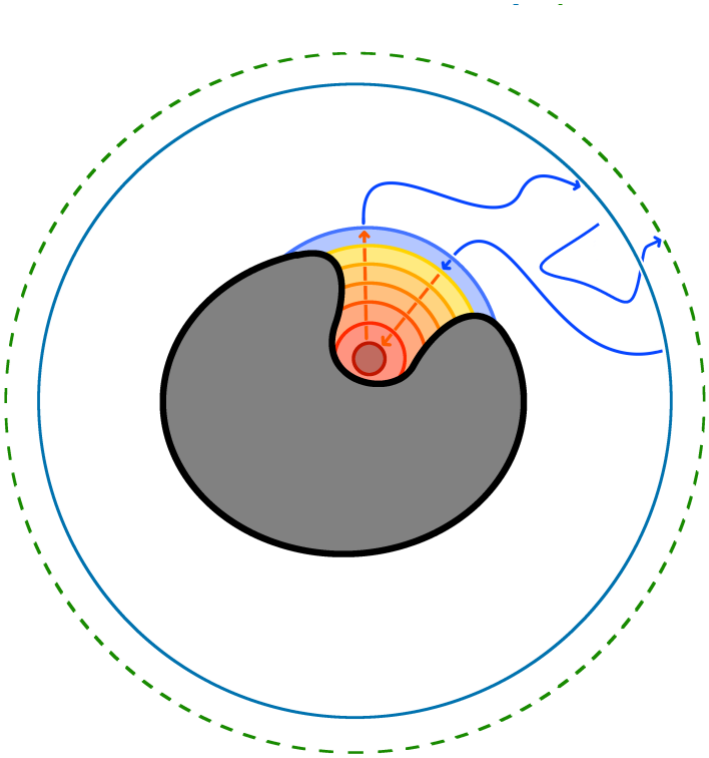


Figure 2.1 A cartoon schematic of trypsin (grey shape) with the concentric spherical milestones (orange and blue circular curves) surrounding the binding site.
Also, the *b*- and *q*-surfaces are represented as the outer blue and dashed green curves, respectively, that sit away from the molecule. Blue arrows represent BD trajectories, and orange arrows represent MD trajectories. Any surface with a blue arrow coming from or going to it represents the starting or ending surface for BD trajectories, respectively. Similarly, a surface with an orange arrow coming from or going to it represents the starting or ending surface for MD simulations, respectively.

By sampling transition statistics and times between the milestones using numerous short simulations, one can construct a transition kernel **K** that represents the transition probabilities and an incubation time vector $\langle \boldsymbol{t} \rangle$ that represents the average times of a system traversing the milestones.[95,96] The transition kernel **K** is a square matrix whose elements are constructed according to the following formula:

$$K_{ij} = {}^{n_{i\to j}}\!/_{\sum_k n_{i\to k}} \tag{2.6}$$

Where $n_{i\to j}$ is the number of trajectories that begin at a given milestone *i* and end at an adjacent milestone *j*. And the incubation time vector $\langle \boldsymbol{t} \rangle$ has elements that are constructed according to the following formula:

$$\langle \boldsymbol{t} \rangle_i = {}^{\sum_l \boldsymbol{t}_l}\!/_{\sum_k \boldsymbol{n_{i\to k}}} \tag{2.7}$$

Where t_l is the time of the *l*'th successful forward trajectory starting at milestone *i*, and $n_{i\to k}$, as before, is the number of trajectories beginning at milestone *i* and ending at milestone *k*. Therefore, $\langle t \rangle_i$ represents the average time spent by the system after crossing *i* and before crossing any other milestone.

In order to compute a free energy profile along the milestones, we must first obtain the stationary flux vector $\mathbf{q_{stat}}$ along the milestones by computing the principle eigenvector of **K**.

$$\mathbf{K} \cdot \mathbf{q_{stat}} = \mathbf{q_{stat}} \tag{2.8}$$

Then $\mathbf{q_{stat}}$ must be multiplied elementwise by $\langle \boldsymbol{t} \rangle$ to find the stationary probability vector $\mathbf{p_{stat}}$.

$$p_{stat,i} = q_{stat,i} \cdot \langle t \rangle_i \tag{2.9}$$

Finally, $p_{stat,i}$ relates to the relative free energy ΔG_i at milestone *i* according to the following:

$$\Delta G_i = -RT \ln\left(p_{stat,i}/p_{stat,ref}\right) \tag{2.10}$$

Where the index of $p_{stat,ref}$ is any reference state, such as the lowest energy, bound state. The value of $p_{stat,ref}$ is found by applying equation 2.9 to the chosen reference state.

To compute the k_{on}, we utilize the formula that is also used in BD theory:[85]

$$k_{on} = k(b)\beta \tag{2.11}$$

Where *k(b)* is computed using the following formula:

$$\boldsymbol{k(b)} = \left[\int_{b}^{\infty} \frac{\boldsymbol{e^{W(r)/k_BT}}}{\boldsymbol{4\pi r^2 D(r)}}\boldsymbol{dr}\right]^{-1} \tag{2.12}$$

The value *k(b)* represents the rate constant at which the ligand particles are crossing the *b*-surface, *W(r)* and *D(r)* are the potential of mean force and diffusion coefficient, respectively, that the ligand experiences at a distance *r* from the center of the receptor beyond the *b*-surface.[85] *D(r)* is computed by generating a Rotne-Prager diffusion tensor to approximate the hydrodynamics of a two body interaction in a viscous medium.[97] The value *k(b)* is computed automatically in BrownDye.

To find β, which represents the proportion of ligands crossing the *b*-surface that continue on to bind to the receptor, a starting probability vector $\mathbf{q_0}$ must be obtained in BD simulations by running a large number of conventional BD simulations where ligand molecules are started on a *b*-surface surrounding a receptor molecule. As the simulations run, and the proportion of trajectories that touch the outermost milestone(s) that encompasses a binding site on the biomolecule, rather than escaping to an infinite distance, are counted. In this case, $\mathbf{q_0}$ becomes:

$$\mathbf{q_0} = \left[0, \ldots, 0, q_{0,i}, 0, \ldots, 0, q_{0,j}, 0, \ldots, 0, q_{0,\infty}, 0, \ldots, 0\right]^T \tag{2.13}$$

Where *i* and *j* are the indices of one or more of these outermost site-encompassing milestones, $q_{0,i}$, $q_{0,j}$, are the probabilities that a BD trajectory started on the *b*-surface descend and touch these milestones, and $q_{0,\infty}$ is the probability that a trajectory diffuses away to an infinite distance. All the

entries in $\mathbf{q_0}$ must be normalized such that their sum equals a value of one. An "infinity" state in both vector $\mathbf{q_0}$ and in matrix **K**, represents the condition in which the ligand has escaped to an infinite distance from the receptor.

Next the transition matrix **K** must be modified to a new matrix $\widehat{\mathbf{K}}$ such that the milestones representing the bound and "infinity" states are sink states. That is, they all must have a probability of one that they transition only to themselves, and a zero probability to transition to anything else.

$$\widehat{K}_{ii} = \mathbf{1}\ \boldsymbol{if}\ \boldsymbol{i}\ \textbf{is a bound state, or the "infinity" state} \tag{2.14}$$

$$\widehat{K}_{ij} = \mathbf{0}\ \textbf{if also}\ \boldsymbol{i} \neq \boldsymbol{j} \tag{2.15}$$

Once $\widehat{\mathbf{K}}$ and $\mathbf{q_0}$ are properly defined, we compute the static flux vector $\mathbf{q}_\infty$.[44]

$$\mathbf{q}_\infty = \lim_{a\to\infty} \widehat{\mathbf{K}}^{a} \cdot \mathbf{q_0} \tag{2.16}$$

Finally, we obtain β:

$$\boldsymbol{\beta} = \sum_i \boldsymbol{q}_{\infty,i} \tag{2.17}$$

Where *i* is the index of one of the bound states.

To compute the k_{off}, we must return to the initial definition of matrix **K** as specified in equation 2.6. But it must be modified by introducing a "draining" state *i* by changing **K** into a draining matrix $\widetilde{\mathbf{K}}$ according to the following:

$$\widetilde{K}_{ij} = \mathbf{0}, \forall \boldsymbol{j} \tag{2.18}$$

That is, once we have decided that *i* is the draining state, we set that entire column of the matrix $\widetilde{\mathbf{K}}$ to zeros, while all other columns are kept the same as they were in **K**. In the SEEKR implementation, the outermost non-infinite milestone is considered to be the draining state. Then, we compute a mean first passage time (MFPT) τ:

$$\boldsymbol{\tau} = \mathbf{p_0}\left(\mathbf{I} - \widetilde{\mathbf{K}}^{\boldsymbol{T}}\right)^{-\mathbf{1}}\langle \mathbf{t} \rangle \tag{2.19}$$

Where $\mathbf{p_0}$ is a starting distribution of probabilities along each milestone, and $\widetilde{\mathbf{K}}^{\boldsymbol{T}}$ is the transpose of matrix $\widetilde{\mathbf{K}}$. We set $p_{0,i}$ to be 1 if *i* was a bound state, and set $p_{0,i}$ to be equal to 0 otherwise. The

MFTP τ is equivalent to a residence time of the ligand within the binding site, and can be related to the k_{off} according to the following relation:

$$\boldsymbol{k_{off} = \frac{1}{\tau}} \quad (2.20)$$

2.4 Materials and Methods

2.4.1 Description of the SEEKR package

SEEKR is a collection of scripts and files designed to automate the preparation and analysis of ligand-receptor kinetic calculations that use a multi-scale MD/BD/milestoning framework. SEEKR does not run the simulations themselves, but instead relies on the well-established NAMD[98] and BrownDye[99] programs. In this case, SEEKR is more of a specialist interface or tool that automates the cumbersome process of preparing, running, and analyzing a particular type of multi-scale milestoning calculation so that researchers will be able to run them more easily than if the process were done manually.

SEEKR programs are classified into three general categories:

1. Preparation: These scripts and modules accept input from the user in order to construct all the necessary files needed by both NAMD and BrownDye to run their respective simulations. The files are organized into a file tree whose branches represent the various independent milestones, which simulation method is being used (MD or BD), and the various stages of the calculations. When run, the user will have all the required files arranged and poised for simulation and milestoning calculations.
2. Running: Other scripts aid the user in running the MD and BD simulations locally and on supercomputers. For instance, SEEKR contains a script to prepare the submission of the computationally-intensive MD simulation jobs to a SLURM supercomputer queue, and when the allotted time runs out, the script prepares all the necessary

resubmission files for one, some, or all of the milestones with a single command. Other scripts use previous BD trajectory output to prepare and run ensembles of BD simulations from first hitting point distributions (FHPD).

3. Analysis: When all the simulations are complete, the user can run an analysis script that descends into the file tree, gathering all the simulation output. It then combines this information to construct the milestoning model, and performs all the milestoning and error calculations, providing the user with kinetic and thermodynamic information, including k_{on}, k_{off}, and the free energy profile. It also has the option to perform convergence analysis on these values. Additional analysis scripts can be utilized to generate a single file containing the ligand equilibrium distribution or FHPD of each milestone for easy visualization.

The Python scripts have been tested using Python 2.7 and can be safely run in any version of Python 2 at version 2.7 or later. The remainder of the scripts are written in TCL, particularly those interfacing with NAMD, which has a TCL-based interface. SEEKR also uses the Numpy, Scipy, and MDAnalysis python libraries. The Adaptive Poisson-Boltzmann Solver (APBS)[100] is used to generate the electrostatic potential maps for input to BrownDye, and the AmberTools program LEaP[101] is also used to prepare structures for MD simulation.

2.4.2 Trypsin structure preparation and SEEKR creation of milestoning structures

Atomic coordinates of the trypsin-benzamidine system were obtained from the high resolution crystal structure PDB ID: 3PTB.[102] Hydrogens were added using Molprobity with ring flips allowed.[103,104] The system was then further prepared using LEaP with the Amber forcefield, ff14SB.[105] Disulfide bonds were added manually. The appropriate protonation states of ASP, GLU, and HIS residues at a pH of 7.7 were determined using PROPKA.[106,107] This pH was selected to

align with the experimental conditions of Guillian and Thusius.[108] The structure was then solvated in a truncated octahedron of TIP4Pew[109,110] waters and eight Cl- ions were added to neutralize the overall charge. The benzamidine ligand was parameterized using Antechamber with the GAFF force field.[110,111] The total size of the system was approximately 23,000 atoms. To allow for relaxation from the crystallographic starting structure, the benzamidine ligand was removed and a 20 ns simulation of the apo structure was performed at a constant temperature of 298 K using the Langevin thermostat and a constant pressure of 1 atm using the Langevin piston with a damping coefficient of 5 ps^{-1}. A representative structure from this simulation was then used as the SEEKR input structure to generate all the necessary inputs for the MD simulations to be run using NAMD, and the BD simulations using Browndye.

The benzamidine bound-state coordinates were defined from the center of mass of the alpha carbons of residues 190, 191, 192, 195, 213, 215, 216, 219, 220, 224, 228 of PDB: 3PTB because these residues form the binding pocket in the bound-state crystal structure by manual inspection. Spherical milestones were defined with radii of 1, 1.5, 2, 3, 4, 6, 8, 10, 12, 14 Å, with the origin being the bound state coordinates defined above. This spacing of the milestones was chosen to facilitate the simulation of transitions between milestones while still ensuring the Markov assumptions required by formal milestoning theory. Ten copies of the apo structure were generated, each with the benzamidine ligand inserted on one of the ten spherical milestones (Figure 2.2A). Water molecules that clashed with the ligand structure were removed. The first nine milestones correspond to the MD simulation regime, with the innermost milestone (1 Å) representing the bound state, as the center of mass of the bound benzamidine ligand falls well within the 1 Å sphere that defines this milestone (Figure 2.2B). Furthermore, in a ~170 ns unrestrained MD simulation with the ligand in the bound pose, the 1 Å sphere contained the center of mass of the ligand over

71% of the simulation. The tenth and outermost milestone (14 Å) corresponds to the BD simulation regime. The distribution along any milestone where BD was started was constructed by first running conventional BD simulations and obtaining the distribution of hitting points along that milestone.

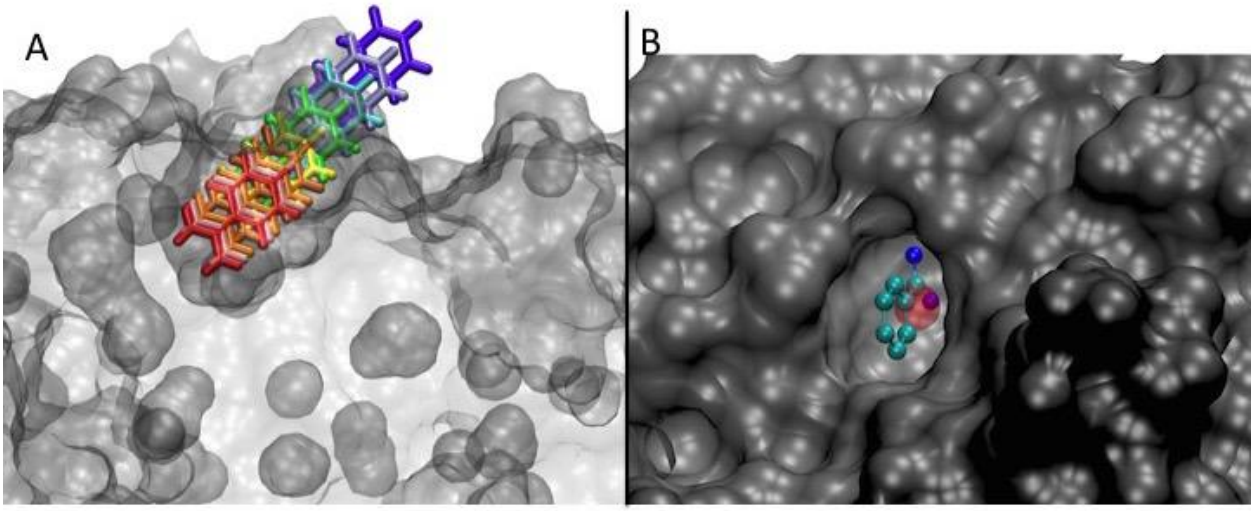


Figure 2.2 The benzamidine binding site of trypsin
A) before beginning the simulations, benzamidine has been placed along each of the milestones in gradually increasing distances from the center of the binding site on trypsin. B) The center-of-mass of the benzamidine molecule in the trypsin 3PTB crystal structure lies within the lowest 1Å milestone (red sphere), which we define as the bound state

The *b*-surface is a relatively large spherical shell that encloses the entire receptor molecule, with a radius of sufficient size that the entire surface sits well out into the bulk solvent where forces between the ligand and receptor would be largely unaffected by molecular orientation, and are therefore centrosymmetric.

2.4.3 MD simulations

A modified version of NAMD 2.11 was used for all MD calculations. The numerous MD inputs, including input files, integrator parameters, boundary conditions, temperature and pressure controls, etc. are either defined by the user or set by SEEKR to default values. Relevant settings and procedures implemented for each milestone in the MD regime are described here.

For each milestone system generated by SEEKR as described above, the solvent molecules were allowed to relax around the newly placed benzamidine ligand by minimizing for 5000 steps with both the ligand and receptor restrained. The solvent was then further relaxed through a series of 2 ps heating simulations, where the temperature was increased from 298 K to 350 K and then cooled back to 298 K in 10 K increments, keeping the atoms of the ligand and receptor restrained. Following this relaxation of the solvent, an equilibrium distribution of the ligand on the milestone surface was obtained from 1 μs of constant volume simulation at a temperature of 298 K where a harmonic spring force of 90 kcal•mol^{-1}•Å^{-2} was imposed to restrain the ligand at the appropriate radius from the binding site center for each milestone to generate an equilibrium distribution (Figure 2.3A). This is also known as the umbrella sampling stage. From this equilibrium distribution, a FHPD (Figure 2.3B) was obtained by selecting 4700 position and velocity configurations from times 60 ns – 1 μs of the equilibrium trajectory and allowing them to propagate backwards in time by reversing their velocities at constant energy and volume (reverse stage). Any trajectories that struck another milestone before re-crossing the milestone from which they originated were counted as part of the FHPD. All members of the FHPD were then brought back to their original positions and velocities and subsequently allowed to propagate forward in time at constant energy and volume (forward stage). When a simulation crossed its starting milestone again, it was then monitored for transitions to adjacent milestones and the incubation time for these transitions was also recorded. Once a trajectory crossed an adjacent milestone, the simulation was terminated. Any trajectories in this forward stage that crossed adjacent milestones before re-crossing their starting milestone were rejected. The 1, 1.5, 2, 4, and 10 Å milestones produced results with significantly fewer transitions than the other milestones. Therefore, to improve the robustness of our statistics, we performed additional reverse and forward simulations where 10

more trajectories were initiated at random Maxwell-Boltzmann velocities from each equilibrium distribution point, in addition to the one described above (a total of 470,000 reversals for each of these milestones), increasing the number of transitions observed. For each milestone, successful forward stage statistics were inserted into the transition kernel **K** and incubation time vector $\langle \boldsymbol{t} \rangle$.

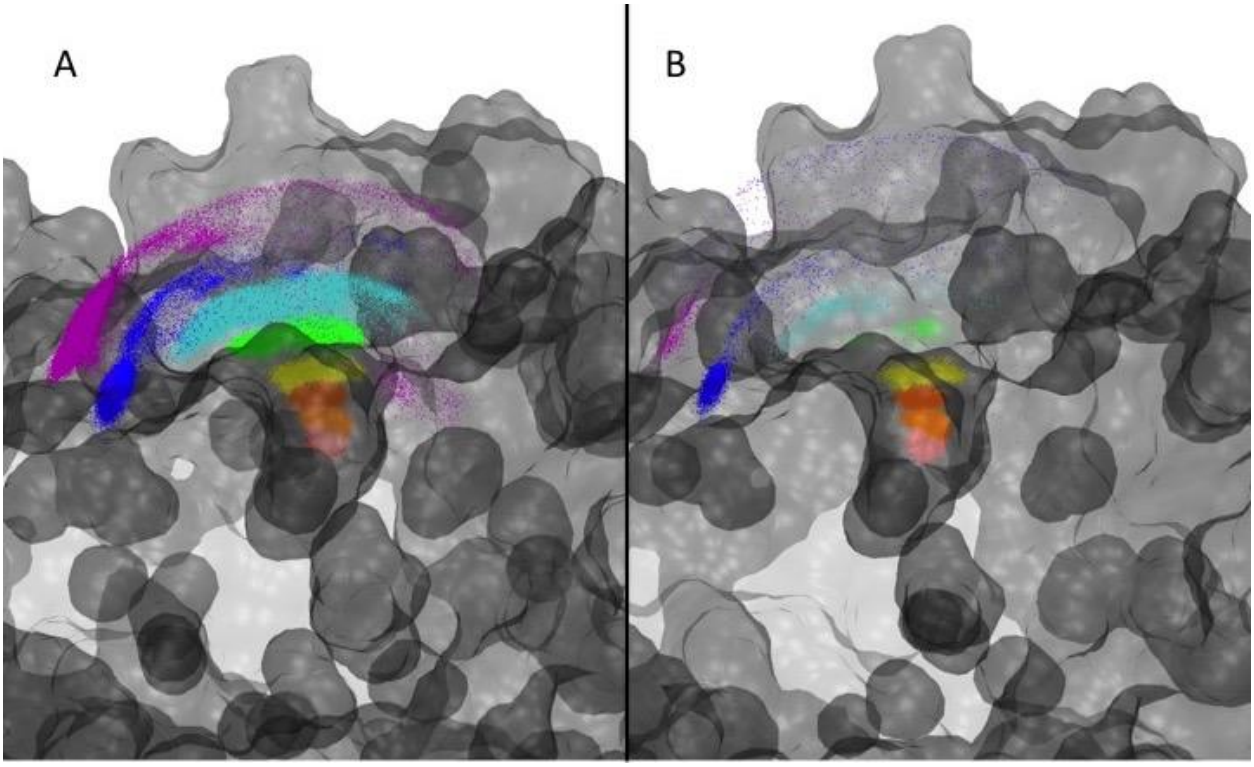


Figure 2.3 Benzamidine conformational sampling during MD simulations
Panel A: The equilibrium distribution of the center of mass of benzamidine generated along all of the milestones from 2 Å (red) to 12 Å (green) at the end of the umbrella sampling. No umbrella sampling is performed for the BD stages, so there are no points representing the 14 Å milestone. Panel B: The FHPD of benzamidine centers of mass generated from the equilibrium distribution that succeeded in the reverse stage. The milestones between 1 Å (red) and 12 Å (green) were generated during the MD simulations. In addition, the blue distribution at 14 Å represents the FHPD obtained from the BD simulation. This FHPD is used to start forward stage trajectories for generating milestoning statistics.

2.4.4 BD simulations

All BD calculations were conducted with BrownDye, a software package specializing in the rigid-body diffusion of two biological molecules in an implicit solvent.[99] The electric potential map used as input for the BD simulation was calculated with the APBS version 1.4. All BD inputs,

as well as the necessary APBS inputs for creation of the electrostatics map, are user defined in the SEEKR input file or generated as SEEKR default values.

In an attempt to recreate the ionic conditions used in the experiment,[108] a nonlinear APBS calculation was run at 298 K, with a solvent dielectric of 78 and a solute dielectric of 2, with the following ions: Ca^{2+} at a concentration of 0.02 mM with a charge of +2.0 *e* and a radius of 1.14 Å, Cl^- at a concentration of 0.10 mM with a charge of -1.0 *e* and a radius of 1.67 Å, and tris at a concentration of 0.06 mM with a charge of +1.0 *e* and a radius of 4.0 Å.[112] At the specified concentrations, these ions generate a Debye length of 8 Å, which is used as input to BrownDye. Both the *b*-surface BD simulations and BD trajectories starting from a milestone ran with a solvent dielectric 78 and a solute dielectric of 2, at 298 K. We ran three additional sets of BD simulations at different ionic concentrations to examine the effect of ionic strength in the BD simulations on the k_{on}. Therefore, three additional simulations were run: one with an ion concentration of zero, another with half of the ion concentrations of the experimental procedure, and another with double the ion concentration of the experimental procedure. Although an electrolyte solution technically has a Debye length equal to infinity, we approximated the Debye length with a value of 99 Å in the BrownDye program.

For each k_{on} calculation, we performed 10^6 BD simulations initiated at random points distributed on the *b*-surface, which were used to construct the vector $\mathbf{q_0}$ in equation 2.13. Once these simulations completed, the trajectories that successfully reached the outermost milestone were used as that milestone's FHPD. From that FHPD, an additional 10^6 BD trajectories were run until reaching the second-outermost milestone or escaping to the *q*-surface. These statistics were also included in the transition kernel **K** and incubation time vector $\langle \boldsymbol{t} \rangle$.

2.4.5 Milestoning calculations

Using the statistics obtained from all the milestones in both the MD and BD regimes, the SEEKR software was used to construct the milestoning model and compute the k_{on}, k_{off}, ΔG_{bind}, and other quantities of interest. Additional scripts used to generate some of the figures and data are also included in the SEEKR package. Error estimates were computed according to our previously defined procedure.[50]

The vast majority of the procedure outlined in the Materials and Methods section is automated within the SEEKR software package.

2.5 Results

Using the MD/BD/milestoning methodology through the SEEKR interface yielded a k_{on} of $2.1\pm0.3\bullet10^{7}$ $M^{-1}s^{-1}$ for the trypsin-benzamidine system. This value deviates from the experimentally measured k_{on} for the same system at $2.9\bullet10^{7}$ $M^{-1}s^{-1}$ by a factor of ~1.5 (no experimental error margins were reported). We also estimate a k_{off} of 83 ± 14 s^{-1}, which is within an order of magnitude of the experimentally determined value of 600 ± 300 s^{-1} though our value is slower than expected. similar phenomenon is observed in other computational k_{off} estimations of this system. An examination of the effect of ionic concentration on the k_{on} convergence of the rate constants as a function of the length of umbrella sampling performed is provided in the SI. Using equation 2.5, we obtain a ΔG_{bind} estimate of -7.3 ± 0.2 kcal•mol^{-1} from a K_d of $4.3\pm1.2\bullet10^{-6}$ M compared to the experimental ΔG_{bind} of -6.71 ± 0.05 kcal•mol^{-1}, computed at 298 K using equation 2.5 and an experimental K_d of $1.2\pm0.1\bullet10^{-5}$ M.[108]

In addition, we obtained a relative free energy at each of the milestones along the binding pathway using the vector $\mathbf{p}_{\mathbf{stat}}$ in combination with equation 2.10. This free energy profile is displayed in Figure 2.4.

Aside from the predicted thermodynamic and kinetic quantities, we used the trajectories generated during the SEEKR run to make other observations about the system during the binding and unbinding process.

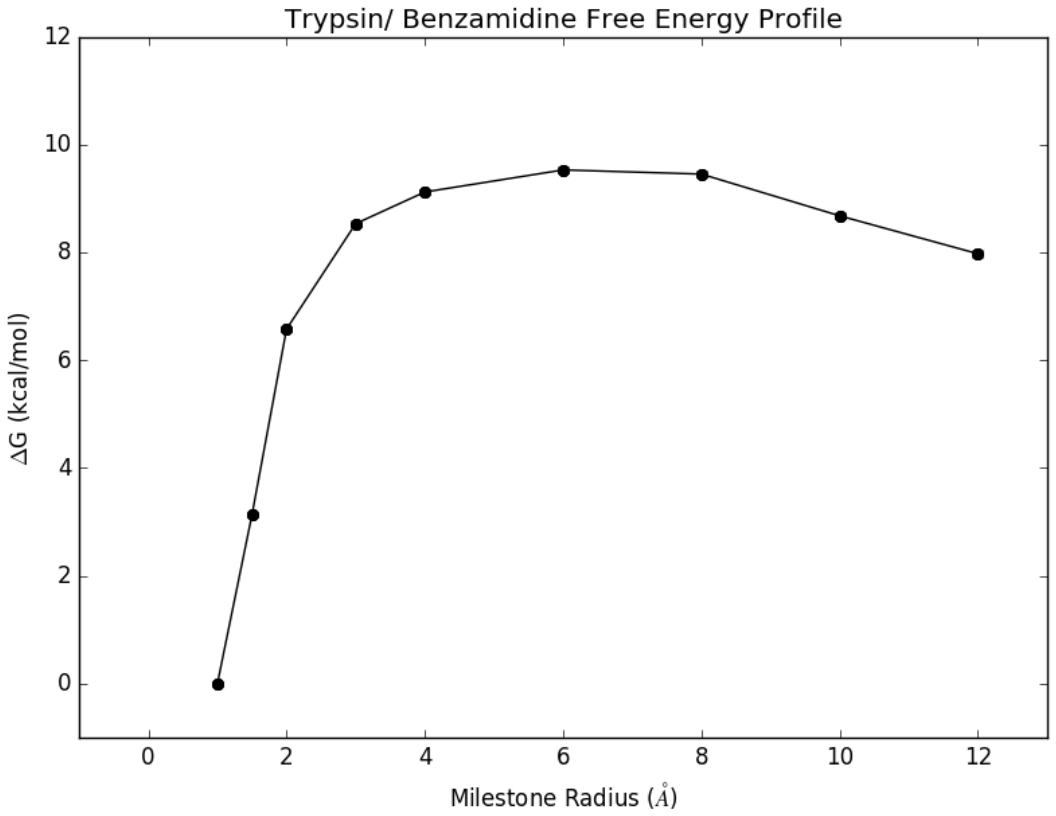


Figure 2.4 The free energy profile of benzamidine along each of the milestones leading to the binding site. The free energy barrier peaks around the milestone located at 6 Å.

By removing the benzamidine molecule and the solvent, we used POVME2[55] to provide pocket volume measurement and characterization during the course of the MD runs. The same

origin and radius of the inclusion region that defined the binding pocket were used for all umbrella sampling trajectories. The pocket itself remains relatively rigid when the benzamidine is deep in the binding site during the umbrella sampling stage, however, more variation in volume was observed when the benzamidine was constrained to a milestone nearer to the entrance of the opening of the binding site (Figure 2.5).

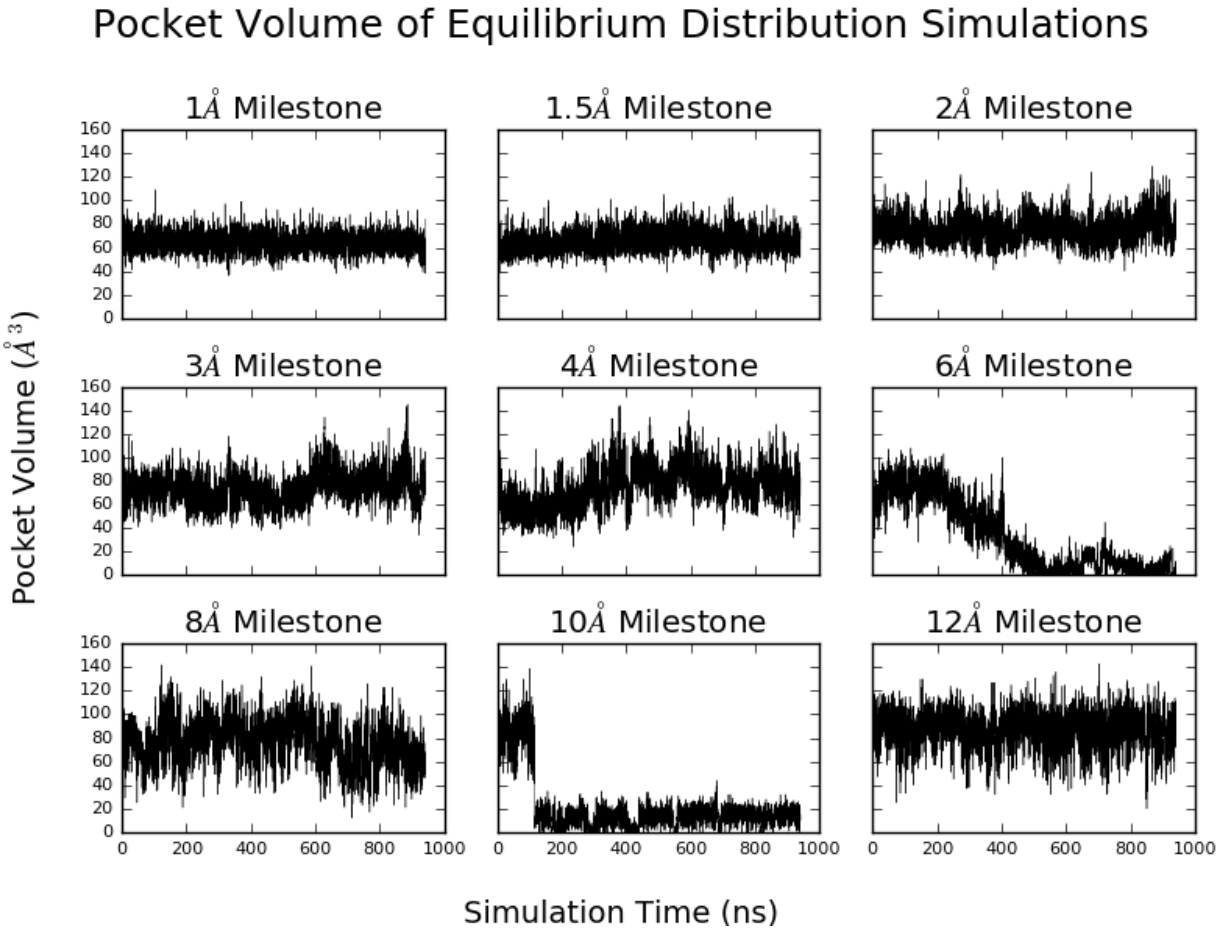


Figure 2.5 The volume of the S1 binding site with benzamidine restrained to the milestones as computed using the POVME2 program.
Stabilization of the binding site pocket volume is observed as the ligand moves closer to the binding site.

Closer analysis of the umbrella sampling trajectories for the 6, 10, and 12 Å milestones in conjunction with the POVME data indicates sampling of multiple conformations of the trypsin S1 binding pocket (Figure 2.6). The binding pocket conformation is primarily dependent on the

motion of two loops; the loop containing TRP215 and the loop containing ASP189, a critical residue for benzamidine recognition. The opening and closing of the S1 pocket is greatly influenced by the orientation of TRP215 When oriented downward as in Figure 2.6A, the S1 pocket is open. This is the conformation observed in the crystal structure 3PTB with benzamidine bound. When TRP215 rotates upwards as in Figure 2.6B, the binding pocket is closed, and pocket volume significantly decreases. The dramatic change in pocket volume for the 10 Å milestone also occurs when TRP215 moves to close the S1 binding site.

We also observe the formation of an S1* pocket, that results from the motion of these two loops. This pocket provides an alternate binding pathway, in which benzamidine can approach ASP189 from a different orientation. These observations are in agreement with the study of Plattner and Noé[31] where these results were observed through several hundred independent MD trajectories totaling over 100 μs of aggregate simulation time.

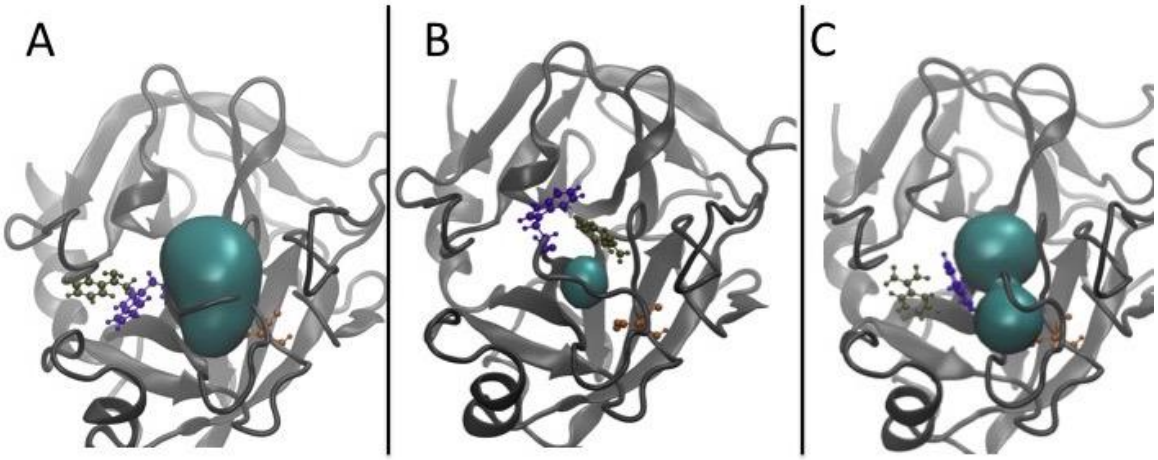


Figure 2.6 Dynamics of the apo trypsin S1 binding pocket umbrella sampling simulations.
Pocket conformations are significantly influenced by the motions of the loop containing TRP215 (violet) and the loop containing ASP189 (orange), which is important for benzamidine recognition. Benzamidine is shown in tan. POVME calculated volumes are shown in cyan. **A)** The open S1 pocket, where TRP215 is pointed in a downward orientation. **B)** Closed conformation of the S1 pocket as a result of TRP215 rotating to an upward pointing conformation. **C)** Formation of the S1* pocket where benzamidine can approach via an alternate pathway and interact with ASP189 from a different angle.

We also observed significant positional and rotational sampling by the benzamidine along most of the milestones during the umbrella sampling stages. This information can provide an idea for the likelihood of pathways that benzamidine follows on its route to binding. Figure 3A shows the equilibrium distribution along each of the milestones, and figure 3B shows the FHPD for each of the milestones. Figure 2.7 shows the angle between a vector pointing along the amidine group and a vector pointing out from the opening of the binding site as a function of time during the equilibrium simulations. Several flips are observed in all but the lowest milestones, where benzamidine rotation was restricted because these milestones are located deep within the binding pocket. The 10 Å also experiences a decrease in rotational sampling because benzamidine is interacting extensively with TRP215 and thus adopts an orientation that favors stacking of the aromatic rings.

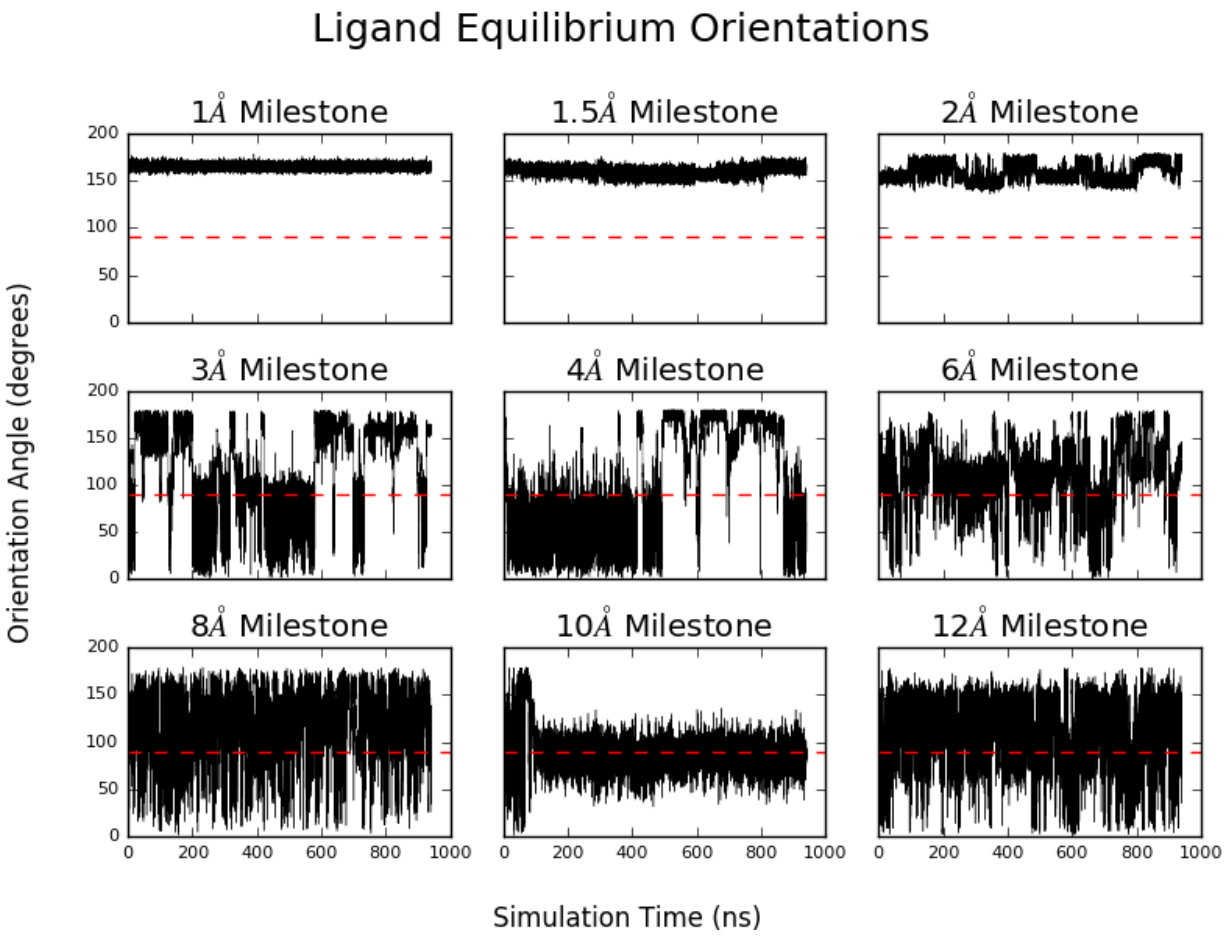


Figure 2.7 The angle of benzamidine along the center-of-mass/amidine axis compared to a vector pointing outward from the binding site.
An angle larger than 90° represents a conformation where the amidine group is pointing toward the binding site. Several flips were observed in all milestones above 2 Å, implying that the orientation of the ligand is well sampled along all of the milestones except for those deepest in the binding pocket, where the orientation found in the crystal structure is preferred, and the amidine group is pointing down into the site.

The crystal structure of the trypsin/benzamidine complex shows the amidine group pointing downward toward the binding site (Figure 2.2B). This structural feature is confirmed by our own simulations, and a relatively narrow arrangement of ligand orientations are observed along the lowest milestone.

The entire calculation cost approximately 1.4 million CPU hours on the Stampede supercomputer and local machines, with a total MD cost of approximately 19 μs of simulation.

2.6 Discussion

Compared to the experimental k_{on}, our estimated k_{on} is slower by about a factor of 1.3, but falls well within an order of magnitude. We attempted to closely recreate the experimental ionic conditions within our simulations, which has a pronounced effect on the k_{on} (details in the SI). Our k_{on} of 2.2±0.3•10^7 $M^{-1}s^{-1}$is much closer to the experimental value of 2.9•10^7 $M^{-1}s^{-1}$ than the k_{on}s obtained by Buch et. al.[81] (15±2•10^7 $M^{-1}s^{-1}$) and comparable to what was obtained by Plattner et. al.[31] (6.4±1.6•10^7 $M^{-1}s^{-1}$), although ours was obtained with significantly less computational resources, smaller by an order of magnitude. Our result is also very close to what was obtained by Tiwary et. al.[84] (1±1•10^7 $M^{-1}s^{-1}$). Our estimated k_{off} of 83±14 s^{-1} is within an order of magnitude of the experimental k_{off}, far closer than the value obtained by Buch et. al. (9.5±3.3•$10^4$$s^{-1}$), and comparable to the values obtained by Plattner et. al. (131±109 s^{-1}), Teo et. al.[83] (260±240 s^{-1}), and Tiwary et. al (9.1±2.5 s^{-1}). To our knowledge, this is the first successful estimate of k_{off} using a hybrid MD/BD/milestoning model.

An advantage of our approach is that both k_{off} and k_{on} can be determined from the same calculation. We can use our calculated k_{off} and k_{on} values in equation 2.5 to obtain an entirely computationally-determined dissociation constant K_D of 3.8±0.8•10^{-6} M, and by extension a free energy of binding ΔG_{bind} estimate of -7.4±0.2 kcal•mol^{-1}. This is in good agreement with the experimental K_D of 1.2•10^{-5} M, which when put through equation 2.5 at a temperature of 298 K, yields a free energy of -6.7 kcal•mol^{-1}.

The accurate determination of kinetics using milestoning requires the proper generation of equilibrium and FHPD distributions. It is important to ensure adequate sampling in the generation of equilibrium distributions. Figure 2.3A shows the equilibrium distribution of benzamidine center-of-mass along the 1 Å to 12 Å milestones in the MD regime. The benzamidine appears to

have explored all solvent-accessible regions along the milestones. Along with positional sampling, the observed diversity of benzamidine orientation in Figure 2.7 indicates that the ligand orientational degree of freedom is well-sampled in all but the lowest milestones. In addition to the ligand, it is important that receptor conformations that may affect ligand binding are also well sampled. By using POVME2, we observed conformational states that have been observed in other studies such as the S1* pocket (figure 6).[31] We do not however observe any complete binding events via the S1* pocket, presumably as a result of our simplified spherical milestoning model. This may provide some explanation as to why our calculated rates are somewhat slower than experiment, as we do not capture this alternate pathway. However, we may reasonably assume that we are capturing most of the effects of slower receptor conformational changes and subsequently, that our kinetics predictions are reasonable.

While, of course, verification of SEEKR as a computational kinetics and thermodynamics estimator will need to be performed on additional systems, this similarity between experimental and theoretical free energies and rate constants in our accessible and highly parallel framework is encouraging.

2.7 Conclusions

In this work, we use our multi-scale MD/BD/milestoning methodology to examine ligand-protein binding events with a larger, more complex, and more drug-like ligand than in our previous work. Furthermore, we present the first successful k_{off} calculation to within one order of magnitude of experiment using this approach. Using the obtained values of k_{on} and k_{off}, and entirely computational estimate of K_D and ΔG_{bind} in good agreement with experiment were obtained. These results are further evidence that the MD/BD/milestoning methodology can be successfully applied

to the investigation of binding and unbinding kinetics in receptor-ligand systems. We also present the SEEKR software package, which automates much of the preparation, submission, and analysis of these types of calculations. We have made SEEKR freely available and open-source on Github, and hope that it will be used and improved by the community to run predictive multi-scale milestoning calculations. SEEKR downloads, tutorials, and the user guide may be found at http://amarolab.ucsd.edu/seekr.

2.8 Acknowledgements

We would like to thank Jim Philips, Wen Ma, Jamie Schiffer, Gary Huber, Robert Malmstrom, Rob Swift, J. Andrew McCammon, and Carlos Simmerling for their assistance to the SEEKR project. We dedicate this work to the memory of Klaus Schulten, a pioneer who inspired so many.

Chapter 2 , in full, is a reprint of the material as it appears in: "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* 2017, *121* (15), 3597–3606." The dissertation author was a primary coinvestigator and author of this work.

2.9 Supporting Information

2.9.1 Convergence of k_{on} and k_{off} values:

The k_{on} calculation is fairly well converged in the 47000 umbrella sampling frames. Although k_{off} is not as well converged within that span of time, fluctuations are less than one order of magnitude.

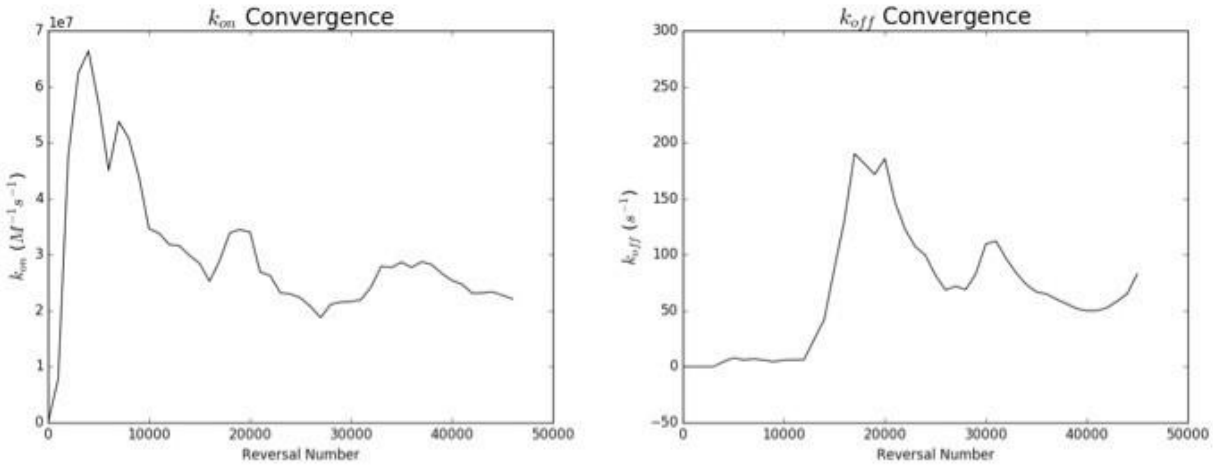


Figure 2.8 Convergence of rate constants as a function of umbrella sampling length.
Fluctuations in the rate constant for the k_{on} are fairly well converged. In contrast, the k_{off} is likely to require additional umbrella sampling to fully converge, but both kinetic values seem to have converged to within one order of magnitude given the amount of umbrella sampling.

2.9.2 Sensitivity of the system to ionic concentration

In order to investigate the sensitivity of this system to the ionic concentrations within the BD stage simulations, we also computed the k_{on} at various ionic concentrations. The dependences of concentration on the k_{on} and k_{off} are listed in table S1. If ionic strength is neglected, then the k_{on} decreases by about a factor of three.

Table 2.1 The effect of ion concentration on the computed k_{on} and k_{off}.

Ion strength factor	TrisHCl concentration (M)	$CaCl_2$ concentration (M)	Debye Length (Å)	k_{on} ($M^{-1}s^{-1}$)	k_{off} (s^{-1})	ΔG_{bind} (kcal/mol)
0	0	0	∞	$6.3\pm0.8\bullet10^{6}$	83±14	-6.7±0.1
0.5	0.03	0.01	12	$2.1\pm0.3\bullet10^{7}$	83±14	-7.4±0.1
1	0.06	0.02	8	$2.1\pm0.3\bullet10^{7}$	83±14	-7.4±0.1
2	0.12	0.04	6	$2.2\pm0.3\bullet10^{7}$	83±14	-7.4±0.1

The reason that ionic strength must be accounted for so carefully in this system is that the k_{on} between trypsin and benzamidine shows high dependence on ionic strength in our BD simulations (table S1). This is likely because the k_{on} is close to the diffusion limit, and electrostatic forces are screened by the dissolved ions. One surprising observation, however, is that increased ionic strength in the BD simulations (smaller Debye length) actually increases the computed k_{on}. This result suggests that the benzamidine experiences repulsive forces during its approach to binding with trypsin, likely contributing to the low free energy barrier to entry observed in the profile in figure 4 of the main text, and that these repulsive forces are shielded by higher ionic strength, and are thus electrostatic in nature. Another possibility is that the high ionic strength is shielding attractive hot spots on the surface of trypsin that compete with the active site for binding. Electrostatic maps show large regions of positive electrostatic fields surrounding the majority of trypsin. These positive fields likely repel the positively charged benzamidine (figure S2). It is clear, however, that ion concentration in kinetics calculations of this system must be carefully chosen in order to properly reproduce experimentally observed values.

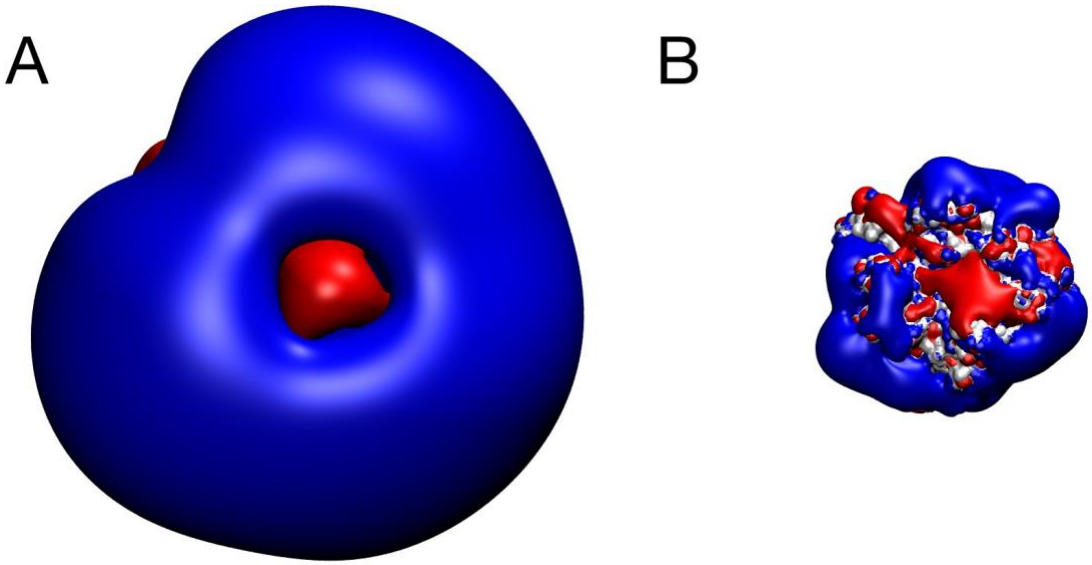


Figure 2.9 The electrostatic potentials around trypsin.
In panel A, the isosurface is drawn in blue at 0.01 kT/e, and in red at -0.01 kT/e. In panel B, the isosurface is drawn in blue at 1 kT/e, and in red at -1 kT/e. The large positive field surrounding trypsin provides a reasonable explanation for why we observe a faster k_{on} at higher salt concentrations.

Chapter 3

Quantitative Ranking of Ligand Binding Kinetics with a Multiscale Milestoning Simulation Approach

3.1 Abstract

Efficient prediction and ranking of small molecule binders by their kinetic (k_{on} and k_{off}) and thermodynamic (ΔG) properties can be a valuable metric for drug lead optimization, as these quantities are often indicators of *in vivo* efficacy. We have previously described a hybrid molecular dynamics, Brownian dynamics, and milestoning model, Simulation Enabled Estimation of Kinetic Rates (SEEKR), that can predict k_{on}'s, k_{off}'s, and ΔG's. Here we demonstrate the effectiveness of this approach for ranking a series of seven small molecule compounds for the model system, β-cyclodextrin, based on predicted k_{on}'s and k_{off}'s. We compare our results using SEEKR to experimentally determined rates as well as rates calculated using long-timescale molecular dynamics simulations and show that SEEKR can effectively rank the compounds by k_{off} and ΔG with reduced computational cost. We also provide a discussion of convergence properties and sensitivities of calculations with SEEKR to establish "best practices" for its future use.

3.2 Main

Molecular binding processes are ubiquitous in biology and serve as the fundamental basis for biological complexity. For the drug discovery community, engineering pharmacologically active small molecules is of particular importance. Traditionally, the paradigm for lead optimization is to select for leads with the greatest affinity for a protein, or other, target of interest. However, recent evidence suggests that the kinetics of binding may also be a useful metric for lead selection. It is now thought that both residence times and association rates are key determinants of *in vivo* efficacy for many drugs.[66,113–116] Similar to computational predictions of binding thermodynamics, molecular simulations can be used to compute binding kinetics.[1,5,117] Methods such as Brownian Dynamics (BD) have been used effectively for estimating molecular association rates.[85–87,99] Molecular dynamics (MD) simulations which explicitly represent all atoms and forces, can also be used to predict binding kinetics. Due to significantly increased model complexity, MD is limited by sampling. Nevertheless, owing to software improvements and the development of commodity hardware such as GPUs and specialty hardware such as Anton,[118,119] "brute force" calculation of binding kinetics with MD is now a possibility.[70–72,120,121] To improve upon "brute force" sampling statistics, many sampling strategies employ force biases or other statistical mechanical techniques to predict both association and dissociation rates of many systems. This includes methods such as: Markov State Models,[31,81,122,123] metadynamics,[84,124,125] milestoning,[29,35,36,126–128] and other techniques.[83,129–134]

Our previous work uses a multiscale MD, BD, and Milestoning approach for the calculation of both association and dissociation rates of receptor-ligand complexes.[50,51] Our implementation, "Simulation Enabled Estimation of Kinetic Rates" (SEEKR) is a freely available software

package[1] that automates the preparation, simulation, and analysis of these multiscale milestoning calculations using existing softwares: NAMD[98] for MD simulations and BrownDye[99] for BD simulations. Milestoning theory provides the glue for the multiscale scheme by providing a strategy to subdivide, simulate, and subsequently statistically reconnect small regions of simulation space called "milestones"[29,50,51,92,94,95,135–141] This approach reduces the compute time required to simulate transition events, is embarrassingly parallel, and is agnostic to the simulation modality used. This allows us to use atomically detailed, yet computationally expensive, fully flexible MD simulations in milestones near the binding site where these interactions are critical for understanding the binding and unbinding, and BD simulations far from the binding site where rigid body dynamics provides a sufficient description at significantly reduced computational cost. For a more thorough description of milestoning theory and the calculation of kinetic quantities, such as k_{on} and k_{off}, we refer the reader to the existing literature. [29,50,51,92,94,95,135–141]

The effectiveness of the SEEKR scheme for the calculation of k_{on} and k_{off} values has been demonstrated for multiple protein-ligand systems.[50,51] However, it has not yet been used for rank ordering sets of compounds by kinetic (k_{on} and k_{off}) and thermodynamic (ΔG) values, as would be done in pharmaceutical discovery settings. Here we use SEEKR to estimate k_{on}, k_{off}, and ΔG for a model host-guest system, β-cyclodextrin with seven ligands representing diverse chemical groups (Figure 3.1), using two forcefields for β-cyclodextrin GAFF[110,111] and Q4MD.[142]

We compare the SEEKR estimates with previously published "brute force" (long-timescale) MD predictions[121] and experimental results.[143–148] Using this model system we examine both the accuracy and efficiency of SEEKR compared to long-timescale MD. We further explore the reduction in computational effort required for SEEKR estimates as well as discuss the convergence properties and sensitivity of SEEKR calculations to establish "best practices" for its future use.

SEEKR calculations and the long timescale MD simulations struggle to reproduce both the values and rank ordering of the experimentally determined k_{on}'s (Figure 3.2).

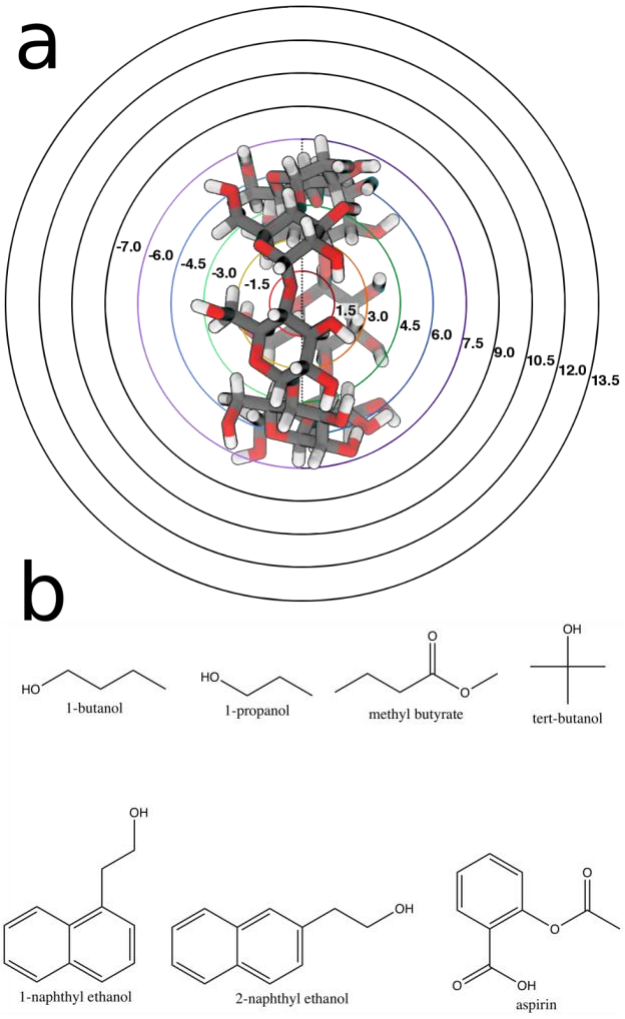
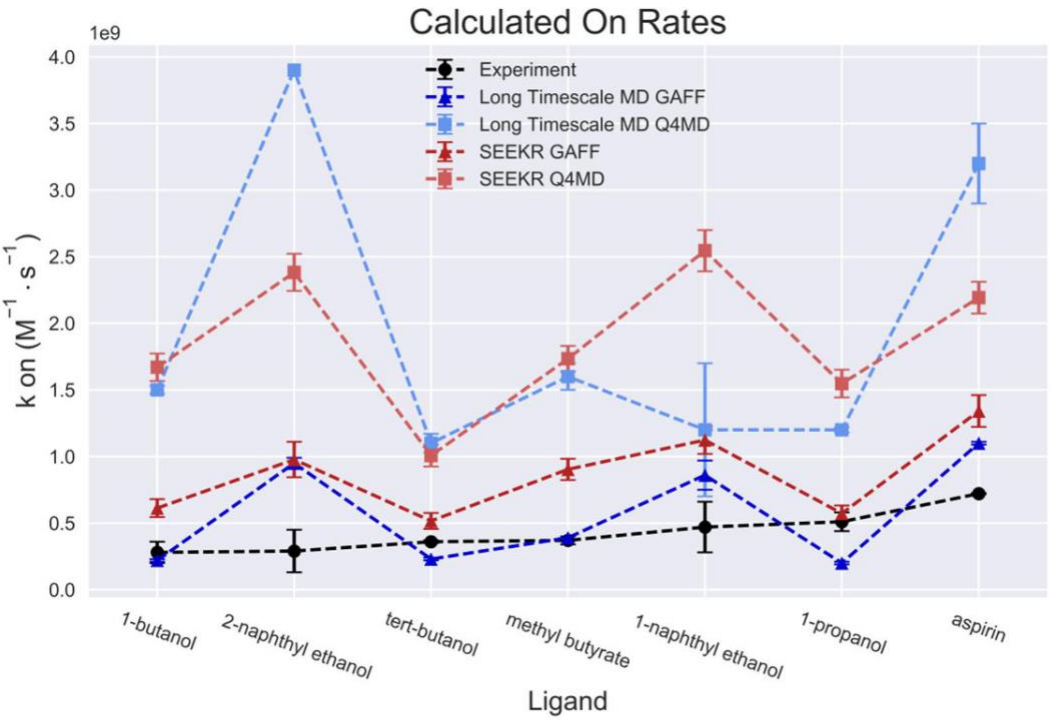


Figure 3.1 Structures for the *β*-cyclodextrin model system
a.) *β*-cyclodextrin with milestones spaced at 1.5 Å increments and b.) the seven ligands used in this study. The top four ligands are known to bind more weakly while the bottom three are known to bind more tightly.

However, similar qualitative results are seen with the SEEKR calculations and long timescale MD calculations using the same forcefield. On rates calculated using Q4MD are approximately one order of magnitude faster than experimental rates, while the GAFF forcefield produces rates closer to the experimental values, differing by approximately a factor of 3 or less. Both methods fail to effectively order the ligands by increasing k_{on}, as demonstrated



(a)

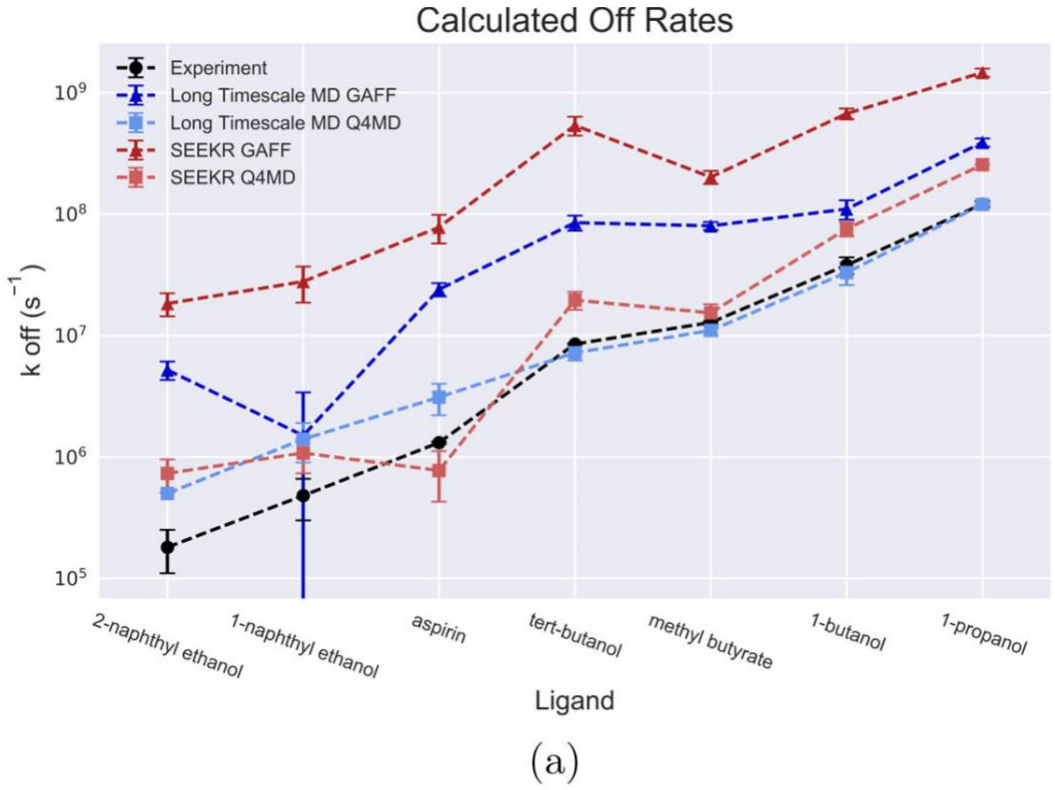
Method	Kendall	Spearman
SEEKR GAFF	-0.20 ± 0.31	-0.29 ± 0.40
SEEKR Q4MD	0.14 ± 0.29	0.14 ± 0.38
Long Timescale MD GAFF	0.24 ± 0.26	0.25 ± 0.29
Long Timescale MD Q4MD	0.00 ± 0.28	-0.05 ± 0.37

(b)

Figure 3.2 On rate results
a) Experimental and calculated on rates for SEEKR GAFF and Q4MD forcefields as well as long timescale MD with both forcefields. b) Calculated rank correlation coefficients. Errors are determined with a bootstrapping analysis.

by low or negative Kendall and Spearman rank correlation coefficients. As the values of all the experimental rates have limited variability (all within half an order of magnitude), the sensitivity of the methods as well as the errors associated with the calculations and experiments makes differentiation and ordering challenging.

Unlike the experimental k_{on}'s, k_{off}'s for the seven guest molecules span multiple orders of magnitude, making them a better target for ranking the compounds with SEEKR. Again, off rates calculated with SEEKR are in good agreement with the long timescale MD simulations using the same forcefield (Figure 3.3). Rates calculated using the GAFF forcefield are consistently faster than experiment by approximately one order of magnitude. This trend is seen in both



(a)

Method	Kendall	Spearman
SEEKR GAFF	0.90 ± 0.06	0.96 ± 0.04
SEEKR Q4MD	0.81 ± 0.09	0.93 ± 0.05
Long Timescale MD GAFF	0.81 ± 0.09	0.93 ± 0.04
Long Timescale MD Q4MD	1.00 ± 0.05	1.00 ± 0.03

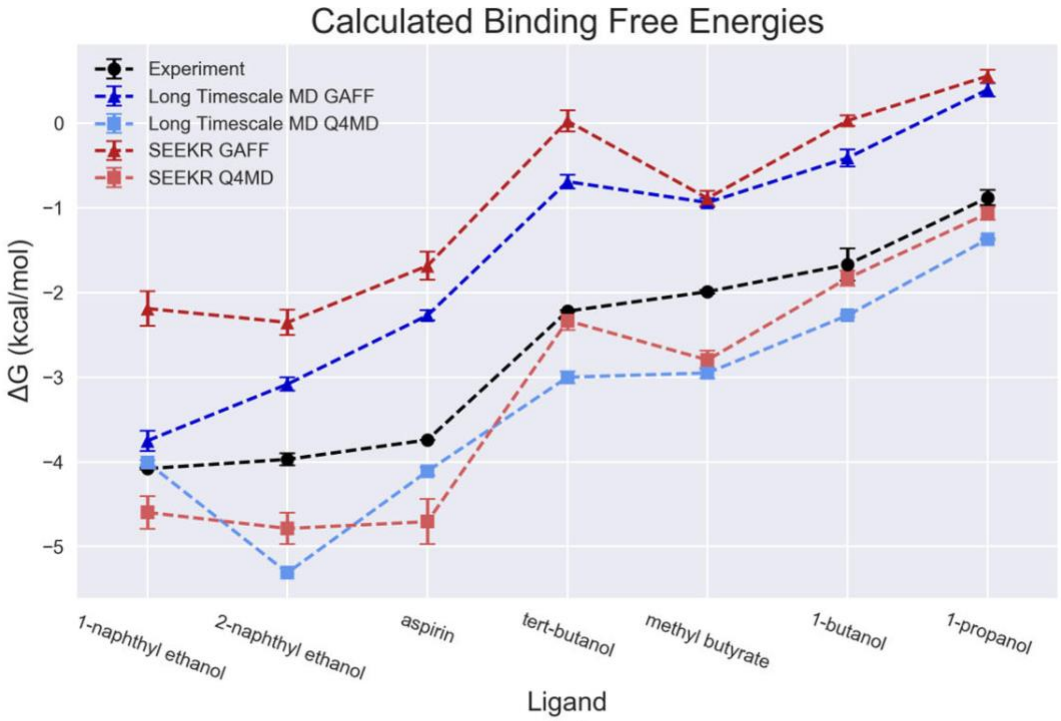
(b)

Figure 3.3 Off rate results
a) Experimental and calculated off rates for SEEKR GAFF and Q4MD forcefields as well as long timescale MD with both forcefields. b) Calculated rank correlation coefficients. Errors are determined with a bootstrapping analysis.

the long timescale MD and SEEKR, but is more pronounced in the SEEKR calculations. The Q4MD forcefield, however, more accurately reproduces the magnitude of the experimental values with both SEEKR and long timescale MD. SEEKR calculations with both Q4MD and GAFF forcefields were effective for ranking the compounds by increasing off rates, as evidenced by high rank correlation values. The smaller magnitudes of the Q4MD values potentially contribute to this forcefield's difficulty to differentiate between compounds with similar rates, where the larger values associated with the GAFF forcefield allow for more variability in the rate value without changing the overall ordering. Both the GAFF and Q4MD forcefields successfully differentiate the three tighter binding compounds from the four weaker binding, with the tighter binding compounds all having slower off rates and a difference of one order of magnitude between the fastest tightly binding compound and the slowest weakly binding compound. This suggests that SEEKR could be useful for identifying and separating long residence time ligands from shorter residence time ligands and then further discriminating the compounds through ranking by k_{off}.

An additional benefit of kinetics calculations with SEEKR is that binding free energies can also be obtained from the same simulations (Figure 3.4). Binding free energies calculated using the rate constants are most heavily influenced by the k_{off} for these ligands, as this value is more variable, where the k_{on}'s for all ligands are more similar. Therefore, similar trends are observed for the calculated binding free energies as were observed for the off rates. Binding free energies can also be calculated using the stationary probabilities for each milestone, rather than the rate constants, and produce similar results. The GAFF forcefield consistently underestimates the binding free energies in both SEEKR and the long timescale MD, resulting from the consistent underestimation of the magnitudes of the k_{off}'s. The magnitudes of the binding free energies calculated using Q4MD are in much better agreement with the experimental values, differing by 1

kcal or less. SEEKR with both Q4MD and GAFF successfully differentiates the three known tighter binding compounds from the four weaker binding compounds. SEEKR can also further discriminate ligands by its effective ranking by binding free energies, demonstrated by high rank correlation values.



(a)

Method	Kendall	Spearman
SEEKR GAFF	0.88 ± 0.08	0.96 ± 0.05
SEEKR Q4MD	0.73 ± 0.10	0.89 ± 0.06
Long Timescale MD GAFF	0.90 ± 0.07	0.96 ± 0.04
Long Timescale MD Q4MD	0.87 ± 0.11	0.94 ± 0.06

(b)

Figure 3.4 Binding free energy results
a) Experimental and calculated binding free energies for SEEKR GAFF and Q4MD forcefields as well as long timescale MD with both forcefields. b) Calculated rank correlation coefficients. Errors are determined with a bootstrapping analysis.

A key aspect of future development of the SEEKR software is the systematic development of methodological best-practices as well as the elucidation of the sensitivity of calculated kinetic

parameters to various SEEKR input conditions. While it is possible to determine the kinetics for small systems like β-cyclodextrin using conventional long timescale MD simulations, increasing system size soon makes this inefficient or even impossible. The convergence of k_{on} and k_{off} were assessed by calculating the rate constants as a function of the reversal trajectory number at increasing intervals of 50 reversal numbers (with 10 trajectories initiated for each reversal number). The reversal number is a direct measure of the equilibrium simulation length, as reversals are initiated from evenly spaced configurations of the equilibrium distribution. In general, both the on and off rates appear converged in less than the maximum number of reversals available. Approximately half the total reversals (4000 of 8000) were sufficient for obtaining reasonably converged rate constants. This suggests that the total simulation cost to obtain a similar result could be as little as 2 μs per ligand, rather than 3.8 μs.

Convergence of the rate constant is a highly complicated quantity dependent on the transition probabilities as well as the incubation times obtained from each milestone. Therefore, a more detailed analysis of the convergence of these quantities on a per milestone level can provide further insight into the overall convergence of a rate calculation within SEEKR. Figure 3.5a,b shows the convergence of k_{on} and k_{off}, respectively, as a function of the number of reversals launched for the representative system of Q4MD *β*-cyclodextrin with aspirin. Reversal number is directly related to the length of equilibrium sampling, the current bottleneck of a SEEKR calculation. While both values appear to converge in fewer than the maximum number of reversals, the dramatic change in k_{off} after reversal number 400 is of note. Further analysis of the per-milestone transition counts (Figure 3.5c) and incubation times (Figure 3.5d) revealed that this change in k_{off} was due to poor initial sampling of the -1.5 Å milestone, which once sampled decreased the overall k_{off}.

This was only observed for the aspirin ligand, one of the bulkiest ligands, where it was extremely unlikely to observe transitions outward from the primary face due to steric effects. Evaluation of these convergence properties on a per milestone basis is a valuable diagnostic tool; identifying which milestones contribute most to the mean first passage time (and therefore k_{on} and k_{off}) and milestones where the ligand spends only a short time, providing detailed molecular insight into the binding and unbinding processes. Furthermore, this analysis is also useful during the simulation process, as the convergence of each milestone can be assessed "on the fly" and individual simulations can be terminated or extended accordingly for each milestone.

We also explore the sensitivity of the calculated rate constants to the milestoning model construction. In particular, the appropriate spacing of milestones is critical for the calculation. Milestones must not be spaced so close such that the velocity of the system cannot decorrelate between transitions.[95,139] This assumption is typically valid for molecular dynamics simulations, as velocities typically decorrelate on the subpicosecond timescale.[46] However, if milestones are spaced far apart, transitions will require much longer simulations and milestioning sampling efficiency is lost.

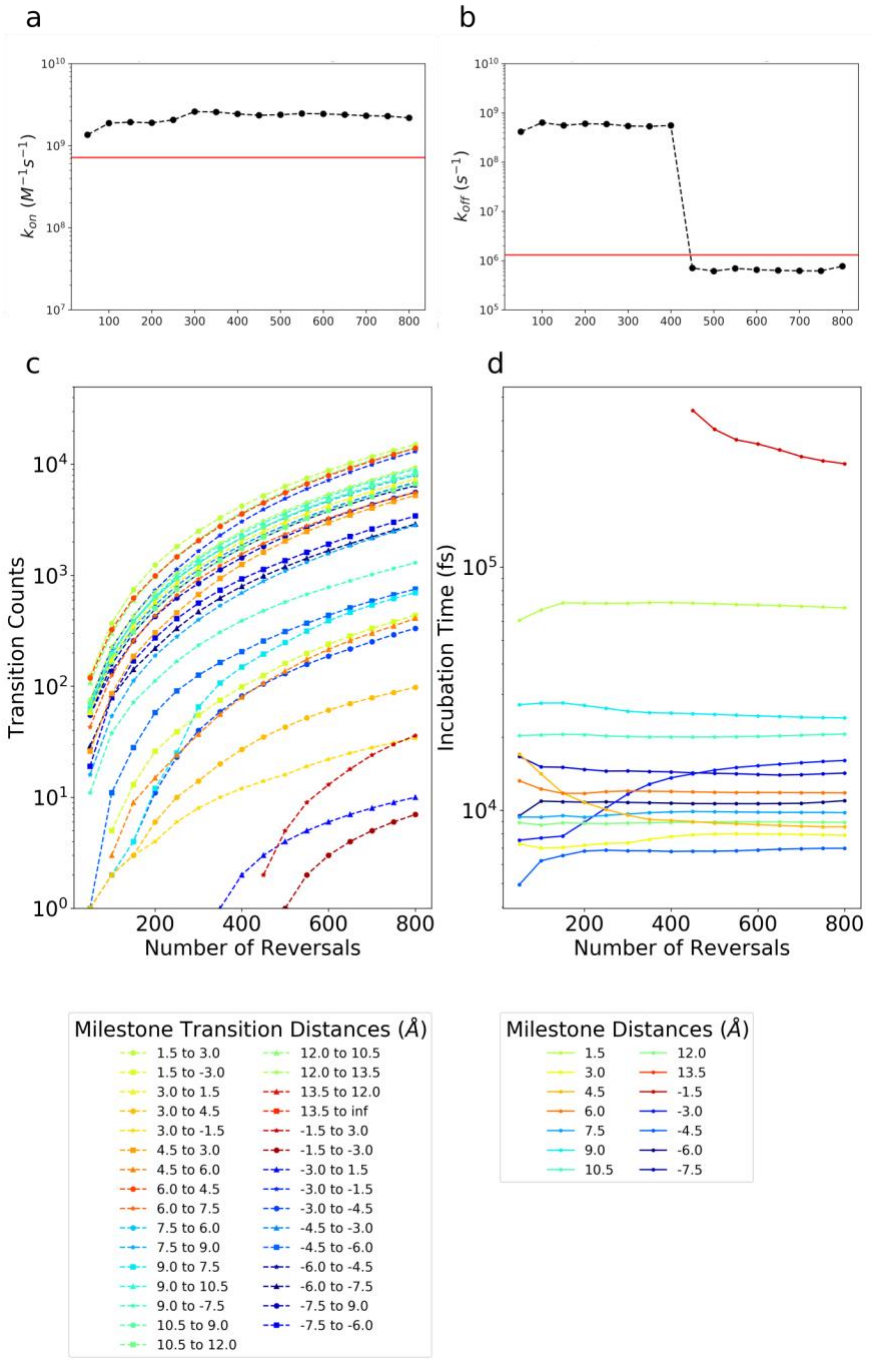


Figure 3.5 Convergence analysis for a representative ligand, aspirin, and β-cyclodextrin with the Q4MD forcefield.

Convergence of a) k_{on}, b) k_{off}, c) transition counts, and d) incubation times for each milestone are plotted as a function of the number of reversals.

For our systems, the incubation times of all milestones are on the order of multiple picoseconds or greater, which is longer than the sub-picosecond timescale typically necessary for decorrelation.[95] When the milestone spacing was doubled to 3 Å, simulation efficiency was dramatically reduced, such that few to no transitions between milestones were observed, precluding the calculation of rate constants. These observations suggest that the 1.5 Å spacing used in our simulations was appropriate for the calculation of the desired kinetic parameters.

Our milestoning model differentiates the two faces of the cyclodextrin ring and therefore defines two bound states, corresponding to each face. Investigation into the effect of this on the resulting rate constants revealed that it had only minimal effects. When the two bound states were combined into a single milestone, only small changes to the rate were observed, within the error of both calculations. Furthermore, when the two faces were not differentiated with unique milestones, minimal change in the calculated rate constants was observed.

It is also important to note that our milestoning model did not explicitly resolve the ligand orientation in any way, and therefore any ligand orientational sampling was achieved entirely through simulation. This resulted in some ligand orientations being unsampled in the deepest milestones where the orientation was sterically restricted to the starting conformation on that milestone. While this is a limitation that will be addressed in future developments of SEEKR, it also highlights that a relatively simplistic model was able effectively calculate kinetic parameters with good agreement to experimental values.

The simplicity of this model has many advantages. The bound state is defined naturally as the innermost milestone and all other milestones can be defined at the same time, including what defines the unbound state. The long timescale MD employed a more empirical definition of the bound state where the ligand was only considered bound when the COM of the ligand was within

7.5 Å of the COM of the β-cyclodextrin for at least 1.0 ns. Similarly the ligand was considered unbound when it left this 7.5 Å bound state for at least 1.0 ns. With the SEEKR approach, minimal prior knowledge of the system is required, as binding and unbinding are determined only from the milestone surfaces. No time cutoff is required, as short excursions that do not result in full binding and unbinding events are captured naturally in the milestoning model. The simplicity of SEEKR milestoning calculation setup, in conjunction with the ability to monitor convergence for each milestone and terminate simulations accordingly, makes this approach well-suited for calculations with multiple ligands as would be necessary in a drug discovery setting.

We present the first successful ranking of a set of seven guest molecules with the *β*cyclodextrin host using the SEEKR hybrid MD/BD/Milestoning approach. SEEKR effectively reproduces both the magnitudes and rankings of the experimental off rates[143–148] and binding free energies, two quantities of interest in typical drug discovery campaigns.[66,113–115] SEEKR also successfully differentiates the known longer residence time and tighter binding compounds from the weaker binding compounds. Our results are also in good agreement with previously conducted long timescale MD simulations for the same set of ligands.[121] In particular SEEKR and long timescale MD simulations using the same forcefield (GAFF or Q4MD) exhibited similar deviations from the experimental values, with GAFF producing consistently faster off rates than experiment and Q4MD producing consistently faster on rates. In general both methods and both forcefields struggled to reproduce the experimental on rate ranking, as all ligands had very similar k_{on}'s. The SEEKR method requires less simulation time (3.8 μs per ligand) than the long timescale MD approach (4.5 - 11 μs per ligand). Furthermore, convergence analysis of the SEEKR calculations suggests that comparable results could be achieved with as little as 2 μs per ligand. In addition, SEEKR's milestoning approach makes these calculations highly parallel, as the

simulations on each milestone are completely independent from all other milestones. We also provide an analysis of the sensitivity of the SEEKR calculations to various factors such as sampling, milestone spacing, and the construction of the milestoning model with the intention of putting forth "best practices" for the use of SEEKR. SEEKR's effectiveness at ranking compounds for this small model system suggest that it is well suited for ranking compounds of more complex protein-ligand and protein-drug systems, where the efficiency and enhanced sampling advantages of our multiscale MD/BD/milestoning approach will be more apparent.

3.3 Acknowledgement

We thank Zhiye Tang and Chia-en Chang for sharing structures and parameters as well as helpful discussions. This work was funded in part by the Director's New Innovator Award Program NIH DP2-OD007237, the National Bio- medical Computation Resource (NBCR) NIH P41-GM103426, and the National Science Foundation through XSEDE supercomputing resources provided via TG- CHE060073 to R.E.A. B.R.J. and C.T.L. also acknowledge support from the NIH Molecular Biophysics Training Program (T32-GM008326).

Chapter 3 , in full, is a reprint of the material as it appears in: "Jagger, B. R.; Lee, C. T.; Amaro, R. E. Quantitative Ranking of Ligand Binding Kinetics with a Multiscale Milestoning Simulation Approach. *J. Phys. Chem. Lett.* 2018, *9* (17), 4941–4948." The dissertation author was a primary investigator and author of this work.

3.4 Supporting Information

3.4.1 System Preparation

GAFF[110,111] forcefield parameters for the seven guest molecule along with both GAFF and Q4MD-CD[142] parameterizations of β-cyclodextrin were obtained from Tang and Chang.[121] For comparison we use identical structure and parameterizations as those used in their study. These initial structures were used by the SEEKR software for preparation of the milestoning simulations. The preparation procedure was the same for each of the seven guest molecules and followed standard SEEKR protocols.[51] All systems were solvated with TIP3P waters.[149]

3.4.2 Preparation of milestoning simulations with SEEKR

The bound state of the host-guest complex was defined as the center of mass (COM) of the *β*-cyclodextrin. The guest molecule was considered to be bound when its COM was within 1.5 Å of the bound state coordinates. From this bound state, spherical milestones were defined in increasing 1.5 Å increments from 1.5 Å to 13.5 Å. Furthermore, spherical milestones of radius 7.5 Å and less, were divided into two half-spheres to better capture the asymmetries between the two faces. When the ligand is less than 7.5 Å away from the bound state, it is trapped on a particular face due to the size of the host molecule. For milestone distances greater than 7.5 Å, the ligand was found to freely sample both faces, and therefore a single spherical milestone was sufficient for sampling host-guest interactions. In total, 14 unique milestones were defined. In practice, this was achieved via post-processing the simulations on each face to identify any trajectories that crossed from one face to the other, and modifying the transitions accordingly in the milestoning model. In addition, two simulations were conducted for the outermost milestones (one with the ligand initiated on each face), and these were then combined into a single milestone (with double the sampling) for milestones that were not restricted to a particular face.

The first 13 milestones correspond to the MD region, while the 14th and outermost milestone corresponds to the BD region. The standard SEEKR preparation protocol[51] was then used to generate the coordinate, parameter, and simulation files necessary for a milestoning calculation. For each of the MD milestones, a copy of the apo β-cyclodextrin structure was generated and the guest molecule was then placed at the appropriate radius from the bound state coordinates. Any water molecules that clashed with the guest molecule were removed. The guest distribution for the BD milestone was constructed by first running a conventional BD simulation where trajectories terminated at the appropriate distance for the milestone surface (13.5 Å).

3.4.3 MD Simulations

A modified version of NAMD 2.12 was used for all MD simulations.[98] For all 13 milestones in the MD region, the standard SEEKR procedure for minimization, equilibration and simulation was followed. First, 5000 steps of minimization were performed to allow for relaxation, particularly of solvent, around the newly placed guest molecule. Further relaxation of the solvent was achieved by a series of 2 ps heating simulations that gradually increased the temperature from 298 K to 350 K and then cooled back to 298 K. Host and guest atoms were constrained during these heating simulations to ensure that the guest remained on the appropriate milestone surface. To obtain the equilibrium distribution of the guest molecule on each milestone, 200 ns of constant volume simulation was performed. A 90$\text{kcal} \cdot \text{mol}^{-1} \cdot \text{Å}^{-2}$ harmonic restraint was used to hold the COM of the guest molecule at the appropriate distance from the binding site. To minimize any bias of the arbitrary guest starting conformation, the first 40 ns of each simulation were discarded and therefore were not included as part of the equilibrium distribution. From these trajectories, position and velocity configurations were selected every 0.2 ns, resulting in a total of 800 configurations per milestone. To obtain the first hitting point distribution (FHPD) of each

milestone, 10 independent and unrestrained simulations were initiated from these equilibrium configurations. Each simulation was propagated backwards in time by reversing its velocity at constant energy and volume (a total of 8000 reversals for each milestone). Only trajectories that struck an adjacent milestone before recrossing the milestone on which they originated were included as part of the FHPD for that milestone. To obtain the transition probabilities and times necessary for the calculation of kinetic parameters, all members of the FHPD were brought back to their starting position and velocity and new unrestrained simulations were then initiated from each configuration. These simulations were propagated forward in time at constant energy and volume. Once a simulation crossed its starting milestone again, it was monitored for crossing of adjacent milestones. When an adjacent milestone was crossed, the simulation was stopped and the transition and incubation time were recorded. Although more of the equilibrium simulation trajectories could likely have been used without biasing the results based on the starting conformation, the 8000 reversal trajectories that resulted from the 160 ns used were more than sufficient to sample the transitions between the milestones, resulting in hundreds of observed transitions between each milestone. In total, 2.6 μs of equilibrium sampling (160 ns for 16 milestones) were used and approximately 570 ns of FHPD sampling for a total of 3.2 μs of simulation used in the milestoning model. The total cost per ligand (including simulation discarded for equilibration) was therefore ~3.8 μs.

3.4.4 BD Simulations

All BD simulations were performed using the BrownDye software package.[99] Electrostatic potentials of the host and guest molecules used as inputs for the BD simulation were calculated with APBS version 1.4.[100] To match experimental conditions, APBS calculations and BD simulations were carried out with a solvent dielectric of 78, a solute dielectric of 2, and zero ionic

concentration. An initial series of simulations were initiated from the b-surface, a sphere that encloses the entire host molecule and has sufficient radius for the guest molecule to be situated in bulk solvent such that forces between the host and guest are centrosymmetric. 10^6 independent BD simulations were initiated from random points on the b-surface and were propagated until the guest either contacted the outermost milestone (13.5 Å) or escaped. Trajectories that successfully contacted the 13.5 Å milestone were used as the FHPD for this milestone. Another series of 10^6 BD simulations were initiated from the FHPD and propagated until contacting the second-outermost milestone (12.0 Å) or escaping to the q-surface. This procedure is automated by SEEKR.

3.4.5 Milestoning Calculations

Statistics from all milestones in the MD and BD regions were extracted using SEEKR and combined to construct a transition kernel as well as an incubation time vector. These are the two key quantities for the calculation of kinetic parameters in milestoning theory.[94] As described previously, a post-simulation analysis was performed to account for ligand transitions between the two faces of the β-cyclodextrin ring for milestones of radius 7.5 Å or less. The analysis portion of SEEKR was then used to compute the desired kinetic quantities, k_{on} and k_{off}, as well as the free energy of binding ΔG_{bind}.

3.4.6 Milestone Convergence Plots

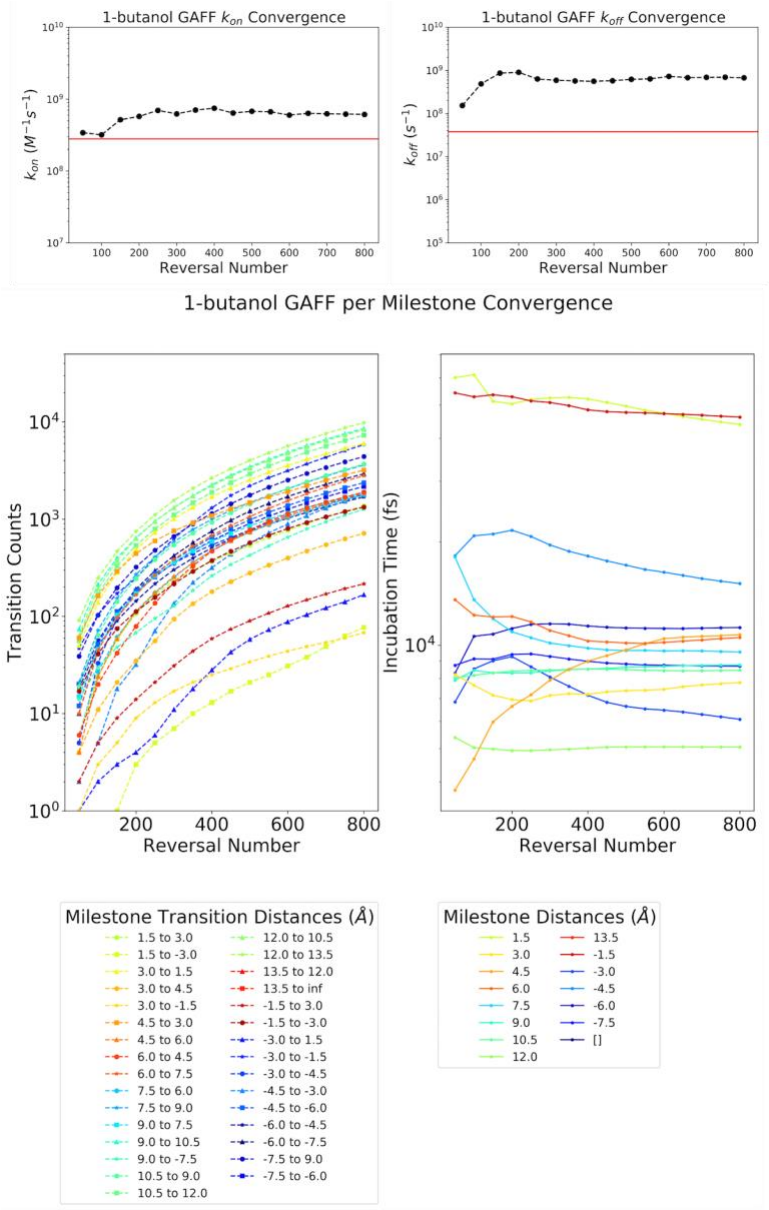


Figure 3.6 1-butanol GAFF per milestone convergence plot

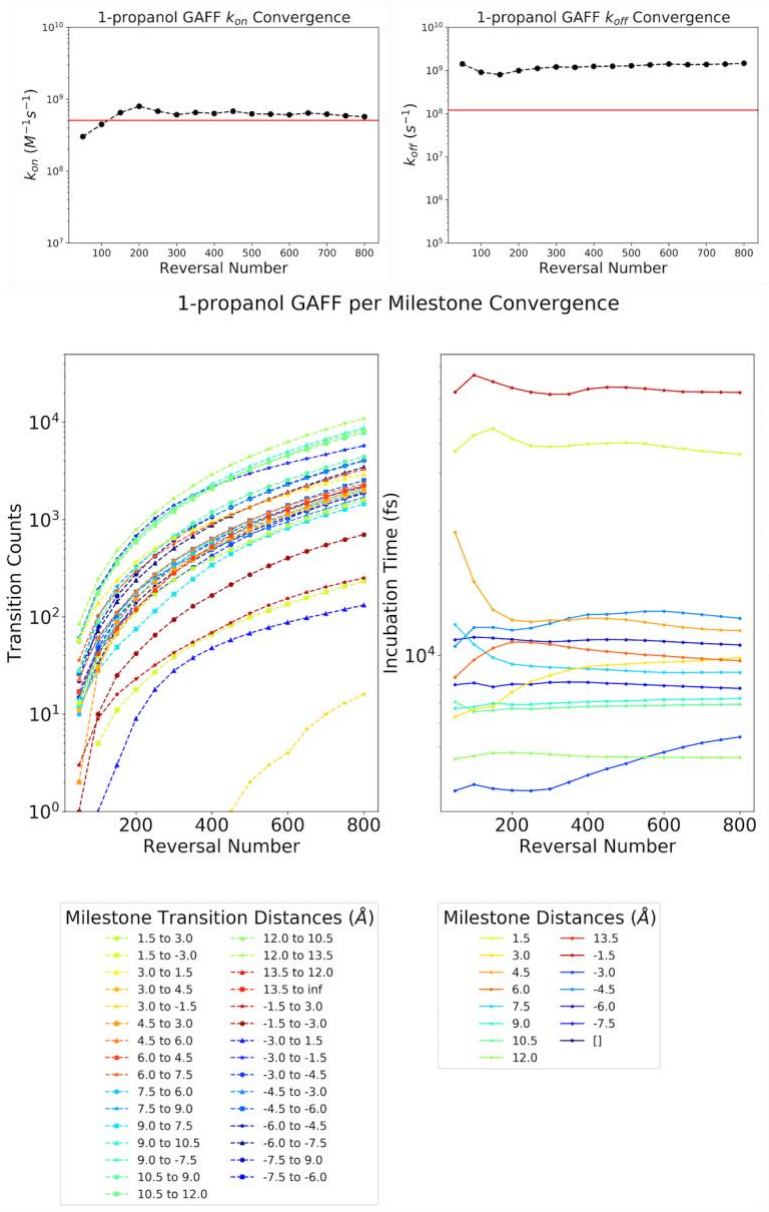


Figure 3.7 1-propanol GAFF per milestone convergence plot

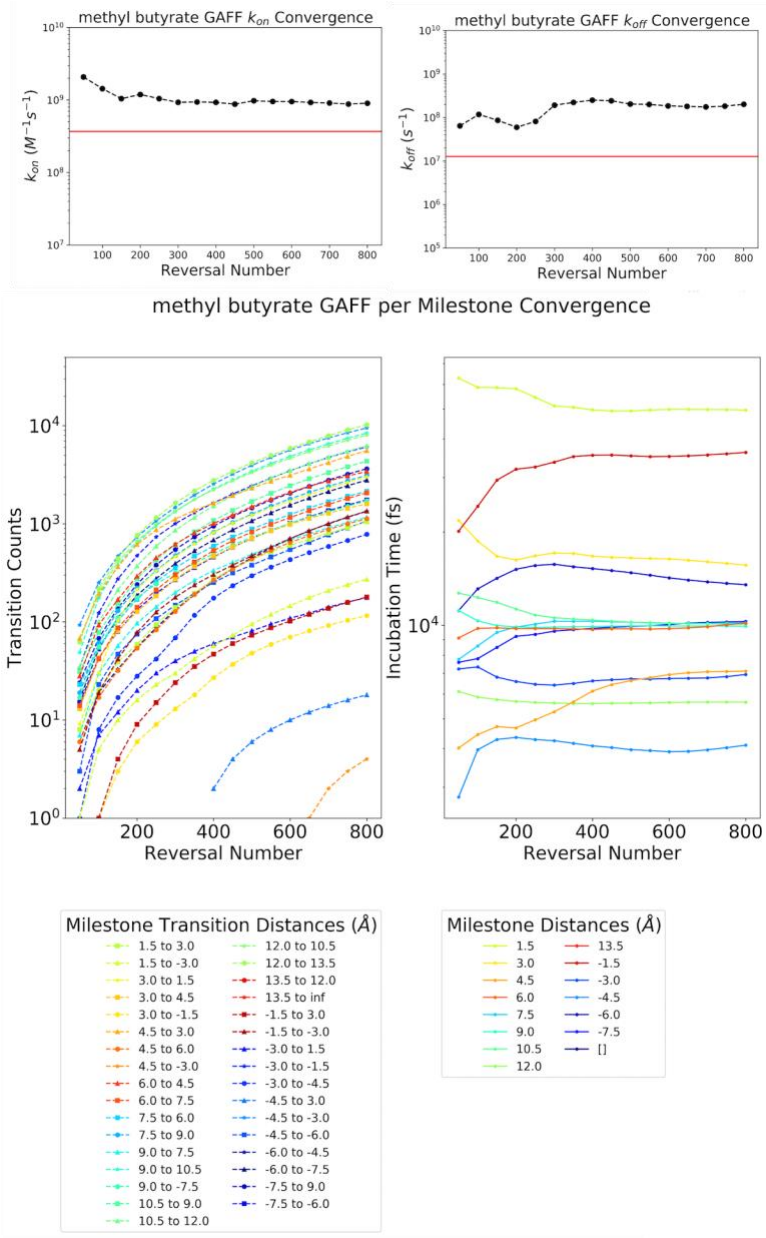


Figure 3.8 methyl butyrate GAFF per milestone convergence plot

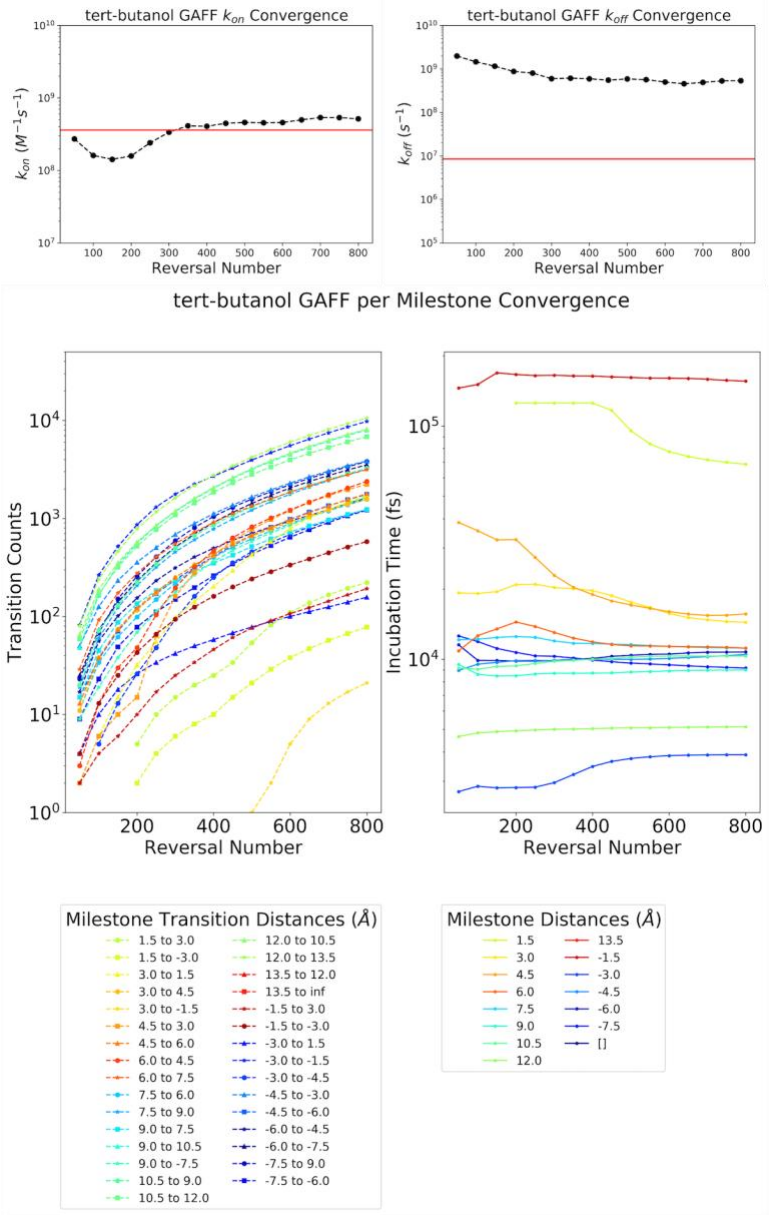


Figure 3.9 tert butanol GAFF per milestone convergence plot

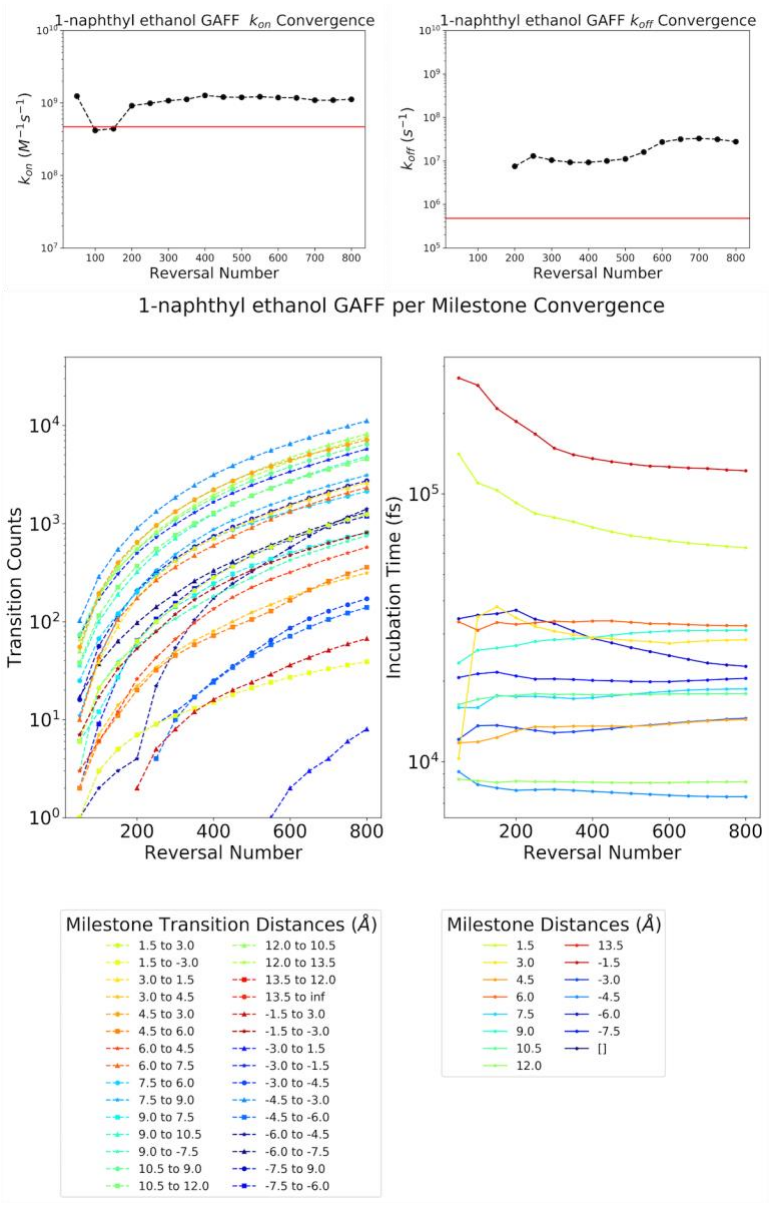


Figure 3.10 1-naphthyl ethanol GAFF per milestone convergence plot

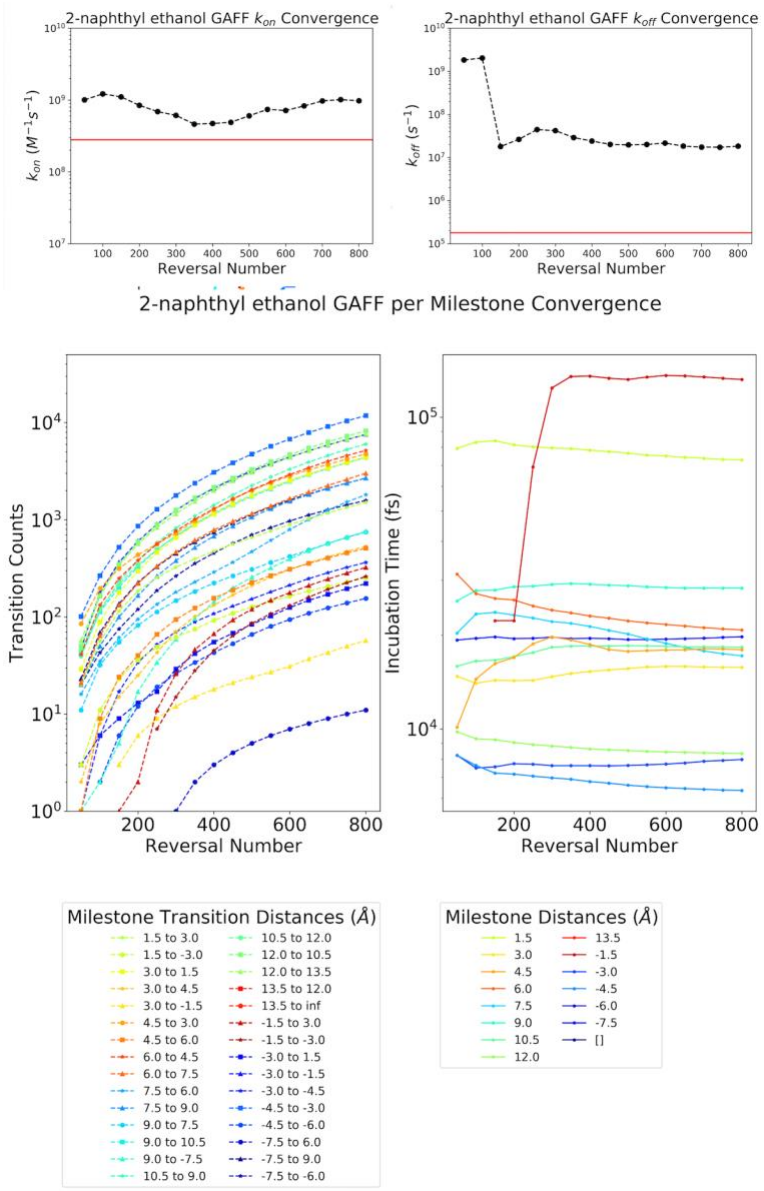


Figure 3.11 2-naphthyl ethanol GAFF per milestone convergence plot

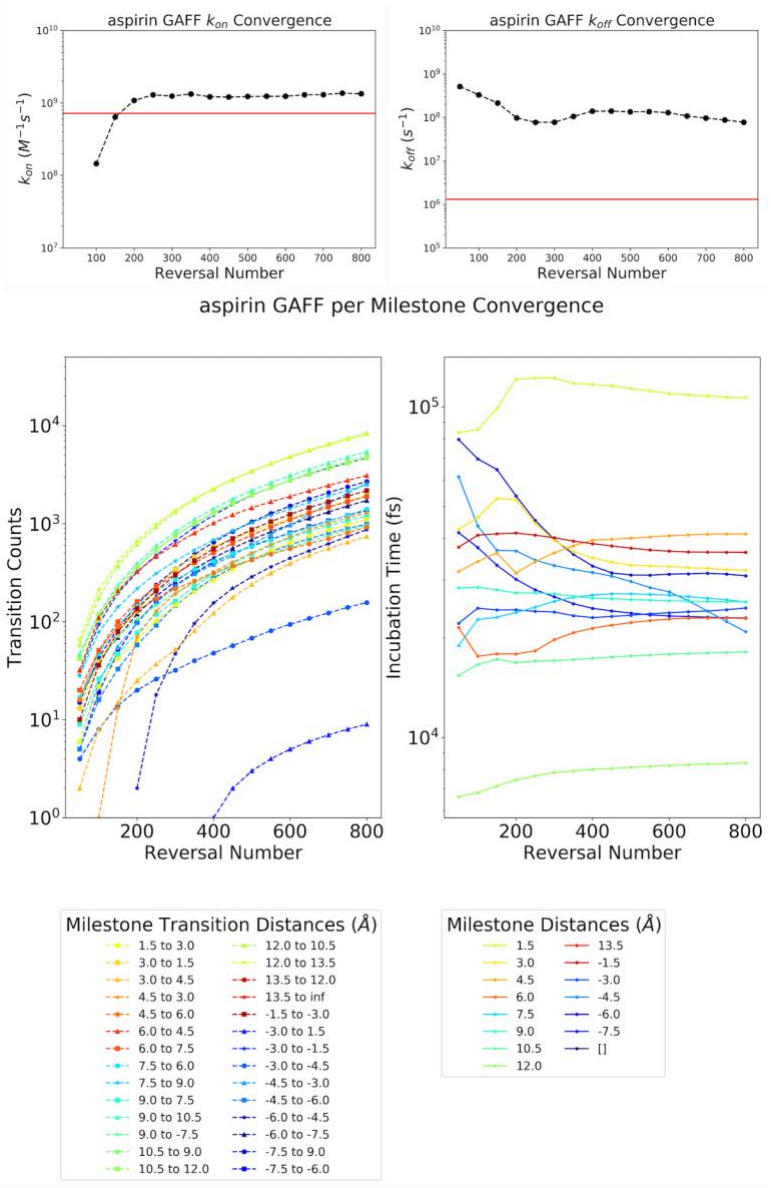


Figure 3.12 aspirin GAFF per milestone convergence plot

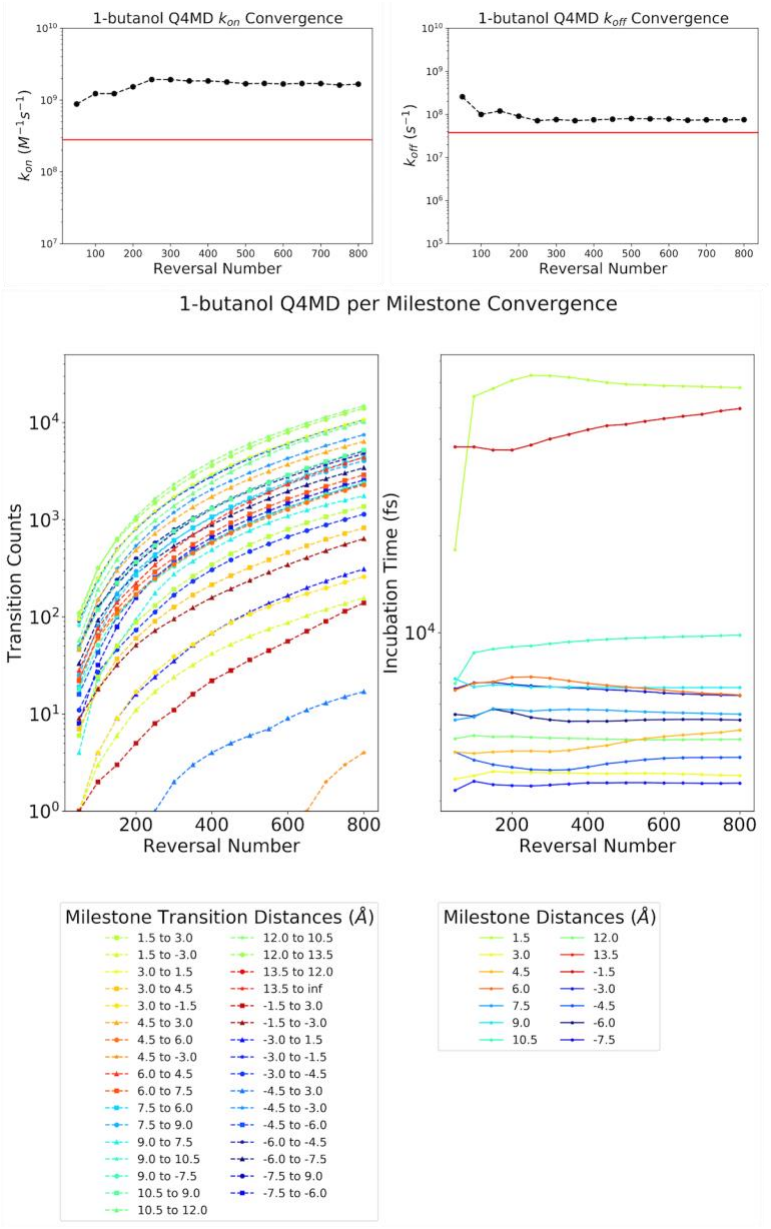


Figure 3.13 1-butanol Q4MD per milestone convergence plot

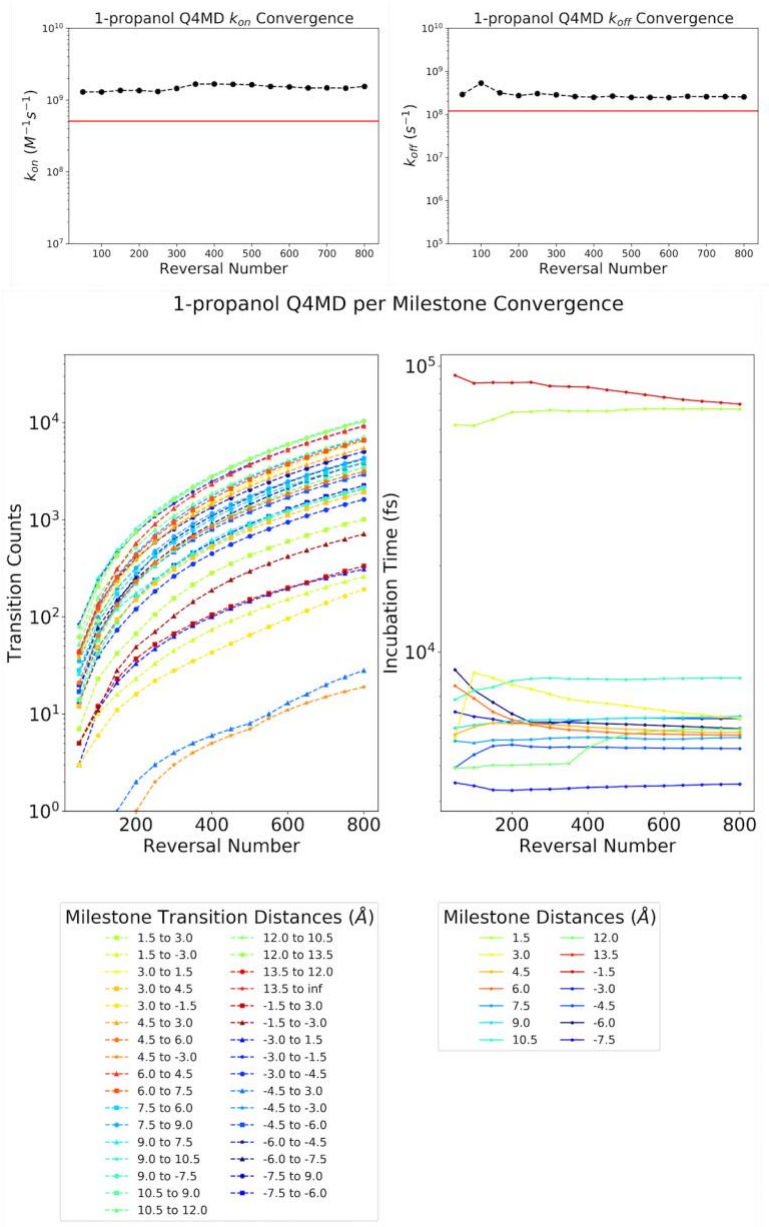


Figure 3.14 1-propanol Q4MD per milestone convergence plot

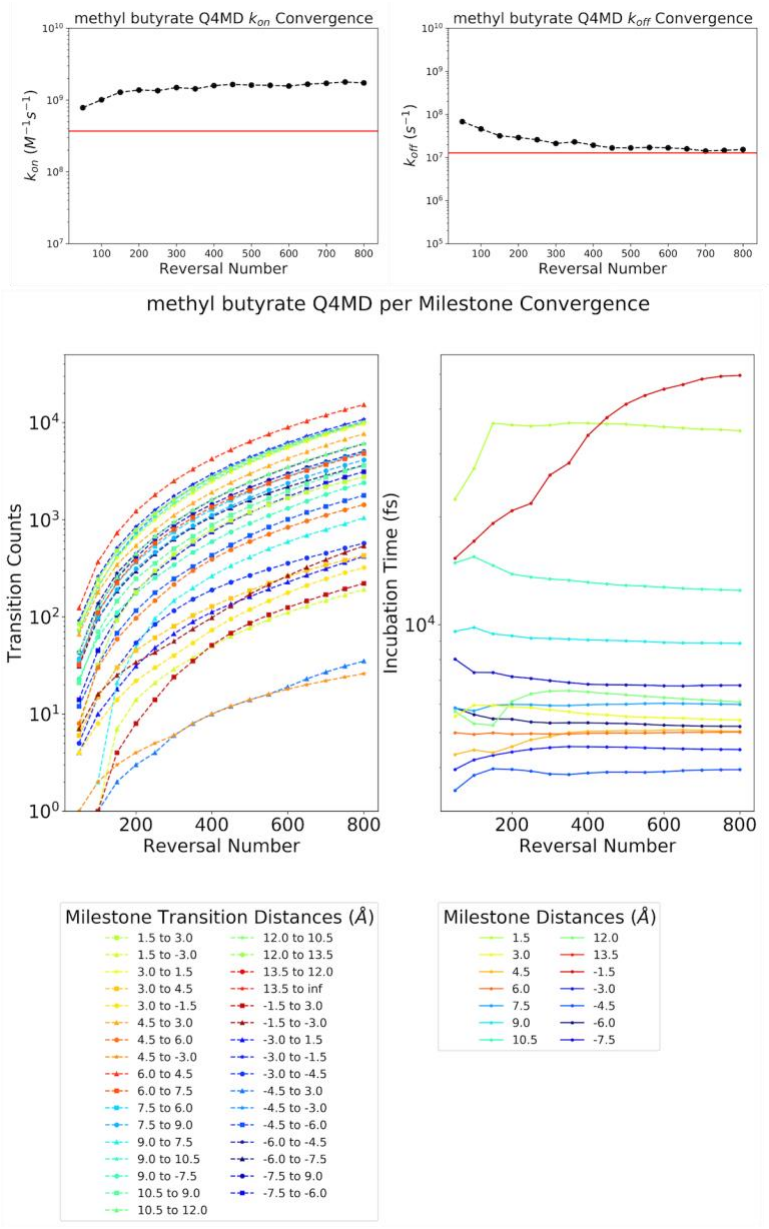


Figure 3.15 methyl butyrate Q4MD per milestone convergence plot

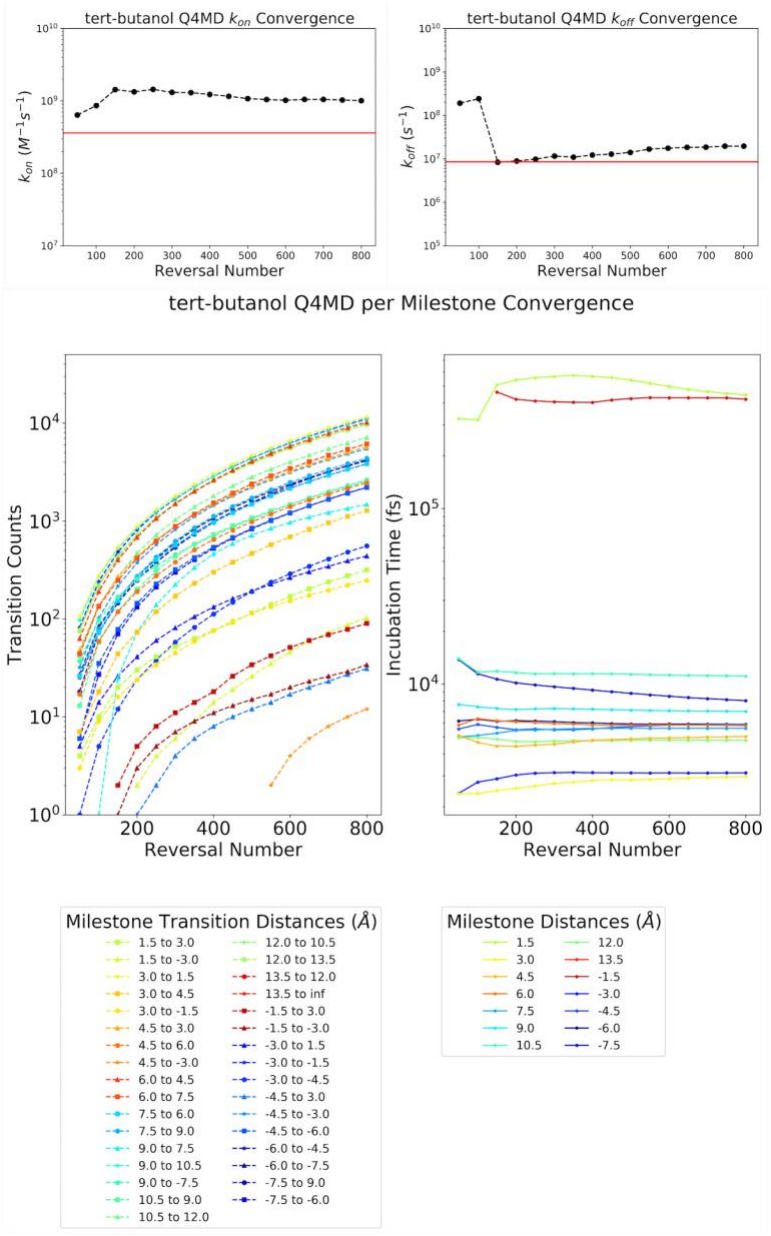


Figure 3.16 tert butanol Q4MD per milestone convergence plot

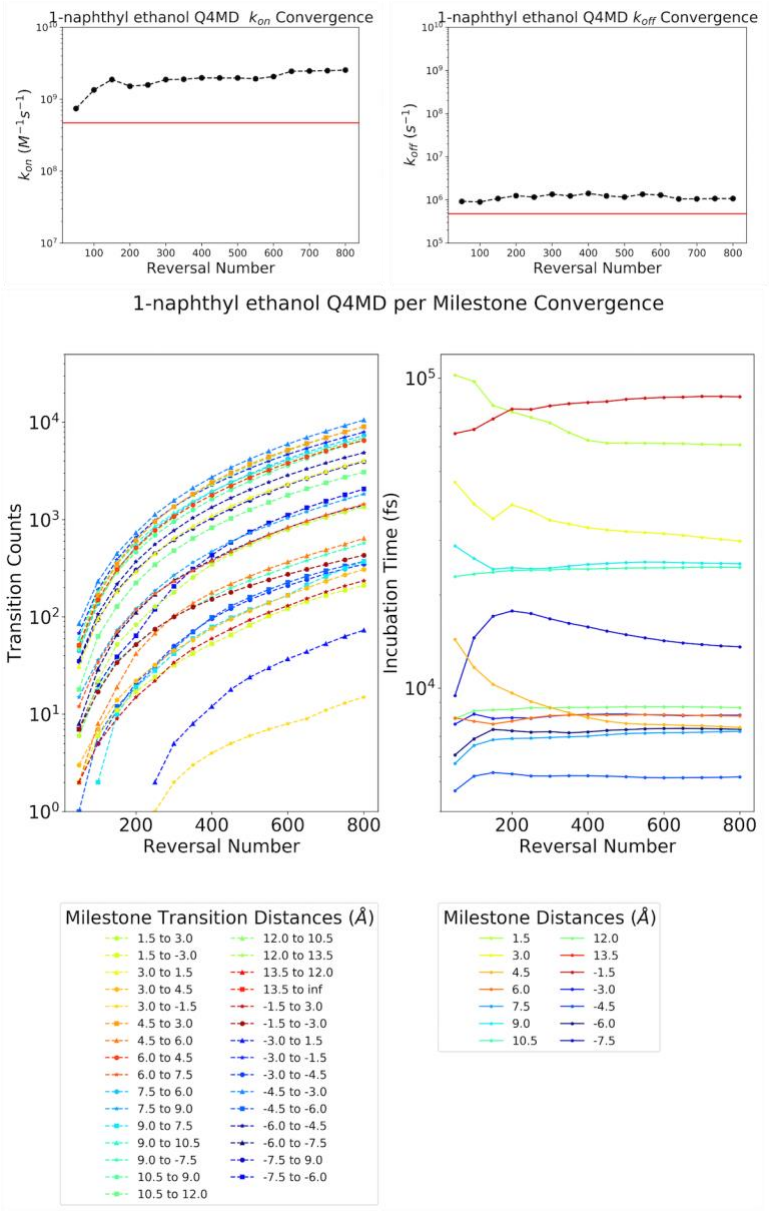


Figure 3.17 1-naphthyl ethanol Q4MD per milestone convergence plot

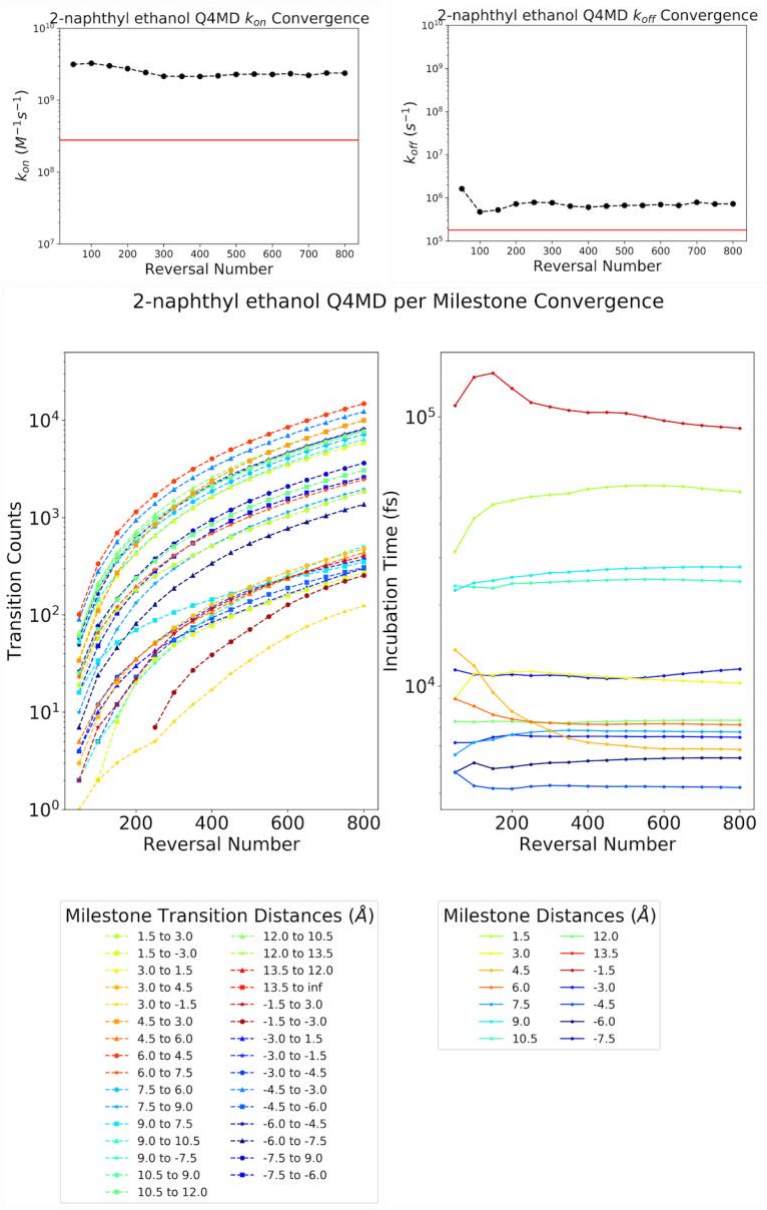


Figure 3.18 2-naphthyl ethanol Q4MD per milestone convergence plot

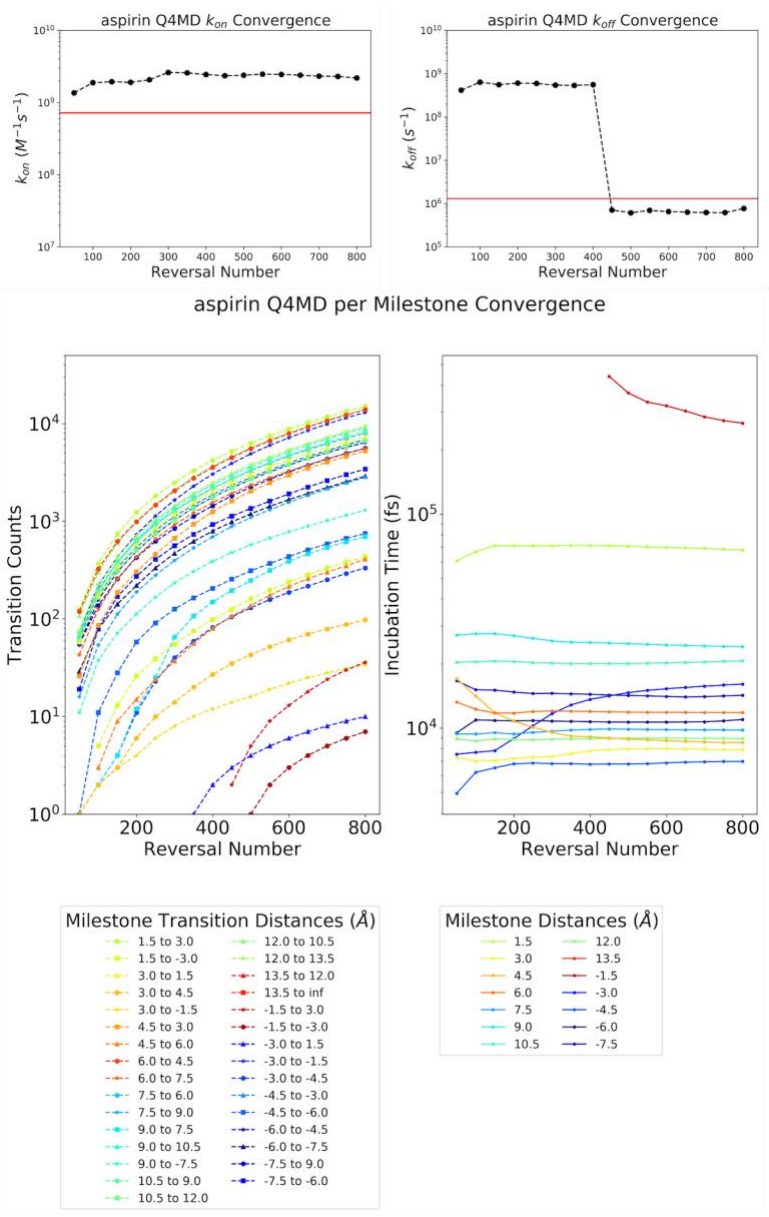


Figure 3.19 aspirin Q4MD per milestone convergence plot

Chapter 4

Predicting Ligand Binding Kinetics Using a Markovian Milestoning with Voronoi Tessellation Multiscale Approach

4.1 Abstract

Accurate and efficient computational predictions of ligand binding kinetics can be useful to inform drug discovery campaigns, particularly in the screening and lead optimization phases. Simulation Enabled Estimation of Kinetic Rates, SEEKR, is a multiscale molecular dynamics, Brownian dynamics, and milestoning simulation approach for calculating receptor-ligand association and dissociation rates. Here we present the implementation of a Markovian milestoning with Voronoi tessellations approach that significantly reduces the simulation cost of calculations as well as further improving their parallelizability. The new approach is applied to a host-guest system to assess its effectiveness for rank-ordering compounds by kinetic rates and to the model protein system, trypsin, with the noncovalent inhibitor benzamidine. For both applications, we demonstrate that the new approach requires up to a factor of 10 less simulation time to achieve results with comparable or increased accuracy.

4.2 Introduction

Historically drug discovery campaigns have focused on equilibrium metrics, such and binding affinity, to inform screening and lead optimization of prospective compounds. However, kinetic parameters of binding, such and the on rate (k_{on}) and the off rate (k_{off}), are receiving increased attention as effective predictors of a compound's *in vivo* efficacy.[113,116] Of particular interest is the residence time ($1/k_{off}$) of compounds, which accounts for the effects of protein conformational flexibility on binding, unbinding, and rebinding as well as other factors.[66,114,115,150] Furthermore, the Kinetics for Drug Discovery Consortium database reports that only 0.4% of compounds uploaded with experimentally measured kinetics have diffusion-controlled association rate constants, suggesting k_{on} may also be an informative parameter to aid in lead optimization and the prediction of efficacy.[113] Ligand binding kinetics (k_{on} and k_{off}) are determined by a combination of effects such as: protein conformational flexibility, ligand induced receptor conformational changes, binding site water rearrangements, and drug rebinding, all of which influence the potency as well as selectivity of prospective compounds. Multiple compounds can have the same equilibrium binding affinity, yet corresponding values of k_{on} and k_{off} can vary by orders of magnitude. The additional level of detail afforded by knowledge of both the association and dissociation rate can therefore be critical for rationalizing why some compounds have efficacy, while others do not, aiding the lead optimization effort and reducing the high attrition rates currently associated with lack of *in vivo* efficacy .

Computational binding kinetics predictions have the potential to reduce the time and cost associated with experimental synthesis, assay development, and testing of many candidate compounds.[1,151] In particular, molecular simulation approaches are attractive for the structural, dynamical, and mechanistic insights they can provide of the drug binding/unbinding pathways in

addition to predicting rate constants.[5] Brownian dynamics (BD) simulations are routinely used to efficiently estimate protein-ligand association rates and identify binding pathways.[42,85,87,93,152] Atomistic molecular dynamics (MD) simulations can also be used to study ligand binding and unbinding, however the increased model complexity necessitated by MD makes it limited by sampling. Hardware and software improvements such as exascale computing, the Anton super computer, increasingly powerful graphical processing units (GPUs), and volunteer distributed computing have made the study of binding kinetics with brute-force type approaches possible.[8,70,72,73,81,118,119,153] Generally, these approaches are limited to a small number of compounds and observe only a few association events and no dissociation events. Brute force MD simulations can access timescales on the order of milliseconds; however drug molecules often have residence times on the order of seconds, minutes, or even longer. As such, sampling remains the foremost limitation for these approaches. Furthermore, for simulation-based techniques to be useful in a drug discovery campaign, they must be able to provide predictions for 10s-100s of compounds in a reasonable timeframe. To overcome these challenges, many MD-based approaches have been developed that utilize biasing forces or other statistical mechanical techniques to access the timescales needed to predict binding and unbinding kinetics.[5,29] These include methods such as Markov State Models (MSMs), [31,76,81,122,123,154] metadynamics,[19,25,84,124,125,155] milestoning,[35,36,121,126,127,156] and others.[18,83,129–131,157] Additionally, multiscale methods exist that integrate MD with other approaches such as quantum mechanics or BD, or continuum approaches to better predict kinetic parameters by improving either accuracy or scalability.[14,45,53,61,117,158,159]

One such multiscale approach is the MD/BD/milestoning methodology "Simulation Enabled Estimation of Kinetic Rates" (SEEKR) which we develop and have shown to be effective

for the calculation of both k_{on} and k_{off} as well as the rank ordering of compounds by their rates.[50–52] Milestoning theory facilitates the division of simulation space into smaller regions called milestones that can be simulated independently and in parallel.[94,95,139,160] SEEKR uses atomistic, fully flexible MD simulations for milestones close to the binding site where these interactions are critical for describing the binding/unbinding process. Rigid body BD simulations are used in regions far from the binding site to dramatically reduce the computational cost, while still providing a sufficient description of the binding process, which is primarily diffusive in these regions. SEEKR is a freely available software package that automates the preparation, simulation and analysis of these binding kinetics calculations using the existing software NAMD[98] for MD simulations and Browndye[99] for BD simulations. While the effectiveness of SEEKR was previously demonstrated for predicting kinetic rates as well as rank ordering compounds, these calculations required a significant computational cost that would make the screening of many compounds challenging. It was therefore necessary to develop improvements to this methodology to reduce the amount of MD simulation required as well as improve the parallelizability of calculations.

Here we present a new implementation of SEEKR which utilizes the theory of Markovian Milestoning with Voronoi Tesselations (MMVT).[140,161] This new approach overcomes the primary sampling bottleneck associated with our previous implementation; obtaining an equilibrium distribution on each milestone. Instead, trajectories are confined to a Voronoi cell with the use of a reflective boundary condition. Figure 4.1 shows a general schematic of an MMVT SEEKR model which combines MD and BD simulations.

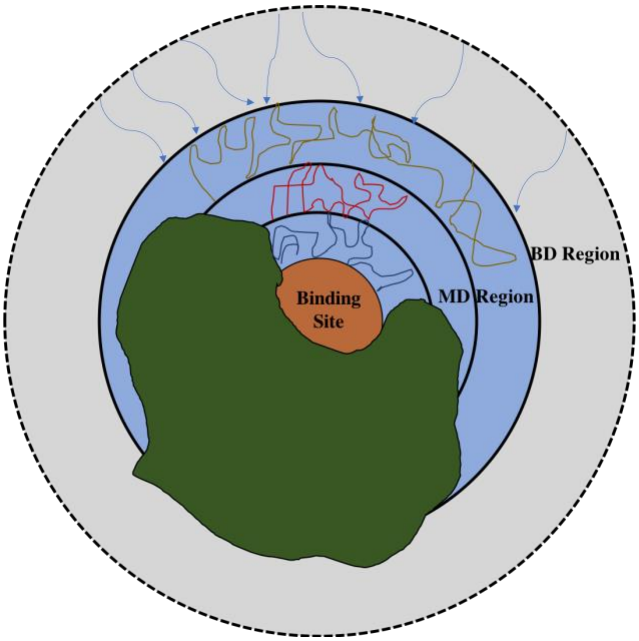


Figure 4.1 Cartoon depiction of a MMVT SEEKR rate calculation using spherical milestones representing radial distances from the binding site (black circles).
The blue shaded regions are treated with MD simulations, while the grey shaded region employs computationally less expensive BD simulations. MD trajectories (colored lines) are confined to a particular cell with the use of a reflective boundary condition when a milestone is touched. Many BD trajectories (blue arrows) efficiently simulate the association of the ligand from large distances. Milestoning theory enables the statistics from many independent cells and both simulation modalities to be combined for the calculation of binding and unbinding rates.

We test this new implementation on a model host-guest system: β-cyclodextrin with seven small molecule ligands, as well as the model protein system: trypsin with the noncovalent inhibitor benzamidine. The accuracy and efficiency of the MMVT SEEKR results are directly compared to experimentally measured kinetics, the previous SEEKR implementation, and other simulation approaches for each system. MMVT SEEKR produces results that are in agreement with experimental measurements and comparable to the previous SEEKR implementation for both model systems, while benefiting from up to a 10-fold reduction in simulation cost. Finally, we discuss convergence estimates for the sampling of each milestone as a way to further reduce the simulation cost by adaptively terminating or extending individual simulations.

4.3 Methods

4.3.1 MMVT SEEKR package

The MMVT SEEKR package is a series of python scripts (python 3.7 or later), freely available on Github, that automates the preparation, running, and analysis of all simulations necessary for ligand binding kinetics calculations. MMVT SEEKR utilizes user-defined inputs of structures and model parameters to generate all files necessary for a SEEKR calculation. Files are organized into a filetree with branches for each independent milestone. MMVT SEEKR uses the freely available softwares NAMD[98] for MD simulations and Browndye[99] for BD simulations and generates all necessary input files for running these simulations in the appropriate portions of the filetree. MMVT SEEKR uses the Colvars collective variable module of NAMD to define and monitor milestones during MD simulation, with any collective variable defined in the module able to be used for milestoning in MMVT SEEKR.[162] The appropriate colvar input files are created by the SEEKR preparation scripts. The SEEKR package also includes an analysis module containing functions to extract results from the simulation outputs, calculate rates, assess simulation convergence, perform error analysis, and easily plot relevant quantities. The module is designed to be imported into a Jupyter notebook, and a sample notebook and tutorial are included in the distribution.

4.3.2 Markovian Milestoning with Voronoi Tessellations: theory and implementation

Our previous implementation of SEEKR employed a traditional milestoning procedure where short trajectories were initiated on each milestone and run only until they touched another milestone.[50,51,94,95] The primary challenge of this approach is that one must know the correct probability distribution from which to reinitialize new trajectories on each milestone, called the first hitting point distribution (FHPD). The FHPD is obtained by first running a long, harmonically

restrained trajectory to sample the equilibrium distribution on each milestone. Position/velocity configurations from this equilibrium distribution are then used to launch new trajectories which are propagated backward in time by reversing the velocity of the system. Only equilibrium configurations which touch another milestone before touching the milestone on which they started are included as part of the FHPD. This procedure is computationally expensive, in particular, long-timescale (microsecond) trajectories must be run for each milestone in order to adequately sample all configurations of the equilibrium distribution. This creates a computational bottleneck for the SEEKR method as this portion of the method has limited parallelizability. The MMVT procedure proposed by Vanden-Eijnden and Venturoli overcomes this barrier by eliminating the requirement of initializing all trajectories from a configuration in the FHPD.[140] Instead, milestones are defined as the edges of a Voronoi tessellation and trajectories are confined to a Voronoi cell with the use of a reflective boundary condition. Here we will briefly describe the key aspects of this theory necessary for our implementation and refer the reader to the original paper for a deeper theoretical description and the paper of Maragliano et. al. for an implementation employing restraining potentials.[140,161]

The central assumption of MMVT is that the evolution of the system through time can be described as a continuous-time Markov-jump process between milestone states with the rate matrix, **Q,** having off-diagonal elements q_{ij} for $i \neq j$ and diagonal elements $q_{ii} = -\Sigma_{j \neq i} q_{ij}$ where i and j correspond to the starting and ending milestone indices. From a maximum likelihood estimation of **Q**, the off-diagonal elements ($i \neq j$) are defined as

$$q_{ij} = \begin{cases} \frac{N_{ij}}{R_i} & if\ R_i \neq 0 \\ 0 & if\ R_i = 0 \end{cases} \tag{4.1}$$

where N_{ij} is the number of transitions between milestone i and milestone j and R_i is the total time spent having last touched milestone i. The quantities N_{ij} and R_i can be estimated from independent simulations confined to Voronoi cells as described below.

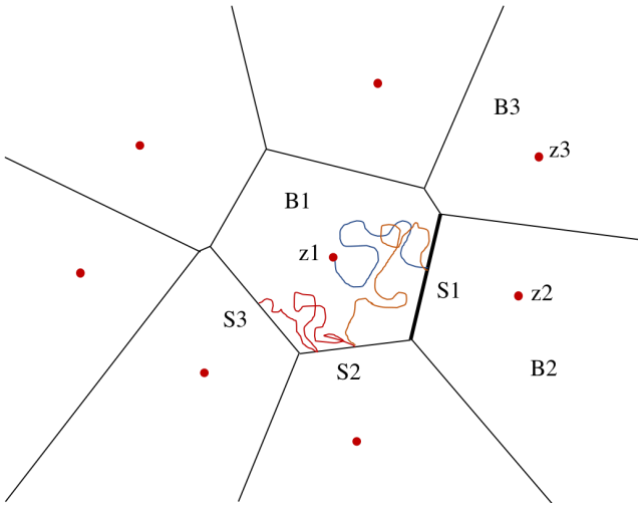


Figure 4.2 Sample Voronoi tessellation from the red generating points, z.
The edges of the cells define the milestones. Milestone S1 (thick line) represents the shared boundary between cells B1 and B2. The colored lines represent a hypothetical trajectory confined to cell B1 using reflective boundary conditions. Changes in color correspond to successful transitions between milestones for a single, continuous trajectory. The yellow portion is a transition from milestone S1 to S2 and the red portion from S2 to S3. The same simulation procedure is conducted independently in each Voronoi cell.

The definition of the Voronoi cells can be generalized from Cartesian space to collective variable space (i.e. bond distances, angles, etc.) with the collective variables denoted as $\boldsymbol{\theta}(x) = (\theta_1(x),\ldots, \theta_M(x))$. A set of generating points, $z_\alpha \in \mathbb{R}^M$, with $\alpha = 1,2,\ldots,\Lambda$, define a unique partition of configuration space, Ω, into Voronoi cells (Figure 4.2). The cell B_α from generating point z_α is the region

$$B_\alpha = \{x \in \Omega: \|\theta(x) - z_\alpha\| < \|\theta(x) - z_\beta\| \text{ for all } \beta \neq \alpha\} \quad (4.2)$$

The milestones are therefore defined as the common boundary, or edges, of adjacent cells. Independent simulations can then be carried out in each of the cells, propagated by the appropriate dynamical integrator, with the addition of a collision rule at the cell boundaries to keeps the trajectory confined to the appropriate cell. This collision rule is defined as

$$x_\alpha(t + \Delta t) = \begin{cases} x_\alpha^* & if\ x_\alpha^* \in B_\alpha \\ x_\alpha(t) & otherwise \end{cases} \tag{4.3}$$

and

$$v_\alpha(t + \Delta t) = \begin{cases} v_\alpha^* & if\ x_\alpha^* \in B_\alpha \\ -v_\alpha(t) & otherwise \end{cases} \tag{4.4}$$

These boundary conditions, in essence, result in the velocity of the system being reversed whenever a trajectory collides with a boundary in order to keep the system inside the appropriate cell. The underlying justification for this rule is that, from time reversibility, every trajectory leaving the cell has a statistically indistinguishable trajectory entering the cell at the same point, but with its velocity reversed. Therefore, the correct Boltzmann-Gibbs distribution is maintained within the cell, as long as some thermal bath ensures that the trajectory does not perfectly retrack itself when its velocity is reversed upon collision with a boundary. Importantly, this procedure eliminates the need to determine the FHPD and equilibrium distribution on each milestone.

From the simulations confined to a Voronoi cell, B_α, with edges (or milestone) indices i and j, one can obtain the quantities N_{ij}^α and R_i^α, where N_{ij}^α is the number of times a trajectory collides with a milestone after having last touched a different milestone and R_i^α is the total time the simulation spends having last touched milestone i. These two quantities can be related to the quantities N_{ij} and R_i needed for the determination of **Q** by weighting the cell specific values by the equilibrium probability of that cell:

$$N_{ij} = T \sum_{\alpha=1}^{\Lambda} \pi_\alpha \frac{N_{ij}^\alpha}{T_\alpha} \tag{4.5}$$

$$R_i = T \sum_{\alpha=1}^{\Lambda} \pi_\alpha \frac{R_i^\alpha}{T_\alpha} \tag{4.6}$$

Here T_α is the total simulation time in cell α and T is the reciprocal sum of time spent in all cells, which ensures dimensional consistency. The equilibrium probability, π_α, can then be computed by solving the system of equations defined by 4.7 and 4.8.[163]

$$\sum_{\beta=1,\beta\neq\alpha}^{\Lambda} \pi_\beta k_{\beta,\alpha} = \sum_{\beta=1,\beta\neq\alpha}^{\Lambda} \pi_\alpha k_{\alpha,\beta} \ , \tag{4.7}$$

$$\sum_{\alpha=1}^{\Lambda} \pi_\alpha = 1 \tag{4.8}$$

Where we assume that the flux in and out of each cell is zero at steady state for the unrestrained system. The quantity $k_{\alpha,\beta}$ is defined as:

$$k_{\alpha,\beta} = \frac{N_{\alpha,\beta}}{T_\alpha} \tag{4.9}$$

Where $N_{\alpha,\beta}$ is the total number of collisions with the common boundary of cells B_α and B_β. It is important to note that the key quantities to determine the rate matrix **Q** are N_{ij}^{α} and R_i^{α}, which can be obtained independently for each Voronoi cell. This independence facilitates the embarrassingly parallel nature of the MMVT SEEKR simulations. Furthermore, the convergence of these key quantities can be monitored as an estimate of the convergence of sampling for a particular MMVT cell, which will be discussed in further detail in section 4.3.4. The off rate can then be approximated as the reciprocal of the mean first passage time (MFPT) from the bound state to the outermost milestone using the standard expression for the MFPT in a continuous-time Markov chain

$$\hat{Q}T^N = -\mathbf{1} \tag{4.10}$$

$\hat{Q}$ is the N-1 by N-1 matrix obtained by deleting the last row and column of Q and **-1** is the unit vector in $\mathbb{R}^{N-1}$. T^N is the a vector with entries T_i^N that are the MFPTs from milestone i to milestone

N. It has been shown that the MFPTs computed from MMVT are exact if optimal milestones are used, as is the case for traditional milestoning simulations.[95,140]

For SEEKR, the MMVT algorithm is implemented directly in the NAMD configuration file using the existing TCL interface. The colvars module is used to monitor the milestones defined by existing collective variables (with no biasing force used).[162] The existing NAMD TCL commands "checkpoint" and "revert" as well as the "rescalevels" command are used to facilitate the reflective boundary conditions needed for the MMVT algorithm when the monitored collective variable crosses a predefined milestone boundary. To improve calculation efficiency, boundary crossings can be checked after a user-defined number of steps, rather than every step of the simulation. All transition events are output in the simulation output file for post-processing with the SEEKR analysis package.

4.3.3 Incorporating Brownian dynamics simulations to calculate k_{on}

BD simulations are extremely useful for efficiently simulating the portion of the association process where the ligand is far from the binding site and therefore the atomistic detail of MD is not required to obtain an adequate description of the process. Instead, solvent is approximated by a dielectric and solute molecules (receptor and ligand) are treated as rigid or semi-rigid bodies with dynamics propagated according to the general equation of Brownian motion. The Northrup Allison McCammon (NAM) method can be used to estimate k_{on} from BD simulations[85,152] using the equation

$$k_{on} = k_b \beta \tag{4.11}$$

where k_b is the rate of diffusion to a spherical surface of radius b (b surface) from the receptor calculated by

$$k_b = 4\pi \left[\int_b^\infty \frac{e^{\left(\frac{U(r)}{k_b T}\right)}}{r^2 D(r)} dr \right]^{-1} \quad (4.12)$$

Where U(r) is the potential energy between the receptor and ligand at distance r, k_bT is the Boltzmann constant times temperature, and D(r) is the diffusion coefficient. β from equation 4.11 is the probability that a ligand on the b surface will continue on to react, rather than escaping to an infinite distance. In practice, k_b is calculated automatically by the Browndye software used by SEEKR for the BD simulations.[99] Traditionally the value of β is calculated from many BD simulations, however in the SEEKR implementation, we calculate this probability from a combination of MD and BD simulations. BD simulations are first conducted from the b surface to the outermost milestone. Successful trajectories from this simulation are a FHPD on the outermost milestone. Subsequent BD simulations are then carried out from each point in this FHPD until they successfully touch the second outermost milestone or escape to infinity. The rate matrix, Q, constructed from the MMVT portion of the model can then be converted to a transition probability matrix, K, and modified to include the probability of binding/escape determined from the BD simulations. β is then calculated as the stationary flux, q_{stat}, of the bound state milestone by solving the equation

$$q_{stat} (\boldsymbol{I} - \boldsymbol{K}) = 0 \quad (4.13)$$

Where **I** is the identity matrix. The values of k_b and β can then be used to calculate k_{on} with equation 4.11. While the transition probabilities are obtained differently for the MMVT implementation, the calculation of k_{on} described here is the same as in the original implementation of SEEKR.

4.3.4 Error analysis simulation convergence estimates

The statistical error associated with the calculation of k_{on} and k_{off} was estimated using a Markov chain Monte Carlo (MCMC) procedure based on the procedure detailed by Noé in 2008 that was modified to sample the rate matrix rather than the transition kernel.[75] Each nonzero entry of the rate matrix, q_{ij} is sampled by pulling a new value from the appropriate gamma distribution with parameters N_{ij} and $1/R_i$ which is accepted or rejected based on a Metropolis criteria. The standard deviation of the rate constants calculated from many iterations of the MCMC procedure is used as an estimate of the statistical error of the calculation. The convergence of the MCMC calculated rate constants is monitored to ensure the rate matrix has been sufficiently sampled. Finally, it should be noted that the average calculated MCMC rate constant and the maximum likelihood estimate described in section 2.2 should converge with sufficient sampling, which can also be monitored as a measure of convergence of the simulations.

The convergence of sampling in each Voronoi cell is essential for determining the amount of simulation needed to accurately calculate the rate constants of intertest. As described in section 2.2, the two key quantities necessary to construct the rate matrix, Q, are $N_{ij}^{\alpha}/T_{\alpha}$ and R_i^{α}/T_{α}, which are independently obtained from simulations in each Voronoi cell. As a result, these two quantities can be monitored as a function of simulation time in each cell to estimate the convergence of sampling from the simulations. The SEEKR analysis package described in section 2.1 contains functions to extract and plot these quantities, providing the user with qualitative, visual estimates of the convergence of sampling. A quantitative metric, however, is desirable, as it can then be utilized to provide a more rigorous and reproduceable metric for convergence that is transferrable between systems. In the SEEKR package, we have implemented a sliding window root mean square deviation (RMSD) function to provide such an estimate. The quantities $N_{ij}^{\alpha}/T_{\alpha}$ and R_i^{α}/T_{α}

are calculated for user-defined strided portions of the data. A window of user-defined length is then moved through the data, grouping it into samples. For each sample, the RMSD from the average value is calculated. A user defined cutoff is specified as a percentage of the magnitude of the quantity (i.e 5% of the magnitude of the value). If the RMSD of each quantity remains below this cutoff for a user-defined length of simulation (i.e. 100ns) then the sampling in that cell is considered converged. Leaving these parameters to be specified by the user allows the convergence estimate to be adaptable to the particular application being considered. This allows the user to balance strictness of convergence, required level of accuracy, and amount of simulation time invested based on the particular question being answered. For example, a rank-ordering application may not require the same strictness of convergence as trying to estimate the true value of the rate constant within experimental error. By estimating the convergence of each Voronoi cell independently, sampling can be adaptively terminated or extended on a cell by cell basis. Allowing more simulation time to be devoted to difficult to sample areas, while eliminating excess simulation in easier to sample regions. We note that a metric such as this could also be used in the future to monitor convergence "on-the-fly" during simulations, rather than after a portion of simulation is run.

4.4 Results and Discussion

4.4.1 Host-guest molecule rank ordering

We assessed the effectiveness of the our new MMVT SEEKR implementation for rank-ordering compounds by their binding/unbinding rates. The model host-guest system β-cyclodextrin with seven different guest molecules was studied (Figure 4.3), as this was the same system studied with the original SEEKR implementation. Therefore, it was possible to directly

compare the accuracy and efficiency of the new MMVT approach to the previous implementation[52] as well as to brute-force MD simulation calculated rates[121] and experimentally measured kinetics.[143–145,147,148]

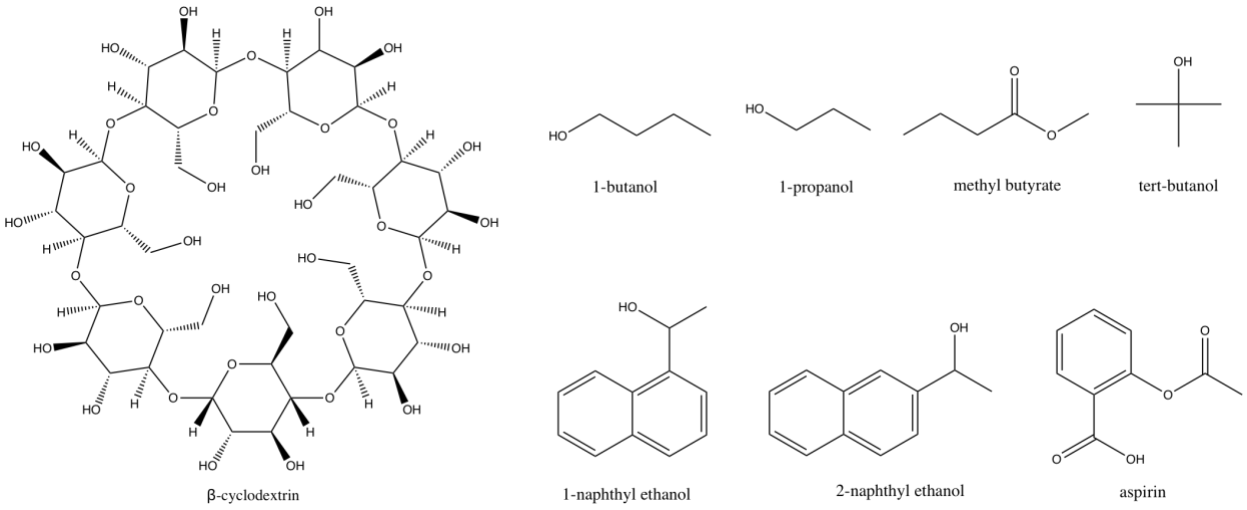


Figure 4.3 Structures of β-cyclodextrin and the seven ligands tested

System and simulation details can be found in section S1 of the Supporting Information. A one dimensional pseudo-Voronoi tessellation was generated using anchor points that resulted in milestones being placed between the center of mass (COM) of the host and the COM of the guest from 1.5 Å to 13.5 Å in 1.5 Å increments. Many short (~20-50 ns) MMVT MD simulations were carried out for each Voronoi cell for a combined total of ~560 ns of simulation per ligand (exact simulation lengths are presented in Table 4.2). Additionally, BD simulations were performed for the 13.5 Å milestone as described in section 4.3.3 with additional details in section 4.7.1. Both k_{on} and k_{off} as well as the binding free energy were calculated for each ligand. The convergence estimates described in section 4.3.4 were also used to determine the minimum simulation necessary to produce a converged result for each cell and the rates were recalculated using only that portion of the data. A sliding window of 30 samples was used pulled from the data with a stride of 1 ns

after skipping the first 10 ns. Cells were considered converged when values remained less than 5% of the average value for 20 windows (20 ns). Rate constants calculated with this minimum RMSD cutoff produced results consistent with the data from the full simulations, while benefitting from an additional ~20% reduction in simulation time. The values of the rate constants and binding free energies calculated using both methods as well as the brute force MD simulations and experiment are presented in Table 4.3 - Table 4.5. Figure 4.4 shows the calculated values for a) k_{on} and b) k_{off} and c) ΔG ordered by increasing magnitude of the experimentally measured value. The values calculated with MMVT SEEKR are in good agreement with values calculated from the previous SEEKR implementation. The values of k_{on} remain approximately one order of magnitude faster than experiment. As in the previous implementation, rank-ordering by k_{on} was not possible due to the limited variation in the experimental and computed values which are all near the diffusion limit.[40] The rank-ordering of ligands by k_{off} was improved with MMVT SEEKR; incorrectly ordering only two ligands, rather than three. The k_{off} values calculated with MMVT SEEKR were consistently faster than the experimental rates, which was also observed in the original SEEKR implementation. In addition, the binding free energy of each ligand can be determined because k_{on} and k_{off} are known. Calculated binding free energies are also in good agreement with experiment (Figure 4.4c), with only the value for methyl butyrate differing from the experimental value by more than 1 kcal/mol. As the on rates for these compounds are similar, the binding free energy is primarily dominated by the values of k_{off}. As such the rank-ordering is also similar to that for $k_{off;}$ incorrectly ordering methyl butyrate and tert-butanol, but also misordering 2-naphthylethanol as a result of a faster k_{on}. The MMVT SEEKR method was able to produce comparable predictions of k_{on} and k_{off} and binding free energy to the original SEEKR implementation, with improved rank-ordering of the ligands by k_{off} and comparable rank-ordering by free energy. Furthermore, MMVT

SEEKR benefits from a roughly 10-fold reduction in the amount of simulation used to achieve this result. The minimum simulation estimates, which produce nearly identical results to the full simulation data, save an additional 20% (~100ns) on the total MMVT simulation cost.

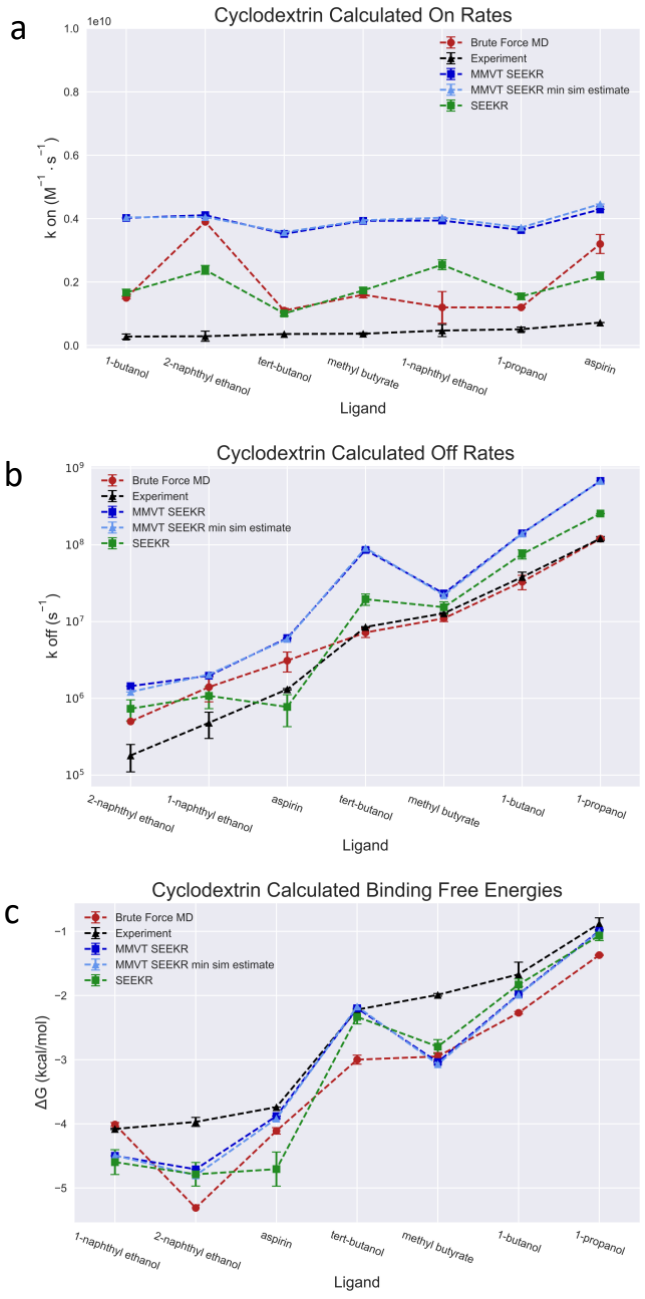


Figure 4.4 Comparison of results for cyclodextrin
Comparison of experimentally measured values[143–145,147,148] (black), brute force MD[121] (red) original SEEKR implementation (green) MMVT SEEKR[52] (dark blue) and MMVT SEEKR minimum simulation estimates for a) k_{on} b) k_{off} and c) binding free energy. Compounds are ordered by increasing experimentally measured values.

4.4.2 Trypsin-benzamidine application

We also tested the MMVT SEEKR method on the well-studied model system trypsin with the noncovalent inhibitor, benzamidine. This system has been used as a benchmark by many simulation-based approaches, including the original SEEKR implementation.[31,51,81,83,84,130] The increased complexity resulting from protein dynamics as well as a k_{off} value multiple orders of magnitude slower than those tested in the host-guest systems serve as a test for the efficiency and accuracy of the new MMVT SEEKR implementation. Detailed system and simulation details can be found in section 4.7.2. The same collective variable representing the distance of the ligand from the binding site was used as in the original SEEKR implementation.[51] Voronoi Cells were generated from anchor points that resulted in milestones with distances of 1.0, 2.0, 3.0, 4.0, 6.0, 8.0, 10.0, 12.0, and 13.0 Å (Figure 4.5). We note that the MMVT algorithm samples the regions between milestones differently than the previous implementation, and therefore the spacing of milestones is not identical to the previous implementation to account for sampling challenges associated with large energy barriers and to ensure that the Markov assumption remains valid. Minimum simulation estimates of k_{on} and k_{off} and binding free energy were obtained using a stride of 2 ns after skipping the first 20 ns, an RMSD sample window of 200 ns and a cutoff of 5% for at least 100 ns.

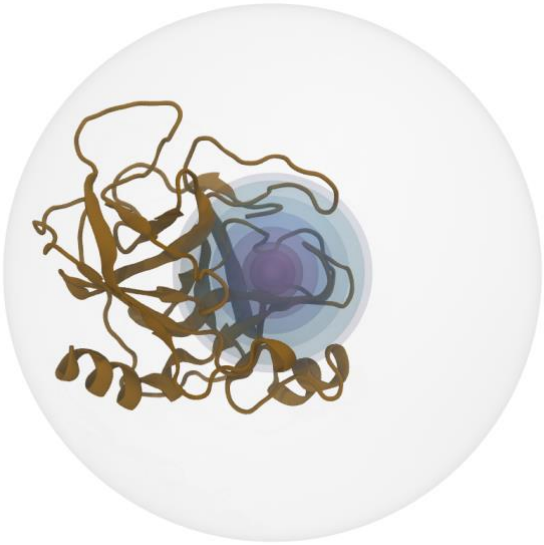


Figure 4.5 Trypsin milestone depiction
Structure of trypsin (cartoon) with milestones drawn as colored spheres. The outermost (grey) sphere represents the BD simulations from the "b surface" described in section 2.3. Many MD milestones are placed close to the binding site, while the BD region covers a much larger portion of the system.

MMVT SEEKR effectively reproduces the experimentally measured on and off rates and binding free energy (Table 2.1).[108] Using the full 4.4 μs of simulation data, the MMVT SEEKR result more closely reproduces experiment than the original SEEKR result, which required 19 μs; a ~4-fold reduction in simulation time. Furthermore, the MMVT minimum simulation estimate produces a comparable result to the original SEEKR implementation, requiring only 2.9 μs of simulation. This minimum simulation estimate saves an additional 35% of simulation from the full MMVT data, a ~7 fold reduction in simulation from the original implementation. The enhancement in sampling provided by the MMVT SEEKR approach is evident, as it predicts residence times that are over 1000 times longer than the simulation time invested. Statistically robust estimates of such residence times on the order of a millisecond would likely pose a significant challenge and expense for a brute force simulation approach, highlighting the value of the SEEKR approach for both its enhancement in sampling as well as parallelism.

Table 4.1 Trypsin-benzamidine calculated rates and binding free energies, simulation time, and experimentally measured values

Method	k_{off} (s^{-1})	Residence Time (μs)	k_{on} (M^{-1} s^{-1})	ΔG (kcal/mol)	Simulation Time (μs)
Experiment[108]	600 ± 300	1700	2.9×10^7	-6.7 ± 0.05	
SEEKR[51]	83 ± 14	12000	$(2.1 \pm 0.3) \times 10^7$	-7.4 ± 0.1	19
MMVT SEEKR	174 ± 9	5750	$(1.2 \pm 0.05) \times 10^8$	-7.9 ± 0.04	4.4
MMVT SEEKR minimum simulation estimate	62 ± 6	16000	$(1.7 \pm 0.1) \times 10^8$	-8.8 ± 0.07	2.9

4.5 Conclusion

We have presented a new MMVT algorithm implemented in the SEEKR package for calculating receptor-ligand binding and unbinding rate constants as well as binding free energies. The results of the two applications we have described here demonstrate that MMVT SEEKR is effective for both rank-ordering compounds by their kinetics as well as reproducing the magnitude of experimentally measured kinetics. MMVT SEEKR benefits from a significant reduction in simulation cost compared to the previous SEEKR implementation by eliminating the need to determine equilibrium distributions and FHPDs for each milestone. We have also described a method for estimating the convergence of sampling for each Voronoi cell and adaptively extending or terminating simulations accordingly. This convergence estimate was shown to further reduce the simulation cost of MMVT SEEKR calculations while retaining accuracy. The MMVT algorithm can also be used to construct and simulate models with multiple dimensions of

milestones, unlike the one dimensional models used in this study. Additional milestones could be useful for improving sampling of other slow degrees of freedom that may exist in a more complicated system. This may be particularly important when studying larger drug molecules with longer residence times and more complicated binding/unbinding mechanisms. The improvements to efficiency, as well as the embarrassingly parallel nature of milestoning simulations, make MMVT SEEKR well suited for use in future prospective studies for larger systems of pharmaceutical relevance.

4.6 Acknowledgement

We thank Zhiye Tang and Chia-en Chang for sharing structures and parameters for the cyclodextrin study. We are also grateful to Christopher T. Lee, J. Andrew McCammon and Gary Huber for insightful discussions.

Chapter 4, in full, has been submitted for publication and is presented as it may appear in: "Jagger, B. R.; Ojha, A. A.; Amaro, R. E. Predicting Ligand Binding Kinetics Using a Markovian Milestoning with Voronoi Tessellation Multiscale Approach. *J. Chem. Theory Comput. Submitted.* The dissertation author was a primary investigator and author of this work.

4.7 Supporting Information

4.7.1 Host Guest Simulations

Structures and parameters are the same as those used in our previous study which were obtained from the brute force MD study of Tang and Chang.[52,121] The cyclodextrin molecule was parameterized with the specialized Q4MD-CD forcefield[142] and all ligands were parameterized with the Generalized Amber forcefield (GAFF).[110,111] Systems were solvated with TIP3P waters. Milestones were defined using a collective variable measuring the distance between the center of

mass of the host and guest molecules. Milestones were placed at 1.5 Å increments from 1.5 to 13.5 Å, as in the previous study. The SEEKR software was used to prepare starting structures and simulation input files for each milestone. The starting structure for each MMVT anchor was obtained by first performing a constant volume MD simulation at 298 K with a Langevin damping coefficient of 5/ps and a 2 fs timestep. An 8 Å cutoff and a PME grid spacing of 1.0 Å were used. A harmonic restraint of 90 kcal/mol was used to pull the ligand from the bound state starting structure to the appropriate distance for that anchor. This simulation was run for 2 ns to allow for a short equilibration after the ligand reached the target distance. From this equilibrated structure many short (~20 ns) MMVT simulations were carried out. Simulations were conducted using the same parameters as the equilibration simulations but without the harmonic restraint. The reflective boundary conditions were implemented as described in the main text, with boundary crossings checked every 10 steps. A combined total of ~560 ns of simulation from all milestones was performed for each of the seven ligands. Total simulation time used as well as minimum simulation estimate times are shown in Table 4.2.

Table 4.2 Total simulation time and minimum estimated simulation time used for each ligand.

	total sim time (ns)	min sim estimate time (ns)
1-butanol	560	420
1-naphthylethanol	560	434
1-propanol	546	420
2-naphthylethanol	560	448
aspirin	562	450
methyl butyrate	560	420
tert-butanol	560	420

BD simulations were carried out using the Browndye software package.[99] Electrostatic potentials of the host and guest molecules used as inputs for the BD simulations were calculated using the Adaptive Poisson Boltzmann Solver (APBS) version 1.4. Experimental conditions were matched in the BD simulations using a solvent dielectric of 78, a solute dielectric of 2, and zero ionic concentration. 10^6 independent simulations were initiated from the b surface (as described in section 2.3 of the main text) to generate a FHPD on the 13.5 Å milestone. An additional 10^6 simulations were initiated from points in this FHPD to collect transition statistics from the 13.5 Å milestone to the 12 Å milestone or the escape state (q surface). Association and dissociation rates as well as binding free energies for each ligand are presented in Table 4.3 - Table 4.5.

Table 4.3 Experimental off rates and calculated values using brute force MD and various SEEKR approaches.

	Experiment[143–145,147,148]		**Brute Force MD**[121]		**SEEKR**[52]		**MMVT SEEKR**		**MMVT SEEKR Min Sim**	
	k_{off} (s^{-1})	**Error (s^{-1})**	**k_{off} (s^{-1})**	**Error (s^{-1})**	**k_{off} (s^{-1})**	**Error (s^{-1})**	**k_{off} (s^{-1})**	**Error (s^{-1})**	**k_{off} (s^{-1})**	**Error (s^{-1})**
1-butanol	3.80E+07	6.00E+06	3.30E+07	7.00E+06	7.57E+07	1.04E+07	1.41E+08	6.61E+03	1.40E+08	6.99E+03
1-naphthylethanol	4.80E+05	1.80E+05	1.40E+06	5.00E+05	1.08E+06	3.44E+05	1.97E+06	8.17E+02	2.04E+06	8.13E+02
1-propanol	1.21E+08	7.00E+06	1.20E+08	2.00E+06	2.55E+08	2.24E+05	6.80E+08	1.78E+04	6.79E+08	1.63E+04
2-naphthylethanol	1.80E+05	7.00E+04	5.00E+05	--	7.31E+05	2.24E+05	1.44E+06	7.37E+02	1.21E+06	6.64E+02
aspirin	1.31E+06	3.00E+04	3.10E+06	9.00E+05	7.72E+05	3.45E+05	6.10E+06	1.46E+03	5.93E+06	1.39E+03
methyl butyrate	1.28E+07	3.00E+05	1.10E+07	1.00E+06	1.54E+07	2.69E+06	2.32E+07	2.75E+03	2.20E+07	2.79E+03
tert-butanol	8.50E+06	1.00E+05	7.20E+06	1.00E+06	1.96E+07	3.29E+06	8.57E+07	5.48E+03	9.09E+07	5.22E+03

Table 4.4 Experimental on rates and calculated values using brute force MD and various SEEKR approaches.

	Experiment[143–145,147,148]		Brute Force MD[121]		SEEKR[52]		MMVT SEEKR		MMVT SEEKR Min Sim	
	k_{on} (M^{-1} s^{-1})	Error (M^{-1} s^{-1})	k_{on} (M^{-1} s^{-1})	Error (M^{-1} s^{-1})	k_{on} (M^{-1} s^{-1})	Error (M^{-1} s^{-1})	k_{on} (M^{-1} s^{-1})	Error (M^{-1} s^{-1})	k_{on} (M^{-1} s^{-1})	Error (M^{-1} s^{-1})
1-butanol	2.80E+08	8.00E+07	1.50E+09	3.00E+07	1.67E+09	1.03E+08	4.02E+09	8.67E+04	4.04E+09	8.77E+04
1-naphthylethanol	4.70E+08	1.90E+08	1.20E+09	5.00E+08	2.55E+09	1.55E+08	3.94E+09	4.14E+05	4.03E+09	3.84E+05
1-propanol	5.10E+08	7.00E+07	1.20E+09	2.00E+07	1.55E+09	1.04E+08	3.64E+09	5.01E+04	3.72E+09	5.20E+04
2-naphthylethanol	2.90E+08	1.60E+08	3.90E+09	--	2.38E+09	1.39E+08	4.11E+09	3.97E+05	4.05E+09	4.80E+05
aspirin	7.21E+08	4.00E+06	3.20E+09	3.00E+08	2.19E+09	1.20E+08	4.29E+09	2.62E+05	4.46E+09	2.27E+05
methyl butyrate	3.70E+08	3.00E+07	1.60E+09	1.00E+08	1.73E+09	9.59E+07	3.93E+09	1.83E+05	3.95E+09	1.75E+05
tert-butanol	3.60E+08	1.00E+07	1.10E+09	7.00E+07	1.01E+09	8.23E+07	3.52E+09	1.23E+05	3.57E+09	1.15E+05

Table 4.5 Experimental binding free energy and calculated values using brute force MD and various SEEKR approaches.

	Experiment[143–145,147,148]		Brute Force MD[121]		SEEKR[52]		MMVT SEEKR		MMVT SEEKR Min Sim	
	ΔG (kcal/mol)	Error (kcal/mol)	ΔG (kcal/mol)	Error (kcal/mol)	ΔG (kcal/mol)	Error (kcal/mol)	ΔG (kcal/mol)	Error (kcal/mol)	ΔG (kcal/mol)	Error (kcal/mol)
1-butanol	-1.67	0.19	-2.27	0.02	-1.83	8.90E-02	-1.98	3.05E-05	-1.99	3.22E-05
1-naphthylethanol	-4.08	0.01	-4.01	0.03	-4.60	1.93E-01	-4.50	2.53E-04	-4.49	2.42E-04
1-propanol	-0.88	0.09	-1.37	0.01	-1.07	3.98E-02	-0.99	1.75E-05	-1.01	1.64E-05
2-naphthylethanol	-3.97	0.07	-5.31	--	-4.79	1.84E-01	-4.71	3.08E-04	-4.80	3.32E-04
aspirin	-3.74	0	-4.11	0.05	-4.71	2.67E-01	-3.88	1.46E-04	-3.92	1.42E-04
methyl butyrate	-1.99	0.02	-2.95	0.06	-2.80	1.09E-01	-3.04	7.54E-05	-3.07	7.95E-05
tert-butanol	-2.22	0.01	-3.00	0.07	-2.33	1.11E-01	-2.20	4.31E-05	-2.17	3.90E-05

4.7.2 Trypsin-benzamidine simulations

The system used in this study was the same as described in our original SEEKR study.[51] Atomic coordinates were obtained from the protein data bank structure 3PTB.[102] Hydrogens were added using Molprobity with ring flips allowed.[103,104] The system was prepared with LEaP using the Amber ff14SB forcefield[105] with protonation states of titratable residues assigned using the

PROPKA package[106,107] according to the experimental conditions.[108] The system was solvated with TIP4Pew waters in a truncated octahedron, as in the previous study.[109,110] Chloride anions were added to neutralize the system total charge. The benzamidine ligand was parameterized using Antechamber and the GAFF forcefield.[110,111] The same equilibrated structure from the previous study was used as the starting structure for this study.

Milestones were defined using a collective variable measuring the distance between the center of mass of the alpha carbons of residues representing the binding site (190, 191, 192, 195, 213, 215, 216, 219, 220, 224, and 228 of PDB: 3PTB) and the center of mass of the benzamidine ligand. Milestones were placed at distances of 1.0, 2.0, 3.0, 4.0, 6.0, 8.0, 10.0, 12.0, and 13.0 Å. The 13.0 Å milestone is also the BD milestone. The SEEKR software was used to prepare starting structures and simulation input files for each milestone. The starting structure for each MMVT anchor was obtained by first performing a constant volume MD simulation at 298 K with a Langevin damping coefficient of 5/ps and a 2 fs timestep. An 8 Å cutoff and a PME grid spacing of 1.0 Å were used. A harmonic restraint of 90 kcal/mol was used to pull the ligand from the bound state starting structure to the appropriate distance for that anchor. This simulation was run for 2 ns to allow for a short equilibration after the ligand reached the target distance. From this equilibrated structure many short (~20 ns) MMVT simulations were carried out. Simulations were conducted using the same parameters as the equilibration simulations but without the harmonic restraint. The reflective boundary conditions were implemented as described in the main text, with boundary crossings checked every 10 steps. A combined total of ~4.4 μs of simulation from all milestones was performed. Experimental conditions were matched in the BD simulations and APBS calculations, using a solvent dielectric of 78, a solute dielectric of 2, Ca^{2+} ions at a concentration of 0.02 mM with a charge of +2.0 *e* and a radius of 1.14 Å, Cl^- ions at a concentration of 0.10 mM

with a charge of −1.0 *e* and a radius of 1.67 Å, and tris at a concentration of 0.06 mM with a charge of +1.0 e and a radius of 4.0 Å. 10^6 independent simulations were initiated from the b surface (as described in section 4.3.3) to generate a FHPD on the 13.0 Å milestone. An additional 10^6 simulations were initiated from points in this FHPD to collect transition statistics from the 13.0 Å milestone to the 12 Å milestone or the escape state (q surface). The BD results were then extracted and incorporated with the MD statistics for the calculation of k_{on}.

References

1. De Vivo, M., Masetti, M., Bottegoni, G. & Cavalli, A. Role of Molecular Dynamics and Related Methods in Drug Discovery. *J. Med. Chem.* **59,** 4035–4061 (2016).

2. Amaro, R. E., Baudry, J., Chodera, J., Demir, Ö., McCammon, J. A., Miao, Y. & Smith, J. C. Ensemble Docking in Drug Discovery. *Biophys. J.* **114,** 2271–2278 (2018).

3. Cournia, Z., Allen, B. & Sherman, W. Relative Binding Free Energy Calculations in Drug Discovery: Recent Advances and Practical Considerations. *J. Chem. Inf. Model.* **57,** 2911–2937 (2017).

4. Callegari, D., Ranaghan, K. E., Woods, C. J., Minari, R., Tiseo, M., Mor, M., Mulholland, A. J. & Lodola, A. L718Q mutant EGFR escapes covalent inhibition by stabilizing a non-reactive conformation of the lung cancer drug osimertinib. *Chem. Sci.* **9,** 2740–2749 (2018).

5. Bruce, N. J., Ganotra, G. K., Kokh, D. B., Sadiq, S. K. & Wade, R. C. New approaches for computing ligand–receptor binding kinetics. *Curr. Opin. Struct. Biol.* **49,** 1–10 (2018).

6. Bernetti, M., Masetti, M., Rocchia, W. & Cavalli, A. Kinetics of Drug Binding and Residence Time. *Annu. Rev. Phys. Chem.* **70,** 143–171 (2019).

7. Ribeiro, J. M. L., Tsai, S.-T., Pramanik, D., Wang, Y. & Tiwary, P. Kinetics of Ligand–Protein Dissociation from All-Atom Simulations: Are We There Yet? *Biochemistry* **58,** 156–165 (2019).

8. Lee, C. T. & Amaro, R. E. Exascale Computing: A New Dawn for Computational Biology. *Comput. Sci. Eng.* **20,** 18–25 (2018).

9. Amaro, R. E. & Mulholland, A. J. Multiscale methods in drug design bridge chemical and biological complexity in the search for cures. *Nat. Rev. Chem.* **2,** 148 (2018).

10. Huggins, D. J., Biggin, P. C., Dämgen, M. A., Essex, J. W., Harris, S. A., Henchman, R. H., Khalid, S., Kuzmanic, A., Laughton, C. A., Michel, J., Mulholland, A. J., Rosta, E., Sansom, M. S. P. & van der Kamp, M. W. Biomolecular simulations: From dynamics and mechanisms to computational assays of biological activity. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **9,** e1393 (2019).

11. Voice, A., Tresadern, G., van Vlijmen, H. & Mulholland, A. Limitations of Ligand-Only Approaches for Predicting the Reactivity of Covalent Inhibitors. *J. Chem. Inf. Model.* **59,** 4220–4227 (2019).

12. Haldar, S., Comitani, F., Saladino, G., Woods, C., van der Kamp, M. W., Mulholland, A. J. & Gervasio, F. L. A Multiscale Simulation Approach to Modeling Drug–Protein Binding Kinetics. *J. Chem. Theory Comput.* **14,** 6093–6101 (2018).

13. Woods, C. J., Shaw, K. E. & Mulholland, A. J. Combined quantum mechanics/molecular mechanics (QM/MM) simulations for protein-ligand complexes: free energies of binding of water molecules in influenza neuraminidase. *J Phys Chem B* **119,** 997–1001 (2015).

14. Lonsdale, R., Fort, R. M., Rydberg, P., Harvey, J. N. & Mulholland, A. J. Quantum Mechanics/Molecular Mechanics Modeling of Drug Metabolism: Mexiletine N-Hydroxylation by Cytochrome P450 1A2. *Chem. Res. Toxicol.* **29,** 963–971 (2016).

15. Bolnykh, V., Olsen, J. M. H., Meloni, S., Bircher, M. P., Ippoliti, E., Carloni, P. & Rothlisberger, U. Extreme Scalability of DFT-Based QM/MM MD Simulations Using MiMiC. *J. Chem. Theory Comput.* **15,** 5601–5613 (2019).

16. Sokkar, P., Boulanger, E., Thiel, W. & Sanchez-Garcia, E. Hybrid Quantum Mechanics/Molecular Mechanics/Coarse Grained Modeling: A Triple-Resolution Approach for Biomolecular Systems. *J. Chem. Theory Comput.* **11,** 1809–1818 (2015).

17. Ranaghan, K. E., Shchpanovska, D., Bennie, S. J., Lawan, N., Macrae, S. J., Zurek, J., Manby, F. R. & Mulholland, A. J. Projector-Based Embedding Eliminates Density Functional Dependence for QM/MM Calculations of Reactions in Enzymes and Solution. *J Chem Inf Model* **59,** 2063–2078 (2019).

18. Kokh, D. B., Amaral, M., Bomke, J., Grädler, U., Musil, D., Buchstaller, H.-P., Dreyer, M. K., Frech, M., Lowinski, M., Vallee, F., Bianciotto, M., Rak, A. & Wade, R. C. Estimation of Drug-Target Residence Times by τ-Random Acceleration Molecular Dynamics Simulations. *J. Chem. Theory Comput.* **14,** 3859–3869 (2018).

19. Gobbo, D., Piretti, V., Di Martino, R. M. C., Tripathi, S. K., Giabbai, B., Storici, P., Demitri, N., Girotto, S., Decherchi, S. & Cavalli, A. Investigating Drug–Target Residence Time in Kinases through Enhanced Sampling Simulations. *J. Chem. Theory Comput.* **15,** 4646–4659 (2019).

20. Schuetz, D. A., Bernetti, M., Bertazzo, M., Musil, D., Eggenweiler, H.-M., Recanatini, M., Masetti, M., Ecker, G. F. & Cavalli, A. Predicting Residence Time and Drug Unbinding Pathway through Scaled Molecular Dynamics. *J. Chem. Inf. Model.* **59,** 535–549 (2019).

21. Valsson, O., Tiwary, P. & Parrinello, M. Enhancing Important Fluctuations: Rare Events and Metadynamics from a Conceptual Viewpoint. *Annu. Rev. Phys. Chem.* **67,** 159–184 (2016).

22. Cavalli, A., Spitaleri, A., Saladino, G. & Gervasio, F. L. Investigating Drug–Target Association and Dissociation Mechanisms Using Metadynamics-Based Algorithms. *Acc. Chem. Res.* **48,** 277–285 (2015).

23. Morando, M. A., Saladino, G., D'Amelio, N., Pucheta-Martinez, E., Lovera, S., Lelli, M., López-Méndez, B., Marenchino, M., Campos-Olivas, R. & Gervasio, F. L. Conformational Selection and Induced Fit Mechanisms in the Binding of an Anticancer Drug to the c-Src Kinase. *Sci. Rep.* **6,** 24439 (2016).

24. Invernizzi, M. & Parrinello, M. Making the Best of a Bad Situation: A Multiscale Approach to Free Energy Calculation. *J. Chem. Theory Comput.* **15,** 2187–2194 (2019).

25. Bernetti, M., Masetti, M., Recanatini, M., Amaro, R. E. & Cavalli, A. An Integrated Markov State Model and Path Metadynamics Approach To Characterize Drug Binding Processes. *J. Chem. Theory Comput.* **15,** 5689–5702 (2019).

26. McCarty, J. & Parrinello, M. A variational conformational dynamics approach to the selection of collective variables in metadynamics. *J. Chem. Phys.* **147,** 204109 (2017).

27. Brotzakis, Z. F., Limongelli, V. & Parrinello, M. Accelerating the Calculation of Protein–Ligand Binding Free Energy and Residence Times Using Dynamically Optimized Collective Variables. *J. Chem. Theory Comput.* **15,** 743–750 (2019).

28. Dibak, M., del Razo, M. J., De Sancho, D., Schütte, C. & Noé, F. MSM/RD: Coupling Markov state models of molecular kinetics with reaction-diffusion simulations. *J. Chem. Phys.* **148,** 214107 (2018).

29. Elber, R. A new paradigm for atomically detailed simulations of kinetics in biophysical systems. *Q. Rev. Biophys.* **50,** e8 (2017).

30. Narayan, B., Fathizadeh, A., Templeton, C., He, P., Arasteh, S., Elber, R., Buchete, N.-V. & Levy, R. M. The transition between active and inactive conformations of Abl kinase studied by rock climbing and Milestoning. *Biochim. Biophys. Acta - Gen. Subj.* 129508 (2019). doi:10.1016/j.bbagen.2019.129508

31. Plattner, N. & Noé, F. Protein conformational plasticity and complex ligand-binding kinetics explored by atomistic simulations and Markov models. *Nat. Commun.* **6,** 7653 (2015).

32. Taylor, B. C., Lee, C. T. & Amaro, R. E. Structural basis for ligand modulation of the CCR2 conformational landscape. *Proc. Natl. Acad. Sci.* **116,** 8131–8136 (2019).

33. Wu, H., Paul, F., Wehmeyer, C. & Noé, F. Multiensemble Markov models of molecular thermodynamics and kinetics. *Proc. Natl. Acad. Sci.* **113,** E3221–E3230 (2016).

34. Olsson, S., Wu, H., Paul, F., Clementi, C. & Noé, F. Combining experimental and simulation data of molecular processes via augmented Markov models. *Proc. Natl. Acad. Sci.* **114,** 8265–8270 (2017).

35. Yu, T.-Q., Lapelosa, M., Vanden-Eijnden, E. & Abrams, C. F. Full Kinetics of CO Entry, Internal Diffusion, and Exit in Myoglobin from Transition-Path Theory Simulations. *J. Am. Chem. Soc.* **137,** 3041–3050 (2015).

36. Bucci, A., Yu, T.-Q., Vanden-Eijnden, E. & Abrams, C. F. Kinetics of O 2 Entry and Exit in Monomeric Sarcosine Oxidase via Markovian Milestoning Molecular Dynamics. *J. Chem. Theory Comput.* **12,** 2964–2972 (2016).

37. Zuckerman, D. M. & Chong, L. T. Weighted Ensemble Simulation: Review of Methodology, Applications, and Software. *Annu. Rev. Biophys.* **46,** 43–47 (2017).

38. Piana, S., Robustelli, P., Tan, D., Chen, S. & Shaw, D. E. Development of a force field for the simulation of single-chain proteins and protein-protein complexes. *J Chem Theory Comput* (2020).

39. Rizzi, A., Murkli, S., McNeill, J. N., Yao, W., Sullivan, M., Gilson, M. K., Chiu, M. W., Isaacs, L., Gibb, B. C., Mobley, D. L. & Chodera, J. D. Overview of the SAMPL6 host–guest binding affinity prediction challenge. *J. Comput. Aided. Mol. Des.* **32,** 937–963 (2018).

40. Berg, O. G. & von Hippel, P. H. Diffusion-controlled macromolecular interactions. *Annu. Rev. Biophys. Biophys. Chem.* **14,** 131–160 (1985).

41. Schreiber, G. Kinetic studies of protein-protein interactions. *Curr. Opin. Struct. Biol.* **12,** 41–7 (2002).

42. Northrup, S. H. & Erickson, H. P. Kinetics of protein-protein association explained by Brownian dynamics computer simulation. *Proc. Natl. Acad. Sci. U. S. A.* **89,** 3338–3342 (1992).

43. Huber, G. & McCammon, J. A. Brownian Dynamics Simulations of Biological Molecules. *Trends Chem.* **1,** 727–738 (2019).

44. Luty, B. A., El Amrani, S. & McCammon, J. A. Simulation of the bimolecular reaction between superoxide and superoxide dismutase: synthesis of the encounter and reaction steps. *J. Am. Chem. Soc.* **115,** 11874–11877 (1993).

45. Chang, C.-E. A., Trylska, J., Tozzini, V. & Andrew McCammon, J. Binding Pathways of Ligands to HIV-1 Protease: Coarse-grained and Atomistic Simulations. *Chem. Biol. Drug Des.* **69,** 5–13 (2007).

46. Huang, Y. M., Kang, M. & Chang, C. A. Switches of hydrogen bonds during ligand-protein association processes determine binding kinetics. *J. Mol. Recognit.* **27,** 537–548 (2014).

47. Huang, Y. M., Raymundo, M. A. V., Chen, W. & Chang, C. A. Mechanism of the Association Pathways for a Pair of Fast and Slow Binding Ligands of HIV-1 Protease. *Biochemistry* **56,** 1311–1323 (2017).

48. Schneider, J., Korshunova, K., Musiani, F., Alfonso-Prieto, M., Giorgetti, A. & Carloni, P. Predicting ligand binding poses for low-resolution membrane protein models: Perspectives from multiscale simulations. *Biochem. Biophys. Res. Commun.* **498,** 366–374 (2018).

49. Alfonso-Prieto, M., Navarini, L. & Carloni, P. Understanding Ligand Binding to G-Protein Coupled Receptors Using Multiscale Simulations. *Front. Mol. Biosci.* **6,** (2019).

50. Votapka, L. W. & Amaro, R. E. Multiscale Estimation of Binding Kinetics Using Brownian Dynamics, Molecular Dynamics and Milestoning. *PLOS Comput. Biol.* **11,** e1004381 (2015).

51. 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).

52. Jagger, B. R., Lee, C. T. & Amaro, R. E. Quantitative Ranking of Ligand Binding Kinetics with a Multiscale Milestoning Simulation Approach. *J. Phys. Chem. Lett.* **9,** 4941–4948 (2018).

53. Zeller, F., Luitz, M. P., Bomblies, R. & Zacharias, M. Multiscale Simulation of Receptor–Drug Association Kinetics: Application to Neuraminidase Inhibitors. *J. Chem. Theory Comput.* **13,** 5097–5105 (2017).

54. Alnæs, M. S., Blechta, J., Hake, J., Johansson, A., Kehlet, B., Logg, A., Richardson, C., Ring, J., Rognes, M. E. & Wells, G. N. The FEniCS Project Version 1.5. *Arch. Numer. Softw.* **3,** (2015).

55. Kerr, R. A., Bartol, T. M., Kaminsky, B., Dittrich, M., Chang, J.-C. J., Baden, S. B., Sejnowski, T. J. & Stiles, J. R. Fast Monte Carlo Simulation Methods for Biological Reaction-Diffusion Systems in Solution and on Surfaces. *SIAM J. Sci. Comput.* **30,** 3126–3149 (2008).

56. Schöneberg, J., Ullrich, A. & Noé, F. Simulation tools for particle-based reaction-diffusion dynamics in continuous space. *BMC Biophys.* **7,** 11 (2014).

57. Roberts, E., Stone, J. E. & Luthey-Schulten, Z. Lattice microbes: High-performance stochastic simulation method for the reaction-diffusion master equation. *J. Comput. Chem.* **34,** 245–255 (2013).

58. Hoffman, M., Frohner, C. & Noé, F. ReaDDy 2: Fast and flexible software framework for interacting-particle reaction dynamics. *PLoS Comput. Biol.* **15,** e1006830 (2019).

59. Lee, C. T., Laughlin, J. G., de La Beaumelle, N. A., Amaro, R. E., McCammon, J. A., Ramamoorthi, R., Holst, M. J. & Rangamani, P. 3D mesh processing using GAMer 2 to enable reaction-diffusion simulations in realistic cellular geometries. *bioRxiv* 534479 (2019). doi:10.1101/534479

60. Lee, C. T., Laughlin, J. G., Moody, J. B., Amaro, R. E., McCammon, J. A., Holst, M. J. & Rangamani, P. An Open Source Mesh Generation Platform for Biophysical Modeling Using Realistic Cellular Geometries. *bioRxiv* 765453 (2019). doi:10.1101/765453

61. Aboelkassem, Y., McCabe, K. J., Huber, G. A., Regnier, M., McCammon, J. A. & McCulloch, A. D. A Stochastic Multiscale Model of Cardiac Thin Filament Activation Using Brownian-Langevin Dynamics. *Biophys. J.* **117,** 2255–2272 (2019).

62. Solernou, A., Hanson, B. S., Richardson, R. A., Welch, R., Read, D. J., Harlen, O. G. & Harris, S. A. Fluctuating Finite Element Analysis (FFEA): A continuum mechanics software tool for mesoscale simulation of biomolecules. *PLOS Comput. Biol.* **14,** e1005897 (2018).

63. Malhotra, S., Träger, S., Dal Peraro, M. & Topf, M. Modelling structures in cryo-EM maps. *Curr. Opin. Struct. Biol.* **58,** 105–114 (2019).

64. Carpenter, T. S. & Lightstone, F. C. An Electrostatic Funnel in the GABA-Binding Pathway. *PLoS Comput Biol* **12,** 1004831 (2016).

65. Bar-Even, A., Noor, E., Savir, Y., Liebermeister, W., Davidi, D., Tawfik, D. S. & Milo, R. The Moderately Efficient Enzyme: Evolutionary and Physicochemical Trends Shaping Enzyme Parameters. *Biochemistry* **50,** 4402–4410 (2011).

66. 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).

67. Jorgensen, W. L. Foundations of Biomolecular Modeling. *Cell* **155,** 1199–1202 (2013).

68. Held, M. & Noé, F. Calculating kinetics and pathways of protein–ligand association. *Eur. J. Cell Biol.* **91,** 357–364 (2012).

69. Swegat, W., Schlitter, J., Krüger, P. & Wollmer, A. MD Simulation of Protein-Ligand Interaction: Formation and Dissociation of an Insulin-Phenol Complex. *Biophys. J.* **84,** 1493–1506 (2003).

70. Dror, R. O., Pan, A. C., Arlow, D. H., Borhani, D. W., Maragakis, P., Shan, Y., Xu, H. & Shaw, D. E. Pathway and Mechanism of Drug Binding to G-Protein-Coupled Receptors. *Proc. Natl. Acad. Sci.* **108,** 13118–13123 (2011).

71. Shan, Y., Eastwood, M. P., Zhang, X., Kim, E. T., Arkhipov, A., Dror, R. O., Jumper, J., Kuriyan, J. & Shaw, D. E. Oncogenic Mutations Counteract Intrinsic Disorder in the EGFR Kinase and Promote Receptor Dimerization. *Cell* **149,** 860–870 (2012).

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

73. Pan, A. C., Borhani, D. W., Dror, R. O. & Shaw, D. E. Molecular Determinants of Drug-Receptor Binding Kinetics. *Drug Discov. Today* **18,** 667–673 (2013).

74. Chodera, J. D. & Noé, F. Probability distributions of molecular observables computed from Markov models. II. Uncertainties in observables and their time-evolution. *J. Chem. Phys.* **133,** 105102 (2010).

75. Noé, F. Probability distributions of molecular observables computed from Markov models. *J. Chem. Phys.* **128,** 244103 (2008).

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

77. Sarich, M., Noé, F. & Schütte, C. On the Approximation Quality of Markov State Models. *Multiscale Model. Simul.* **8,** 1154–1177 (2010).

78. Pande, V. S., Beauchamp, K. & Bowman, G. R. *Everything you wanted to know about Markov State Models but were afraid to ask. Methods* **52,** 99–105 (2010).

79. Lane, T. J., Bowman, G. R., Beauchamp, K., Voelz, V. A. & Pande, V. S. Markov State Model Reveals Folding and Functional Dynamics in Ultra-Long MD Trajectories. *J. Am. Chem. Soc.* **133,** 18413–18419 (2011).

80. Schütte, C., Noé, F., Lu, J., Sarich, M. & Vanden-Eijnden, E. Markov state models based on milestoning. *J. Chem. Phys.* **134,** 204105 (2011).

81. Buch, I., Giorgino, T. & De Fabritiis, G. Complete Reconstruction of an Enzyme-Inhibitor Binding Process by Molecular Dynamics Simulations. *Proc. Natl. Acad. Sci.* **108,** 10184–10189 (2011).

82. Pan, A. C., Sezer, D. & Roux, B. Finding Transition Pathways Using the String Method with Swarms of Trajectories. *J. Phys. Chem. B* **112,** 3432–3440 (2008).

83. Teo, I., Mayne, C. G., Schulten, K. & Lelièvre, T. Adaptive Multilevel Splitting Method for Molecular Dynamics Calculation of Benzamidine-Trypsin Dissociation Time. *J. Chem. Theory Comput.* acs.jctc.6b00277 (2016). doi:10.1021/acs.jctc.6b00277

84. Tiwary, P., Limongelli, V., Salvalaglio, M. & Parrinello, M. Kinetics of Protein–Ligand Unbinding: Predicting Pathways, Rates, and Rate-Limiting Steps. *Proc. Natl. Acad. Sci.* **112,** 201424461 (2015).

85. Northrup, S. H., Allison, S. A. & McCammon, J. A. Brownian Dynamics Simulation of Diffusion-Influenced Bimolecular Reactions. *J. Chem. Phys.* **80,** 1517–1524 (1984).

86. Zhou, H. On the Calculation of Diffusive Reaction Rates Using Brownian Dynamics Simulations. *J. Chem. Phys.* **92,** 3092–3095 (1990).

87. McCammon, J. A., Northrup, S. H. & Allison, S. A. Diffusional Dynamics of Ligand-Receptor Association. *J. Phys. Chem.* **90,** 3901–3905 (1986).

88. Lindert, S., Kekenes-Huskey, P. M. & McCammon, J. A. Long-Timescale Molecular Dynamics Simulations Elucidate the Dynamics and Kinetics of Exposure of the Hydrophobic Patch in Troponin C. *Biophys. J.* **103,** 1784–1789 (2012).

89. Cheng, Y., Suen, J. K., Zhang, D., Bond, S. D., Zhang, Y., Song, Y., Baker, N. A., Bajaj, C. L., Holst, M. J. & McCammon, J. A. Finite Element Analysis of the Time-Dependent Smoluchowski Equation for Acetylcholinesterase Reaction Rate Calculations. *Biophys. J.*

92, 3397–3406 (2007).

90. Boras, B. W., Hirakis, S. P., Votapka, L. W., Malmstrom, R. D., Amaro, R. E. & McCulloch, A. D. Bridging scales through multiscale modeling: a case study on protein kinase A. *Front. Physiol.* **6,** (2015).

91. Votapka, L. W. Numerical and Computational Solutions for Biochemical Kinetics, Druggability, and Simulation. (2016).

92. Kirmizialtin, S. & Elber, R. Revisiting and computing reaction coordinates with directional milestoning. *J. Phys. Chem. A* **115,** 6137–6148 (2011).

93. Ermak, D. L. & McCammon, J. A. Brownian dynamics with hydrodynamic interactions. *J. Chem. Phys.* **69,** 1352–1360 (1978).

94. Faradjian, A. K. & Elber, R. Computing Time Scales from Reaction Coordinates by Milestoning. *J. Chem. Phys.* **120,** 10880–10889 (2004).

95. Vanden-Eijnden, E., Venturoli, M., Ciccotti, G. & Elber, R. On the assumptions underlying milestoning. *J. Chem. Phys.* **129,** 174102 (2008).

96. Votapka, L. W., Lee, C. T. & Amaro, R. E. Two Relations to Estimate Membrane Permeability Using Milestoning. *J. Phys. Chem. B* **120,** 8606–8616 (2016).

97. Skolnick, J. Perspective: On the importance of hydrodynamic interactions in the subcellular dynamics of macromolecules. *J. Chem. Phys.* **145,** 100901 (2016).

98. Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R. D., Kalé, L. & Schulten, K. Scalable Molecular Dynamics with NAMD. *J. Comput. Chem.* **26,** 1781–1802 (2005).

99. Huber, G. A. & McCammon, J. A. Browndye: A Software Package for Brownian Dynamics. *Comput. Phys. Commun.* **181,** 1896–1905 (2010).

100. Baker, N. A., Sept, D., Joseph, S., Holst, M. J. & McCammon, J. A. Electrostatics of nanosystems: Application to microtubules and the ribosome. *Proc. Natl. Acad. Sci.* **98,** 10037–10041 (2001).

101. Pearlman, D. A., Case, D. A., Caldwell, J. W., Ross, W. S., Cheatham, T. E., DeBolt, S., Ferguson, D., Seibel, G. & Kollman, P. AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Comput. Phys. Commun.* **91,** 1–41 (1995).

102. Marquart, M., Walter, J., Deisenhofer, J., Bode, W. & Huber, R. The geometry of the reactive site and of the peptide groups in trypsin, trypsinogen and its complexes with inhibitors. *Acta Crystallogr. Sect. B* **39,** 480–490 (1983).

103. Davis, I. W., Leaver-Fay, A., Chen, V. B., Block, J. N., Kapral, G. J., Wang, X., Murray, L. W., Arendall, W. B., Snoeyink, J., Richardson, J. S. & Richardson, D. C. MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res.* **35,** W375–W383 (2007).

104. Chen, V. B., Arendall, W. B., Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, J. S., Richardson, D. C. & IUCr. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **66,** 12–21 (2010).

105. Maier, J. A., Martinez, C., Kasavajhala, K., Wickstrom, L., Hauser, K. E. & Simmerling, C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **11,** 3696–3713 (2015).

106. Dolinsky, T. J., Nielsen, J. E., McCammon, J. A. & Baker, N. A. PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. *Nucleic Acids Res.* **32,** W665–W667 (2004).

107. Dolinsky, T. J., Czodrowski, P., Li, H., Nielsen, J. E., Jensen, J. H., Klebe, G. & Baker, N. A. PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. *Nucleic Acids Res.* **35,** W522–W525 (2007).

108. Guillain, F. & Thusius, D. The Use of Proflavin as an Indicator in Temperature-Jump Studies of the Binding of a Competitive Inhibitor to Trypsin. *J. Am. Chem. Soc.* **92,** 5534–5536 (1970).

109. Horn, H. W., Swope, W. C., Pitera, J. W., Madura, J. D., Dick, T. J., Hura, G. L. & Head-Gordon, T. Development of an improved four-site water model for biomolecular simulations: TIP4P-Ew. *J. Chem. Phys.* **120,** 9665–9678 (2004).

110. Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. A. Development and testing of a general amber force field. *J. Comput. Chem.* **25,** 1157–1174 (2004).

111. Wang, J., Wang, W., Kollman, P. A. & Case, D. A. Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graph. Model.* **25,** 247–260 (2006).

112. Schindler, P., Robinson, R. A. & Bates, R. G. Solubility of tris(hydroxymethyl)aminomethane in water-methanol solvent mixtures and medium effects in the dissociation of the protonated base. *J. Res. Natl. Bur. Stand. Sect. A Phys. Chem.* **72A,** 141 (1968).

113. Schuetz, D. A., de Witte, W. E. A., Wong, Y. C., Knasmueller, B., Richter, L., Kokh, D. B., Sadiq, S. K., Bosma, R., Nederpelt, I., Heitman, L. H., Segala, E., Amaral, M., Guo, D., Andres, D., Georgi, V., Stoddart, L. A., Hill, S., Cooke, R. M., De Graaf, C., Leurs, R., Frech, M., Wade, R. C., de Lange, E. C. M., IJzerman, A. P., Müller-Fahrnow, A. & Ecker, G. F. Kinetics for Drug Discovery: an Industry-Driven Effort to Target Drug Residence Time. *Drug Discov. Today* **22,** 896–911 (2017).

114. Lu, H. & Tonge, P. J. Drug–Target Residence Time: Critical Information for Lead Optimization. *Curr. Opin. Chem. Biol.* **14,** 467–474 (2010).

115. Copeland, R. A. The Drug-Target Residence Time Model: A 10-Year Retrospective. *Nat. Rev. Drug Discov.* **15,** 87–95 (2016).

116. Swinney, D. C. Opinion: Biochemical Mechanisms of Drug Action: What Does it Take for Success? *Nat. Rev. Drug Discov.* **3,** 801–808 (2004).

117. Amaro, R. E. & Mulholland, A. J. Bridging Biological and Chemical Complexity in the Search for Cures: Multiscale Methods in Drug Design. *Nat Rev Chem* **2,** 0148 (2018).

118. Shaw, D. E., Bowers, K. J., Chow, E., Eastwood, M. P., Ierardi, D. J., Klepeis, J. L., Kuskin, J. S., Larson, R. H., Lindorff-Larsen, K., Maragakis, P., Moraes, M. A., Dror, R. O., Piana, S., Shan, Y., Towles, B., Salmon, J. K., Grossman, J. P., Mackenzie, K. M., Bank, J. A., Young, C., Deneroff, M. M. & Batson, B. Millisecond-Scale Molecular Dynamics Simulations on Anton. in *Proc. Conf. High Perform. Comput. Networking, Storage Anal. - SC '09* 1 (ACM Press, 2009). doi:10.1145/1654059.1654126

119. Shaw, D. E., Grossman, J. P., Bank, J. A., Batson, B., Butts, J. A., Chao, J. C., Deneroff, M. M., Dror, R. O., Even, A., Fenton, C. H., Forte, A., Gagliardo, J., Gill, G., Greskamp, B., Ho, C. R., Ierardi, D. J., Iserovich, L., Kuskin, J. S., Larson, R. H., Layman, T., Lee, L. S., Lerer, A. K., Li, C., Killebrew, D., Mackenzie, K. M., Mok, S. Y. H., Moraes, M. A., Mueller, R., Nociolo, L. J., Peticolas, J. L., Quan, T., Ramot, D., Salmon, J. K., Scarpazza, D. P., Ben Schafer, U., Siddique, N., Snyder, C. W., Spengler, J., Tang, P. T. P., Theobald, M., Toma, H., Towles, B., Vitale, B., Wang, S. C. & Young, C. Anton 2: Raising the Bar for Performance and Programmability in a Special-Purpose Molecular Dynamics Supercomputer. in *Int. Conf. High Perform. Comput. Networking, Storage Anal. SC* **2015-Janua,** 41–53 (IEEE, 2014).

120. Pan, A. C., Borhani, D. W., Dror, R. O. & Shaw, D. E. Molecular determinants of drug–receptor binding kinetics. *Drug Discov. Today* **18,** 667–673 (2013).

121. Tang, Z. & Chang, C. A. Binding Thermodynamics and Kinetics Calculations Using Chemical Host and Guest: A Comprehensive Picture of Molecular Recognition. *J. Chem. Theory Comput.* **14,** 303–318 (2018).

122. Wu, H., Paul, F., Wehmeyer, C. & Noé, F. Multiensemble Markov Models of Molecular Thermodynamics and Kinetics. *Proc. Natl. Acad. Sci.* **113,** E3221–E3230 (2016).

123. Doerr, S. & de Fabritiis, G. On-the-Fly Learning and Sampling of Ligand Binding by High-Throughput Molecular Simulations. *J. Chem. Theory …* **10,** 2064–2069 (2014).

124. Mollica, L., Theret, I., Antoine, M., Perron-Sierra, F., Charton, Y., Fourquez, J.-M., Wierzbicki, M., Boutin, J. A., Ferry, G., Decherchi, S., Bottegoni, G., Ducrot, P. & Cavalli, A. Molecular Dynamics Simulations and Kinetic Measurements to Estimate and Predict Protein–Ligand Residence Times. *J. Med. Chem.* **59,** 7167–7176 (2016).

125. 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).

126. Ma, P., Cardenas, A. E., Chaudhari, M. I., Elber, R. & Rempe, S. B. The Impact of Protonation on Early Translocation of Anthrax Lethal Factor: Kinetics from Molecular Dynamics Simulations and Milestoning Theory. *J. Am. Chem. Soc.* jacs.7b07419 (2017). doi:10.1021/jacs.7b07419

127. Kirmizialtin, S., Nguyen, V., Johnson, K. A. & Elber, R. How Conformational Dynamics of DNA Polymerase Select Correct Substrates: Experiments and Simulations. *Structure* **20,** 618–627 (2012).

128. Ma, W. & Schulten, K. Mechanism of Substrate Translocation by a Ring-shaped ATPase Motor at Millisecond Resolution. *J. Am. Chem. Soc.* **137,** 3031–3040 (2015).

129. Dickson, A. & Lotz, S. D. Ligand Release Pathways Obtained with WExplore: Residence Times and Mechanisms. *J. Phys. Chem. B* **120,** 5377–5385 (2016).

130. Dickson, A. & Lotz, S. D. Multiple Ligand Unbinding Pathways and Ligand-Induced Destabilization Revealed by WExplore. *Biophys. J.* **112,** 620–629 (2017).

131. Lotz, S. D. & Dickson, A. Unbiased Molecular Dynamics of 11 min Timescale Drug Unbinding Reveals Transition State Stabilizing Interactions. *J. Am. Chem. Soc.* jacs.7b08572 (2018). doi:10.1021/jacs.7b08572

132. Chiu, S. H. & Xie, L. Toward High-Throughput Predictive Modeling of Protein Binding/Unbinding Kinetics. *J. Chem. Inf. Model.* **56,** 1164–1174 (2016).

133. Wong, C. F. Steered Molecular Dynamics Simulations for Uncovering the Molecular Mechanisms of Drug Dissociation and for Drug Screening: A Test on the Focal Adhesion Kinase. *J. Comput. Chem.* (2018). doi:10.1002/jcc.25201

134. Tran, D. P., Takemura, K., Kuwata, K. & Kitao, A. Protein-Ligand Dissociation Simulated by Parallel Cascade Selection Molecular Dynamics. *J. Chem. Theory Comput.* **14,** 404–417 (2018).

135. Cardenas, A. E. & Elber, R. Computational study of peptide permeation through membrane: searching for hidden slow variables. *Mol. Phys.* **111,** 3565–3578 (2013).

136. Cardenas, A. E., Shrestha, R., Webb, L. J. & Elber, R. Membrane Permeation of a Peptide: It Is Better to be Positive. *J. Phys. Chem. B* **119,** 6412–6420 (2015).

137. Bello-Rivas, J. M. & Elber, R. Exact milestoning. *J. Chem. Phys.* **142,** 094102 (2015).

138. Shalloway, D. & Faradjian, A. K. Efficient Computation of the First Passage Time Distribution of the Generalized Master Equation by Steady-State Relaxation. *J. Chem. Phys.* **124,** 054112 (2006).

139. West, A. M. A., Elber, R. & Shalloway, D. Extending Molecular Dynamics Time Scales with Milestoning: Example of Complex Kinetics in a Solvated Peptide. *J. Chem. Phys.* **126,** 145104 (2007).

140. Vanden-Eijnden, E. & Venturoli, M. Markovian milestoning with Voronoi tessellations. *J. Chem. Phys.* **130,** 194101 (2009).

141. Májek, P. & Elber, R. Milestoning without a Reaction Coordinate. *J. Chem. Theory Comput.* **6,** 1805–1817 (2010).

142. Cézard, C., Trivelli, X., Aubry, F., Djedaïni-Pilard, F. & Dupradeau, F.-Y. Molecular dynamics studies of native and substituted cyclodextrins in different media: 1. Charge derivation and force field performances. *Phys. Chem. Chem. Phys.* **13,** 15103 (2011).

143. Fukahori, T., Nishikawa, S. & Yamaguchi, K. Kinetics on Isomeric Alcohols Recognition by α- and β-Cyclodextrins Using Ultrasonic Relaxation Method. *Bull. Chem. Soc. Jpn.* **77,** 2193–2198 (2004).

144. Fukahori, T., Kondo, M. & Nishikawa, S. Dynamic Study of Interaction between β-Cyclodextrin and Aspirin by the Ultrasonic Relaxation Method. *J. Phys. Chem. B* **110,** 4487–4491 (2006).

145. Nishikawa, S., Fukahori, T. & Ishikawa, K. Ultrasonic relaxations in aqueous solutions of propionic acid in the presence and absence of beta-cyclodextrin. *J. Phys. Chem. A* **106,** 3029–3033 (2002).

146. Nishikawa, S. & Kondo, M. Kinetic Study for the Inclusion Complex of Carboxylic Acids with Cyclodextrin by the Ultrasonic Relaxation Method. *J. Phys. Chem. B* **110,** 26143–26147 (2006).

147. Rekharsky, M. V & Inoue, Y. Complexation Thermodynamics of Cyclodextrins. *Chem. Rev.* **98,** 1875–1918 (1998).

148. Barros, T. C., Stefaniak, K., Holzwarth, J. F. & Bohne, C. Complexation of Naphthylethanols with β-Cyclodextrin. *J. Phys. Chem. A* **102,** 5639–5651 (1998).

149. 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).

150. Tummino, P. J. & Copeland, R. A. Residence time of receptor-ligand complexes and its effect on biological function. *Biochemistry* **47,** 5481–92 (2008).

151. Romanowska, J., Kokh, D. B., Fuller, J. C. & Wade, R. C. in 211–235 (John Wiley & Sons, Ltd, 2015). doi:10.1002/9783527673025.ch11

152. Luty, B. A., McCammon, J. A. & Zhou, H. X. Diffusive reaction rates from Brownian dynamics simulations: Replacing the outer cutoff surface by an analytical treatment. *J.*

Chem. Phys. **97,** 5682–5686 (1992).

153. Huang, D. & Caflisch, A. The free energy landscape of small molecule unbinding. *PLoS Comput. Biol.* **7,** e1002002 (2011).

154. Bowman, G. R., Pande, V. S. & Noé, F. *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation*. *Springer* **797,** (Springer Netherlands, 2014).

155. Mollica, L., Decherchi, S., Zia, S. R., Gaspari, R., Cavalli, A. & Rocchia, W. Kinetics of protein-ligand unbinding via smoothed potential molecular dynamics simulations. *Sci. Rep.* **5,** 11539 (2015).

156. Tang, Z., Chen, S.-H. & Chang, C. A. Transient States and Barriers from Molecular Simulations and the Milestoning Theory: Kinetics in Ligand–Protein Recognition and Compound Design. *J. Chem. Theory Comput.* acs.jctc.9b01153 (2020). doi:10.1021/acs.jctc.9b01153

157. Miao, Y., Huang, Y. M., Walker, R. C., McCammon, J. A. & Chang, C. A. Ligand Binding Pathways and Conformational Transitions of the HIV Protease. *Biochemistry* **57,** 1533–1541 (2018).

158. Haldar, S., Comitani, F., Saladino, G., Woods, C., Van Der Kamp, M. W., Mulholland, A. J., Gervasio, F. L., Van Der Kamp # §, M. W. & †department, ‡ *. A Multiscale Simulation Approach to Modelling Drug-Protein Binding Kinetics A Multiscale Simulation Approach to Modelling Drug-Protein Bind-ing Kinetics. *Just Accept. Manuscr.* • (2018). doi:10.1021/acs.jctc.8b00687

159. Jagger, B. R., Kochanek, S. E., Haldar, S., Amaro, R. E. & Mulholland, A. J. Multiscale simulation approaches to modeling drug–protein binding. *Curr. Opin. Struct. Biol.* **61,** 213–221 (2020).

160. Bello-Rivas, J. M. & Elber, R. Simulations of thermodynamics and kinetics on rough energy landscapes with milestoning. *J. Comput. Chem.* **37,** 602–613 (2016).

161. Maragliano, L., Vanden-Eijnden, E. & Roux, B. Free Energy and Kinetics of Conformational Transitions from Voronoi Tessellated Milestoning with Restraining Potentials. *J. Chem. Theory Comput.* **5,** 2589–2594 (2009).

162. Fiorin, G., Klein, M. L. & Hénin, J. Using collective variables to drive molecular dynamics simulations. *Mol. Phys.* **111,** 3345–3362 (2013).

163. Vanden-Eijnden, E. & Venturoli, M. Revisiting the finite temperature string method for the calculation of reaction tubes and free energies. *J. Chem. Phys.* **130,** (2009).