UNIVERSITY OF CALIFORNIA,
IRVINE

Improved Methods of Simulation and Analysis For Stochastic Processes in Cell Biology

DISSERTATION

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

DOCTOR OF PHILOSOPHY
in Chemical Engineering

by

Brian Kelly Chu

Dissertation Committee:
Assistant Professor Elizabeth L. Read, Chair
Associate Professor Jun Allard
Professor Frank G. Shi

2019

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to thank Elizabeth L. Read for guidance and support throughout the last five years. It has been a privilege to work together and pursue research in the field of systems biology. I am also grateful to Jun Allard for his mentorship on the spatial systems studies. Finally, I would like to mention my colleagues in the Read Lab, especially Margaret Tse, who really set a great example for me as a researcher.

CURRICULUM VITAE

Brian Chu

EDUCATION

Doctor of Philosophy in Chemical Engineering	**2019**
University of California, Irvine	Irvine, CA
Master of Science in Chemical Engineering	**2014**
University of California, Irvine	Irvine, CA
Bachelors of Science in Chemical Engineering	**2013**
University of California, Irvine	Irvine, CA

RESEARCH EXPERIENCE

Undergraduate Research Assistant	**2012-2013**
University of California, Irvine	Irvine, CA
Graduate Research Assistant	**2013-2018**
University of California, Irvine	Irvine, CA

TEACHING EXPERIENCE

Teaching Assistant	**2014-2018**
University of California, Irvine	Irvine, CA
CBEMS 149a	**Winter 2014**
CBEMS 230	**Fall 2015**
CBEMS 135	**Fall 2016**
CBEMS 140b	**Winter 2017**
CBEMS 45c	**Spring 2017**
CBEMS 45c	**Spring 2018**

REFEREED JOURNAL PUBLICATIONS

Stimuli-Responsive Materials: The Role of Electrostatics and Temperature on Morphological Transitions of Hydrogel Nanostructures Self-Assembled by Peptide Amphiphiles Via Molecular Dynamics Simulations **2013**
Advanced Healthcare Materials

Iris W. Fu, Cade B. Markegard, **Brian K. Chu**, Hung D. Nguyen

The Role of Hydrophobicity on Self-Assembly by Peptide Amphiphiles via Molecular Dynamics Simulations. **2014**
Langmuir

Iris W. Fu, Cade B. Markegard, **Brian K. Chu**, Hung D. Nguyen

A Tail of Two Peptide Amphiphiles: Effect of Conjugation with Hydrophobic Polymer on Folding of Peptide Sequences **2014**
Biomacromolecules

Brian K. Chu, Iris W. Fu, Cade B. Markegard, Hung D. Nguyen

DNA-Binding Kinetics Determines Mechanism of Noise-Induced Switching in Gene Networks **2015**
Biophysical Journal

Margaret J. Tse, **Brian K. Chu**, Mahua Roy, and Elizabeth L. Read

Markov State Models of Gene Regulatory Networks **2017**
BMC Systems Biology

Brian K. Chu, Margaret J. Tse, Royce R. Sato, and Elizabeth L. Read

Rare Event Sampling of Epigenetic Landscapes and Phenotype Transitions **2018**
PLOS Computational Biology

Margaret J. Tse, **Brian K. Chu**, Cameron P. Gallivan, and Elizabeth L. Read

Hydrodynamics of transient cell-cell contact: The role of membrane permeability and active protrusion length (in review) **2019**

PLOS Computational Biology
Kai Liu, **Brian Chu**, Jay Newby, Elizabeth Read, John Lowengrub, Jun Allard

Simulation of Rare Events in the Diffusion of Molecules on Crowded Cell Surfaces (in preparation) **2019**

Brian Chu, Robert Taylor, Jay Newby, Rob Jun Allard, Elizabeth Read

ABSTRACT OF THE DISSERTATION

Improved Methods of Simulation and Analysis For Stochastic Processes in Cell Biology

By

Brian Kelly Chu

University of California, Irvine, 2019

Assistant Professor, Elizabeth L. Read, Chair

Stochasticity (that is, randomness) is an inherent property of many biological systems. For example, gene expression is stochastic, resulting in random fluctuations of mRNA and protein copy numbers in the cell. In cell differentiation, there is evidence that the phenotype of the cell can be driven toward an entirely different type of cell due to noise. Stochastic fluctuations are also important in the spatio-temporal dynamics of molecular interactions within the cell, affecting processes such as cell activation and signal transduction. To gain a better understanding of biological systems, computer simulations of biomolecular processes in the cell are increasingly utilized to complement experiments, quantify mechanistic hypotheses, and predict the effect of perturbations. Stochastic models, in particular, can be prohibitively expensive to simulate and difficult to analyze. In this work, we develop and extend methods of stochastic simulation and analysis that are applicable to a variety of cell biological systems. We focus on two specific application areas: The first is development of a method to analyze gene regulatory network models that have multiple, metastable states. The method enables a simplified, quantitative representation of complex phenotype landscapes and transitions. Second is the development of improved simulation methods for spatial stochastic systems. This work focuses on rare events in reaction-diffusion systems and found several extensions to currently-employed simulation methods which improve simulation efficiency.

1. Introduction

1.1 Stochastic biological models

Stochasticity is prevalent in a host of biochemical systems involving gene expression [1], cellular heterogeneity[2], and spatial heterogeneity[3]–[5]. While deterministic modeling can predict the static average behavior at the population level, systems that are inherently metastable can dynamically switch between states and exhibit multistable distributions. Experimental evidence of distributions with multiple peaks is observed in [6]–[8] using various single cell analysis techniques which allows for study of populations at the resolution of a single cell.

In deterministic modeling, a standard analysis method of solving for the attractor states is solving for the fixed points in the system. In stochastic modeling, it is more complex due to the metastability and heterogeneity of the attractor states. Intrinsic noise drives transitions between attractors and the states can be belong partially to more than one attractor state. This complexity along with the increased computational power required for stochastic methods makes it a challenge to apply in practice and analyze the noisy results, which can impact the statistical errors.

Models can make predictions about the system behavior in a multitude of regimes not covered by experiments. They are also useful as probing the unknown mechanisms of the poorly understood processes that experimental data cannot explain. For example, a model can serve as a means to test a hypothesis; the assumptions of the unknown mechanisms made by the model can be verified by comparing the simulation results to experimental results[9]. Models do not have to be fully accurate in order to be useful; often simple models can be build our intuition of complex processes[10].In this thesis, the models that we use to model biological processes are based on stochastic methods since noise and random fluctuations are key drivers of their dynamics and system state transitions.

1.2 Gene regulatory networks

A gene regulatory network is a description of the interactions between genes which governs the cell's state or phenotype. Mapping the interactions is a difficult process due to the sheer number of genes in the networks and the combinatorial explosion of the possible number of states. A variety of networks have been mapped, such as the heart [11], plants[12], and various animals[13].

A greater understanding of cell fate can be gained by incrementally mapping gene regulatory networks from low level to high level. From the thousands of molecular entities at the organism level, the key ones can be identified and established their respective roles in the cell fate process. The molecular entities can be grouped into 'sub-circuits' or 'modules' according to the process that they are involved. By piecing together how each module effects one other, a mechanistic understanding can be learned of how the cell fate process works.

High-throughput technologies, such as DNA microarrays or the ChIP-on-chip technique, have allowed us to acquire massive amounts of genomic data. The current challenge is how to utilize this data to make predictions or learn more about the network behavior. Two major computational approaches are top-down and bottom-up. The top-down approaches uses data inference to reverse engineer a trained model which typically does not provide insight on detailed molecular mechanisms, but is useful in gene network reconstruction. The bottom-up approach is to build a dynamic model, which does include some molecular detail, such as kinetics of Transcription Factor-DNA interactions, and fits parameters for those processes from the data where possible. This thesis uses the latter approach.

1.3 Modeling of gene regulatory networks

Gene regulatory networks are complicated due to the vast number of interactions and molecular entities at play. Therefore, different modeling approaches are selected based on the system size/complexity. The least expensive/simple modeling approach are Boolean models [14]. The entities are abstracted into being a highly expressing ON state or a low expressing OFF state. This greatly reduces the number of possible states and the computational work required to train the model. A more accurate method, continuous models, such as ordinary differential equations and stochastic differential equations, can model the dynamic behavior of molecular concentrations[15]. Often these types of models assume gene expression levels depend solely on the regulatory proteins [16]. Stochastic models, which treat entities as discrete quantities, are able to account for the transient switching of gene expression while continuous models cannot. The transient switching of gene expression is important for cellular functions such as priming cells to develop along differentiation pathways[17], [18] and bistable cytokine expression levels in immune cells[19]. The stochastic formalism that we have applied in Chapter 2 is the Chemical Master Equation (CME), which describes the time evolution of the probability density vector of all possible cell state configurations. It is very accurate and detailed but can only be applied to the simplest of systems due to the combinatorial explosion effect when scaling up. In Chapter 2, we extend the CME framework by combining with Markov State Models methodology, which is a coarse graining method to extract the slow time scale dynamics from the system that was modeled by the CME framework.

1.4 Receptor-ligand binding at cell-cell interfaces

Another area in which stochastic models have been utilized to shed light on cellular processes is in the spatio-temporal dynamics of biochemical interactions that govern signal transduction and cell activation. For example, in T-cells, the binding of a receptor to its ligand produces cellular responses, such as proliferation, differentiation, and survival. In order for the binding to occur, two cell membranes must be able to form an interface at close contact. Repulsive forces and steric interactions between the two membranes prevent the surfaces from forming interfaces. Steric interactions can be attributed to large surface molecules which can physically block the two surfaces due their long ectodomain length [20], [21], such as CD45, a molecule that play a dual role in T-cell activation[22]. One of the steps of the kinetic segregation model of receptor triggering [23], a theory on how close contacts form between T-cell surfaces, suggests that CD45 molecules must first evacuate the region of binding. Chapter 3 deals with calculating the mean first passage times associated with this evacuation event using simplified spatial simulations. Repulsive forces can stem from the hydrodynamics of the extracellular fluid between the two membranes [24], [25]. When attempting to come into close contact, a “thin layer effect”[26] appears, where the repulsion force grows stronger as the two membranes come closer together. The aim of chapter 4 is to characterize the mean first passage time associated with reaching various displacements of contact separation using simplified stochastic models.

1.5 Rare event sampling

1.5.1 Rare events

Rare events are defined as an event of interest (often, a transition to a target state) which occurs very infrequently relative to other intrinsic timescales of the system. They are tied to biological processes that may be metastable in nature (gene regulatory networks fall into this category as shown in Chapter 2) or transient (certain significant events pertaining to cell-cell interfaces described in Chapter 3 and 4). Of gene regulatory networks, embryonic stem cells exhibit dynamic heterogeneity through expression levels of its transcription factors prior lineage commitment [17], [27], [28]. In order to quantify the transition rates and probabilities of the embryonic cell network, a methodology able to identify metastable states and quantify their transitions is needed. It also must be computationally efficient due to the presence of rare events in these systems. Such a framework for capturing the transitions between metastable states as well as the rare states was developed by Tse et. al [29]. The underlying algorithm which allowed for efficient sampling is Weighted Ensemble Method [30], which is used extensively in this thesis and is described in detail in the next section.

1.5.2 Weighted Ensemble sampling

The primary rare event sampling method used for Chapters 3 and 4 is the Weighted Ensemble (WE) Method, which partitions the state space into bins and records the flux of probability from one initial state to a final state [30]. In biological processes ranging from different scales such as molecular [31], atomistic [32], [33], coarse grained models[34], and well-mixed/spatial resolved cellular level[35], [36], WE has been used to calculate the rate constant and gain insight on rare transitions of interest with greater computational efficiency compared to brute force simulation.

Other examples of techniques which that can also be used the obtain the rate constant in rare event systems include milestoning [37], transition path sampling[38] , forward flux sampling[39]. These algorithms and WE share the characteristic of dividing the space up in terms of some progress coordinate, a degree of freedom which is highly indicative of the current state of the system.

There have been many variants of the Weighted Ensemble Algorithms [35], [40]–[45] which have reported greater efficiency compared to the original Huber and Kim algorithm [30]. For example, Bhatt and coworkers have used an accelerated steady state attainment procedure [42] and Dickson and coworkers have improved the binning procedure using a hierarchy of bins that is built adaptively throughout the simulation[43]. The original algorithm does not place strict restrictions on number of replicas per bin unlike the version implementation by Donovan and coworkers [35]. The reorganization of replicas is solely determined the ideal weight of a replica with a bin. Replicas above the splitting threshold and below the merging threshold are candidates for modification. WESTPA, a comprehensive package which implements the WE method designed for general use for a multitude of simulation packages that use different dynamics such as molecular dynamics (GROMACS,NAMD,AMBER) and cell-modeling (BioNetGen, MCell), has the options to specify the splitting/merging thresholds in addition to the target replica number [46].

Although it has been used to solve a wide variety of rare event problems, complex systems with multiple degrees of freedom pose an issue with weighted ensemble in the control of the number of bins the coordinate to choose to define your bins. It may happen that the slow degrees of freedom are correlated so it is sufficient to describe the bins with a one-dimensional coordinate

[47], [48]. WExplore [43] was designed to handle multiple slow independent degrees of freedom and was shown to handle up to 50 degrees of freedom. In the last chapter, an incremental approach for high dimensional progress coordinates is suggested and explained.

2. Markov State Models of Gene Regulatory Networks

2.1 Introduction

Gene regulatory networks (GRNs) often have dynamics characterized by multiple attractor states. This multistability is thought to underlie cell fate-decisions. According to this view, each attractor state accessible to a gene network corresponds to a particular pattern of gene expression, i.e., a cell phenotype. Bistable network motifs with two possible outcomes have been linked to binary cell fate-decisions, including the lysis/lysogeny decision of bacteriophage lambda[18], the maturation of frog oocytes[49] and a cascade of branch-point decisions in mammalian cell development (reviewed in [50]). Multistable networks with three or more attractors have been proposed to govern diverse cell fate-decisions in tumorigenesis [51], stem cell differentiation and reprogramming [52]–[54], and helper T cell differentiation[55]. More generally, the concept of a rugged, high-dimensional epigenetic landscape connecting every possible cell type has emerged [56]–[58]. Quantitative models that can link molecular-level knowledge of gene regulation to a global understanding of network behavior have the potential to guide rational cell-reprogramming strategies. As such, there has been growing interest in the development of theory and computational methods to analyze global dynamics of multistable gene regulatory networks.

Gene expression is inherently stochastic[18], [59]–[61], and fluctuations in expression levels can measurably impact cell phenotypes and behavior. Numerous examples of stochastic phenotype transitions have been discovered, which diversify otherwise identical cell-populations. This spontaneous state-switching has been found to promote survival of microorganisms or cancer cells in fluctuating environments[62]–[64], prime cells to follow alternate developmental fates in higher eukaryotes [27], [65], and generate sustained heterogeneity (mosaicism) in a homeostatic mammalian cell population[66]. These findings have motivated theoretical studies of stochastic state-switching in gene networks, which have shed light on network parameters and topologies that promote the stability (or instability) of a given network state[66]. Characterizing the global stability of states accessible to a network is akin to quantification of the "potential energy" landscape of a network. Particularly, with the advent of stem-cell reprogramming techniques, there has been renewed interest in a quantitative reinterpretation of Waddington's classic epigenetic landscape[67], in terms of underlying regulatory mechanisms[57], [68].

A number of mathematical frameworks exist for modeling and analysis of stochastic gene regulatory network (GRN) dynamics (reviewed in [69], [70]), including probabilistic Boolean Networks, Stochastic Differential Equations, and stochastic biochemical reaction networks (i.e., Chemical Master Equations). Of these, the Chemical Master Equation (CME) approach is the most complete, in that it treats all biomolecules in the system as discrete entities, fully accounts for stochasticity due to molecular-level fluctuations, and propagates dynamics according to chemical rate laws. The CME is analytically intractable, but trajectories can be simulated by Monte Carlo methods such as the Stochastic Simulation Algorithm (SSA)[71]. Alternatively, methods for reducing the dimensionality of the CME, enabling numerical approximation of network behavior by matrix methods, have been developed[72]–[76].

Analysis of multistability and global dynamics of discrete, stochastic GRN models remains challenging. Multistability is generally assessed by plotting multi-peaked steady-state probability distributions (obtained either from long stochastic simulations [52], [77], [78] or from approximate CME solutions[76], [79], [80]), projected onto one or two user-specified system coordinates. However, even small networks generally have more than two dimensions along which dynamics may be projected, meaning that inspection of steady-state distributions for a given projection may underestimate multistability in a network. For example, the state-space of a GRN may comprise different activity-states of promoters and regulatory sites on DNA, the copy-number of mRNA transcripts and encoded proteins, and the activity- or multimer-states of multiple regulatory molecules or proteins. Furthermore, while steady-state distributions give a global view of system behavior, they do not directly yield dynamic information of interest, such as the lifetimes of attractor states.

In this chapter, we present an approach for analyzing multistable dynamics in stochastic GRNs based on a spectral clustering method widely applied in Molecular Dynamics[81]–[83]. The output of the approach is a Markov State Model (MSM)—a coarse-grained model of system dynamics, in which a large number of system states (i.e., "microstates") is clustered into a small number of metastable (that is, relatively long-lived) "macrostates", together with the conditional probabilities for transitioning from one macrostate to another on a given timescale. The MSM approach identifies clusters based on separation of timescales, i.e., systems with multistability exhibit relatively fast transitions among microstates within attractor basins and relatively slow inter-basin transitions. By neglecting fast transitions, the size of the system is vastly reduced. Based on its utility for visualization and analysis of Molecular Dynamics, the potential application of the MSM framework to diverse dynamical systems, including biochemical networks, has been discussed[84].

Biochemical reaction networks present an unexplored opportunity for the MSM approach. Herein, we applied the method to small GRN motifs and analyzed their global dynamics using two frameworks: the quasipotential landscape (based on the log-transformed stationary probability distribution), and the MSM. The MSM approach distilled network dynamics down to the essential stationary and dynamic properties, including the number and identities of stable phenotypes encoded by the network, the global probability of the network to adopt a given phenotype, and the likelihoods of all possible stochastic phenotype transitions. The method revealed the existence of network states and processes not readily apparent from inspection of quasipotential landscapes. Our results demonstrate how MSMs can yield insight into regulation of cell phenotype stability and reprogramming. Furthermore, our results suggest that, by delivering systematic coarse-graining of high-dimensional (i.e., many-species) dynamics, MSMs could find more general applications in Systems Biology, such as in signal-transduction, evolution, and population dynamics.

2.2 Methods

2.2.1 Gene regulatory network motifs

We studied two common GRN motifs that are thought to control cell fate-decisions. The full lists of reactions and associated rate parameters for each network are given in the Supplement. Both motifs consist of two mutually-inhibiting genes, denoted by *A* and *B*. In the Exclusive Toggle Switch (ETS) motif, each gene encodes a transcription factor protein; the protein forms homodimers, which are capable of binding to the promoter of the competing gene, thereby repressing its expression. One DNA-promoter region controls the expression of both genes; when a repressor is bound, it excludes the possibility of binding by the repressor encoded by the competing gene. Therefore, the promoter can exist in three possible binding configurations, $P_{00,}$ $P_{10,}$ and P_{01}, denoting the unbound, a_2-bound, or b_2-bound states, respectively. Production of new protein molecules (including all processes involved in transcription, translation, and protein synthesis) occurs at a constant rate, which depends on the state of the promoter. When the gene is repressed, the encoded protein is produced at a low rate, denoted g_0. When the gene is not repressed, protein is produced at a high rate, g_1. For example, when the promoter state is P_{10} the *a* protein is produced at rate g_1, and the *b* protein is produced at g_0. When the promoter is unbound, neither gene is repressed, causing both proteins to be produced at rate g_1.

In the Mutual Inhibition/Self-Activation (MISA) motif, each homodimeric transcription factor also activates its own expression, in addition to repressing the other gene. The *A* and *B* genes are controlled by separate promoters, and each promoter can be bound by repressor and activator simultaneously. Therefore, the *A*-promoter can exist in four possible states, $A_{00,}$ $A_{10,}$ A_{01} and A_{11}, denoting unbound, *a_2*-activator bound, *b_2*-repressor bound, and both transcription factors bound, respectively (and similarly for the B-promoter). Proteins are produced at rate *g_1* only when the activator is bound and the repressor is unbound. For example, the A_{10} promoter state allows a protein to be produced at *g_1*. The other three *A* promoter states result in a protein being produced at rate *g_0*. Similarly, the rate of *b* protein production depends only on the binding configuration of the *B*-promoter. In both the ETS and MISA networks, protein dimerization is assumed to occur simultaneously with binding to DNA. All rate parameters are given in Tables S1-2.

2.2.2 Chemical Master Equation

The stochastic dynamics are modeled by the discrete, Markovian Chemical Master Equation, which gives the time-evolution of the probability to observe the system in a given state over time. In vector-matrix form, the CME can be written

$$\frac{d\mathbf{p}(\mathbf{x},t)}{dt} = \mathbf{Kp}(\mathbf{x},t)$$

where $\mathbf{p}(\mathbf{x},t)$ is the probability over the system state-space at time t, and $\mathbf{K}$ is the reaction rate-matrix. The off-diagonal elements K_{ij} give the time-independent rate of transitioning from state $\mathbf{x_i}$ to $\mathbf{x}_j$, and the diagonal elements are given by $K_{ii} = -\sum_{j\neq i} K_{ji}$. We assume a well-mixed system of reacting species, and the state of the system is fully specified by $\mathbf{x} \in \mathbb{N}^S$, a state-vector containing the positive-integer values of all *S* molecular species/configurations. We hereon denote these state-vectors as “microstates” of the system. In the ETS network, $\mathbf{x} = [n_A, n_B, P_{ab}]$, where n_A is the copy-number of a molecules (protein monomers expressed by gene *A*, and similar for *B*), and P_{ab}indexes the promoter binding-configuration. In the MISA network, $\mathbf{x} = [n_A, n_B, A_{ab}, B_{ba}]$, which lists the protein copy numbers and promoter configuration-states associated with both genes.

The reaction rate matrix $\mathbf{K} \in \mathbb{R}^{N \times N}$is built from the stochastic reaction propensities (Supplement Eq. 1)[85], for some choice of enumeration over the state-space with N reachable microstates. In general, if a system of S molecular species has a maximum copy number per species of $n_{\max}$, then $N \sim {n_{\max}}^{S}$. To enumerate the system state-space, we neglect microstates with protein copy-numbers larger than a threshold value, which exceeds the maximum steady-state gene expression rate, g_1/k (where $g_1$is the maximum production rate of protein and k is the degradation rate), as these states are rarely reached. This truncation of the state-space introduces a small approximation error, which we calculate using the Finite State Projection method [86] (Fig. S1).

2.2.3 Stochastic simulations

Stochastic simulations were performed according to the SSA method [71], implemented by the software package StochKit2 [87].

2.2.4 Quasipotential landscape

The steady-state probability $\boldsymbol{\pi}(\mathbf{x})$ over N microstates is obtained from $\mathbf{K}$ as the normalized eigenvector corresponding to the zero-eigenvalue, satisfying $\mathbf{K}\boldsymbol{\pi}(\mathbf{x})=\mathbf{0}$[15]. Quasipotential landscapes were obtained from $\boldsymbol{\pi}(\mathbf{x})$ using a Boltzmann definition, $U(\mathbf{x}) = -\ln(\boldsymbol{\pi}(\mathbf{x}))$[68]. All matrix calculations were performed with MATLAB[88].

2.2.5 Markov State Models: Mathematical background

The last 15 years have seen continual progress in development of theory, algorithms, and software implementing the MSM framework. We briefly summarize the theoretical background here; the reader is referred to other works (e.g., [89]–[93]) for more details.

The MSM is a highly coarse-grained projection of system dynamics over N microstates onto a reduced space of selected size C (generally, $C \ll N$). The C states in the projected dynamics are constructed by clustering together microstates that experience relatively fast transitions among them. The C clusters, also called "almost invariant aggregates"[94], are hereon denoted "macrostates".

The MSM approach makes use of Robust Perron Cluster Analysis [95] (PCCA+), a spectral clustering algorithm that takes as input a row-stochastic transition matrix, $\mathbf{T}(\tau)$ which gives the conditional probability for the system to transition between each pair of microstates within a given lagtime τ . The lagtime determines the time-resolution of the model, as expressed by the transition matrix. Off-diagonal elements T_{ij} give the probability of the system to transition to microstate j within τ, given that it was initialized in i. Diagonal elements T_{ii} give the conditional probability to remain in microstate i over the τ interval, and thus rows sum to 1. $\mathbf{T}(\tau)$ is directly obtained from the reaction rate matrix by [96]:

$$\mathbf{T}(\tau) = \exp(\tau \mathbf{K}^{\mathrm{T}}),$$

(where exp denotes the matrix exponential). The evolution of the probability over discrete intervals of τ is given by the Chapman-Kolmogorov equation,

$$\boldsymbol{p}^T(\mathbf{x}, t + k\tau) = \boldsymbol{p}^T(\mathbf{x}, t)\mathbf{T}^k(\tau).$$

For an ergodic system (i.e., any state in the system can be reached from any other state in finite time), $\mathbf{T}(\tau)$ will have one largest eigenvalue, the Perron root, $\lambda_1 = 1$. The stationary probability is then given by the normalized left-eigenvector corresponding to the Perron eigenvalue,

$\boldsymbol{\pi}^{\mathrm{T}}(\mathbf{x})\mathbf{T}(\tau) = \boldsymbol{\pi}^{\mathrm{T}}(\mathbf{x})$.

If the system exhibits multistability, then the dynamics can be approximately separated into fast and slow processes, with fast transitions occurring between microstates belonging to the same metastable macrostate, and slow transitions carrying the system from one macrostate to another. Then $\mathbf{T}(\tau)$ is nearly decomposable, and will exhibit an almost block-diagonal structure (for an appropriate ordering of microstates) with C nearly uncoupled blocks. In this case, the eigenvalue spectrum of $\mathbf{T}(\tau)$ shows a cluster of C eigenvalues near $\lambda_1 = 1$, denoting C slow processes (including the stationary process), and for $i > C$, $\lambda_i \ll \lambda_C$, corresponding to rapidly decaying processes. The system timescales can be computed from the eigenvalue spectrum according to $t_i = -\tau/\ln|\lambda_i(\tau)|$.

The PCCA+ algorithm obtains fuzzy membership vectors $\boldsymbol{\chi}$=$[\chi_1, \chi_2, \ldots, \chi_C] \in \mathbb{R}^{N\times C}$, which assigns microstates $i \in \{1, \ldots, N\}$ to macrostates $j \in \{1, \ldots, C\}$ according to grades (i.e., probabilities) of membership, $\chi_j(i) \in [0,1]$. The membership vectors satisfy the linear transformation:

$$\boldsymbol{\chi} = \boldsymbol{\psi}\mathbf{B}$$

Where $\boldsymbol{\psi} = [\psi_1, \ldots, \psi_C]$ is the $N \times C$ matrix constructed from the C dominant right-eigenvectors of $\mathbf{T}(\tau)$, and $\mathbf{B}$ is a non-singular matrix whose elements are determined by an optimization procedure. The original PCCA method [94] used the sign structure of the eigenvectors to identify almost invariant aggregates, in an optimization procedure with the objective of maximizing metastability via the trace of the coarse-grained matrix. A new optimization algorithm was introduced in a method known as PCCA+ [95], which improved numerical robustness. The results of this chapter were generated using the PCCA+ implementation of MSMBuilder2 [97].

2.2.6 Construction of Markov State Models and pathway decomposition

The PCCA+ algorithm generates a fuzzy discretization. We convert fuzzy values into a so-called “crisp” partitioning of N states into C clusters, which entirely partitions the space with no overlap, by assigning $\chi_j^{crisp}(i) \in \{0,1\}$. That is, $\chi_j^{crisp}(i) = 1$ if the *j*th element of the row vector $\chi(i)$ is maximal, and 0 otherwise. Transition probabilities are estimated over the C coarse-grained sets by summing over the fluxes, or equivalently[98]:

$$\widetilde{\mathbf{T}}(\tau) = \widetilde{\boldsymbol{D}}^{-1}\boldsymbol{\chi}^T\boldsymbol{D}\mathbf{T}(\tau)\boldsymbol{\chi},$$

where $\widetilde{\mathbf{T}}(\tau) \in \mathbb{R}^{C\times C}$ is the coarse-grained Markov State Model and $\boldsymbol{D}$ is the diagonal matrix obtained from the stationary probability vector, $\boldsymbol{D} = \mathrm{diag}(\pi_1, \dots, \pi_N)$. The coarse-grained probability $\widetilde{\boldsymbol{\pi}}(\mathbf{x})$ is obtained by $\widetilde{\boldsymbol{\pi}}(\mathbf{x}) = \boldsymbol{\chi}^{\mathrm{T}}\boldsymbol{\pi}(\mathbf{x})$, and $\widetilde{\boldsymbol{D}} = \mathrm{diag}(\tilde{\pi}_1, \dots, \tilde{\pi}_C)$.

The Markov State Model is visualized using the PyEmma 2 plotting module[89], where the magnitude of the transition probabilities and steady state probabilities are represented by the thickness of the arrows and size of the circles, respectively.

Upon construction of the Markov State Model, transition-path theory[99]–[101] was applied in order to compute an ensemble of transition paths connecting two states of interest, along with their relative probabilities. This was achieved by applying a pathway decomposition algorithm adapted from Noe, et al. in a study of protein folding pathways[101] (details in Supplement).

2.3 Results

2.3.1 Eigenvalues and eigenvectors of the stochastic transition matrix reveal slow dynamics in gene networks

In order to explore the utility of the MSM approach for analyzing global dynamics of gene networks, we studied common motifs that control lineage decisions. The MISA network motif (Fig. 2.1A, Supplement, and Methods) has been the subject of previous theoretical studies and is thought to appear in a wide variety of binary fate-decisions. [102]–[104]. In the network model, the *A/B* gene pair represents known antagonistic pairs such as Oct4/Cdx2, PU.1/Gata1, and GATA3/T-bet, which control lineage decisions in embryonic stem cells, common myeloid progenitors, and naïve T-helper cells, respectively [56], [105], [106]. In general, a particular cell lineage will be associated with a phenotype in which one of the genes is expressed at a high level, and the other is expressed at a low (repressed) level. The MISA network has been reported to have up to four attractors[51], [52], corresponding to the *A/B* gene pair expression combinations Lo/Lo, Lo/Hi, Hi/Lo, and Hi/Hi. We computed the probability and quasipotential landscape of the MISA network. For a symmetric system with sufficiently balanced rates of activator and repressor binding and unbinding from DNA, four peaks (attractor basins) can be distinguished in the steady state probability (quasipotential) landscape, plotted as a function of protein *a* copy number vs. protein *b* copy number (Fig. 2.1A,B).

The Markov State Model framework has been applied in studies of protein folding, where dynamics occurs over rugged energetic landscapes characterized by multiple long-lived states (reviewed in [91], [107]). Therefore, we reasoned that the approach could be useful for studying global dynamics of multistable GRNs. The method identifies the slowest system processes based on the dominant eigenvalues and eigenvectors of the stochastic transition matrix, $\mathbf{T}(\tau)$, which gives the probability of the system to transition from every possible initial state to every possible destination state within lagtime τ (with τ having units of $k^{-1}$and k being the rate of protein degradation). Inspection of the eigenvalue spectrum of $\mathbf{T}(\tau = 5)$ for the MISA network in Fig. 2.1B reveals four eigenvalues near 1 followed by a gap, indicating four system processes that are slow on this timescale. Decreasing τ to 0.5 reveals a step-structure in the eigenvalue spectrum, suggesting a hierarchy of system timescales. The timescales are related to the eigenvalues according to $t_i = -\tau/\ln|\lambda_i(\tau)|$. The Perron eigenvalue $\lambda_1 = 1$ is associated with the stationary (infinite time) process, and the lifetimes t_2 through t_5 are computed to be {95.6, 49.4, 30.8, 2.6} (in units of k^{-1}). Thus, the first gap in the eigenvalue spectrum arises from a more than ten-fold separation in timescales between t_4 and t_5. The original PCCA method[94] used the sign structure of the eigenvectors to assign cluster memberships. Plotting the left-eigenvectors corresponding to the four dominant eigenvalues in the MISA network is instructive: the stationary landscape is obtained from the first eigenvector (ϕ_1), which is positive over all microstates, while the opposite-sign regions in $\phi_2,\ \phi_3,\ \phi_4$ reveal the nature of the slow processes (Fig. 1D). An eigenvector with regions of opposite sign corresponds to an exchange between those two regions (in both directions, since eigenvectors are sign-interchangeable). For example, the slowest process corresponds to exchange between the $a > b$ and $b > a$ regions of state-space, i.e., switching between *B*-gene dominant and *A*-gene-dominant expression states. Eigenvectors ϕ_3 and ϕ_4 show that somewhat faster timescales are associated with exchange in and out of the Lo/Lo and Hi/Hi basins.

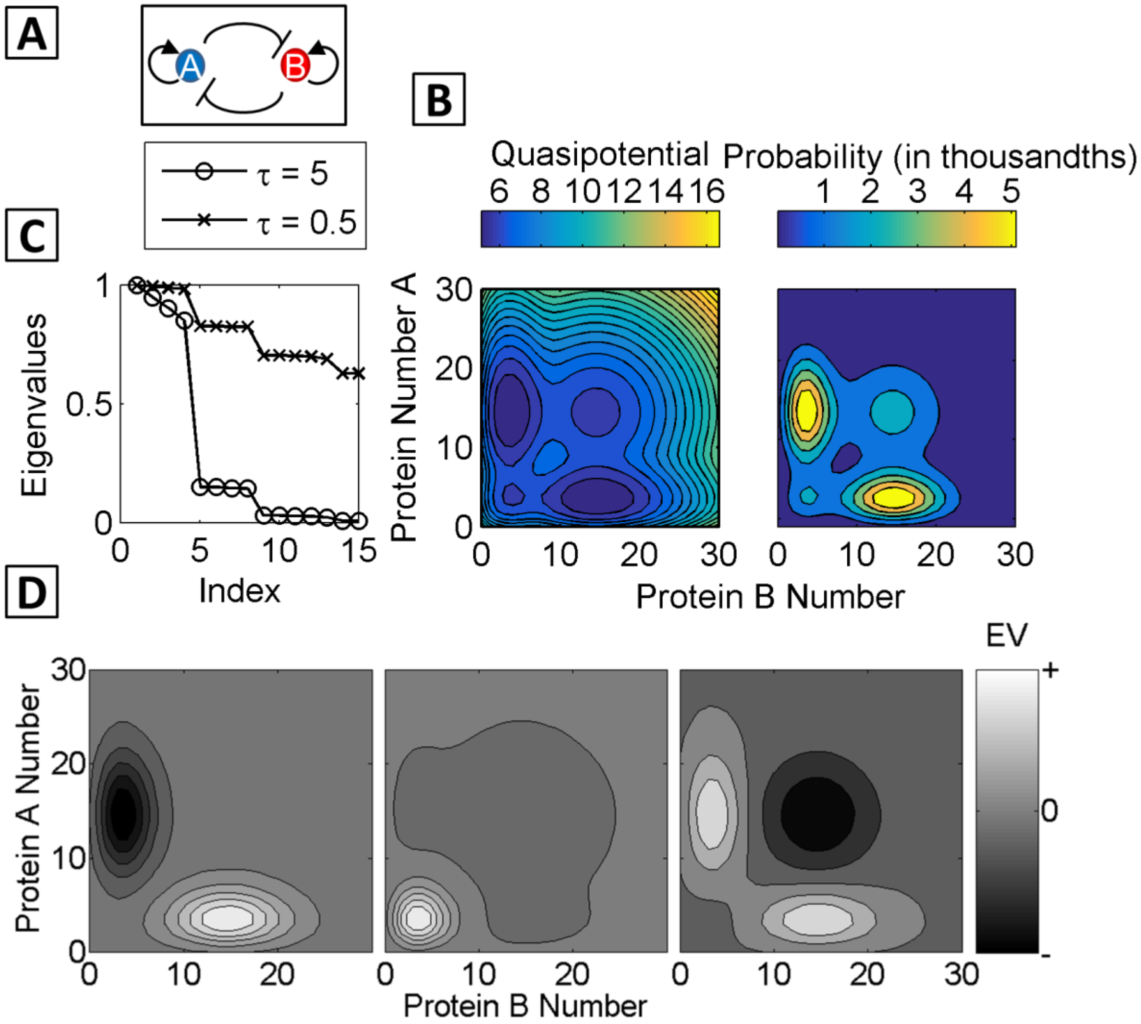


Figure 2.1 Eigenvalue and eigenvector analysis of the Mutual Inhibition/Self Activation (MISA) network

(A) Schematic of the MISA network motif. (B) The fifteen largest eigenvalues of the stochastic transition matrix $\mathbf{T}(\tau)$, indexed in descending order, for τ =5 (circles) and $\tau = 0.5$ (crosses) (time units of inverse protein degradation rate, k^{-1}). Gaps indicate separation between processes occurring on different timescales. Network parameter values are listed in Table S1. (C) The quasipotential landscape (left) and probability landscape (right) for the MISA motif, projected onto the A vs. B protein copy number subspace, showing four visible attractors. Landscapes were obtained from ϕ_1, the eigenvector associated with the largest eigenvalue of $\mathbf{T}(\tau)$. (D) Left to right: second, third, and fourth eigenvectors ($\phi_2,\ \phi_3,\ \phi_4$) of $\mathbf{T}(\tau)$. The sign structure reveals the nature of the slowest dynamical processes (see text).

2.3.2 The Markov State Model approach identifies multistability in GRNs

2.3.2.1 Reduced models of the MISA network.

The MSM framework utilizes a clustering algorithm known as PCCA+ (see Methods and Supplement) to assign every microstate in the system to a macrostate (i.e., a cluster of microstates) based on the slow system processes identified by the eigenvectors and eigenvalues of $\mathbf{T}(\tau)$. Applying the PCCA+ algorithm to the MISA network for the parameter set of Fig. 2.1 resulted in a mapping from $N = 15{,}376$ $(31 \times 31 \times 4 \times 4)$ microstates onto $C = 4$ macrostates (Fig. 2.2). The *N* microstates were first enumerated by accounting for all possible system configurations with $0 \leq a \leq 30$ and $0 \leq b \leq 30$. This enumeration assumes a negligible probability for the system to ever exceed 30 copies of either protein, which introduces a small approximation error of $1E-5$ (details in Fig. S2.1). Because the promoters of each gene can take four possible configurations—that is, two binding sites (for the repressor and activator) that can be either bound or unbound—a total of 16 gene configuration states are possible, giving $N = 15{,}376$ enumerated microstates. Quasipotentials calculated from a long brute force simulation and from ϕ_1 showed agreement (Fig. S2.2). For this parameter set, the four macrostates obtained correspond to the visible peaks (basins) in the probability (quasipotential) landscape (Fig. 2.2A). The average expression levels of proteins in each macrostate indicate the four distinct cell phenotypes (Lo/Lo, Lo/Hi, Hi/Lo, Hi/Hi). The representative gene promoter configurations for each macrostate are shown (Fig. 2.2B). However, for each macrostate cluster there are other possible gene promoter configurations present with lower steady-state probability that are not shown, since every enumerated microstate is assigned to a macrostate.

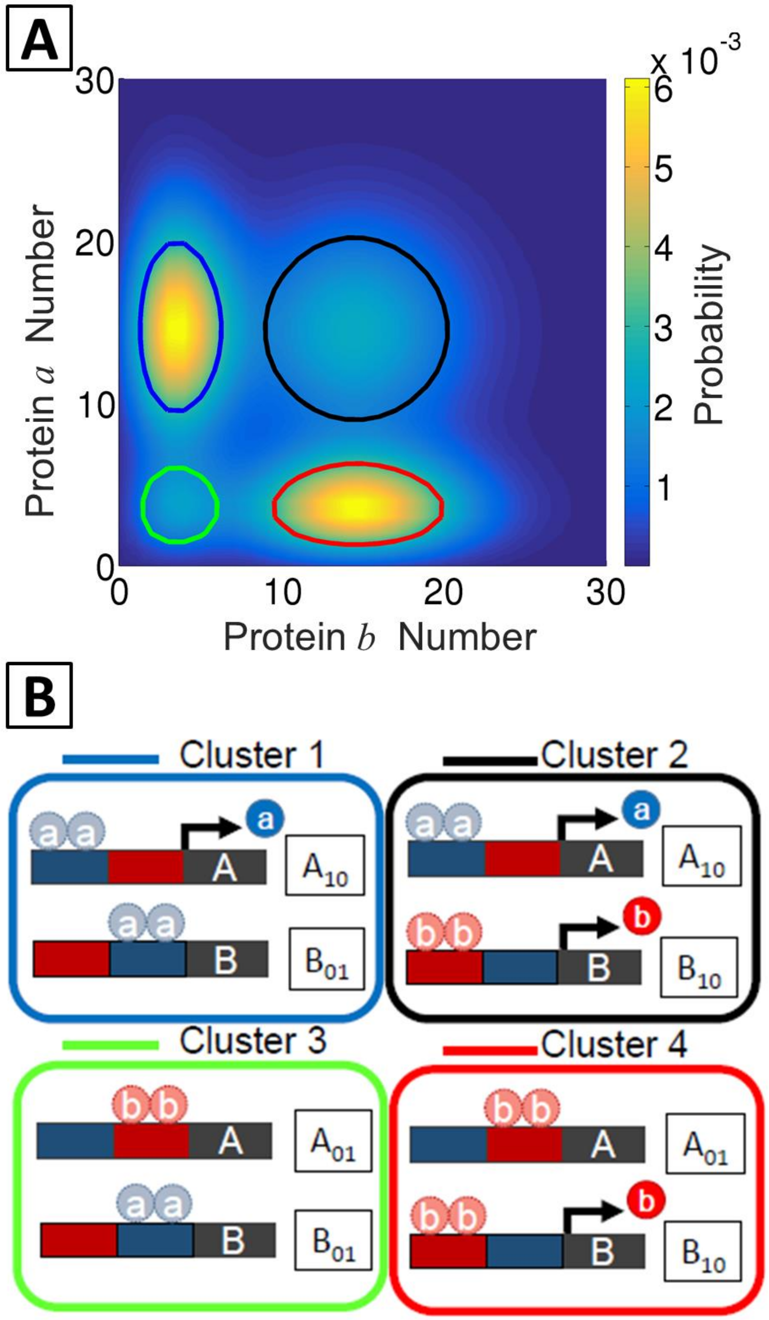


Figure 2.2 Four metastable clusters, or network "macrostates", identified for the MISA network by the Markov State Model approach. (Rate parameters same as Fig. 2.1)

(A) Macrostate centers located by their respective 50% probability contours, corresponding to visible peaks in the probability landscape. B) Schematics of the most probable gene promoter configuration for each metastable cluster.

2.3.2.2 Parameter-dependence of landscapes and MSMs.

To determine whether the MSM approach can robustly identify gene network macrostates, we applied it over a range of network parameters by varying the repressor unbinding rate f_r (all parameters defined in Table S1). Increasing f_r relative to other network parameters modulates the quasipotential landscape by increasing the probability of the Hi/Hi phenotype, in which both genes express at a high level simultaneously (Fig 3B). This occurs as a result of weakened repressive interactions, since the lifetimes of repressor occupancy on promoters are shortened when f_r is increased. The eigenvalue spectra show a corresponding shift: when $f_r = 1\mathrm{E}-3$, four dominant eigenvalues are present. When f_r is increased to $f_r = 1$, the largest visible gap in the eigenvalue spectrum shifts to occur after the first eigenvalue ($\lambda = 1$), indicating loss of multistability on the timescale of τ (here, $\tau = 5$) (Fig. 2.3A). Correspondingly, for this parameter set, the landscape shows only a single visible Hi/Hi basin.

The PCCA+ algorithm seeks C long-lived macrostates, where C is user-specified. We constructed Markov State Models for the MISA network over varying f_r, specifying four macrostates. The MSMs are shown graphically in Fig 3D. The sizes of the circles are proportional to the relative steady-state probability of the macrostate, and the thickness of the directed edges are proportional to the relative transition probability within τ. In agreement with the landscapes, the MSMs over this parameter regime show increasing probability of the Hi/Hi state, as a result of an increasing ratio of transition probability “into” versus “out of” the Hi/Hi state. The locations of the clusters in the state-space (according to 50% probability contours) do not change appreciably. The choice of lagtime τ sets the timescale on which metastability is defined in the system. However, in practice, the PCCA+ seeks an assignment of C clusters regardless of whether C metastable states exist in the system on the τ timescale, and the resulting aggregated macrostates are generally invariant to τ. Thus, for $f_r = 1$, the algorithm locates four macrostates, although the (low-probability) Hi/Lo, Lo/Lo, and Lo/Hi macrostates are likely to experience transitions away, into the Hi/Hi macrostate, within τ. These low-probability states appear in the landscape as shoulders on the outskirts of the Hi/Hi basin. Overall, Fig. 2.3 demonstrates that, for this parameter regime, the quasipotential landscape and the MSM yield similar information on the global system dynamics in terms of the number and locations of attractor states, and their relative probabilities as a function of the unbinding rate parameter f_r. The MSM further provides quantitative information on the probabilities (and thus timescales) of transitioning between each pair of macrostates.

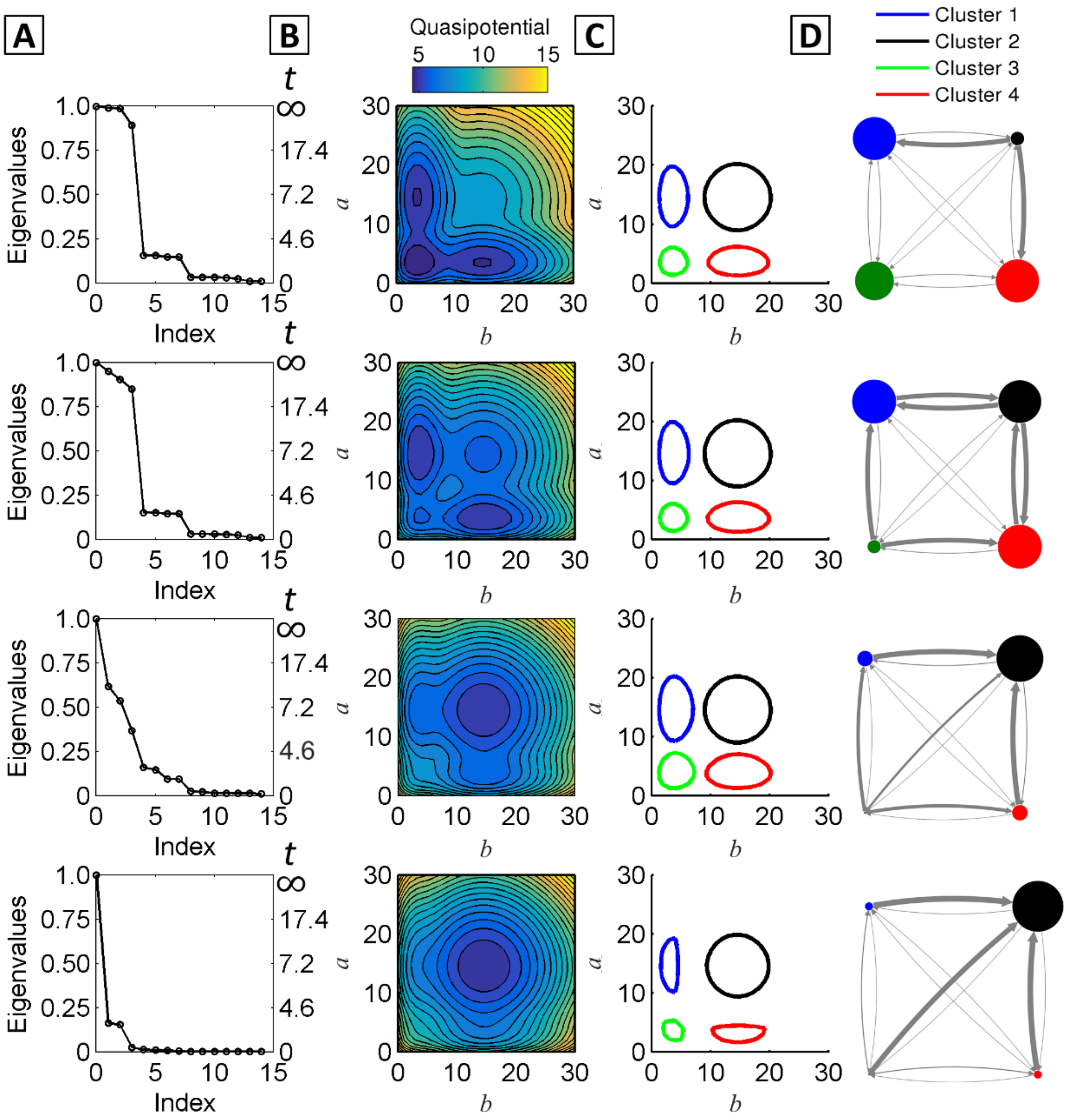


Figure 2.3 Dependence of the MISA network eigenvalues, landscape, and MSM on the repressor unbinding parameter fr

Top to Bottom: increasing $fr = \{1E-3, 1E-2, 1E-1, 1\}$ in units of protein degradation rate, k^{-1} (complete parameter list in Table S1). (A) The eigenvalue spectrum of $\mathbf{T}(\tau)$ for $\tau = 5$, and associated timescales. (B) The quasipotential landscape. (C) The Markov State Model with four macrostates, visualized by the 50% probability contour for each metastable state. (D) The state transition graph. Nodes and edges denote macrostates and transition probabilities, respectively. The size of each node is proportional to the steady-state probability, and edge thickness is proportional to the probability of transition within τ = 5.

2.3.2.3 MSM identifies purely stochastic multistability

Multistability in gene networks is often analyzed within an ordinary differential equation (ODE) framework, by graphical analysis of isoclines and phase portraits, or by linear stability analysis [51], [55]. ODE models of gene networks treat molecular copy numbers (i.e,. proteins, mRNAs) as continuous variables and apply a quasi-steady-state approximation to neglect explicit binding/unbinding of proteins to DNA. Previous studies have shown that such ODE models can give rise to landscape structures that are qualitatively different from those of their corresponding discrete, stochastic networks. For example, multistability in an ODE model of the genetic toggle switch requires cooperativity—i.e., multimers of proteins must act as regulators of gene expression[108]. However, it was found that monomer repressors are sufficient to give bistability in a stochastic biochemical model[109], [110]. We compared the dynamics of the monomer ETS network (shown schematically in Fig. 2.4A) as determined by analysis of the ODEs, along with the corresponding stochastic quasipotential landscape and the MSM. In a small-number regime, the ODEs predict monostability (Fig. 2.4C), while the stochastic landscape shows tristability—that is, three basins corresponding to the Hi/Lo, Hi/Hi, and Lo/Hi expressing phenotypes (Fig. 2.4A). This discrepancy has been shown to occur in systems with small number effects, i.e., extinction at the boundaries[110].

The MSM approach identifies three metastable macrostates for the monomer ETS in this parameter regime, as seen in the eigenvalue spectrum, which shows a gap after the third index. The reduced Markov State Model constructed for this network thus reduces the system from $N = 7{,}803\ (51 \times 51 \times 3)$ microstates to $C = 3$ macrostates (Fig. 2.4B), corresponding to the same Hi/Lo, Hi/Hi, and Lo/Hi attractor phenotypes seen in the quasipotential landscape. Figure 2.4 demonstrates that the MSM approach can accurately identify purely stochastic multistability in systems where continuous models predict only a single stable fixed-point steady state.

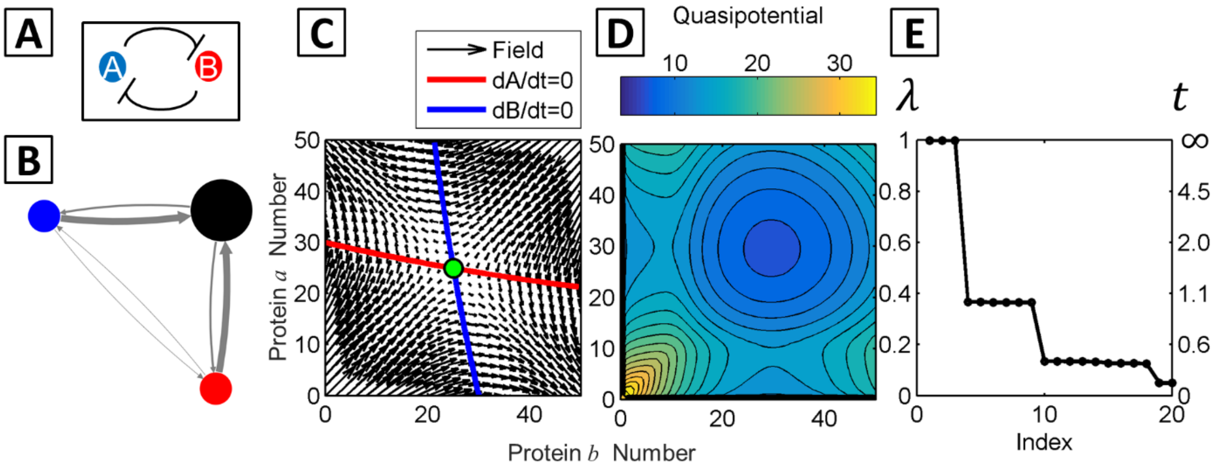


Figure 2.4 Comparison of ODE and MSM analysis of the monomer Exclusive Toggle Switch (ETS) network

(A) Schematic of the ETS network motif. (B) The Markov State Model identifies three macrostates corresponding to the Hi/Lo, Hi/Hi, and Lo/Hi phenotypes. Parameter values are listed in Table S2. (C) The nulllclines and vector field of the deterministic ODEs show a single fixed point steady-state, with both genes expressing at the maximum rate (Hi/Hi phenotype). (B,D,E) The corresponding landscape and MSM show tristability: (D) The quasipotential landscape shows three visible attractors corresponding to the Hi/Lo, Hi/Hi, and Lo/Hi phenotypes. (E) The 20 dominant eigenvalues reveal timescale separation, including a gap after λ_3.

2.3.3 Analyzing global gene network dynamics with the Markov State Model

2.3.3.1 MSM provides good approximation to relaxation dynamics from a given initial configuration

Figs. 1-4 demonstrate the utility of the MSM approach for analyzing stationary properties of networks—that is, for identifying the number and locations of multiple attractors at steady state.

Additionally, the MSM can be used to make dynamic predictions about transitions among macrostates. Dynamics for either the "full" transition matrix (with all system states enumerated up to a maximum protein copy number) or reduced transition matrix (i.e., the MSM) is propagated according to the Chapman-Kolmogorov equation (see Methods and Supplement). We sought to determine the accuracy of the dynamic predictions obtained from the MSM. Applying the methods proposed by Prinz, et al.([90]) (details in Supplement), we compared the dynamics propagated by the fully enumerated transition matrix $\mathbf{T(\tau)}$, which is then projected onto the coarse-grained macrostates, to the dynamics of the coarse-grained system propagated by $\widetilde{\mathbf{T}}(\tau)$ (i.e., the MSM). We thus computed the error in dynamics of relaxation out of a given initial system configuration. The system relaxation from a given initial microstate can also be computed by running a large number of brute force SSA simulations. Relaxation dynamics for the full, brute-force, and reduced MSM methods, applied to the MISA with $f_r = 1E - 2$, all show good agreement (Fig. 2.5 A,B, and C). The error computed between the reduced MSM vs. full dynamics (i.e., $\widetilde{\mathbf{T}}(\tau)$ vs $\mathbf{T(\tau)}$), is maximally $7.8E - 3$, varies over short times, and decreases continuously after time $t = 140$. Alternatively, the error of the MSM can be quantified by comparing the autocorrelation functions of the MSM and brute force simulation[96], [111]. In Figure S2.3, we show that the derived autocorrelation functions of the MSM and brute force, and the relaxation constants τ_r, which describes the amount of time to reach equilibrium, are close in value ($\tau_r = 1E3$, for the MSM, and $\tau_r = 1.1E3$ for the brute force). Overall, these results demonstrate that the most accurate predictions of the coarse-grained MSM can be obtained on long timescales, but dynamic approximations with reasonable accuracy can also be obtained for short timescales.

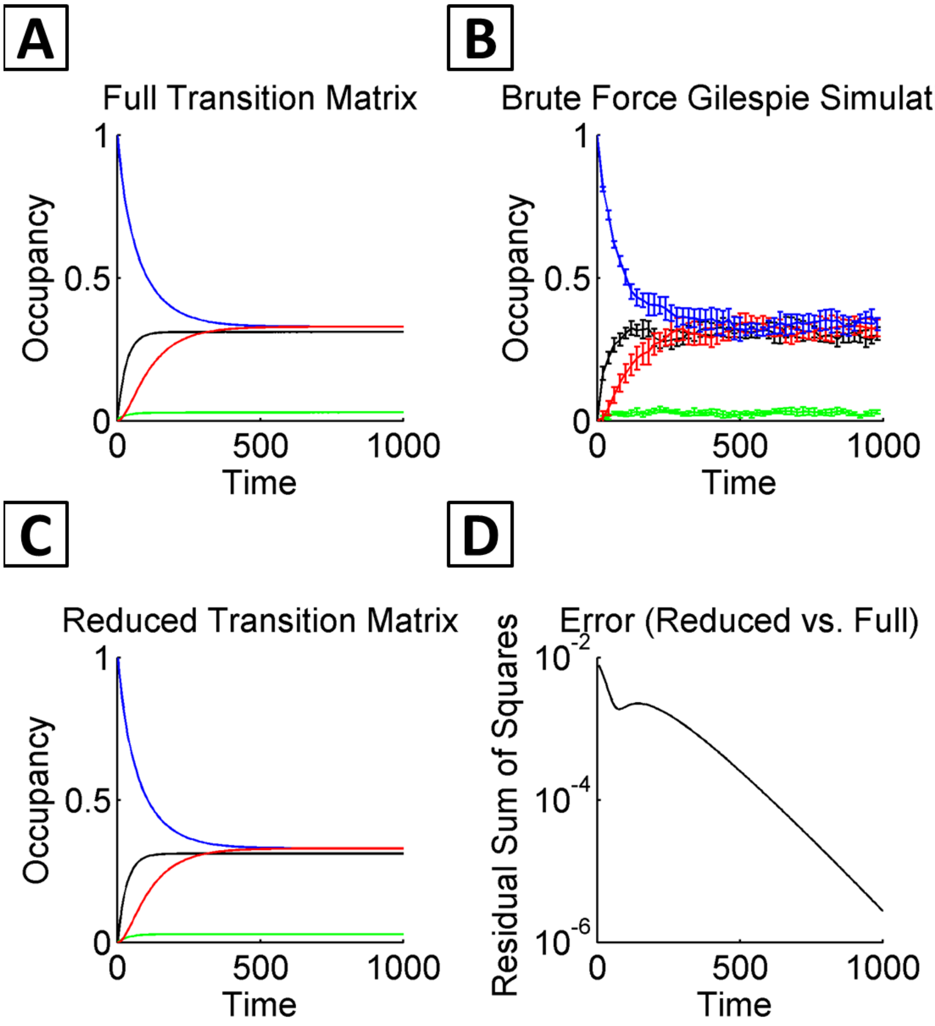


Figure 2.5 MSM approximation error for the MISA motif

Relaxation of the system from a particular initial configuration (see text), as obtained from (A) the full transition matrix, (B) brute force SSA simulation, and (C) the reduced transition matrix obtained from the MSM. Color-coding is according to the macrostates, as in Figs. 1-3: blue, black, red, green correspond to A/B expression phenotypes Hi/Lo, Hi/Hi, Lo/Hi, and Lo/Lo, respectively. (D) Calculated approximation error as a function of time, comparing the reduced MSM to the full CME dynamics. Network parameter values are same as Figs. 2.1, 2.2.

2.3.3.2 Parameter-dependence of MSM error

The accuracy of the MSM dynamic predictions depends on whether inter-macrostate transitions can be treated as memory-less hops. Previous theoretical studies of gene network dynamics found that the height of the barrier separating phenotypic states, and the state-switching time associated with overcoming the barrier, depends on the rate parameters governing DNA-binding by the protein regulators[52], [53], [110], [112]. We reasoned that a larger timescale separation between intra- and inter-basin transitions (corresponding to a larger barrier height separating basins) should result in higher accuracy of the MSM approximation. Thus, we hypothesized that the accuracy of the MSM dynamic predictions should depend on the DNA-binding and unbinding rate parameters. We demonstrated this using the dimeric ETS motif, by computing the error of the MSM approximation for a range of repressor unbinding rates f. We varied the binding kinetics without changing the overall relative strength of repression, by varying f together with the repressor binding rate h, to maintain a constant binding equilibrium ($X_{eq} = \frac{f}{h} = 100$). By varying f and h in this way over eight orders of magnitude, we found that the barrier height and timescale of the slowest system process (t_2) had a non-monotonic dependence on the binding/unbinding parameters. Thus, the fastest inter-phenotype switching was observed in the regime with intermediate binding kinetics, in agreement with previous work[103]. The system also exhibits a shift from three visible basins in the quasipotential landscape in the small f regime to two basins in the large f regime. We performed clustering by selecting $C = 2$ (dashed lines, Fig. 2.6) and $C = 3$ clusters (solid lines, Fig. 2.6), and computed the total error over all choices of system initialization, as well as the error associated with relaxation from a particular system microstate. In general, we find that the 3-state MSM approximation is more accurate than the 2-state partitioning. The 3-state MSM dynamic predictions are highly accurate when the DNA-binding/unbinding kinetics is slow. As such, in this regime the Markovian assumption of memory-less transitions between the three phenotypic states is most accurate. As hypothesized, the accuracy of the MSM approximation is lowest (highest error) when the lifetime t_2 is shortest (intermediate regime, $f = 1$), and the error decreases modestly with further increase in f (i.e., increase in t_2).

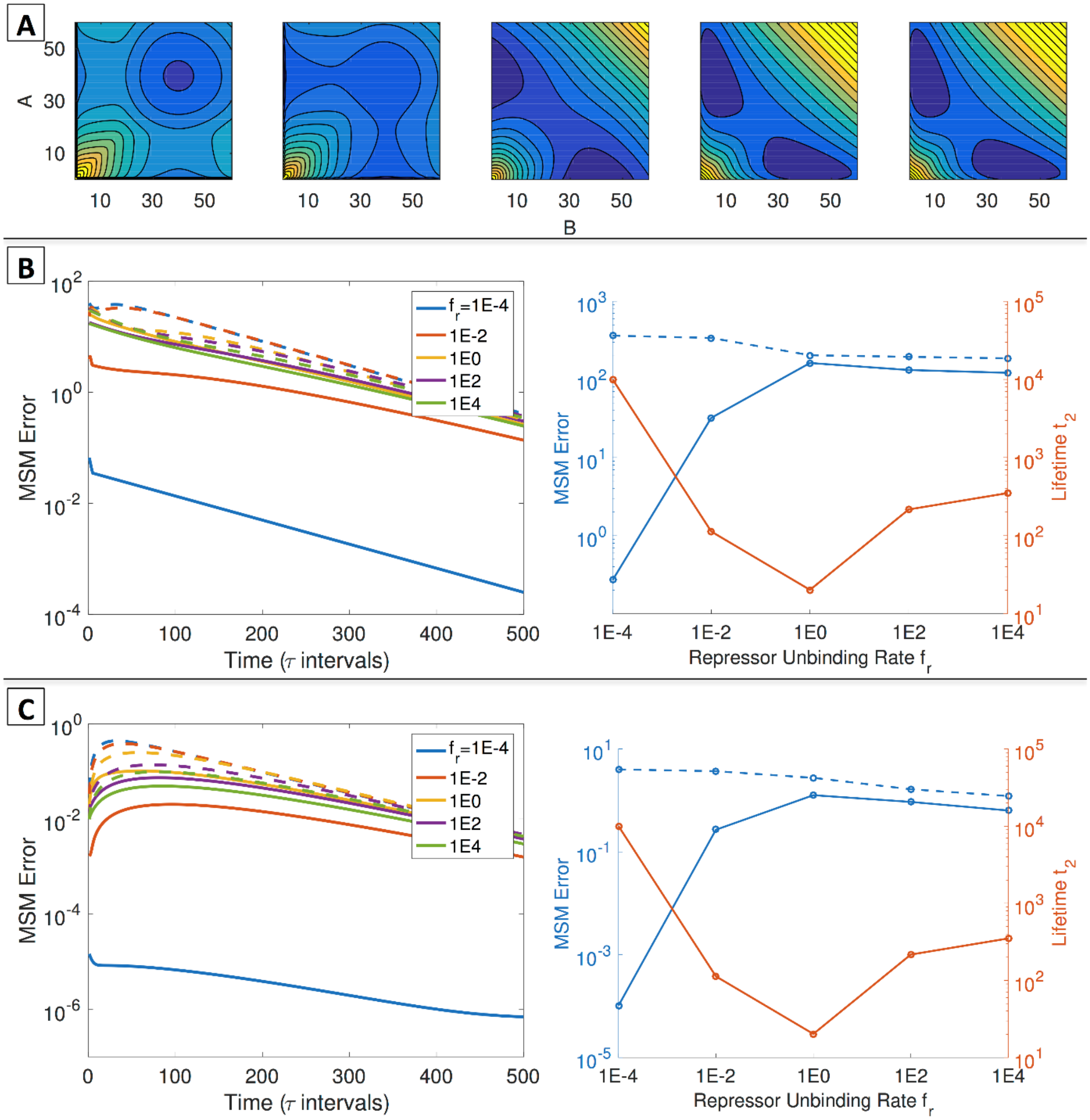


Figure 2.6 The MSM approximation accuracy for the ETS motif depends on rate parameters and number of macrostates in the reduced model

(A) Quasipotential landscape for the exclusive dimeric repressor toggle switch, with increasing DNA-binding rates (left to right: $fr\ = \{1E-4, 1E-2, 1E0, 1E2, 1E4\}$, all parameter values listed in Table 2), demonstrating the dependence of basin number and barrier height on network parameters. (B) Global error of the MSM approximation. Left: Global error as a function of time (in intervals of τ) for different f_r and numbers of macrostates. Solid lines: global error of the 3-state MSM. Dashed lines: global error of the 2-state MSM. Right: Total global error over kτ, $k =$ 0 to 500, for a 3-state (solid blue) or 2-state (dashed blue) MSM. Solid orange line: the longest system lifetime t_2. (C) Error of the MSM approximation when the system is initialized in a particular microstate. Left: Error as a function of time (in intervals of τ) for different adiabaticities and different numbers of macrostates. Solid lines: error of the 3-state MSM. Dashed lines: error of the 2-state MSM. Right: Total error from a particular microstate over kτ where $k = 0$ to 500, for a 3-state (solid blue) or 2-state (dashed blue) MSM. Orange line: the longest system lifetime t_2.

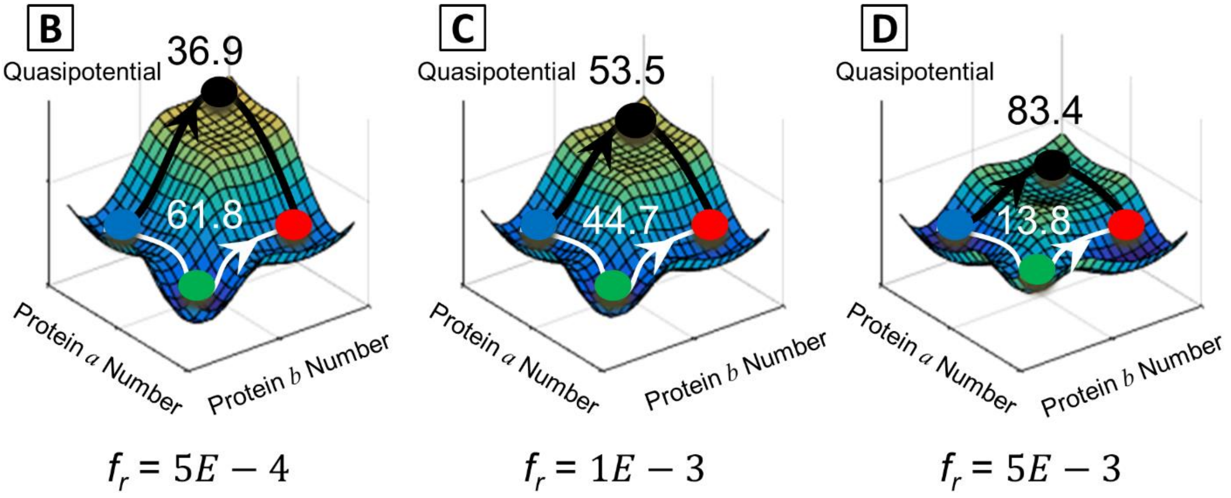
2.3.2.3 Decomposition of state-transition pathways in gene networks using the MSM framework.

Quantitative models of gene network dynamics can shed light on transition paths connecting phenotypic states. The MSM approach coupled with transition path theory[113]–[115] enables decomposition of all major pathways linking initial and final macrostates of interest. This type of pathway decomposition has previously shed light on mechanisms of protein folding[101]. We demonstrate this pathway decomposition on the MISA network, by computing the transition paths linking the polarized *A*-dominant (Hi/Lo) and *B*-dominant (Lo/Hi) phenotypes. Multiple alternative pathways linking these phenotypes are possible: for the 4-state coarse-graining, the system can alternatively transit through the Hi/Hi or Lo/Lo phenotypes when undergoing a stochastic state-transition from one polarized phenotype to the other. Not all possible paths are enumerated since only transitions with net positive fluxes are considered (see Equation S18). The hierarchy of pathway probabilities for successful transitions depends on the kinetic rate parameters (Fig. 2.7A). It could be tempting to intuit pathway intermediates based on visible basins in the quasipotential landscape. However, we found that the steady-state probability of an intermediate macrostate (i.e., the Hi/Hi or Lo/Lo states) does not accurately predict if it serves as a pathway intermediate for successful transitions, because parameter regimes are possible in which successful transitions are likely to transition through intermediates with high potential/low probability (Fig.

2.7C). This occurs because the relative probability of transiting through one intermediate macrostate versus another is based on the balance of probabilities for entering and exiting the intermediate: intermediate states that can be easily reached—but not easily exited—as a result of stochastic fluctuations can act as “trap” states. Therefore, it is shown that the pathway probability cannot be inferred from the steady state probability of the intermediates alone.

A

	Pathway Probability (%)				Stationary Probability (%)			
f_r	🔵→🔴	🔵→🟢→🔴	🔵→⚫→🔴	🔵→⚫→🟢→🔴	🔵	⚫	🟢	🔴
5E-4	1.0	61.8	36.9	2.3E-01	27.4	1.3	44.0	27.4
1E-3	1.5	44.7	53.5	3.3E-01	34.5	3.2	27.8	34.5
5E-3	2.3E-01	13.8	83.4	5.0E-01	37.9	17.9	6.3	37.9
1E-2	2.4E-01	7.3	89.7	5.2E-01	33.0	31.1	2.9	33.0
1.5E-2	2.5E-01	4.9	92.1	5.2E-01	28.8	40.7	1.7	28.8



$f_r = 5E-4$ $f_r = 1E-3$ $f_r = 5E-3$

Figure 2.7 Dependence of stochastic transition paths on the repressor unbinding rate parameter fr in the MISA network (parameter values listed in Table 1)

(A)Table of all possible transition paths starting from the Hi/Lo (blue) and ending in the Lo/Hi (red) macrostate (color coding is same as Figs. 2.1-2.3 and Fig. 2.5). Relative probabilities of traversing a given path are shown, along with the stationary probabilities of the system to be found in a given macrostate. (B-D) Dominant transition paths superimposed on the 3D quasipotential surfaces for $fr = \{5E-4, 1E-3, 5E-3\}$, demonstrating how dominant paths can traverse high-potential areas of the landscape. For example, when $f_r = 1E-3$, (panel C), successful transitions most likely go through the Hi/Hi state (3.2% populated at steady state), though this requires a large barrier crossing. Pathway percentages are superimposed on the landscapes.

2.3.2.4 MSMs can be constructed with different resolutions of coarse-graining

The eigenvalue spectrum of the MISA network shows a step-structure, with nearly constant eigenvalue clusters separated by gaps. These multiple spectral gaps suggest a hierarchy of dynamical processes on separate timescales. A convenient feature of the MSM framework is that it can build coarse-grained models with different levels of resolution by PCCA+, in order to explore such hierarchical processes. We applied the MSM framework to a MISA network with very slow rates of DNA-binding and unbinding ($f_r = 1E-4, h_r = 1E-6$), comparing the macrostates obtained from selecting $C = 4$ versus $C = 16$ clusters. For $\mathbf{T}(\tau = 1)$, a prominent gap occurs in the eigenvalue spectrum between λ_{16} and λ_{17}, corresponding to an almost 30-fold separation of timescales between $t_{16} = 27.8$ and $t_{17} = 0.99$(Fig. 2.8A). Applying PCCA+ with $C = 16$ clusters uncovered a 16-macrostate network with four highly-interconnected subnetworks consisting of four states each (Fig. 2.8C). The identities of the sixteen macrostates showed an exact correspondence to the sixteen possible *A/B* promoter binding configurations. This correspondence reflects the fact that, in the slow binding/unbinding, so-called non-adiabatic regime[116], the slow network dynamics are completely determined by unbinding and binding events that take the system from one promoter configuration macrostate to another, while all fluctuations in protein copy number occur on much faster timescales.

Each subnetwork in the MSM constructed with $C = 16$ corresponds to a single macrostate in the MSM constructed with $C = 4$. Thus, in the $C = 4$ MSM, four different promoter configurations are lumped together in a single macrostate, and dynamics of transitions among them is neglected. Counterintuitively, the locations of the $C = 4$ macrostates do not correspond directly to the four basins visible in the quasipotential landscape (Fig. 2.8B,D). Instead, the clusters combine distinct phenotypes—e.g., the red macrostate combines the *A/B* Lo/Lo and Lo/Hi phenotypes, because it includes the promoter configurations $A_{01}\,B_{10}$ and $A_{11}\,B_{10}$ (corresponding to Lo/Hi expression) and $A_{01}\,B_{00}$ and $A_{11}\,B_{00}$ (corresponding to Lo/Lo expression) (Fig. 2.8B, Table S2.3 and Fig. S2.4). This result demonstrates that the barriers visible in the quasipotential landscape do not reflect the slowest timescales in the system. This occurs because of the loss of information inherent to visualizing global dynamics via the quasipotential landscape, which often projects dynamics onto two system coordinates. In this case, projecting onto the protein *a* and protein *b* copy numbers loses information about the sixteen promoter configurations, obscuring the fact that barrier-crossing transitions can occur faster than some within-basin transitions. Plotting a time trajectory of brute force SSA simulations for this network supports the findings from the MSM: the dynamics shows frequent transitions within subnetworks, and less-frequent transitions between subnetworks, indicating the same hierarchy of system dynamics as was revealed by the 4- and 16-state MSMs (Fig. 2.8E).

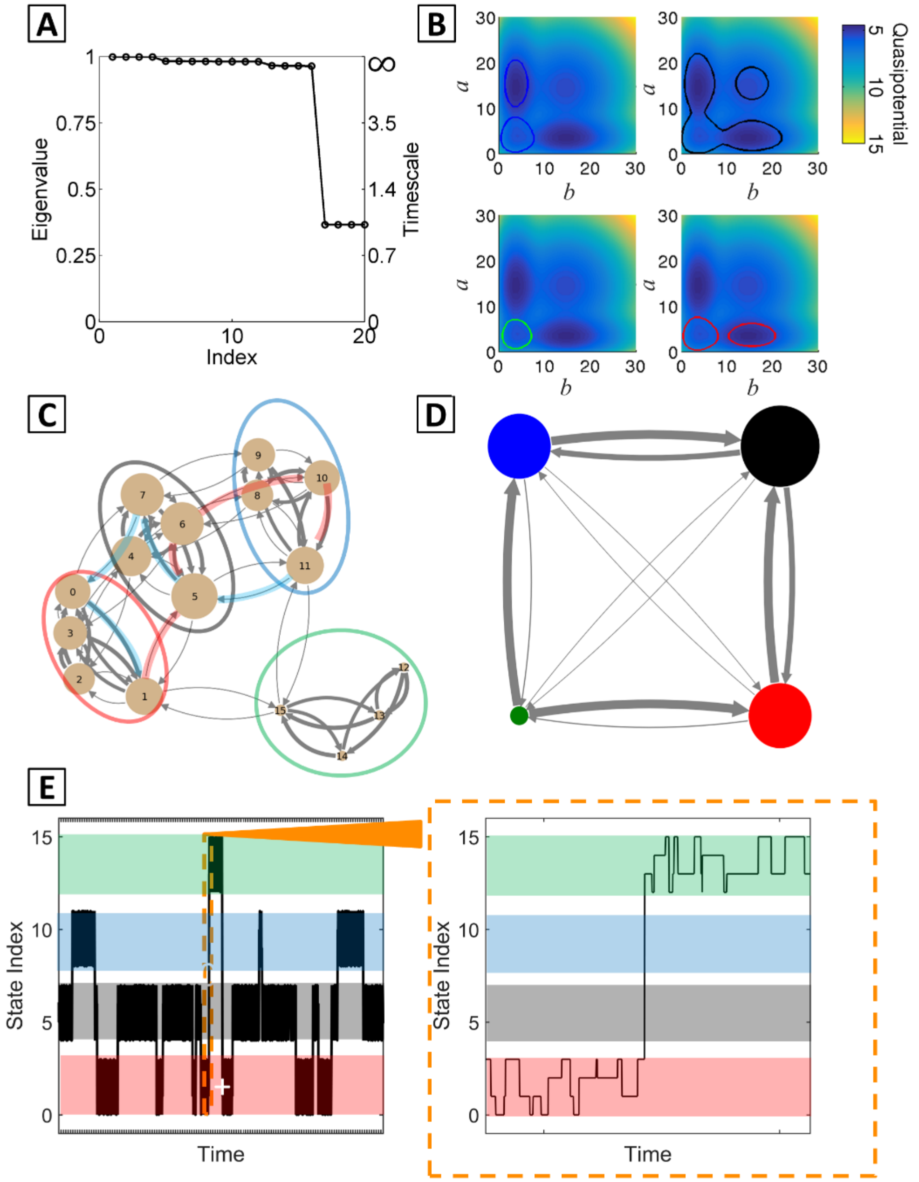


Figure 2.8 Hierarchical dynamics revealed by MSM analysis of the MISA network in the slow DNA-binding/unbinding parameter regime. All network parameters listed in Table 1.

(A) Eigenvalue spectrum of $\mathbf{T}(\tau)$, $\tau = 1$, showing 16 dominant eigenvalues. (B) 4-macrostate MSM: 70% probability contours superimposed onto the quasipotential surface. In this parameter regime, separate attractors in the landscape are kinetically linked in the same subnetwork (see text). (C) 16-macrostate MSM showing 4 highly connected subnetworks (colored ovals). Each macrostate corresponds to a particular promoter binding-configuration (see numbering scheme in Table S2.5). A pair of representative transition paths through the network are highlighted. Red path: most probable forward transition path from macrostate 1 to macrostate 11. Blue path: most probable reverse path from 11 to 1. (D) State transition graph for the 4-macrostate MSM. (E)

Brute force SSA simulation of the MISA network over time. Trajectory is plotted according to the 16-macrostate (promoter configuration) indexing as in panel C and Table S5. Colored panels reflect the four subnetworks/$C = 4$ macrostates. Orange inset: zoomed in trajectory segment, showing a switching event between the red and green subnetworks.

2.3.2.5 Transition path decomposition reveals nonequilibrium dynamics

Mapping the most probable paths forward and backward between macrostate "1" (promoter configuration: $A_{01}B_{00}$) and macrostate "11" (promoter configuration: $A_{00}B_{01}$) revealed that a number of alternative transition paths are accessible to the network, and the paths typically transit between three and five intermediate macrostates. The decomposition shows three paths with significant (i.e., >15%) probability and 12 distinct paths with >1% probability (for both forward and backward transitions, Tables S3-4). The pathway decomposition also reveals a great deal of irreversibility in the forward and reverse transition paths, which is a hallmark of nonequilibrium dynamical systems[117]. For example, the most probable forward and reverse paths both transit three intermediates, but have only one intermediate (macrostate 5) in common (Fig. 2.8C and Table S2.4-5). Thus, the complete process of transitioning away from macrostate 1, through macrostate 11, and returning to 1 maps a dynamic cycle.

2.4 Discussion

Our application of the MSM method to representative GRN motifs yielded dynamic insights with potential biological significance. Decomposition of transition pathways revealed that stochastic state-transitions between phenotypic states can occur via multiple alternative routes. Preference of the network to transition with higher likelihood through one particular pathway depended on the stability of intermediate macrostates, in a manner not directly intuitive from the steady-state probability landscape. The existence of "spurious attractors", or metastable intermediates that act as trap states to hinder stem cell reprogramming, has been discussed previously[58] as a general explanation for the existence of partially reprogrammed cells. By analogy, MSMs constructed in protein folding studies predict an ensemble of folding pathways, as well as the existence of misfolded trap states that reduce folding speed[101]. Our results suggest that multiple partially reprogrammed cell types could be accessible from a single initial cell state. Successful phenotype-transitions can occur predominantly through high-potential (unstable)—and thus difficult to observe experimentally—intermediate cell types. In future applications to specific gene GRNs, the MSM approach could predict a complex map of cell-reprogramming pathways, and thus potentially suggest combinations of targets towards improved safety and efficiency of reprogramming protocols.

Our study revealed that the two-gene MISA network can exhibit complex dynamic phenomena, involving a large number of metastable macrostates (up to 16), cycles and hierarchical dynamics, which can be conveniently visualized using the MSM. The quasipotential landscape has been used recently as a means of visualizing global dynamics and assessing locations and relative stabilities of phenotypic states of interest, in a manner that is quantitative (deriving strictly from underlying gene regulatory interactions), rather than qualitative or metaphorical (as was the case for the original Waddington epigenetic landscape)[67]. However, our study highlights the potential difficulty of interpreting global network dynamics based solely on the steady-state landscape, which is often projected onto one or two degrees of freedom. We found that phenotypically identical cell states—that is, network states marked by identical patterns of protein expression, inhabiting the same position in the projected landscape—can be separated by kinetic barriers, experiencing slow inter-conversion due to slow timescales for update to the epigenetic state (or promoter binding occupancy). Conversely, phenotypically distinct states marked by different levels of protein expression can be kinetically linked, experiencing relatively rapid inter-conversion. This type of stochastic inter-conversion is thought to occur in embryonic stem cells—for example, fluctuations in expression of the Nanog gene have been proposed to play a role in maintaining pluripotency[28], [118]. The hierarchical dynamics revealed by our study supports the idea that the phenotype of a cell could be more appropriately defined by dynamic patterns of regulator or marker expression levels[28], rather than on single-timepoint levels alone. This was seen in the 16-state MSM for the MISA network, where a given expression pattern (e.g., the Lo/Lo attractor) comprised multiple macrostates from separate dynamic subnetworks.

Complex, high-dimensional dynamical systems call for systematic methods of coarse-graining (or dimensionality reduction), for analysis of mechanisms and extraction of information that can be compared with experimental results. In the field of Molecular Dynamics, the complexity of, e.g., macromolecular conformational changes—involving thousands of atomic degrees of freedom and multiple dynamic intermediates—has driven the development of automated methods for prediction and analysis of essential system dynamics from simulations[119], [120]. In that field, coarse-graining has been achieved based on a variety of so-called geometric (structural) or, alternatively, kinetic clustering methods[95], [121]. Noe, et al.[121], discussed that geometric (or structure-based) coarse-graining methods can fail to produce an accurate description of system dynamics when structurally similar molecular conformations are separated by large energy barriers or, conversely, when dissimilar structures are connected by fast transitions, as they found in a study of polypeptide folding dynamics. In such cases, kinetic (i.e., separation-of-timescale-based) coarse-graining methods such as the MSM approach are more appropriate. Our application of the MSMs to GRNs demonstrates how similar complex dynamic phenomena can manifest at the “network”-scale.

The challenge of solving the CME due to the curse-of-dimensionality is well known. The MSM approach is related to other projection-based model reduction methods that aim to reduce the computational burden of solving the CME directly by projecting the rate (or transition) matrix onto a smaller subspace or aggregated state-space with fewer degrees of freedom. Such approaches include the Finite State Projection algorithm[86], and methods based on Krylov subspaces[74], [122], [123], sparse-gridding[124], and separation-of-timescales[75], [125], [126]. The MSM is distinct from other timescale-based approaches in that, rather than partitioning the system into categories of slow versus fast reactions[125] or species[75], or basing categories on physical intuition[126], it systematically groups microstates in such a way that maximizes metastability of aggregated states[107]. The practical benefit of this approach is its capacity to describe a system compactly in terms of long-lived, perhaps experimentally observable, states. Another important distinction between the MSM approach and other CME model reduction methods is that its primary end-goal is *not* to solve the CME per se. Rather, the emphasis in studies employing MSMs has generally been on gaining mechanistic, physical, or experimentally-relevant insights to complex system dynamics[127]–[129]. As such, the approach does not optimally balance the tradeoff between computational expense versus quantitative accuracy of the solution, as other methods have done explicitly[130]. Instead, the method can be considered to balance the tradeoff between accuracy and “human-interpretability”, where decreasing the number of macrostates preserved in the MSM coarse-graining tends to favor the latter over the former.

A potential drawback of the workflow presented in this paper is that it requires an enumeration of the system state-space in order to construct the biochemical rate matrix $\mathbf{K}$. Networks of increased complexity or molecular copy numbers will lead to prohibitively large matrix sizes. Here, we restricted our study to model systems with a relatively small number of reachable microstates (i.e., $\sim 10^4$ microstates permitted tractable computations on desktop computers with MATLAB[88]). However, an advantage of the MSM approach is its use of the stochastic transition matrix $\mathbf{T}(\tau)$ (rather than $\mathbf{K}$), which can be estimated from simulations by sampling transition counts between designated regions of state-space in trajectories of length τ. Systems of increased complexity/dimensionality are generally more accessible to simulations, because the size of the state-space is automatically restricted to those states visited within finite-length simulations. In our group (Tse, *et al.*), we find that the MSM approach interfaces well with SSA simulations of biochemical network dynamics, combined with enhanced sampling techniques [41],[40], [131]. We anticipate that, as in the Molecular Dynamics field, the MSM framework in applications to biochemical networks will prove useful as a tool for post-processing simulation data. Furthermore, we anticipate that the approach could potentially interface with other numerical approximation techniques that have been developed in recent years for reduction of the CME.

A potential challenge for the application of the PCCA+-based spectral clustering method to biochemical networks is that, as open systems, biochemical networks generally do not obey detailed balance. This means that the stochastic transition matrices do not have the property of irreversibility, which was originally taken to be a requirement for application of the PCCA algorithm[95]. However, later work by Roblitz et al.[93] , found that the PCCA+ method also delivers an optimal clustering for irreversible systems. In this study, we found that the PCCA+ method could determine appropriate clusters in GRNs, and could furthermore uncover nonequilibrium cycles, as seen in the irreversibility (distinct forward and backward) of transition paths in the 16-state system. Newer methods of MSM building, which are specifically designed to treat nonequilibrium dynamical systems, have appeared recently[132]. It may prove fruitful to explore these alternative methods in order to identify the most appropriate, general MSM framework for application to various biochemical networks.

2.5 Conclusions

In this work, we present a method for analyzing multistability and global state-switching dynamics in gene networks modeled by stochastic chemical kinetics, using the MSM framework. We found that the approach is able to: (1) identify the number and identities of long-lived phenotypic-states, or network "macrostates", (2) predict the steady-state probabilities of all macrostates along with probabilities of transitioning to other macrostates on a given timescale, and (3) decompose global dynamics into a set of dominant transition pathways and their associated relative probabilities, linking two system states of interest. Because the method is based on the discrete-space, stochastic transition matrix, it correctly identified stochastic multistability where a continuum model failed to find multiple steady states. The quantitative accuracy of the dynamics propagated by the coarse-grained MSM was highest in a parameter regime with slow DNA-binding and unbinding kinetics, indicating that in GRNs the assumption of memory-less hopping among a small number of macrostates is most valid in this regime. By projecting dynamics encompassing a large state-space onto a tractable number of macrostates, the MSMs revealed complex dynamic phenomena in GRNs, including hierarchical dynamics, nonequilibrium cycles, and alternative possible routes for phenotypic state-transitions. The ability to unravel these processes using the MSM framework can shed light on regulatory mechanisms that govern cell phenotype stability, and inform experimental reprogramming strategies. The MSM provides an intuitive representation of complex biological dynamics operating over multiple timescales, which in turn can provide the key to decoding biological mechanisms. Overall, our results demonstrate that the MSM framework—which has been generally applied thus far in the context of molecular dynamics via atomistic simulations—can be a useful tool for visualization and analysis of complex, multistable dynamics in gene networks, and in biochemical reaction networks more generally.

2.6 Supplementary Information

2.6.1 MISA network reactions

2.6.1.1 Protein synthesis and degradation reactions

$$A_{01} \xrightarrow{g0} A_{01} + a$$

$$A_{00} \xrightarrow{g0} A_{00} + a$$

$$A_{10} \xrightarrow{g1} A_{10} + a$$

$$A_{11} \xrightarrow{g0} A_{11} + a$$

$$B_{01} \xrightarrow{g0} B_{01} + b$$

$$B_{00} \xrightarrow{g0} B_{00} + b$$

$$B_{10} \xrightarrow{g1} B_{10} + b$$

$$B_{11} \xrightarrow{g0} B_{11} + b$$

$$a \xrightarrow{k} \emptyset$$

$$b \xrightarrow{k} \emptyset$$

2.6.1.2 Gene repression

$$A_{00} + 2b \xrightarrow{h_r} A_{01}$$

$$A_{01} \xrightarrow{f_r} A_{00} + 2b$$

$$A_{10} + 2b \xrightarrow{h_r} A_{11}$$

$$A_{11} \xrightarrow{f_r} A_{10} + 2b$$

$$B_{00} + 2a \xrightarrow{h_r} B_{01}$$

$$B_{01} \xrightarrow{f_r} B_{00} + 2a$$

$$B_{10} + 2a \xrightarrow{h_r} B_{11}$$

$$B_{11} \overset{f_r}{\rightarrow} B_{10} + 2a$$

2.6.1.3 Gene activation

$$A_{00} + 2a \overset{h_a}{\rightarrow} A_{10}$$

$$A_{10} \overset{f_a}{\rightarrow} A_{00} + 2a$$

$$A_{01} + 2a \overset{h_a}{\rightarrow} A_{11}$$

$$A_{11} \overset{f_a}{\rightarrow} A_{01} + 2a$$

$$B_{00} + 2b \overset{h_a}{\rightarrow} B_{10}$$

$$B_{10} \overset{f_a}{\rightarrow} B_{00} + 2b$$

$$B_{01} + 2b \overset{h_a}{\rightarrow} B_{11}$$

$$B_{11} \overset{f_a}{\rightarrow} B_{01} + 2b$$

2.6.2 Toggle switch network reactions

2.6.2.1 Protein synthesis

$$P_{00} \overset{g_1}{\rightarrow} P_{00} + a$$

$$P_{00} \overset{g_1}{\rightarrow} P_{00} + b$$

$$P_{01} \overset{g_0}{\rightarrow} P_{01} + a$$

$$P_{01} \overset{g_1}{\rightarrow} P_{01} + b$$

$$P_{10} \overset{g_1}{\rightarrow} P_{10} + a$$

$$P_{10} \overset{g_0}{\rightarrow} P_{10} + b$$

2.6.2.2 Protein degradation

$$a \overset{k}{\rightarrow} \emptyset$$

$$b \overset{k}{\rightarrow} \emptyset$$

2.6.2.3 Gene binding

$$P_{00} + 2a \xrightarrow{h} P_{10}$$

$$P_{10} \xrightarrow{f} P_{00} + 2a$$

$$P_{00} + 2b \xrightarrow{h} P_{01}$$

$$P_{01} \xrightarrow{f} P_{00} + 2b$$

2.6.3 List of parameters

Table 1 MISA network parameters I

Figure	k	g_0	g_1	h_a	h_r	f_a	f_r	τ
1	1	4	16	1e-1	1e-4	1	1e-2	5
2	1	4	16	1e-1	1e-4	1	1e-2	5
3	1	4	16	1e-1	1e-4	1	1e-3,1e-2,1e-1,1	5
5	1	4	16	1e-1	1e-4	1	1e-2	5
7	1	4	16	1e-1	1e-4	1	5e-4,1e-3,5e-3,1e-2,1.5e-2	5
8	1	4	16	1e-3	1e-6	1e-2	1e-4	1

The parameters are as follows- k: protein degradation rate, g_0: basal/repressed expression rate, g_1: activated expression rate, h_a: activator binding rate, h_r: repressor binding rate, f_a:activator unbinding rate, f_r: repressor unbinding rate.

Table 2 MISA network parameters II

Figure	k	g_0	g_1	h	f	τ
4	1	0	30	1e-6	1e-4	1
6	1	0	40	1e-6,1e-4,1e0,1e2	1e-4,1e-2,1e0,1e2,1e4	1

The parameters are as follows- k: protein degradation rate, g_0: basal/repressed expression rate, g_1: activated expression rate, h: protein binding rate, f: protein unbinding rate.

2.6.4 Theoretical background

2.6.4.1 Connection between the Master Equation and the Transition Matrix

Herein, we summarize the main previously derived, theoretical results and refer the reader to cited references for further detail. The discrete, Markovian Chemical Master Equation (CME) describes the time-evolution of the probability distributions over the state-space for a well-mixed system of S reacting species with M possible reactions. The state of the system is given by the molecular popular vector $x \epsilon N^S$. Let $p(x,t)$ be the probability that the system is found in the state x at time t, and let $a_\mu(x)dt$ and ν_μ bet the propensity function and stoichiometric transition vector, respectively, of the possible reactions $\mu = (1,2,\ldots,M)$. Then the CME is given by

$$\frac{\partial p(x,t|x_{0,}t_0)}{\partial t} = \sum_{\mu}^{M} [a_\mu(x-\nu_\mu)p(x-\nu_\mu,t|x_0,t_0) - a_\mu(x,t|x_0,t_0)] \qquad (2.1)$$

In vector-matrix form,

$$\frac{dp(\boldsymbol{X},t)}{dt} = \boldsymbol{K}p(\boldsymbol{X},t) \qquad (2.2)$$

where the off-diagonal element K_{ij}, gives the time-independent rate of transitioning from state x_i to x_j, and $K_{ii} = -\sum_{j \neq i} K_{ji}$ (i.e., columns sum to 0). Given the probability at time t, $p(\boldsymbol{X}, t)$, the probability density at a later time (so-called lag time) τ may be computed by:

$$p(\boldsymbol{X}, x + \tau) = \exp(\boldsymbol{K}\tau)\, p(\boldsymbol{X}, t) \tag{2.3}$$

where $\exp(K\tau)$ is the matrix-exponential of $K\tau$. Many biochemical reaction networks are modeled as open systems, such that the molecular state space is technically infinite. That is, a given molecule in the system may occur with any positive integer value for its copy number, though practically only a finite number of states are reachable in finite time. We assume here that the CME is approximated over a finite subspace e.g., via the Finite State Projection Algorithm [86]. The linear system may be recast using the row-stochastic transition matrix [90]

$$\boldsymbol{T}(\tau) = \exp(\boldsymbol{K}^T \tau) \tag{2.4}$$

where the off-diagonal elements $T_{ij}(\tau)$ give the probability that, given the system is in state i (corresponding to molecular population vector x_i), it will then be found in state j after lag time τ. Probability is preserved, and thus rows sum to 1. The relationship between the state reaction matrix $\boldsymbol{K}$ and the stochastic transition matrix $\boldsymbol{T}$ leads to the relationship for the implied timescales:

$$t_i = \frac{-\tau}{ln\lambda_i(\tau)} \tag{2.5}$$

The dynamics is propagated according to the Chapman-Kolmogorov equation, which gives the probability at discrete times $t + k\tau$, with $k = \{1,2,3, \dots\}$ by

$$p^T(X, k\tau + t) = p^T(X, t)[T(\tau)]^k \tag{2.6}$$

2.6.4.2 Building Markov State Models with PCCA+

The goal of the clustering approach is to derive a new coarse-grained stochastic transition matrix with drastically reduced dimensionality relative to the original matrix $\boldsymbol{T}$, and which preserves as accurately as possible the slow system dynamics. Consider a reactive system in which $\boldsymbol{T}$ is defined over a total number of reachable states N (i.e., if a system of S molecular species has a maximum copy number per species of n, then $N{\sim}n^S$). Then the Perron eigenvalue $\lambda_1 = 1$ corresponds to the left-eigenvector π according to

$$\pi^T T = \pi^T \tag{2.7}$$

that is, $\pi^T = (\pi_1, \dots, \pi_N)$ gives the stationary probability over N states. If the reactive system has the property of metastability, then the dynamics can be approximately decomposed into fast and slow processes, with fast transitions occurring within metastable states, and slow transitions carrying the system between metastable states. For sufficient separation of fast and slow timescales, Markovian hopping between metastable states is a good model of global dynamics, and the full dynamics may be projected onto this slow subspace. In this case, for a sufficient ordering of the N states, $\boldsymbol{T}$ has a dominant block-diagonal structure with some number C of weakly coupled blocks. In such systems, a cluster of C eigenvalues can be found near the Perron root, with $\lambda_1 \geq \lambda_i \geq \lambda_C$ corresponding to the slow processes in the system, and all fast processes corresponding to rapidly decaying processes with eigenvalues $\lambda_i < \lambda_C$. The PCCA+ algorithm determines membership vectors χ which assign states $i \in \{1, \dots, N\}$ to clusters of states $j \in \{1, \dots, C\}$ (with $C < N$) where grades of membership are given by $\chi_j(i) \in [0,1]$. (The original PCCA algorithm was based on strict "crisp" assignments $\chi_j(i) \in \{0,1\}$, but the algorithm was not robust numerically). The membership vectors satisfy the linear transformation

$$\chi = \psi B \tag{2.8}$$

Where $\psi = [\psi_1, \dots, \psi_C]$ is the $N \times C$ matrix constructed from C dominant right-eigenvectors, and $\boldsymbol{B}$ is a non-singular matrix whose elements are determined by an unconstrained optimization procedure. The membership vectors determine a projection of the full $N \times N$ transition matrix $\boldsymbol{T}$ onto a reduced (coarse) subspace, and we denote this reduced $C \times C$ transition matrix $\widetilde{\boldsymbol{T}}$. As discussed previously, the optimization procedure can be based on maximizing the metastability, taken to be measured by $trace(\widetilde{\boldsymbol{T}})$. The coarse grained stationary probability is given by the projection

$$\tilde{\pi}^T = \pi^T \chi \tag{2.9}$$

And the transition probabilities in the reduced space are given by

$$\widetilde{\boldsymbol{T}_{\boldsymbol{ii}}}(\tau) = \frac{\langle \chi_i, \boldsymbol{T}(\tau)\chi_i \rangle_\pi}{\tilde{\pi}_i} \tag{2.10}$$

(for the probability to remain in state i within time τ) and

$$\widetilde{\boldsymbol{T}_{\boldsymbol{ij}}}(\tau) = \frac{\langle \chi_i, \boldsymbol{T}(\tau)\chi_j \rangle_\pi}{\tilde{\pi}_i} \tag{2.11}$$

(for the probabilities to jump from state i to state j in τ, where $\langle u. v \rangle_\pi$ is the π-weighted inner product of u and v). Therefore we may write

$$\widetilde{\boldsymbol{T}}(\tau) = \widetilde{D}\langle \chi_{i,} \boldsymbol{T}(\tau)\chi_j \rangle_\pi \tag{2.12}$$

where $\widetilde{D} = diag(\pi_1, \dots, \pi_C)$, or equivalently,

$$\widetilde{\boldsymbol{T}}(\tau) = (\chi^T D \chi)^{-1} \chi^T D \boldsymbol{T}(\tau)\chi \tag{2.13}$$

where $D = diag(\pi_1, \dots, \pi_N)$.

One may compare, on the coarse sub-space, the true system dynamics given by $\boldsymbol{T}$ with the dynamics given by the reduced Markov State Model, $\tilde{T}$. Given an initial probability vector $p(x, t_0)$, the true dynamics would propagate the system according to $p^T(x, k\tau + t_0) = p^T(x, t_0)T(\tau)^k$. By projection,

$$\tilde{p}^T_{true}(x, k\tau + t_0) = [p^T(x, t_0)[\boldsymbol{T}(\tau)]^k \chi \tag{2.14}$$

and

$$\tilde{p}^{T}_{MSM}(x, k\tau + t_0) = [p^T(x, t_0)[\check{\boldsymbol{T}}(\tau)]^k \chi \tag{2.15}$$

The error after k timesteps is thus given by: $\epsilon(k) = \left\| p^T(x, t_0)([T(\tau)]^k \chi - \chi[\tilde{T}(\tau)]^k) \right\|_2$. As discussed previously in [90] the maximum error associated with the projection of $\boldsymbol{T}$ onto $\widetilde{\boldsymbol{T}}$ in the reduced space (independent of initial state) is given by:

$$\epsilon(k) = \left\| [T(\tau)]^k \chi - \chi[\tilde{T}(\tau)]^k \right\|_2 \tag{2.16}$$

2.6.4.3 Transition Path Theory

Transition path theory [101] is used in this study to extract information about the kinetics and mechanism of a transition between a starting set of states A and ending set of states B. First, we calculate the flux between states during the transition between A and B according to (2.17). We obtain a net flux from (2.18) that extracts only the flux that contributes to the transition. Finally in (2.19), flux of each pathway can be used to compute the probability of traversing the pathway. Equation (2.20) and (2.21) show how to reduce the full transition matrix to a reduced representation if it is desirable to perform the pathway probability calculation on the metastable states of the system.

To compute the probabilities of the transition paths, from a starting state A to an ending state B, one must calculate the effective flux f_{ij} for all pairs of states. The transition probability relevant to the A to B state transition T_{ij}, and backward commitor ${q_j}^+$. The forwards and backwards commitors are probabilities of transitioning to B or returning to A respectively. Finally, the flux is obtained by weighting it by its starting state stationary probability.

$$f_{ij} - \pi_i q_i^{-} T_{ij} q_j^{+} \tag{2.17}$$

To get the net flow from A to B, only positive fluxes are considered by taking differences of the fluxes between pairs of states.

$$f_{ij}^{+} = \max\{0, f_{ij} - f_{ji}\} \tag{2.18}$$

After extracting out the pathways by removing their fluxes sequentially from the system, the probability of each pathway is calculated by taking the ratio of the flux of the individual path to the total flux of all paths.

$$p_i = \frac{f_i}{\sum_j f_j} \tag{2.19}$$

The total stationary probability $\tilde{\pi}_i$of a cluster of states is simply the sum of its individual states π_i in the cluster.

$$\tilde{\pi}_i = \sum_{i \in I} \pi_i \tag{2.20}$$

The transition probabilities between clusters $\widetilde{T_{IJ}}$ is found by summing transition probabilities between states T_{IJ} weighted by their stationary probabilities of the states π_i and normalized by the stationary probabilities of the clusters π_I.

$$\tilde{T}_{IJ} = \frac{1}{\pi_I}\sum_{i\epsilon I}\pi_i\sum_{j\epsilon J}T_{ij} \tag{2.21}$$

2.6.5 Pseudocode

Step 1. Define propensity function and stoichiometric transition vectors for all possible reactions in network model (see Section 2.6.4.1, equation (2.1)).

Step 2. Enumerate a reaction matrix in terms of the propensity functions stoichiometric transition vectors (see Section 2.6.4.1, equation (2.2)).

Step 3. Take the matrix exponential of the reaction matrix to obtain a transition matrix (see Section 2.6.4.1, equation (2.4)).

Step 4. Calculate the left/right eigenvectors and eigenvalues (see Section 2.6.4.2).

Step 5. Use PCCA+ algorithm, to solve for macrostate to microstate mapping using the transition matrix, eigenvalues, and eigenvectors (see Section 2.6.4.2).

Step 6. Coarse-grain the transition matrix using the PCCA+ mapping into a Markov State Model of the core states (see Section 4.2, equation (2.13)).

Step 7. Analysis of global dynamics of network model in terms of core states defined by Markov State Model

- Apply pathway decomposition to obtain a probability distribution of possible paths in terms of core states (see Section 2.6.4.3).
- Propagate the dynamics of the core states by simulating the Markov State Model (see Section 4.3, equation (15)).

2.6.6 Supporting figures

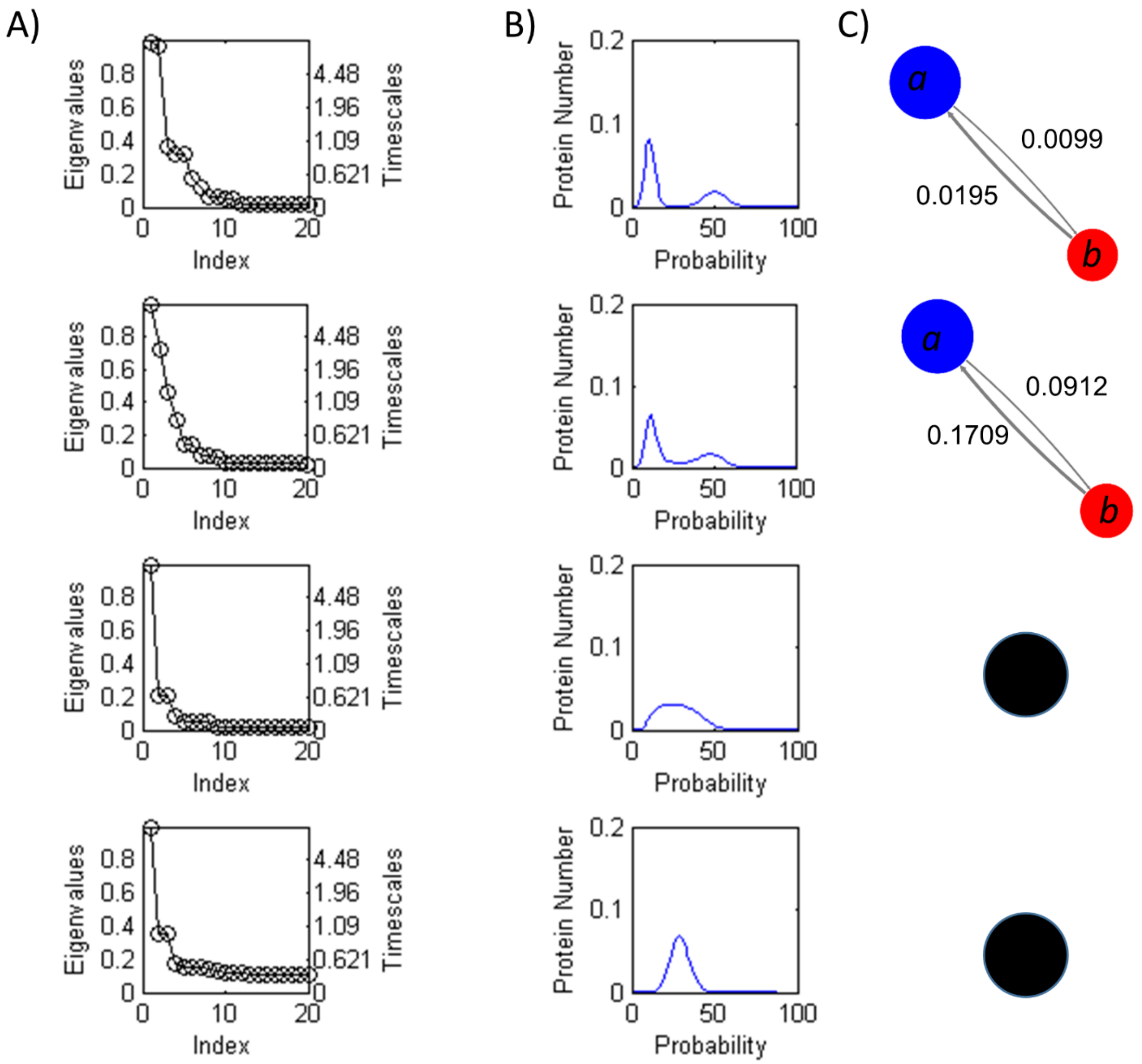


Figure 2.9 Dependence of the Self Regulating Single Gene Network eigenvalues, probability distributions,and state transition graphs on the binding parameter f

The model is defined as in [133], with parameters g_1 = 50, g_0= 10, X_{eq} = 25, $h = f = X_{eq}$, $k =$ 1. Top to Bottom: increasing $f = \{0.01, 0.1, 1, 10\}$ in units of protein degradation rate, k_A. The eigenvalue spectrum of $T(\tau)$ for $\tau = 1$, and associated timescales (B) The steady-state probability distribution, calculated from the CME truncated to $n < 100$, where n is the copy-number of the expressed gene. (C) The state transition graph produced by applying the Markov

State Model framework. A timescale thresholdof t = 2 was used to determine the number of metastable macrostates to retain in the MSM (in time-units of k^{-1}, the inverse protein degradation rate). For $f = 0.01$ and $f = 0.1$, the eigenvalue spectrum reveals a slow system timescales $t_1 = 33.17$ and $t_2 = 3.05$, respectively, thus two macrostates are retained in the MSM (edge labels give the transition probabilities within $\tau = 1$). These macrostates correspond directly to the gene off and gene on states (a and b, respectively). For $f = 1$ and $f = 10$, the eigenvalue spectrum shows some structure, but all implied timescales are shorter than $t = 2$, thus one macrostate is retained in the MSM.

f(k)	$k_{ab}(k)(from\ MSM)$	$k_{ba}(k)(from\ MSM)$	$t_2(k)$	$T_s(switching\ time, units\ k)$
0.01	0.0101	0.0198	33.17	33.34
0.1	0.1058	0.1982	3.05	3.3.4

Table 3 Estimated rate constants from the two-state MSMs for the self-regulating single-gene

Longest system timescale implied by the eigenvalue spectrum of T, and T_s, the switching time.

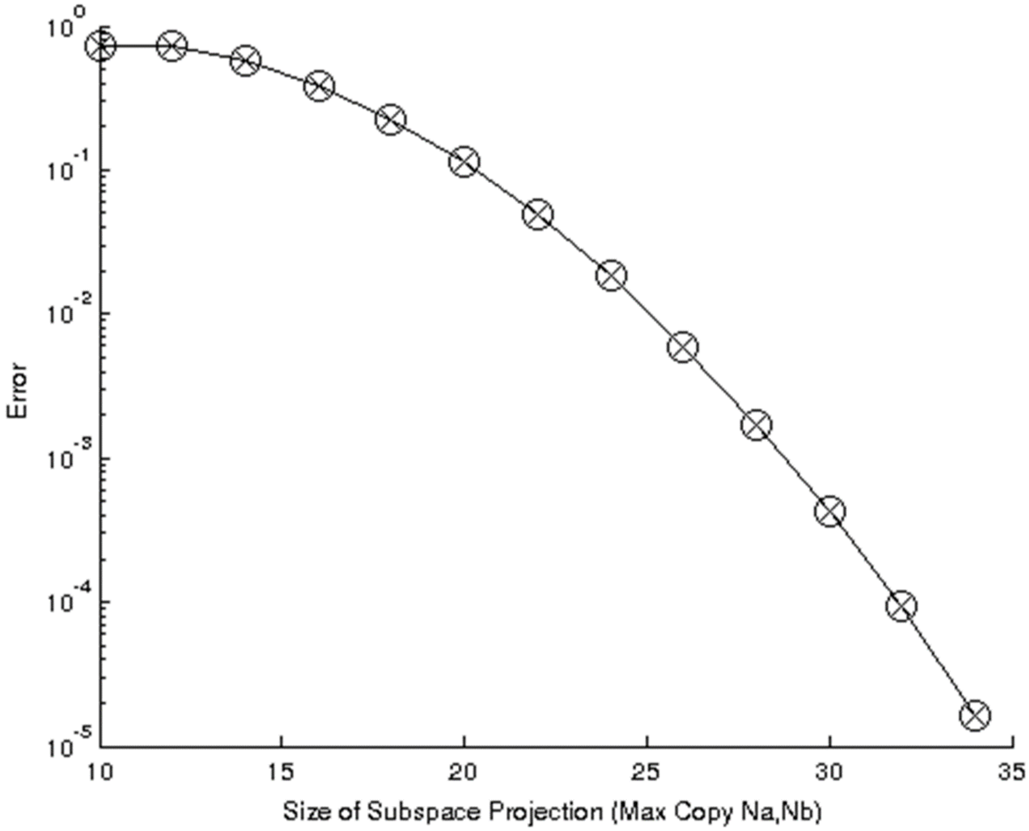


Figure 2.10 MISA network, Parameter Set 1. The truncation error as a function of the size of the subspace projection after 1000 time steps (approximately the relaxation time)

The x and o markers denote two randomly initialized state vectors. The two curves virtually overlap, showing that system has lost memory of its initial starting point. The chosen subspace for all parameter sets are cut off at N = 30 proteins, while the maximum gene expression rate $\frac{g_1}{k}$ is 16.

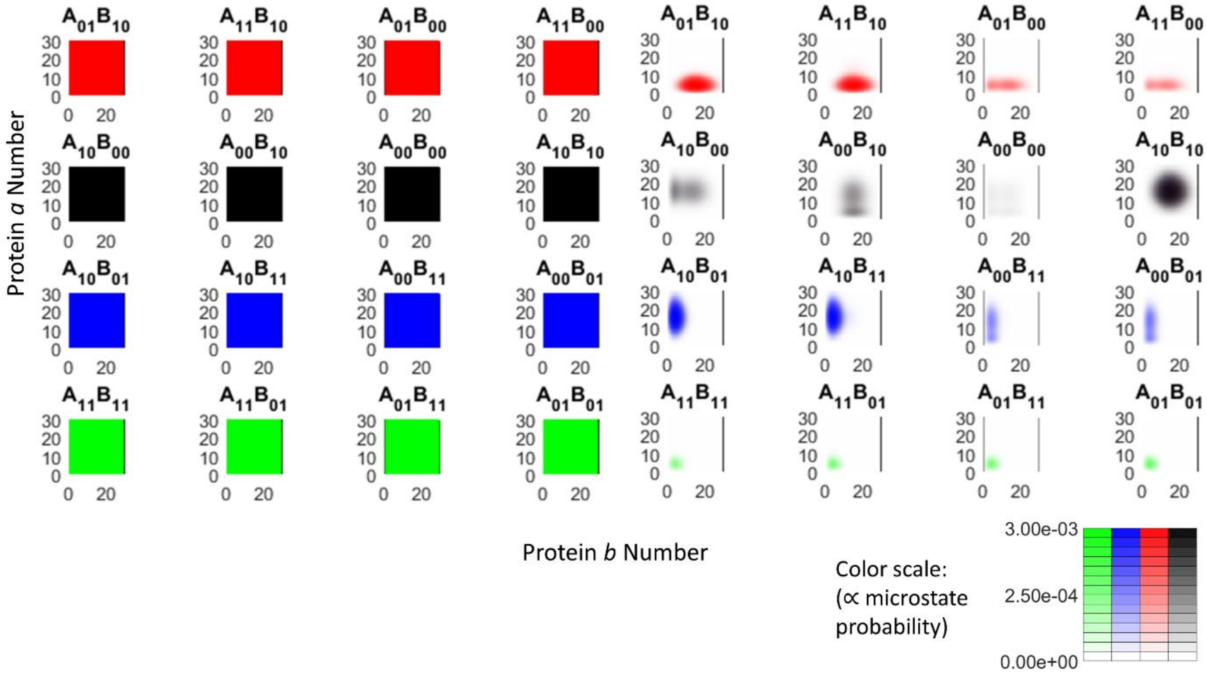
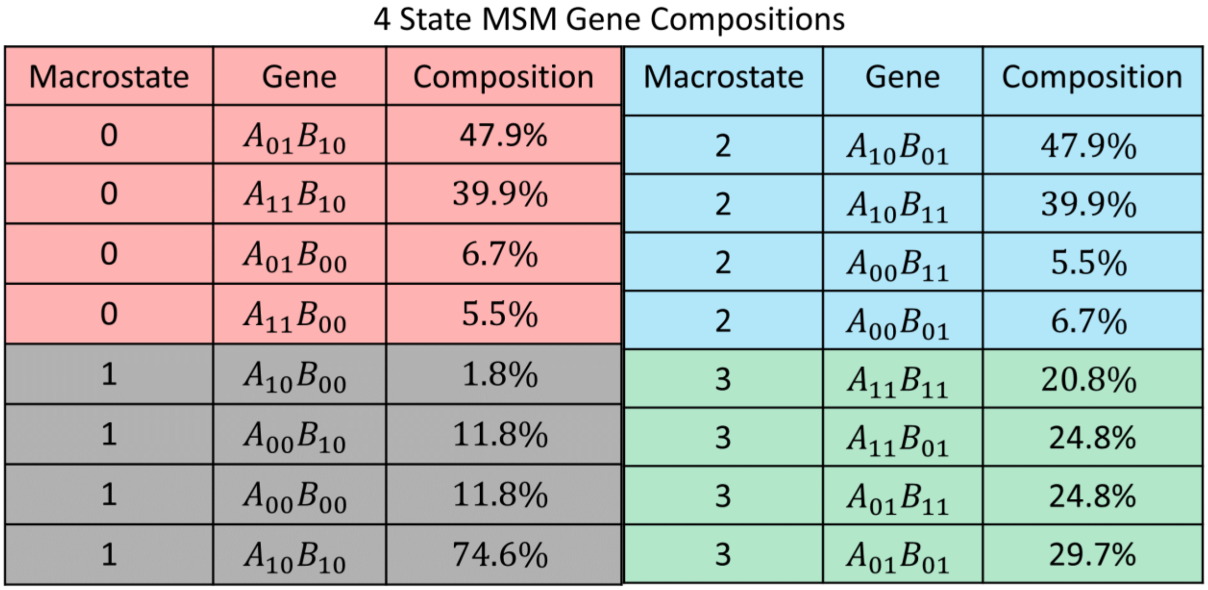


Figure 2.11 MISA network, Parameter Set 1, detailed microstate to macrostate mapping (see Fig.2). (Left)

The entirety of the enumerated state-space (all N=15,376 (31×31×4×4) microstates) is shown, and each microstate is colored according to its macrostate assignment, as determined by the PCCA+ crisp partitioning. Each of the sixteen panels corresponds to a particular DNA promoter binding configuration A_{00}, etc., as defined in MISA reactions, and Fig. 2.2B. For this parameter set, the promoter configuration determined the macrostate assignment exactly. (Right) The color-scale for each microstate is proportional to steady-state probability, such that low-probability-density regions appear white. Each of the promoter configurations shows a distinct pattern of protein expression. Grouping the high-density microstates within a macrostate and projecting onto the protein subspace gives rise to the ellipsoids shown in Fig. 2.2A. The probability density contributed by each gene state to each macrostate is shown in Table 4.



4 State MSM Gene Compositions

Macrostate	Gene	Composition	Macrostate	Gene	Composition
0	$A_{01}B_{10}$	47.9%	2	$A_{10}B_{01}$	47.9%
0	$A_{11}B_{10}$	39.9%	2	$A_{10}B_{11}$	39.9%
0	$A_{01}B_{00}$	6.7%	2	$A_{00}B_{11}$	5.5%
0	$A_{11}B_{00}$	5.5%	2	$A_{00}B_{01}$	6.7%
1	$A_{10}B_{00}$	1.8%	3	$A_{11}B_{11}$	20.8%
1	$A_{00}B_{10}$	11.8%	3	$A_{11}B_{01}$	24.8%
1	$A_{00}B_{00}$	11.8%	3	$A_{01}B_{11}$	24.8%
1	$A_{10}B_{10}$	74.6%	3	$A_{01}B_{01}$	29.7%

Table 4 MISA network, Parameter set 1, Contribution of each promoter configuration to total probability density of each macrostate

Each macrostate contains four promoter configurations.The probability contribution of each configuration to the macrostate (composition) is shown. That is, within each macrostate the probabilities sum to 100%. While more than one configuration contributes significantly to the probability density within a macrostate, the configurations contributing the most probability to the macrostate are termed 'representative' and shown schematically in Figure 2.2B.

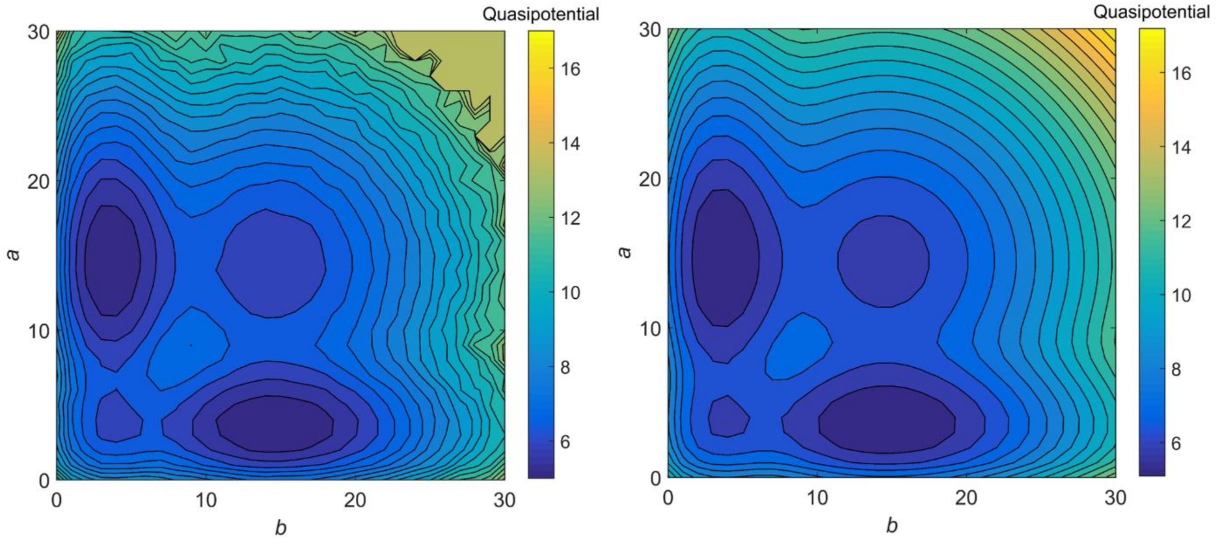


Figure 2.12 MISA network, Parameter Set 1. Brute Force SSA simulation (left) vs Chemical Master Equation (right) potential landscape comparison

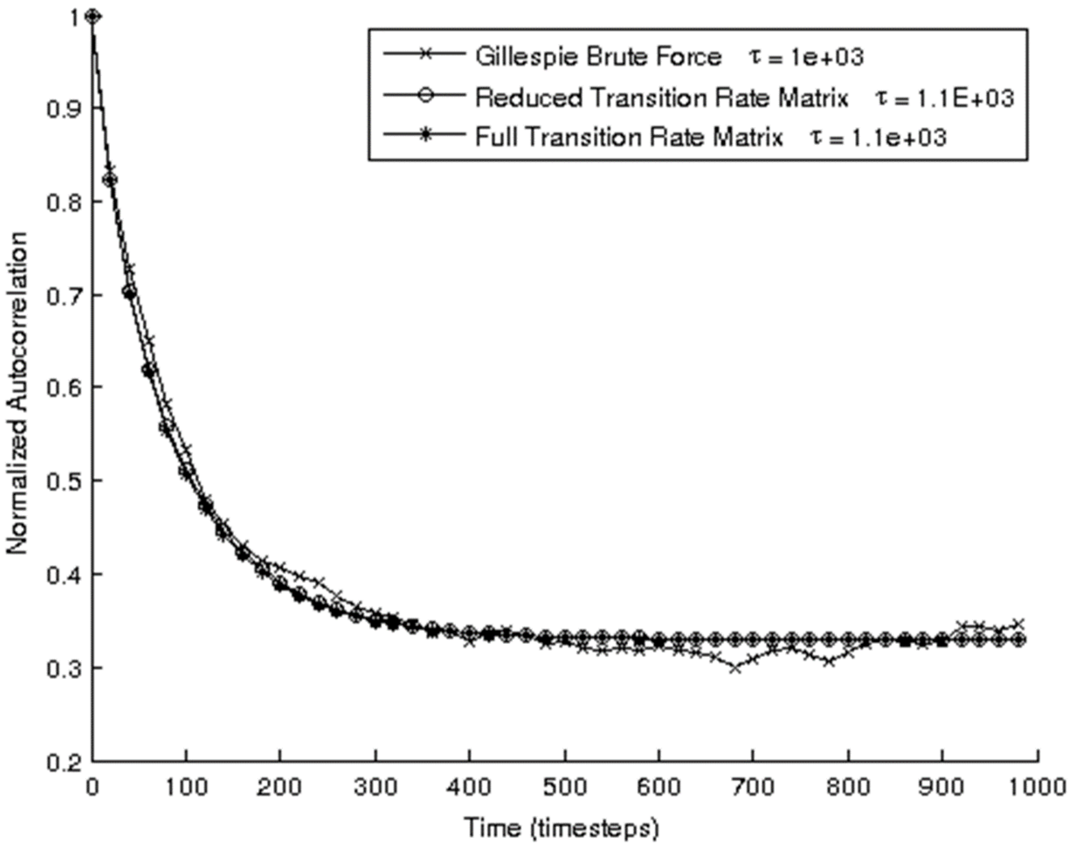


Figure 2.13 MISA network, Parameter Set 1. Autocorrelation function for the polarized state calculated from a simulation trajectory

Using Gilespie Brute Force, Reduced Transition Rate Matrix, and Full Transition Rate Matrix. Relaxation constants, τ_r, are calculated from fitting the curve to single decaying exponential.

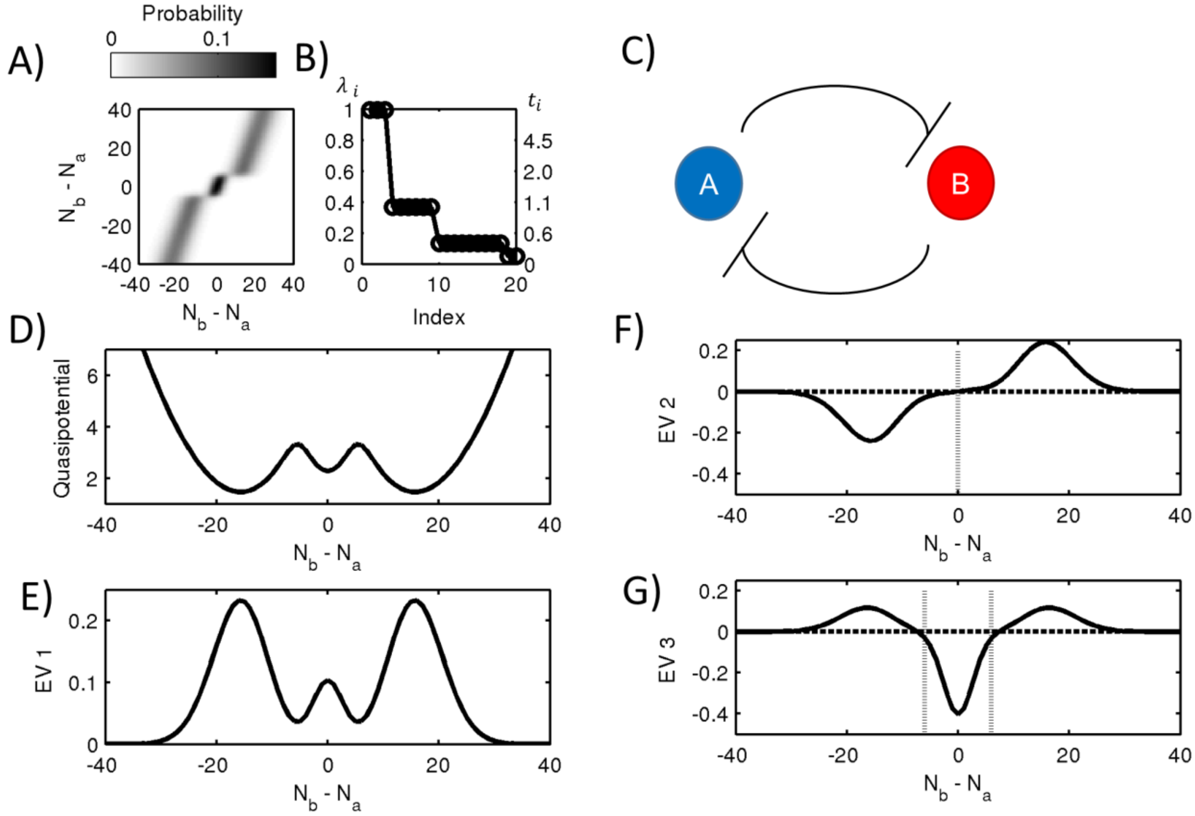


Figure 2.14 ETS network, Parameter Set 1

(A) Transition matrix in terms of number of proteins B (N_b) - number of proteins A (N_a), where probability is showed by the darkness intensity (B) The first twenty eigenvalues in descending order (C) Schematic of exclusive toggle switch (D) The one-dimensional quasipotential in terms of N_b - N_a (E) The first eigenvector in terms of in terms of N_b - N_a (F) The second eigenvector in terms of in terms of N_b - N_a (G) The third eigenvectorin terms of in terms of N_b - N_a

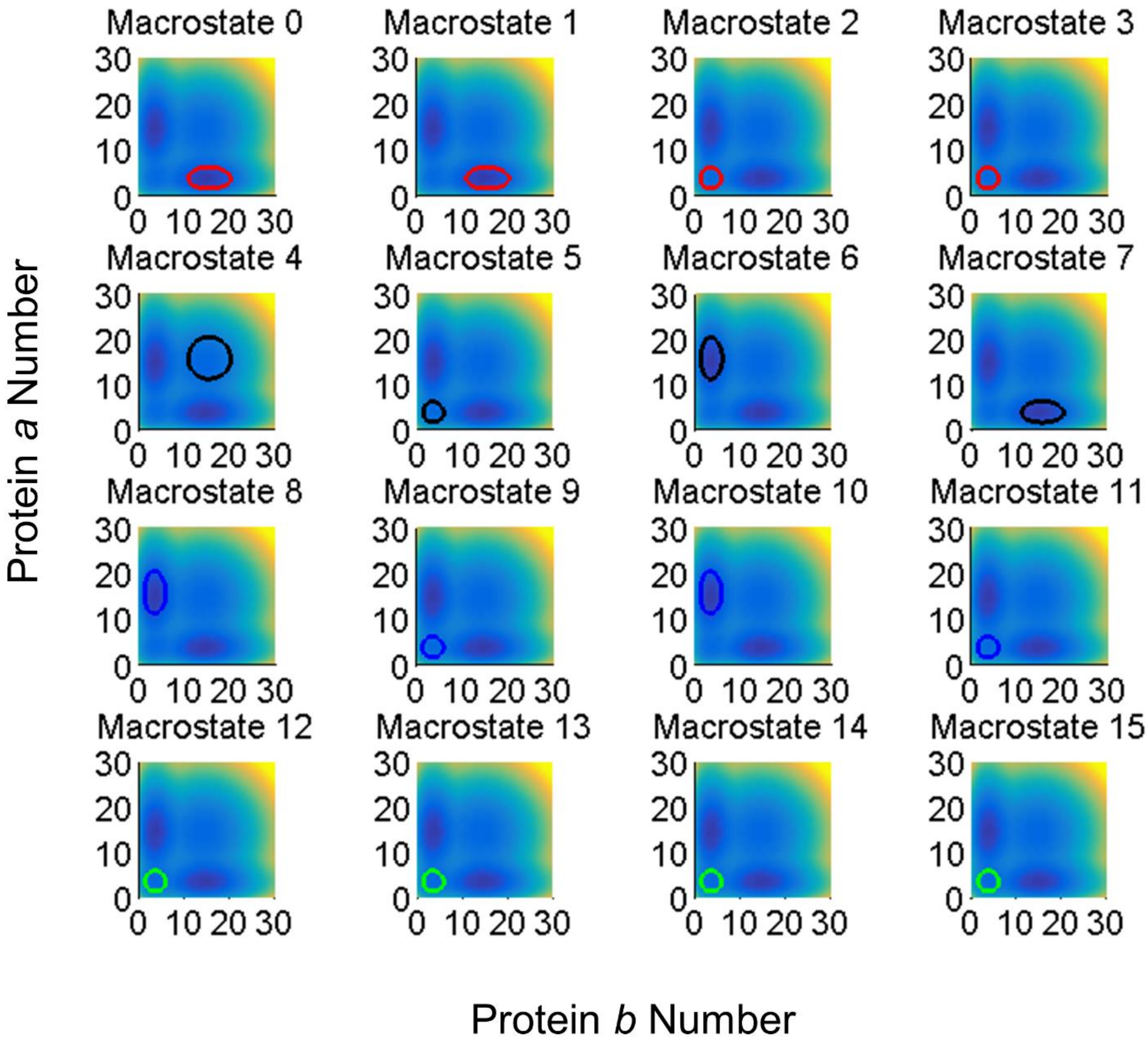
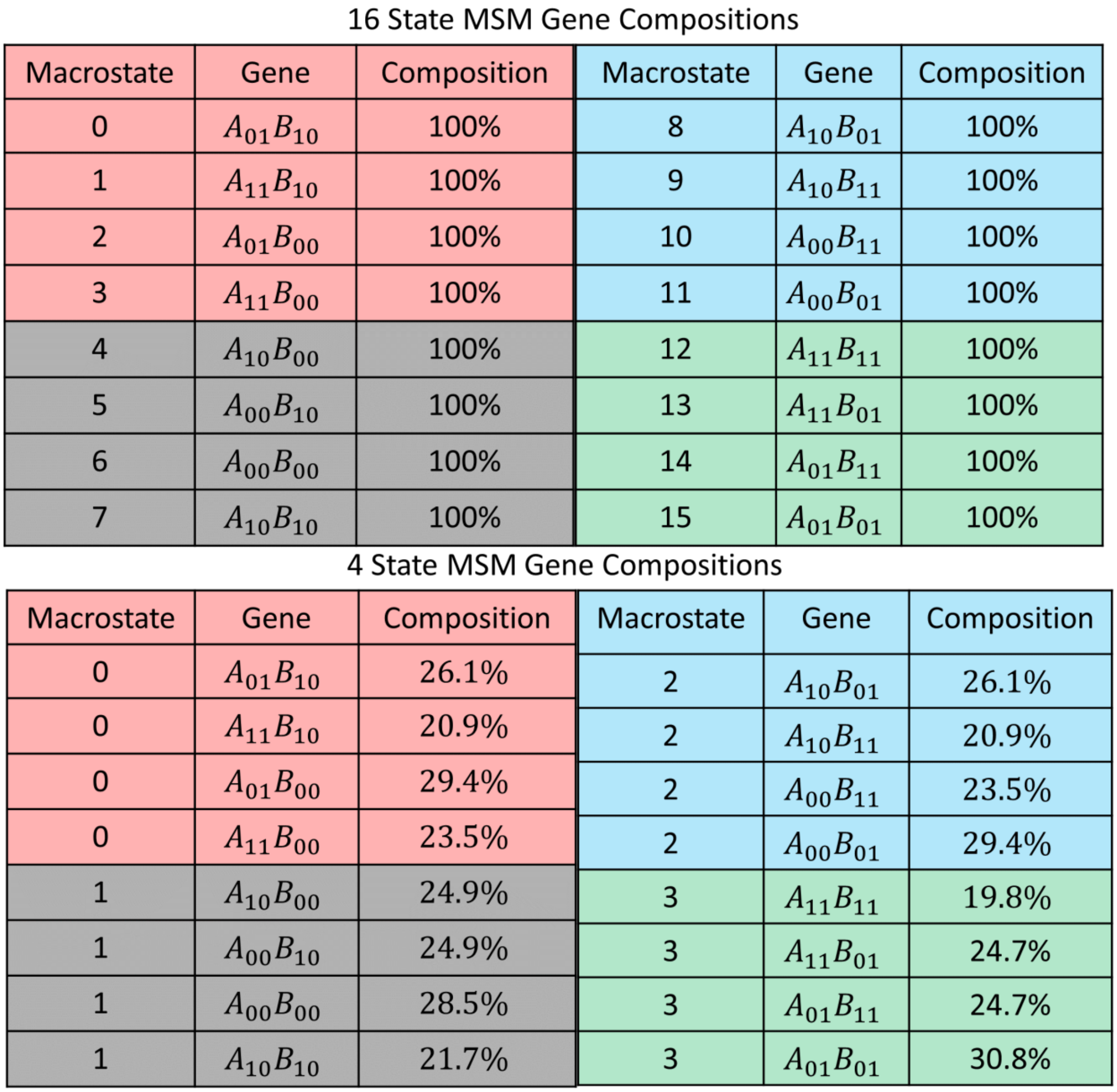


Figure 2.15 MISA network, Parameter Set 8. 16-macrostate MSM: 50% probability contours superimposed onto the dominant eigenvalues

The index of each macrostate is labeled above each plot.



16 State MSM Gene Compositions

Macrostate	Gene	Composition	Macrostate	Gene	Composition
0	$A_{01}B_{10}$	100%	8	$A_{10}B_{01}$	100%
1	$A_{11}B_{10}$	100%	9	$A_{10}B_{11}$	100%
2	$A_{01}B_{00}$	100%	10	$A_{00}B_{11}$	100%
3	$A_{11}B_{00}$	100%	11	$A_{00}B_{01}$	100%
4	$A_{10}B_{00}$	100%	12	$A_{11}B_{11}$	100%
5	$A_{00}B_{10}$	100%	13	$A_{11}B_{01}$	100%
6	$A_{00}B_{00}$	100%	14	$A_{01}B_{11}$	100%
7	$A_{10}B_{10}$	100%	15	$A_{01}B_{01}$	100%

4 State MSM Gene Compositions

Macrostate	Gene	Composition	Macrostate	Gene	Composition
0	$A_{01}B_{10}$	26.1%	2	$A_{10}B_{01}$	26.1%
0	$A_{11}B_{10}$	20.9%	2	$A_{10}B_{11}$	20.9%
0	$A_{01}B_{00}$	29.4%	2	$A_{00}B_{11}$	23.5%
0	$A_{11}B_{00}$	23.5%	2	$A_{00}B_{01}$	29.4%
1	$A_{10}B_{00}$	24.9%	3	$A_{11}B_{11}$	19.8%
1	$A_{00}B_{10}$	24.9%	3	$A_{11}B_{01}$	24.7%
1	$A_{00}B_{00}$	28.5%	3	$A_{01}B_{11}$	24.7%
1	$A_{10}B_{10}$	21.7%	3	$A_{01}B_{01}$	30.8%

Table 5 MISA network, Parameter Set 8. Gene composition for every state

16 state MSM (top) and 4 state MSM (bottom), the color corresponds to its respective substate. Compositions associated with each gene configuration is defined as the percentage of the total stationary probability attributed to that state.

2.6.7 List of Pathways of 16 State MSM

Table 6 Paths from State 1 to 11

State Path	Probability
1,5,6,10,11	24
1,3,6,10,11	21
1,0,7,4,8,9,11	18.5
1,5,11	3.3
1,0,7,9,11	2.8
1,15,11	2.2
1,0,7,4,8,10,11	1.9
1,0,14,9,11	1.8
1,3,13,10,11	1.9
1,0,2,13,8,9,11	1.9

Table 7 Paths from State 11 to 1

State Path	Probability
11,5,7,0,1	23.2
11,9,7,0,1	21.7
11,10,6,4,3,2,1	20.6
11,9,8,4,2,0,1	8.3
11,10,8,4,2,3,1	7.8
11,5,1	3.1
11,10,6,3,1	2.7
11,15,1	2.2
11,10,13,3,1	1.8
11,9,14,0,1	1.7
11,10,8,4,2,0,1	1.5
11,9,8,12,2,0,1	1.1

3. Simulation of Rare Events in the Diffusion of Molecules on Crowded Cell Surfaces

3.1 Introduction

The surface of a cell is crowded, with a high density of cell surface proteins. Some cellular processes require this crowding to be overcome. For example, large surface molecules such as CD45 are known to locally impede the T cell receptor (TCR) from engaging its target on the opposing cell (Figure 3.1). It has been estimated that a region of radius 100 nm laterally between cell surfaces must be cleared[20] of large surface molecules. Large surface molecules (LSM) may inhibit T-cell receptor binding through steric interactions by occupying an important region of interest[23] (Figure 3.1). The steric hindrance of LSM in TCR triggering has been studied in mathematical models [20], [134]–[136] and observed in experiments [21], [137]. Thus, we are interested in the kinetics of this event: how long it takes for the large-surface molecules to evacuate, the mean first passage time, (MFPT) and mechanism, and the intermediate steps required to fully evacuate.

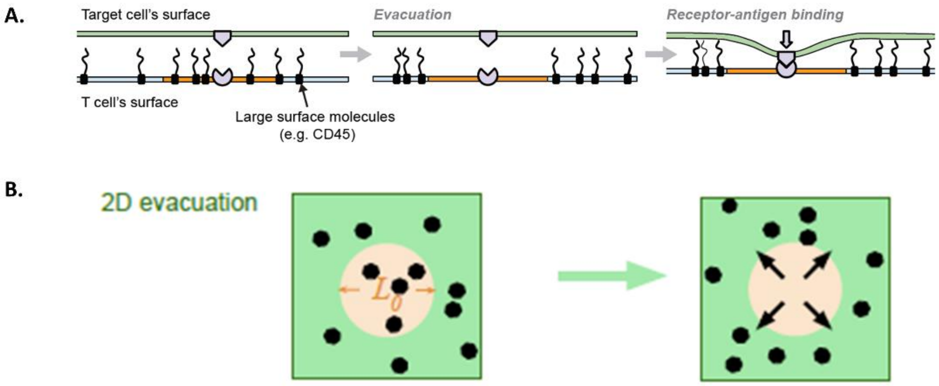


Figure 3.1 Schematic of evacuation event in biological and simulation settings

(A) A Target cell's surface (green) is in proximity of a T cell's surface. In the first snapshot, the region of interest is occupied with large surface molecules, such as CD45. In the second snapshot, large surface molecules diffuse away from the region of interest, which we call evacuation. After evacuation has occurred, the receptor-antigen binding event is able to occur. (B) The process of evacuation is modeled by a 2D box with point particles, representing the large surface molecules. It is considered evacuated when the all point particles leave the region of interest, which has a radius of L_0. [138]

A number of simulation approaches and tools have been developed towards studying the reaction-diffusion of molecules on cell surfaces [139]–[142]. The approaches can be characterized as either based on the reaction diffusion master equation [143] or the Smoluchowski framework[144]. The differences in the two frameworks works lies in the accuracy and computational cost. While the Smoluchowski framework is highly accurate due to explicit modeling of the exact particle positions, it also very computationally costly. The opposite can be said of reaction diffusion master equation approaches, it is less accurate than the Smoluchowski framework since particle positions are modeled according to discretized compartments in the spatial domain, but it is also less computationally costly. Since the position of particles is of high importance to our problem of interest, we adopt the Smoluchowski framework to model the reaction-diffusion of molecules on the cell surface.

Particle-based reaction -diffusion processes can be expensive to simulate. Enhanced sampling techniques can speed computation when the process of interest involves a broad range of timescales. Specialized rare event sampling algorithms can be used to extract kinetic information of events in an efficient manner in a wide range of characteristic length scales and system detail. For the evacuation problem, the level of system detail is at the spatially resolved cell-scale. An algorithm that has been applied at this level of system detail is Weighted Ensemble Sampling in a previous study by Donovan et. al [145]. The results of the paper provided estimates of MFPT for several spatially resolved cell state systems using WESTPA [46], a Weighted Ensemble software package, and Mcell, a 3D spatially resolved cell scale model simulation package. The choice of Weighted Ensemble Sampling is motivated by the minimal assumptions required. For example, another rare event sampling algorithm called Milestoning [37] and Markov State Models have the requirement of memory loss [146], [147]. Rare event sampling algorithms with high generality, such as Weighted Ensemble Sampling and Forward Flux Sampling[39], [148], [149], have the tradeoff of relatively high computational costs in comparison to algorithms with low generality. Overall, adapting algorithms to the specific characteristics of the spatially resolved cell-scale domain has been relatively unexplored and potential challenges remain.

We use the 2D evacuation problem as a model system for expanding the applicability of the Weighted Ensemble algorithm to particle-based reaction diffusion simulations. It is considered a rare event due to MFPT of evacuation is much longer than the diffusion time scale $\frac{L^2}{D}$, where L is the side length of the box and D is the diffusion coefficient. The main goal of this study is to characterize the trend of the MFPT with respect to the particle density. However, scaling to high densities became increasingly difficult due to the rising duration of the typical transition event. Choosing parameters that gave sufficient sampling became absolutely necessary to obtain the exact MFPT estimates and not have overestimation of the MFPT due to insufficient sampling. Our results showed that some strategies of increased sampling were more efficient than others in the convergence of the MFPT. Also, the results were compared to an asymptotic approximation of the 2D evacuation problem and the causes of the differences between were analyzed [138].

3.2 Methods

3.2.1 2D evacuation Brownian Diffusion

The evacuation problem was modeled as a 2D box with particles diffusing freely and evacuation is characterized as the state when all particles leave the circular region (ROI). The units are defined in terms of characteristic time, t^* and characteristic length x^*. The physical time units, x and $t,$ can be recovered by multiplying the dimensionless quantities, X and T, by the characteristic units (equations 3.1 and 3.2). The characteristic units used in the simulations are $0.1\ s$ for t^* and x* is $100\ nm$, which correspond to the physiological conditions found in TCR triggering at a diffusion coefficient of $0.1\ \mu m^2/s$ [150], [151]. A typical biological density is around $955\ \frac{particles}{\mu m^2}$, which corresponds to 30 particles in the region of interest. The parameters are set as follows: the Region of Interest radius, L_0 , is $1\ x^*$, box side length, L is $3\ x^*$and the diffusion coefficient, D is $1\ x^{*2}/t^*$. The Brownian Dynamics 2D box code was written in MATLAB, however it is also possible simulate the dynamics in Smoldyn, an open-source, particle-based spatial stochastic simulator [142].

$$x = Xx^* \tag{3.1}$$

$$t = Tt^* \tag{3.2}$$

The study of the mean first passage time of evacuation of particles from a region of interest (ROI) is a computational challenge as the number of particles in the box grows. With our current system parameters, when the particle density, $\langle\rho\rangle$ becomes greater than 2.55, the brute force cannot sample the evacuated state, within 10^6 simulation time steps of $dt = 10^{-5}t^*$. This can be understood by comparing the sampled brute force distribution and the corresponding binomial distribution at two different values of $\langle\rho\rangle$ (Figure 3.2). For $\langle\rho\rangle = = 1.44\,\frac{particles}{x^{*2}}$, the evacuated state is sampled as well as most of the other states in the brute force distribution. However, for $\langle\rho\rangle = \quad = 11.11\,\frac{particles}{x^{*2}}$ only states near the peak of the distribution are sampled. This can be understood from a coin toss analogy which can also be modeled as a binomial distribution: the evacuated state which has 100 particles outside of the ROI would be analogous to landing heads with a biased coin 100 times.

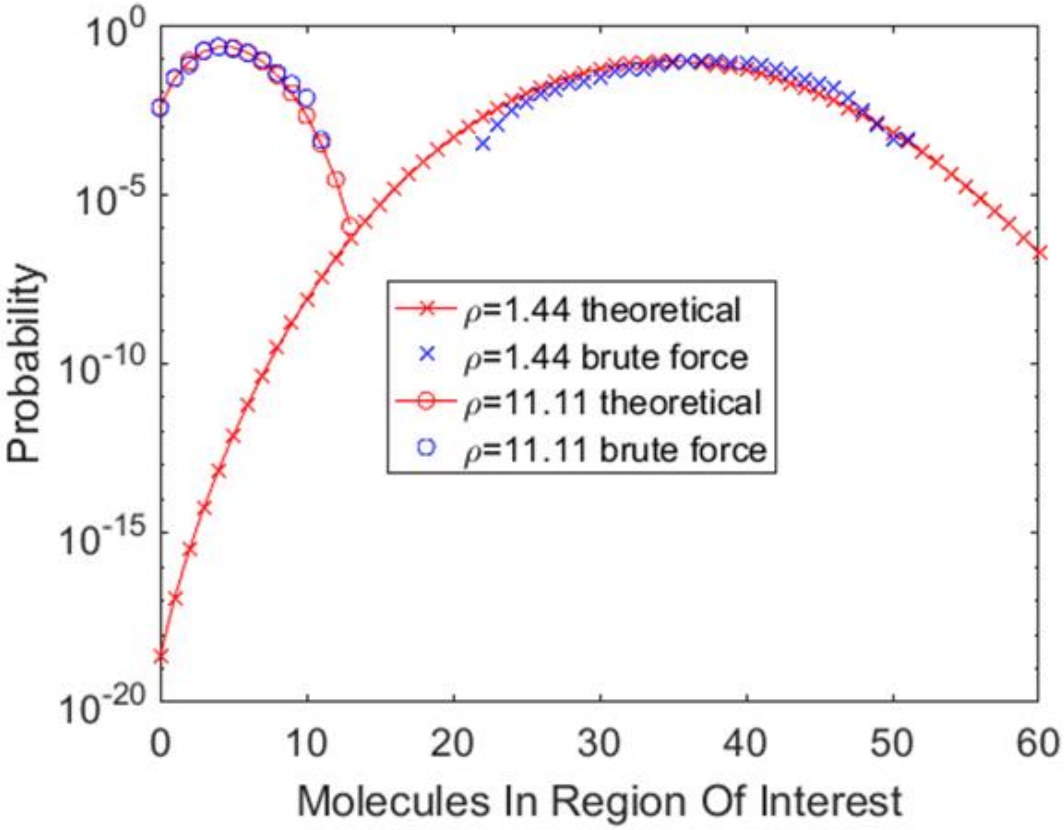


Figure 3.2 Sampling the evacuation event becomes increasingly difficulty at higher particle densities

The stationary equilibrium distributions of particle densities of 1.44 and 11.11 are compared as a function of the molecules in region of interest. The theoretical stationary equilibrium distribution of molecules in region of interest is modeled by a binomial and plotted in red. The blue corresponds to probabilities derived from brute force sampling corresponding to 1 t^* simulation time.

3.2.2 Weighted Ensemble Rare Event Sampling

3.2.2.1 Background

The Weighted Ensemble (WE) method of rare event sampling was introduced to enable efficient simulation of protein-association reactions [30]. A key aspect of the WE algorithm is that computational power is directed toward low-probability regions of state space, which would typically be undersampled in a traditional simulation (Figure 3.3). This is achieved by running a large number of short simulation trajectories, initialized throughout the state-space, and redistributing trajectories from more probable regions to less probable regions. Information from these trajectories is combined to compute key observables (e.g. the steady state distribution or the MFPT for transitions between regions of interest) in a manner that is consistent with the true system dynamics. The WE code was written in MATLAB, but it also possible to use an available open source software that implements WE as a wrapper for any stochastic dynamics, WESTPA [46].

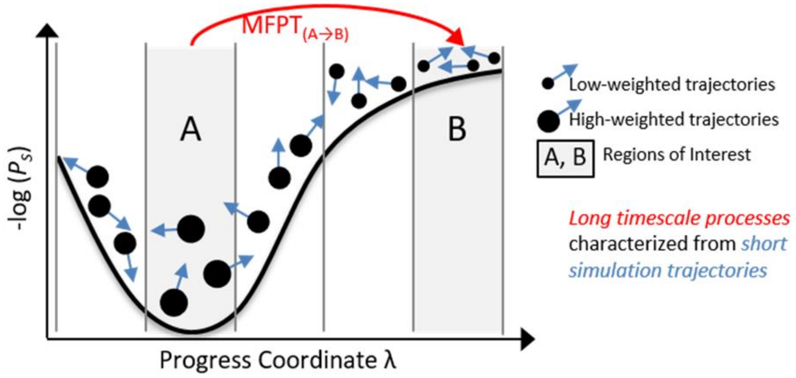


Figure 3.3 Schematic of Weighted Ensemble sampling

The weighted ensemble algorithm is described by a schematic of how the algorithm derives the $MFPT(A \rightarrow B)$. The region A represents a subset of the state space that represents the initial state, while B represents the subset of the state space corresponding to the target state. Trajectories with higher weight reside in the region A**,** while the low weighted trajectories populate the states near region B. The short time dynamics of these trajectories is used to quantify the flux, the rate at which probability crosses over to region B**.**

The algorithm is described as follows: the state of the system is characterized by a single or multidimensional order parameter that either uses the system degrees of freedom directly or can be derived from the degrees of freedom. Choice of the order parameter is crucial to the convergence of the algorithm because even sampling is enforced along this order parameter. An optimal choice of order parameter will approximately match a system progress coordinate to reach the target of interest. However, a bad choice of order parameter will make convergence of the algorithm slower since time will be wasted in irrelevant parts of the configuration space. The state space is divided into bins defined by the order parameter that span the transitions of interest. Initially, the algorithm starts out with one "replica" assigned a weight of 1, which is one of the many simulation trajectories with weights proportional to their probability that will eventually populate the bins. Each bin is required to have a target number of replicas, M_{targ}, occupying it after each iteration of the algorithm. The replicas are propagated forward in time by a constant time step, τ_{we}. The replicas are then either combined if the weight of the particle in a bin, w_i, is too low or duplicated if too large in order to keep the weight of the replicas inside the bin close to the average weight of the replicas inside the bin, $\frac{M_{targ}}{\sum_i w_i}$. Even if a bin contains low probability weights, the resampling procedure will ensure that there are sufficient replicas within the bin and also that the weights of the replicas portray the probability within that bin.

3.2.2.2 Simulation Parameters

The WE parameters for the evacuation problem require careful selection due to the unique qualities of the target of interest, the evacuated state. First off, we chose the order parameter to be $Number\ Inside$, the number of molecules in the region of interest. When a simulation trajectory has $Number\ Inside = 0$, it is considered in the evacuated state; otherwise it is not evacuated. There is no concept of being in a transition region. Our initial strategy was to assign one bin (containing M_{targ} replicas) to each discrete value of the order parameter, $Number\ Inside$. However, we found that this strategy dedicates unnecessarily large computational time to sampling high $Number\ Inside$ states, which are far from the target state and thus provide no benefit for sampling the evacuation event. We then chose an approach wherein the high $Number\ Inside$ states are lumped together into one or two bins (Figure 3.4). We found that weight must be transferred with the smallest possible time interval, $dt,$ as opposed to waiting until the end of τ_{we}, which is too long for the problem currently at hand (see section 3.3.1 for more details).

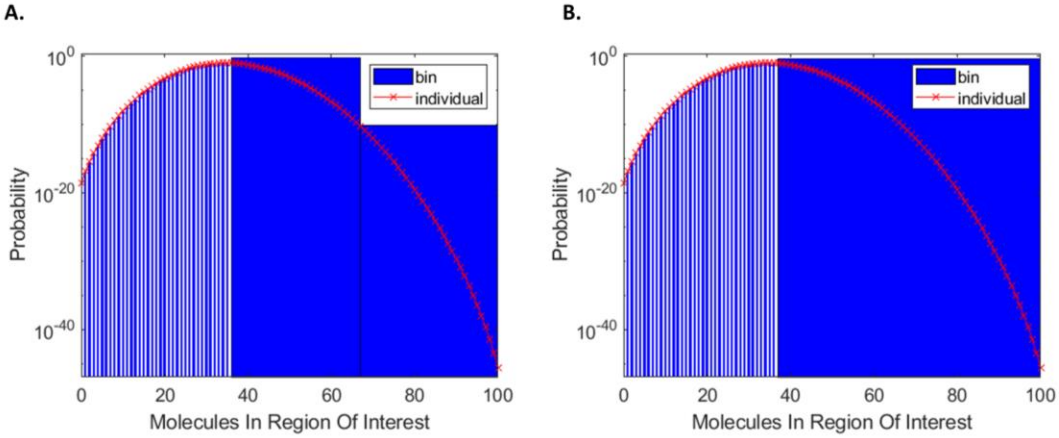


Figure 3.4 Binning schemes used in simulations.

Binning is fine-grained near the target (evacuated) state, and coarse-grained far from the target state. The transition from fine-grained to coarse-grained occurs at the maximum (mode) of the expected (binomial) equilibrium distribution. For $Number\ Inside < max$, each discrete value of $Number\ Inside$ is assigned one bin. For $Number\ Inside > max$, all states are lumped into either (A) two, or (B) one bin(s).

3.2.2.3 WE MFPT Calculation

The following Hill relation [48] (equation 3.3) is used to compute the MFPT from initial state A to target state B (adapted from [145]):

$$MFPT(A \rightarrow B) = \frac{1}{Flux(A \rightarrow B)} \tag{3.3}$$

All trajectories originating from bins designated as initial state A that have entered the bins designated as target state B are summed during that duration of τ_{we}. The sum of weights divided by τ_{we} gives the probability flow per unit time, $Flux(A \rightarrow B)$. The validity of the expression hinges on achieving steady state conditions of the weights within bins, among other things.

3.2.2.4 Estimation of Computational Efficiency

Computational efficiency can be defined in various ways, but it often is used as a measure of the relative speedup of the method compared to brute force. In this manuscript, we describe computational efficiency, $Efficiency\ gain$, as the ratio of the time steps (measured in units of t^*) of the estimated MFPT and the WE total simulation time steps to estimate the MFPT from state A to B.

$$Efficiency\ gain = \frac{Time\ steps(conventional)}{Time\ Steps\ (WE)} \tag{3.4}$$

In order to compute $Time\ steps\ (WE)$, we use: $Time\ steps\ (WE) = WE_{iter} \times N_{bins} \times M_{targ} \times \tau_{we}$. For events that are too rare to sample by conventional simulation, we use the WE-sampled estimate of the MFPT as a lower bound for $Time\ steps\ (conventional)$. This is justified in that obtaining a single “successful” evacuated trajectory with conventional simulation requires, on average, a simulation of length MFPT. Thus, Eqn. 3.4 provides a lower-bound estimate of the true efficiency gain.

3.3 Results

3.3.1 Stop condition is necessary to obtain correct results

Checking for immediate arrival of trajectories to the target state is necessary in order to obtain correct estimates of the MFPT and is especially important when target states are located at the tail of a long distribution. In Figure 3.5, it is shown that there is a significant dependence on τ_{we} without including a “stop condition”, which is to check for the target state after dt, the simulation time step. A “stop condition” only interrupts the dynamics if a trajectory has crossed into the target state by including its weight in the flux estimate and restarting the trajectory randomly according to the current bin distribution, otherwise a trajectory will be simulated for the full τ_{we}. When the “stop condition” is included, the τ_{we} dependence vanishes and is unchanging over the range of τ_{we} used. Our target state in the tail of the distribution is short-lived. The τ_{we} time step is too long to capture an accurate flux estimation to the target state because several crossings may have happened at the resolution of dt that were missed at the resolution of τ_{we}. The longer τ_{we} is, the more the MFPT is overestimated when no “stop condition” is applied as shown in Figure 3.5.

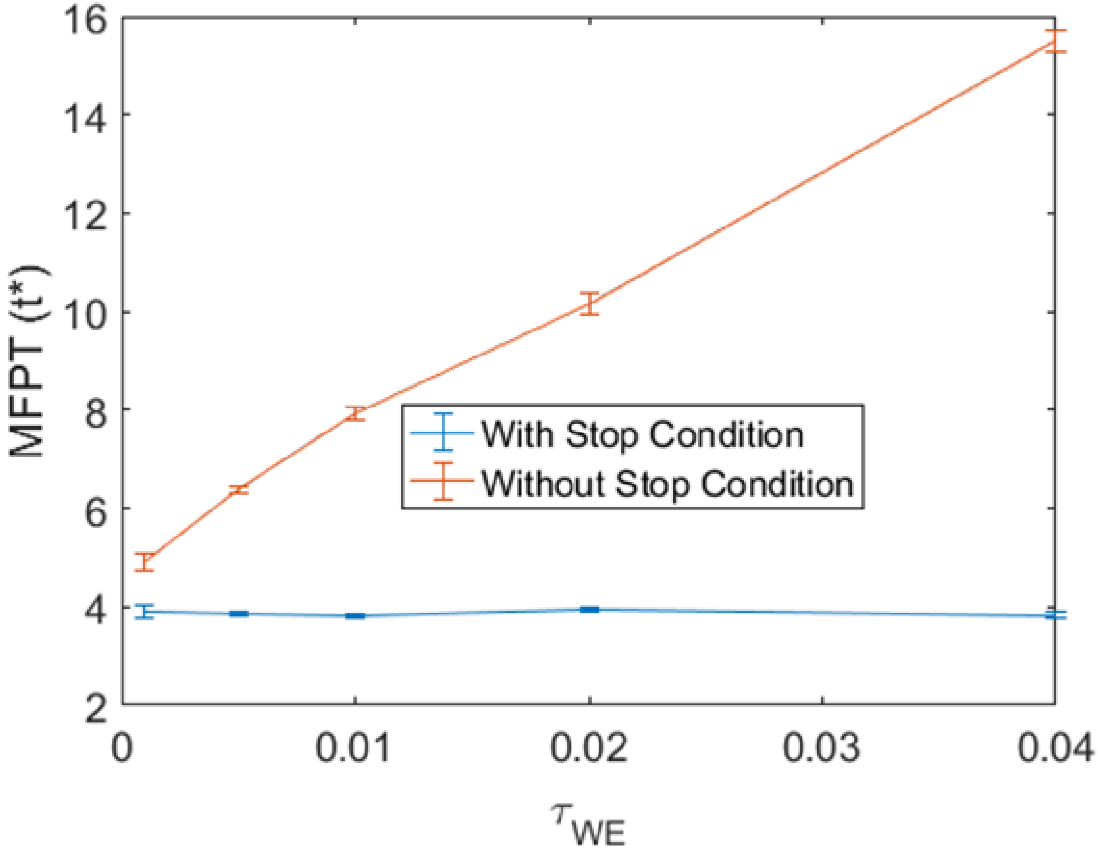


Figure 3.5 A stop condition is required to correctly capture the MFPT

The behavior of the MFPT is plotted as a function of τ_{we} with (blue) and without (red) the stop condition. Twenty simulations were averaged to produce the data points and the error bars (standard deviation). The following parameters were used: $\langle\rho\rangle = 1.44$, $WE_{iter} = 10000$, $M_{targ} = 500$ and $L = 3L_0$. The binning scheme in Figure 3.4a is applied.

3.3.2 Weighted Ensemble enables simulation of rare evacuation events at high particle densities

The challenge of simulating high particle densities is overcome with WE sampling. 20 simulations were run with WE parameters which gave sufficient sampling such that the standard error is kept minimal. At low particle densities, since the computational cost is still manageable (within 24 hours, but still takes longer than WE), the brute force estimates are compared with the WE estimates and are shown to agree in Figure 3.6a. The trend shows that the MFPT scales subexponentially with particle density. Based on the obtained values at the highest density which is roughly equivalent to the physiological density, passive diffusion cannot alone drive evacuation. The efficiency gain, (Eqn. 3.4), is plotted for each value of the particle density. The Weighted Ensemble computational time (as measured in units of the model timestep, t^*) scales with the number of replicas which increases at a linear rate with increasing particle density, while the MFPT itself scales subexponentially. Thus, the efficiency gain (Eqn. 3.4) grows at nearly the same rate as the MFPT.

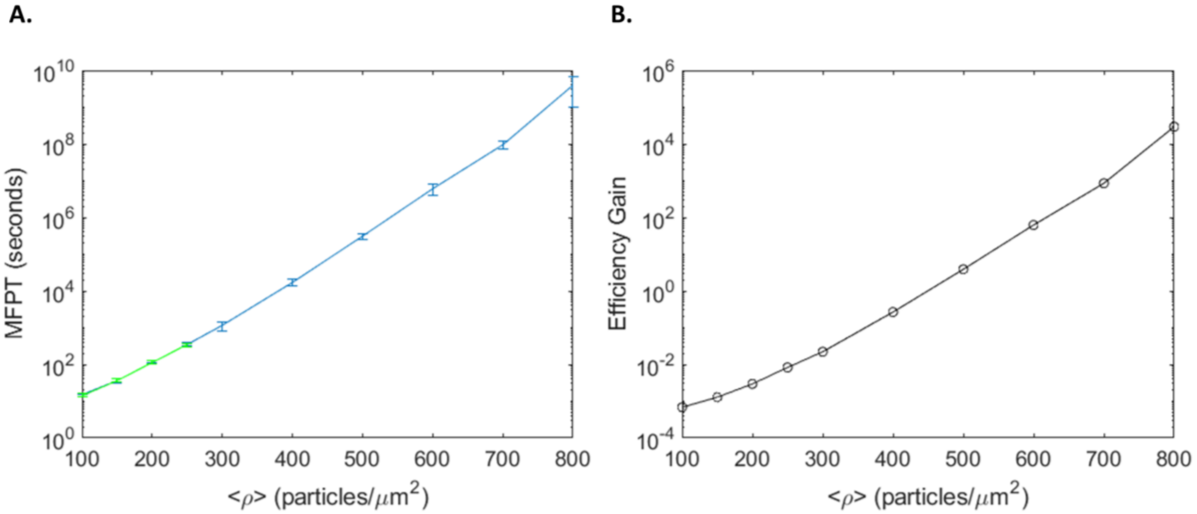


Figure 3.6 Mean First Passage Times of particle density computed by Weighted Ensemble and the associated efficiency gains

Twenty simulations were averaged to produce the data points and the error bars (standard deviation). **(**A) The MFPT as a function of particle density $\langle\rho\rangle$. **(**B) The efficiency gain as a function of particle density $\langle\rho\rangle$. The following parameters were used: $\tau_{we} = 10^{-4}\ seconds$, $WE_{iter} = 10000$, $M_{targ} = 100$, and $L = 6L_0$. The binning scheme in Figure 3.4b is applied.

3.3.3 The MFPT to escape depends strongly on dynamics time step and also the size of the box

The dependence of the MFPT on the dynamics time step and size of the box indicates that the WE estimate may approach the asymptotic approximation in the limit of infinitely large size and small time step. In our WE estimation of MFPT vs N, the values are constantly exceeding the asymptotic result of [138]. We verified from brute force simulations that we are correctly simulating Brownian Dynamics by matching the WE simulations and brute force simulations at various dt and L in Figure 3.7a. There are two factors we have identified that contribute to the difference: the larger predicted MFPT resulting from the finite time step, dt, and the assumption of infinite wall length, L, made from the asymptotic approximation shown in Figure 3.7a. For $\langle\rho\rangle = 2.55\ \frac{particles}{x^{*2}}$, the MFPT is compared between WE and the asymptotic result for a range of dt and L. For dt, the MFPT drops dramatically at $dt = 10^{-3}\ t^*$and for smaller dt, the subsequent decreases are small. For $L = 3L_0$, the dt is varied from $10^{-2}\ t^*$ to $10^{-7}\ t^*$, and for $L = 6L_0$ the dt is varied from $10^{-2}\ t^*$ to $10^{-6}\ t^*$. The value of the MFPT is decreased even further within 50% of the asymptotic MFPT prediction at $L = 6L_0$, though it is still not quite matching the asymptotic approximation which might be due to the finite time step and finite box length.

The dependence of the wall length for increasing particle density is shown alongside the asymptotic approximation. When the wall length is at $L = 3L_0$, the MFPT grows much faster than the asymptotic approximation. At $L = 6L_0$, the MFPT also grows faster than the asymptotic approximation but at a slower rate than $L = 3L_0$.

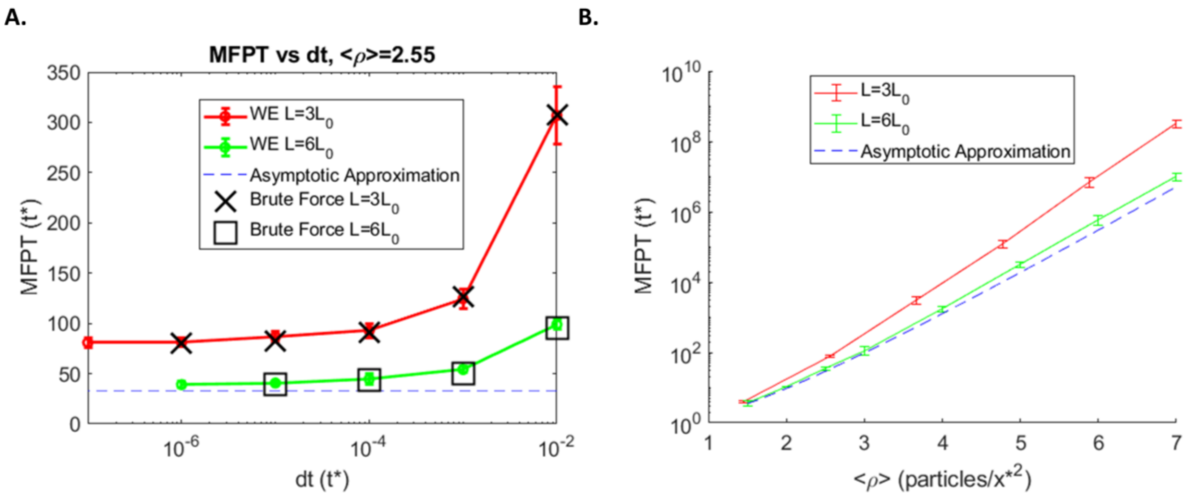


Figure 3.7 MFPT is dependent on box size (i.e., the size of the simulated area, where $L = length\ of\ the\ box$) and Brownian dynamics time step.

Ten to twenty simulations were averaged to produce the data points and the error bars (standard deviation for WE and 95% confidence intervals for brute force). (A) The MFPT plotted as function of dt, the finite step of the dynamics at particle density of 2.55 at a side length of $L = 3L_0$ and $L = 6L_0$. (B) The MFPT for the WE at $L = 3L_0$ and $L = 6L_0$ and the asymptotic approximation as function of particle density $\langle\rho\rangle$. The following parameters were used : $\tau_{we} = 10^{-3}t^*$, $WE_{iter} = 10000$ and $M_{targ} = 100.$ The binning scheme in Figure 3.4b is applied.

3.3.4 The effect of not reintroducing replicas on MFPT and flux

Weighted Ensemble can be carried out in two distinct methods: in the original method, replicas are removed from the system immediately upon reaching the target state and then reintroduced back into the intial state. We call this *reintroduction* method (and in our implementation, “immediate” removal implies checking for evacuation every dt, see 3.3.1). In a second type of WE method, replicas are allowed to freely propagate in an out of the target state. In this approach, it is critical to track flux only from replicas that are *first* transitioning from region A to region B (the target state), i.e., only those replicas that enter B after most recently being in A. The importance of tracking only these one-way fluxes has been discussed previously [48].In order to differentiate replicas that are entering versus exiting the target region, replicas must be labeled or “color-coded” [44]. First, regions of state-space were designated either A (source), B (target) or I (any state not in A or B), and all replicas were designated as having been most recently in either A or B. We will term this method “non-reintroduction”. In the previous sections, we did not explicitly define an intermediate (I) region; all particles were considered to be in the source state (A) until they evacuated, at which point they were removed from the system and reintroduced again into A. We found that, when implementing the non-reintroduction method, the size of the I region (which we also call the “gap”) has an effect on the MFPT estimate. When the size of the gap region is zero (Figure 3.8c), the MFPT is constantly below the reintroduction method (Figure 3.8a). However, when the size of the gap region is set to four (Figure 3.8d), the MFPT matches closely with the reintroduction method (Figure 3.8a). The non-introduction does hold a potential advantage in that it produces much more stable fluxes at higher densities (which can result in smaller error bars as seen at $\langle\rho\rangle = 8$ in Figure 3.8a). At $\langle\rho\rangle = 7$, the fluxes of each variant is plotted as a function of iteration in Figure 3.8b. The two non-reintroduction variants exhibit much more non-zero fluxes as a result of not reintroducing randomly back into the distribution.

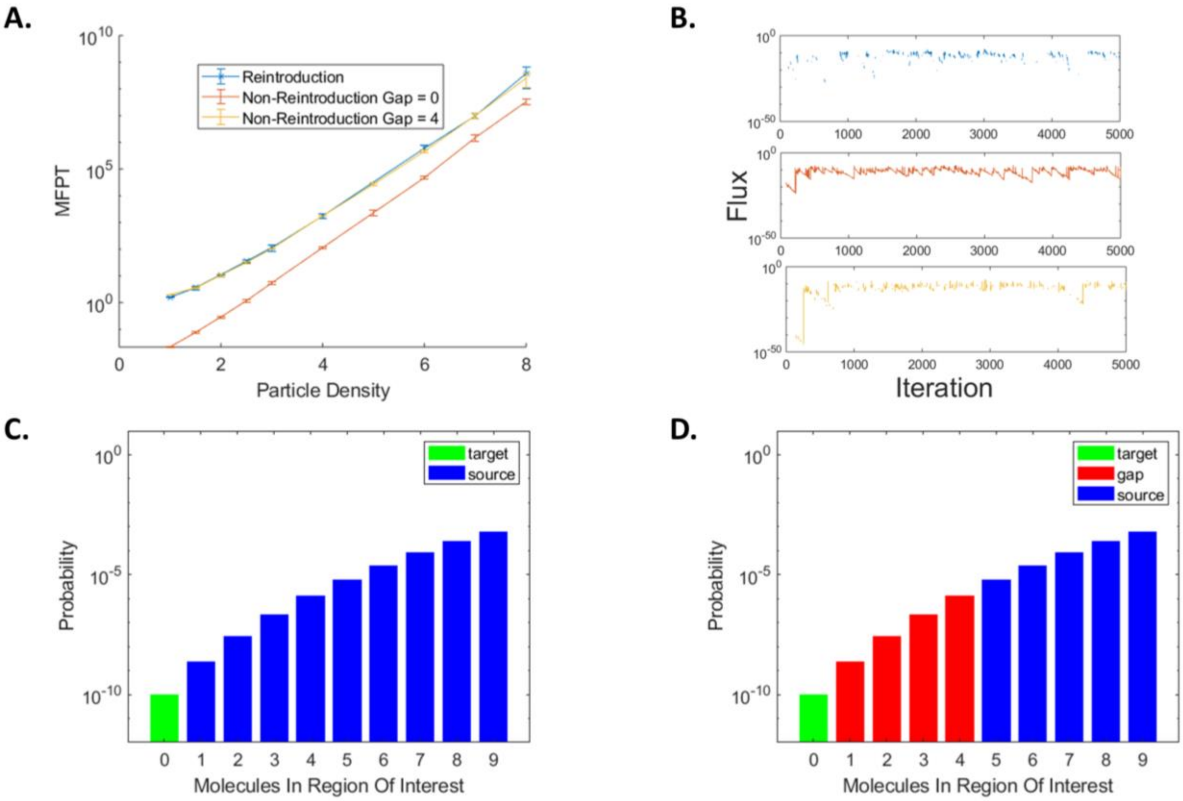


Figure 3.8 Effect of reintroduction versus non-reintroduction on flux and MFPT.

Five to twenty simulations were averaged to produce the data points and the error bars (standard deviation). **(**A) The MFPT as function of particle density $\langle\rho\rangle$ using reintroduction, non-reintroduction with a gap region of size 0 and 4. (B). At $\langle\rho\rangle = 7$, the flux is plotted as a function of the iteration using reintroduction, non-reintroduction with a gap region of size 0 and 4. (C) The visual representation of a gap region of size 0 (no gap) with bins colored by their state. (D) The visual representation of a gap region of size 4 with bins colored by their state. The following parameters were used: $\tau_{we} = 10^{-3}t^{*}$, $WE_{iter} = 10000$, $M_{targ} = 100$ and $L = 6L_0$. The binning scheme in Figure 3.4b is applied.

3.3.5 Optimizing simulation parameters for sufficient sampling becomes more challenging with increasing particle density

Increasing the particle density not only increases the rarity of the evacuated state, but also increases the difficulty of obtaining accurate estimates of the MFPT with limited sampling. The cumulative average flux vs iteration profile for different particle densities are displayed in Figure 3.10a with all other parameters held constant. During the window of simulation, $\langle\rho\rangle = 1.44$ flux quickly converges to steady state while $\langle\rho\rangle = 11.11$ is still increasing. We propose using the cumulative average flux profile as a metric for adequate sampling. There are three WE simulation parameters that can be linked to the amount of sampling, τ_{we}, $M_{targ,}$ and WE_{iter}. We study the effect of varying these simulation parameters in Figures 3.9 and 3.10.

First, τ_{we} is varied from 0.001 to 0.01 and plotted against MFPT in Fig. 3.9a. These simulations use the same total simulation time, thus shorter τ_{we} corresponds to a larger number of simulation iterations. Unlike the smaller density of $\langle\rho\rangle = 1.44$ which exhibited independence from τ_{we} in Figure 3.5 , the MFPT here (density of 11.11) increases approximately one order of magnitude over this range. In Figure 3.9b, we plot the fraction of iterations during which the first bin ($Number\ inside = 1$) was occupied with at least one replica. We focus on this first bin because of its proximity to the target state; most evacuation events occur from this bin (data not shown). We found that $\tau_{we} = 0.001\ t^*$ was small enough for replicas to occupy the first bin nearly every iteration (the time-fraction of occupancy for this first bin adjacent to the target state was $0.9981\ \pm 0.0038$). For $\tau_{we} = 0.01\ t^*$, the time-fraction of occupancy of this bin was $0.1897\ \pm 0.0278$, which is approximately a five-fold decrease compared to $\tau_{we} = 0.001\ t^*$(Figure 3.9b).

The values for M_{targ} and WE_{iter} were chosen from an initial guess that would make the simulations computationally feasible.

The second WE simulation parameter related to the amount of sampling is, M_{targ}, the target number of replicas per bin. It is shown in Figure 3.10b, that for a particle density of, $\langle\rho\rangle = 7$, the MFPT decreases in response to a larger M_{targ} (trend also shown in Figure 3.11 for $\langle\rho\rangle = 11.11$) . Importantly, these simulations share the same number of iterations and the same τ_{we}. Thus, the simulations with more replicas per bin correspond to larger total simulation time.

We also test the third parameter, $WE_{iter,}$ the total number of iterations that the algorithm executes (for equivalent M_{targ} and τ_{we}) . At lower particle densities, the MFPT estimate is shown to have little sensitivity to $WE_{iter,}$ in Figure 3.10c. At $\langle\rho\rangle = 11.11$, the MFPT is overestimated when WE_{iter} is low. But with enough iterations, multiple independent simulations seem to be in agreement.

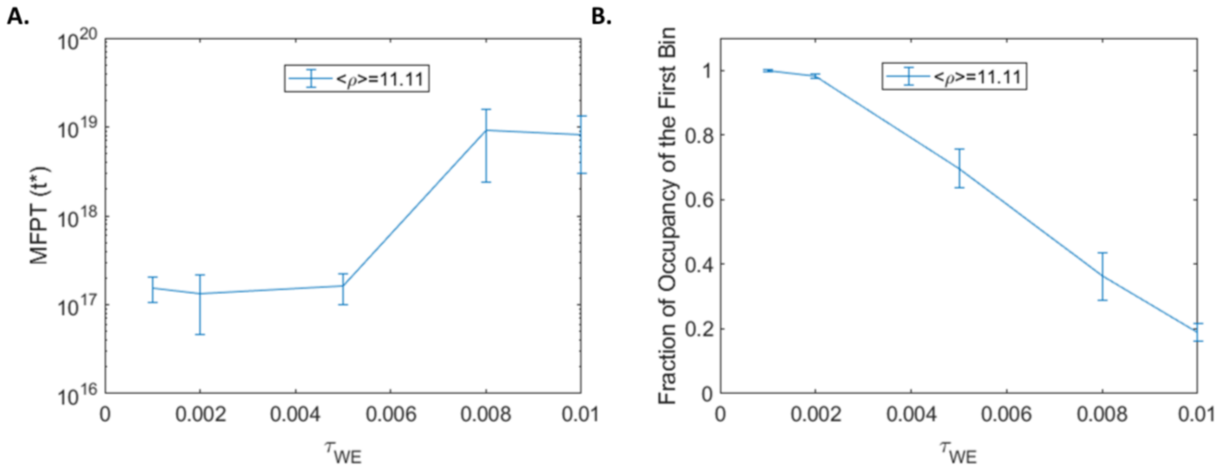


Figure 3.9 Dependence of MFPT estimate and edge-bin occupancy on τ_{we} at a high particle density of ⟨ρ⟩=11.11.

(A) The MFPT is plotted against τ_{we} for the same total simulation time. (B) The time-fraction of occupancy of the first bin adjacent to the target state is plotted against τ_{we}. (Occupancy is defined as fraction of simulation iterations in which the bin with $Number\ inside = 1$ had at least one replica). The following parameters were used: $WE_{iter} = \frac{10\ t^*}{\tau_{we}}$, $M_{targ} = 100.$ and $L = 3L_0$. The binning scheme in Figure 3.4a is applied.

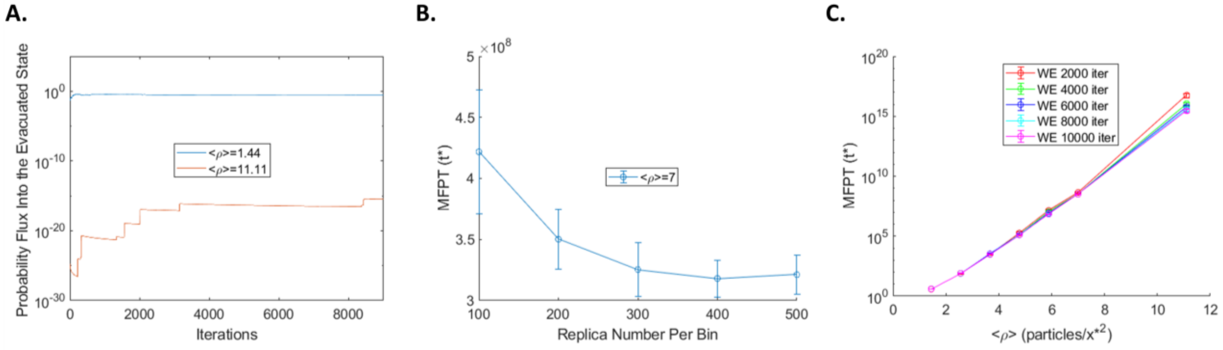


Figure 3.10 Simulation progress and dependence of MFPT estimates on simulation parameters.

(A) The cumulative average probability flux into the evacuated state is plotted against iterations. (B) The MFPT is plotted against replica number per bin. Twenty simulations were averaged to produce the data points and the error bars (standard error of the mean). (C) The MFPT is plotted against particle density for a range of iteration numbers at $M_{targ} = 500$. Twenty simulations were averaged to produce the data points and the error bars (standard error of the mean).The following parameters were used when not specified in the plot or elsewhere: $\tau_{we} = 10^{-3}t^*$, $WE_{iter} = 10000$ and $M_{targ} = 100.$ and $L = 3L_0$. The binning scheme in Figure 3.4a is applied.

3.3.6 The effect of fluctuations/missing values of the flux on the MFPT estimate

In this section, we are interested in the effect of fluctuations/missing values of the flux on the rate of convergence in WE simulations and it is explored with alternative progress coordinates and resampling procedures. A single order parameter was used in all previously shown figures using the collective variable $Number\ Inside$. However, there were sampling issues with scaling to higher particle densities, such as in Figure 3.10a. In Figure 3.11a, two additional progress coordinates are plotted for MFPT vs Replica Number per Bin. The first new progress coordinate uses two dimensions, $Number\ Inside$, the number of particles in the region of interest, and R_{avg}, the average distance from the center of all particles. The second new progress coordinate uses two dimensions, $Number\ Inside$, the number of particles in the region of interest, and R_{max}, the distance of the furthest particle within the region of interest. The intent of adding these additional variables to the progress coordinate is to capture any missing relevant processes that may facilitate flux to the target of interest. The physical intuition behind including distance as a variable is that a particle must be close to the edge of region of interest before leaving it. To compare the performance of the three order parameters, the amount of sampling is increased by the same amount for each and plotted against the MFPT. For the $Number\ Inside$ order parameter, the number of replicas per bin are increased and for the two dimensional order parameters, more bins are added along the new coordinate.

As seen in panel Figure 3.11a, the MFPT for the 2D coordinates converge at roughly the same rate or slower than the 1D coordinate when increasing the sampling. Based on these results, an additional order parameter based on distance does not seem to improve the estimate of flux when compared to the $Number\ Inside$ by itself. The choice of the bin edges for the second distance-based coordinate was based on the probability distributions from previous simulation data. Even with this previous knowledge, the 2D coordinate is more or less at the same efficiency of the 1D coordinate. The noise (fluctuations) of each simulation plotted as the standard deviation divided by the mean of the flux was plotted against the amount of sampling as well in Figure 3.11b for all three progress three coordinates. The noise of the 1D coordinate decreased the fastest with respect to more sampling. However, all three progress coordinates converged to the same MFPT values at high sampling.

The second method of testing the role of fluctuations in the MFPT convergence is changing the resampling procedure. The resampling procedure by Darve and Izaguirre [152] differs from the original formulation by Huber and Kim in that it keeps the number of replicas in each bin at exactly M_{targ} and the replica weight identical within a bin. It was noted by the authors that the statistical error is minimized in this way. In the panel Figure 3.11c, we plot the MFPT vs the replica number per bin and the convergence trend is mostly the same between the original and new resampling procedure. The noise is also plotted for each resampling procedure as a function of sampling amount. It does seem that for most sampling amounts except the initial value, the noise is smaller for the new resampling procedure. It appears that the noise in the fluxes does not seem to correlate with the convergence of the MFPT estimates with respect to sampling.

The noise shown in Figure 3.11b indicates that the $Number\ Inside$ order parameter has the least fluctuations between the three order parameters. However, this is misleading. If we plot the fluxes as a function of iteration at the Replica Number Per Bin of 500 in Figure 3.11e, we can visually see that the $Number\ Inside$ order parameter has significantly more zeroes (missing values) than $(Number\ Inside, R_{\max})$. These zeroes correspond to iterations in which no replicas entered the target (evacuated) state, and thus no flux was recorded. For the $Number\ Inside$ order parameter, at high particle densities, we believe that the high number of zeroes in the fluxes is attributed to the τ_{WE} being shorter than the time it takes for replicas to reach the target state. From section 3.3.5, it was shown that it is necessary to keep τ_{WE} this short in order to occupy the tail of the distribution. A possible explanation for the $(Number\ Inside, R_{\max})$ order parameter having fewer zeroes is that more bins are closer to the target state, thus more replicas are able to transition to the target state within τ_{WE}. Although the flux profile looks sparse for the $Number\ Inside$ order parameter, the MFPT estimates are in good agreement for all three order parameters as shown in Figure 3.11a at the Replica Number Per Bin of 500. Therefore, zeroes in the fluxes appear to be tolerable at high levels of sampling.

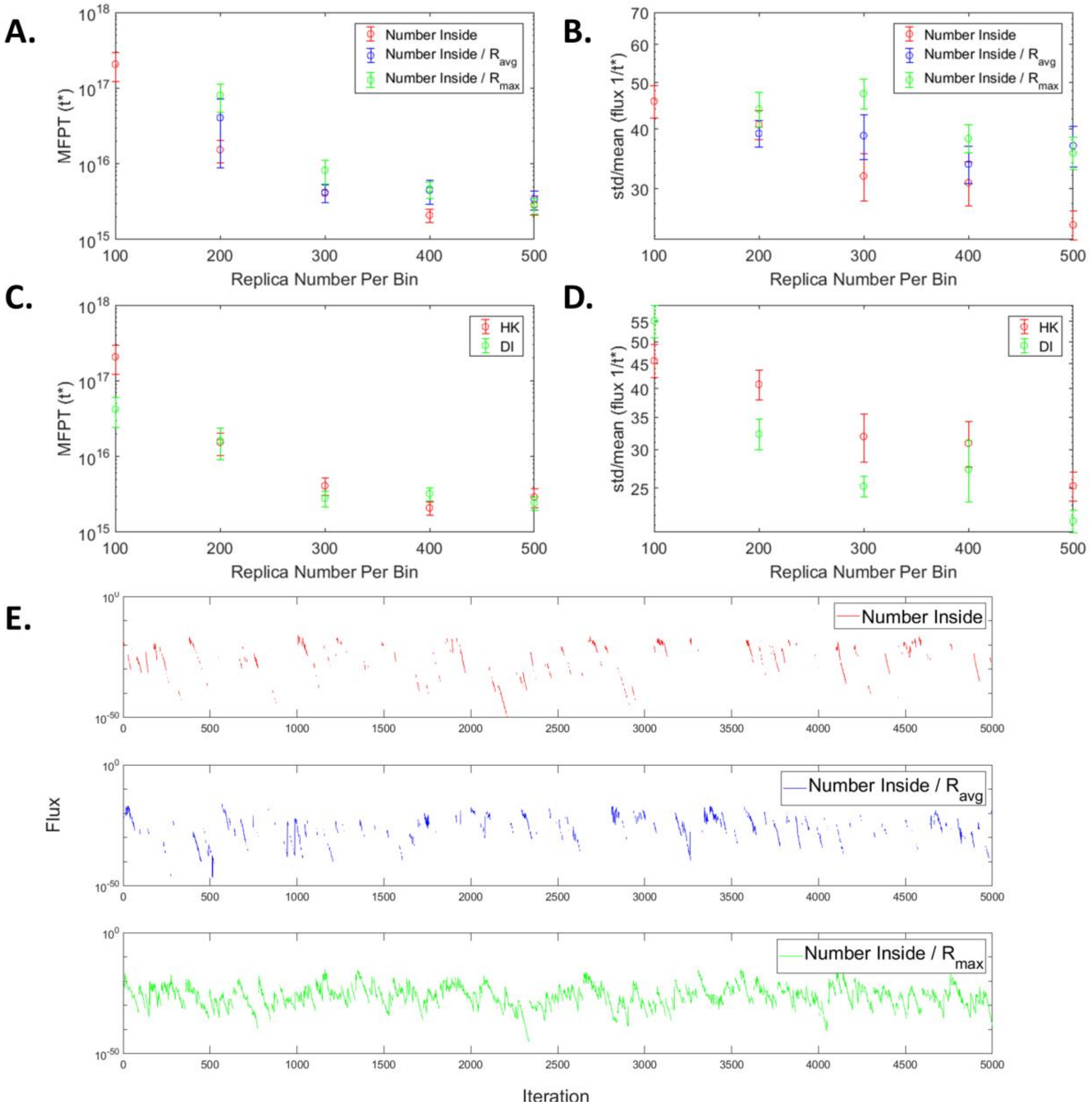


Figure 3.11 The current binning/resampling strategy is compared to alternative ones and performs more efficiently

Twenty simulations were averaged to produce the data points and the error bars (standard deviation). At higher densities, more sampling is required to reach convergence of the MFPT. The effectiveness of alternative order parameters and resampling procedures are tested. (A) The number of bins in the second dimension is increased to match the same amount of sampling at each value of replica number per bin when testing the hybrid order parameters (Number Inside, R_{avg}) and (Number Inside, R_{max}). **(**B) A measure of the noise (std/mean) is plotted for each order parameter. (C) The MFPT is plotted against the amount of sampling for the default (HK) and alternate (DI) resampling procedure (D) A measure of the noise (std/mean) is plotted for each resampling procedure. (E) At the highest amount of sampling of 500 replicas per bin, the flux as a function of iteration is plotted for the three different order parameters. The following parameters was used: $\tau_{we} = 10^{-3}t^*$ and $WE_{iter} = 10000$ and $L = 3L_0$. The binning scheme in Figure 3.4a is applied.

3.4 Discussion

The following work has characterized the MFPT behavior of the evacuation system with respect to various system parameters, Weighted Ensemble parameters, and a previous approximation. The algorithm reached a speed up of 10^4 over conventional simulation for biologically relevant densities. Various challenges were uncovered for designing the algorithm to be robust for highly rare events. Adaptations were required to accurately estimate the flow of probability of states at the long tail of the distribution.

When scaling the system such that event becomes rarer, large bursts of flux which occur infrequently *can* cause simulations to overestimate the MFPT *if* proper testing for convergence is not performed. The bursts were drastically reduced when not reintroducing particles, allowing them to freely propagate in the full state space, but the mean estimates were significantly underestimated. However, we found that adding a gap between the target and initial state fixed the underestimation issue. We speculate that the cause of the overestimation is from the short-lived quality of the target state. When the gap is not present, it is likely that frequent recrossings at the boundary of the source and target state causes the flux to be overestimated. WE parameters, such as τ_{we}, M_{targ}, and WE_{iter} when tuned to increase the sampling amount, aided in the convergence of the MFPT towards the true mean. Much of the dependence we see of MFPT estimates on simulation parameters for high particle densities in Figures 3.9, 3.10, and 3.11 may be ultimately due to inadequate sampling (lack of convergence). A potential solution for slow convergence is the reweighting procedure introduced by Bhatt et al. [42], which was not applied here, but could be included in future extensions of this work.

We demonstrated that zeroes in the fluxes (missing values) will not necessarily result in poor estimation as long as sufficient checks for convergence with respect to the WE simulation parameters are performed. Zeroes in the fluxes can be caused by different reasons and we suspect that the effect of zeroes on the estimation of MFPT will depend on the reason for the zeroes occuring. If the zeroes appear due to sparse bin occupation out at the tail of the distribution, it is likely to cause overestimation of MFPT as seen in Figure 3.9. When the bin occupation is nearly full at every iteration, the zeroes in the fluxes are still present in the simulation as shown in Figure 3.11e. In this case, the effect of the zeroes in the fluxes is probed by comparing the fluxes and MFPT of alternative order parameters. The order parameter $(\text{Number Inside}, R_{max})$ contains an equal amount of sampling as the original order parameter (Number Inside), except that more bins are added in the R_{max} dimension instead of increasing the replica count per bin. By increasing the number of bins, instead of replicas per bin, the number of zeroes in the flux significantly decrease for $(\text{Number Inside}, R_{max})$ because the extra bins in the R_{max} dimension make the replicas “closer” to the target state. In the equally sampled simulations, where (Number Inside), contained many zero fluxes while $(\text{Number Inside}, R_{max})$did not, both MFPT estimates for each order parameter were in good agreement (Figure 3.11a), regardless.

Our simulations agree with brute force at low particle density and seem to overestimate with respect to the asymptotic approximation. We show that the WE estimates may eventually agree with the asymptotic approximation by showing trends in the wall length and time step approaching the asymptotic approximation. After analyzing convergence and comparing to other approximations, the efficiency gains of WE was shown to grow at nearly the same rate as the MFPT with particle density, thereby demonstrating the effectiveness of the algorithm.

The importance of a “stop condition”, where replicas are observed at a much finer interval than τ_{we} and stopped if they reach the target state, is demonstrated when performing WE simulations for the evacuation network. Usually, after every τ_{we}, replicas are checked whether they have reached the target state. This may work in cases where the target state is long-lived, but for calculating target states out in the tail of the distribution, it results in overestimation of the MFPT. The impact of the stop condition was shown by varying the interval of observation of replicas, dt. The concept of a stop condition can be related back to other rare event algorithms such as Forward Flux Sampling [39] and Milestoning [37], which include a short observation time interval to check for whether trajectories reached the next state, in their respective versions of bins.

Our analysis of the effect of 1D and 2D order parameters has demonstrated that they converge roughly the same rate. A possible extension of the current work is to automate the binning procedure to see if the convergence can be improved over using a naïve selection of bin edges. Using an algorithm, such as the WE string method [41], may allow collective variables to be used more efficiently since it is designed to take a high dimensional space and project to a 1D representation. Though in order to use this method, adjustments would have to be made in the calculation of the flux since it relies on a finite time step and the lack of a transition region may be problematic.

Speaking of alternative algorithms, Forward Flux Sampling [39] may overcome some of the limitations of Weighted Ensemble encountered when scaling up. The core of Forward Flux Sampling is to estimate conditional probabilities along interfaces from the initial to final state. The major problem that was encountered with WE is the increasingly infrequent bursts of probability while scaling the particle density. As the particle density is increased, the distance from the evacuated state to the peak of the distribution also increases, which increases the likeliness of probability getting trapped in the probability dense regions. For Forward Flux Sampling, instead of relying on the flux to reach the target state which could take a long time to occur, correct estimation of the MFPT would instead depend on correct estimate on the adjacent bin transition probabilities. The Forward Flux Sampling algorithm on its own would need to be modified appropriately for this problem since there is no transition region which is where the bins are usually place along and the amount of sampling in each bin would have to be sufficient just like WE.

This work will be used as a base platform to simulate more complex reaction diffusion systems since passive diffusion alone is too slow for T-cell receptor ligand contact. Simulators, such as Smoldyn [142], can perform particle diffusion, but also add in realistic biological features, such as membrane geometries, molecule-membrane interactions, and reactions of individual molecules. Since the core of Weighted Ensemble is decoupled from its dynamics, it will be straightforward to swap in different simulators and models.

4. Rare Event Sampling of the Ornstein-Uhlenbeck Process

4.1 Introduction

The study of two or more cells coming into contact with each other to form an interface is a process of interest. For example, the surfaces of T cells require close contact to interact with their ligands and receptors. The formation of these contacts is affected by multiple interactions such as the presence of large surface molecules [20], [21], [137], [153] and the forces of the surrounding fluid [24], [25], [154], [155]. In this section, we are primarily interested in the role of the forces of the surrounding fluid in the formation of these contacts. In a regime dominated by thermal undulations[156], a repulsive interaction that exists between two membranes, close contact becomes a rare event which requires simulation methods, such as Weighted Ensemble Sampling [30], in order to study its behavior.

To better describe the system, the model geometry is explained in detail and visualized in a schematic. One way of modeling the problem may be to employ rigid spheres of radius r_{cell} undergoing a force F in the direction of contact with a cell to cell to distance, z, as shown in Fig. 4.1 a. It is less realistic but it does exhibit the thin layer effect [24], [26], [157], an effect which incurs a much slower time scale as z decreases. A more realistic scenario is described in panel Fig. 4.1 b. where two membranes are modeled instead. The two membranes of radius r_{cell} contain a smaller region of r_{free} where active forces F may act upon a smaller radius a. The two membranes are held at max by a far-field separation of Δz_{∞} by nonspecific adhesion molecules. The current membrane separation distance is denoted by Δz_0.

Using computational fluid dynamics, Liu et. al, [158] has simulated two cell surfaces under the effect of thermal undulations as visualized in Fig 4.1 b. Two time scales were discovered from the fit of the autocorrelation function which is consistent with the thin layer effect. The target state of the required interaction separation displacement was not reached within the simulation time. This meant that reaching target state was likely a rare event and will require a rare event algorithm in order to accurately measure the mean first passage time. Since the complex fluid dynamic simulations are too expensive to access those times, the problem is cast into a much more computationally feasible stochastic model called an Ornstein-Ulhenbeck (OU) process, a stochastic process that has a Gaussian stationary density with a constant mean and variance. In this chapter, we will explore the behavior of the mean first passage time as a function of the separation displacement of a single membrane or two membranes from its mean position. As a control, a single membrane was simulated using the simplification of a one-dimensional OU process. While, the two membrane interface case was modeled as a two-dimensional OU process. The single membrane case was validated with a theoretical solution, but to our knowledge no theoretical solution is available for the interface. Problems were encountered in the interface case when scaling to higher distances due to the slower time scale dominating. Two different approaches are applied tested to improve the estimation of the mean first passage times in at the long tail of the distribution in the interface case.

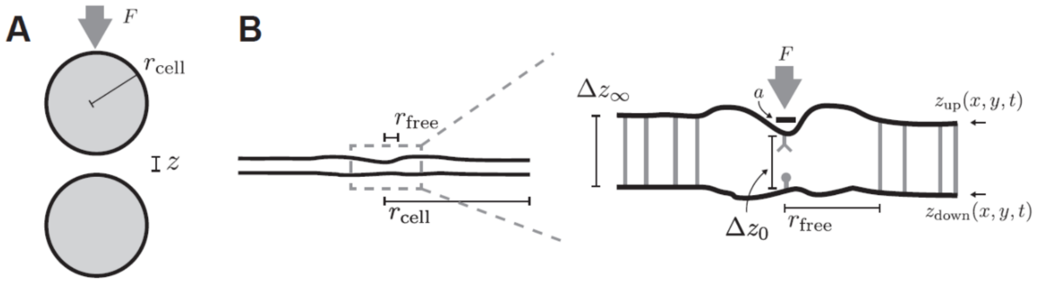


Figure 4.1 Schematic of Close Contact Event in Simulation Setting

(A) Two cells depicted as spheres pushed together by a force F. (B) Assuming the radius of the cell is much larger than the cell-cell separation distance, the cells are modeled by a top and bottom membrane with a force F applied near the ligand and receptor. [158]

4.2 Methods

4.2.1 Ornstein-Ulhenbeck Model

Data acquired from computational fluid dynamics simulations was used to derive the autocorrelation function and stationary distribution. The mean membrane separation, was measured to be, $\langle z_0 \rangle = 70\ nm$. The current membrane separation is denoted as $\Delta {z_0}^*$. The contact displacement, $Z^* = \langle z_0 \rangle - \Delta {z_0}^*$, can be approximated by a 1D Ornstein-Ulenbeck process for the single membrane case:

$$dZ = -\frac{1}{\tau} Z dt + \frac{\sigma}{\sqrt{\tau}} dW \tag{3.1}$$

The parameters σ, the standard deviation, and τ, the timescale, were estimated from the simulation data. W is a wiener process.

The interface case is approximated by a 2D Ornstein-Ulenbeck process:

$$dX = -\frac{1}{\tau_{slow}} X dt + \frac{\sigma}{\sqrt{\tau_{slow}}} dW_1 \tag{3.2}$$

$$dY = -\frac{1}{\tau_{fast}} Y dt + \frac{\sigma}{\sqrt{\tau_{fast}}} dW_1 \tag{3.3}$$

$$Z^* = cX + \sqrt{1 - c^2} Y \tag{3.4}$$

where c, is a parameter which determines the fraction attributed to the slow (τ_{slow}) and fast (τ_{fast}) processes. A trajectory and histogram of the composite process is plotted in Figure 4.2.

The trajectory fluctuates around the mean of zero and a Gaussian distribution is fit to the simulation data.

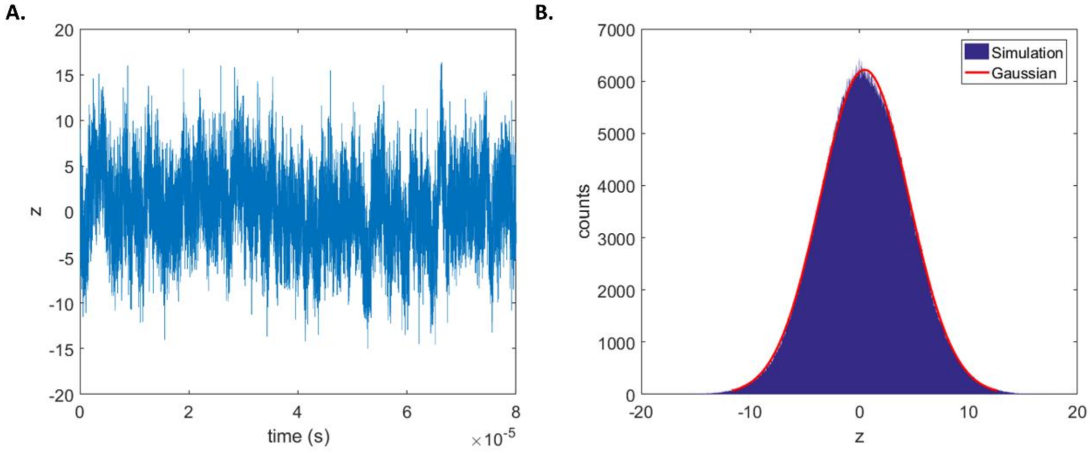


Figure 4.2 Brute force simulation of interface using 2D OU model

(A) The simulation trajectory of z, the OU process variable as function of time in seconds (s).
(B) The histogram of the simulation trajectory in blue and the Gaussian fit as a red outline.

4.2.2 Weighted Ensemble Parameter Design

The Weighted Ensemble (WE) algorithm is applied to solve for the mean first passage times to reach a target Z^*. A previous detailed explanation of WE can be found in chapter 2. Choice of parameters are impactful when applying WE to a new problem. Some of the reasoning and decision making when deciding WE parameters is described in this section. An arbitrary, but reasonable choice, of M_{targ}, was initially chosen (32 for the 1D case and 200 for the 2D case).

The bin edges for replicas were chosen such that sampling is focused in the direction of the target state. The OU process is a stochastic process that fluctuates around zero, but the target state we are interested is in the positive direction. Hence, all negative values of Z^{*}, were assigned to one bin. For the positive values, initially every integer value of Z^{*} was assigned to its own bin. However, we found that the flux contained too many zeros since they were not enough bins, so we ended up using every $0.1\ Z^{*}$ as the bin edges. Since the target state is out in the tail of the distribution, the τ_{we} (1×10^{-8}), was set to be sufficiently small such that the bins at the edges were fully occupied at each iteration. Also, a stop condition to check for the target state at a smaller interval,dt (see Chapter 2) was applied to obtain a more accurate flux due to the transient nature of the target state.

4.3 Results

4.3.1 The mean first passage time to first contact grows much slower in the interface case compared to the single membrane

A 1D OU process was used to model the displacement of the single membrane and the 2D OU process modeled the interface case of two membranes. The progress coordinate was represented by Z^{*}, the target contact displacement, which measures how close the two membranes are coming into contact (the larger the value, the closer the membranes are). The MFPT of the 1D OU process is validated by a theoretical approximation developed in [159]. In Figure 4.3, it is shown that the interface scales much faster than the single membrane case; the time to reach $Z^{*} = 20\ nm$ for the interface case is approximately 500-fold faster than the single membrane case. The critical displacement for binding is around $Z^{*} = 57\ nm$ and based on the super exponential growth, the MFPT to reach that displacement would not be realistic biologically (on the order of seconds). Therefore, other interactions, such as active forces and membrane permeability, are needed to be modeled to accurately model the close contact event.

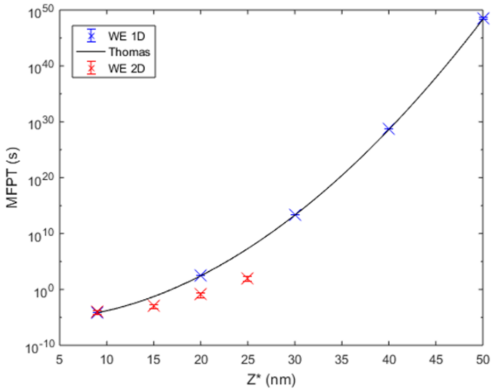


Figure 4.3 1D and 2D prediction of Mean First Passage Times as function of Contact Displacements

Twenty simulations were averaged to produce the data points and the error bars (standard deviation). The mean first passage time (in seconds) is plotted as a function of the contact displacement for two different cases: the 1D single membrane case is plotted in blue along with the theoretical approximation by Thomas et. al and the 2D interface case is plotted in red. The following parameters were used for the 1D case M_{targ} = 32, $\tau_{WE} = 1 \times 10^{-8}$, and $\langle WE_{iter} \rangle = 10{,}000$. A bin edge was assigned at every 0.1 Z^* above. The following WE parameters were used for the 2D case M_{targ} = 200, $\tau_{WE} = 1 \times 10^{-8}$, and $\langle WE_{iter} \rangle = 100{,}000$. A bin edge was assigned at every 0.3 Z^* up to 15 and 0.1 Z^* above. The 2D OU model parameters were: $\tau_{fast} = 8.18 \times 10^{-5}$, $\tau_{slow} = 5.22 \times 10^{-7}$, $\sigma_{\text{fast}} = \sigma_{\text{slow}} = 4.27$ nm. The 1D OU model parameters were: $\tau = 1.05 \times 10^{-6}$ and $\sigma = 3.1385\, nm$

4.3.2 Scaling to higher contact displacement when estimating the mean first passage times becomes challenging in the interface case

4.3.2.1 The composite process of two processes is more difficult to sample than the single pure processes

If we wish to actually measure the MFPT at higher contact displacements, it becomes a computational issue due to the presence of two processes with different timescales. This is exemplified by Figure 4.4 a and 4.4 b where sampling is much better at $Z^* = 9\ nm$ when compared to the $Z^* = 45\ nm$ case. The large spread at intermediate values of c for $Z^* = 45\ nm$ can be attributed to the larger fluctuations of the raw flux at higher Z^* seen in panel Figure 4.4 c and 4.4 d. It is worthy to note that the pure processes can be sampled adequately as shown by their agreement with the theoretical approximations at $c = 0$ and $c = 1$ since only the pure process is present in each.

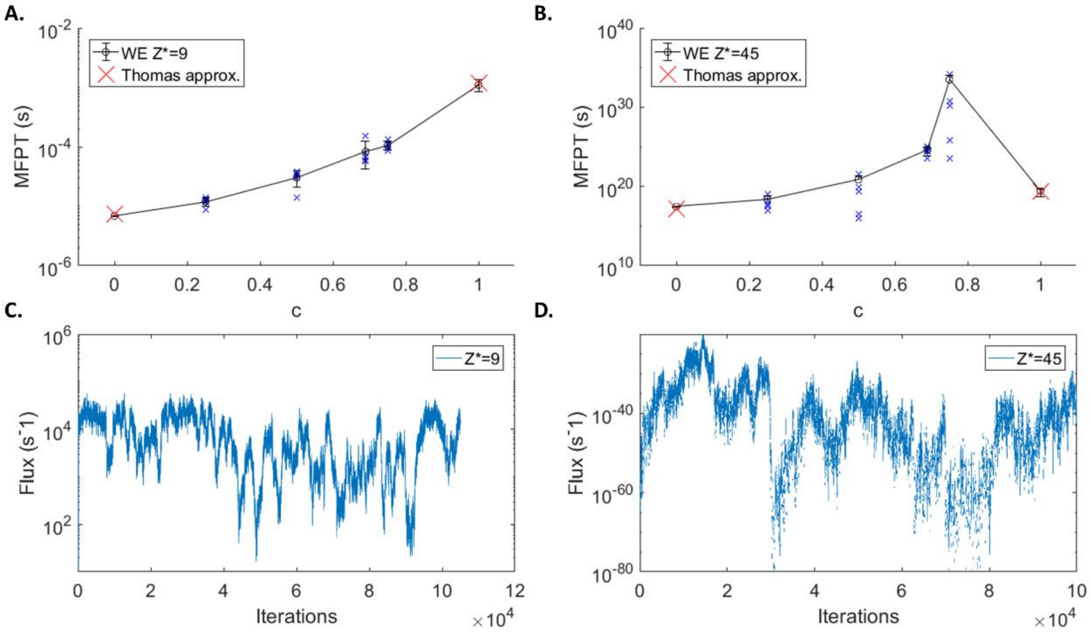


Figure 4.4 Scaling to higher contact displacements is difficult in the 2D case

Ten simulations were averaged to produce the data points and the error bars (standard deviation). (A) The mean first passage time (in seconds) is plotted as a function of the contact displacement for the interface case to reach target state $Z^* = 9$. The Thomas approximation is plotted as validation when there is only one process present. (B) The mean first passage time (in seconds) is plotted as a function of the contact displacement for the interface case to reach target state $Z^* = 45$ and $c = 0.75$. The Thomas approximation is plotted as validation when there is only one process present. (C) The flux, the probability entering the target state per unit time, is plotted as a function of the simulation iteration at $Z^* = 9$. (D) The flux, the probability entering the target state per unit time, is plotted as a function of the simulation iteration at $Z^* = 45$ and $c = 0.75$. The following parameters were used M_{targ} = 200, $\tau_{WE} = 1 \times 10^{-8}$, and$\langle WE_{iter} \rangle = 100{,}000$. A bin edge was assigned at every 0.3 Z^* up to 15 and 0.1 Z^* above. The 2D OU model parameters were: $\tau_{fast} = 8.18 \ \times 10^{-5}$, $\tau_{slow} = 5.22 \ \times 10^{-7}$, $\sigma_{\text{fast}} = \sigma_{\text{slow}}$ = 4.27 nm.

4.3.2.2 The 1D order parameter does not adequately sample the target of interest in the presence of two processes

When using a 1D order parameter for two processes, it is simple and directly corresponds to the observable of interest, however any important events occurring in the individual processes may not be sufficiently sampled. For example, both processes are capable of being in the negative Z^* states while the other is positive which hinders the progress towards reaching the target positive Z^* state. This explains the larger fluctuations in the flux as Z^* is increased further due to the time one or both components spend in the negative Z^* states instead of progressing towards the target state. The sampling especially worsens as the slow component dominates at higher values of c. This makes sense because the slow component will more spend more lengths of time in the negative displacement values and making it difficult for the two processes to “coordinate” (both be positive) to reach the target state.

4.3.3 Using a 2D order parameter which separates the two independent processes results in improved estimation of the mean first passage times

Instead of directly using Z^* as the 1D order parameter, the two components are used as a 2D order parameter. The 1D order and 2D order parameter were used to estimate the MFPT at $Z^* = 25\ nm$ using the same simulation and model parameters excluding the replica count per bin and the spacing between bin centers, which was decreased and increased, respectively, in the 2D case due to the exponential cost of expanding to two dimensions in Figure 3.5. The average replica count and average number of iterations was roughly 15,000 replicas and 70,000 iterations for the 1D case and 10,000 replicas and 40,000 iterations for the 2D case. Even with less sampling than the 1D case, the 2D case is able to produce relatively stable fluxes and MFPT with a lesser spread in the data at larger values of c. An important factor to the increased stability of the 2D order parameter was likely due to the increased emphasis on the positive Z^*values in the placement of the bin centers in both dimensions which is not present in the 1D case. In the 1D case, each component was allowed to freely sample the positive and negative values of Z^*. However, with the 2D parameter, only one bin center was allocated to the negative value of Z^* and the rest were assigned to the positive values.

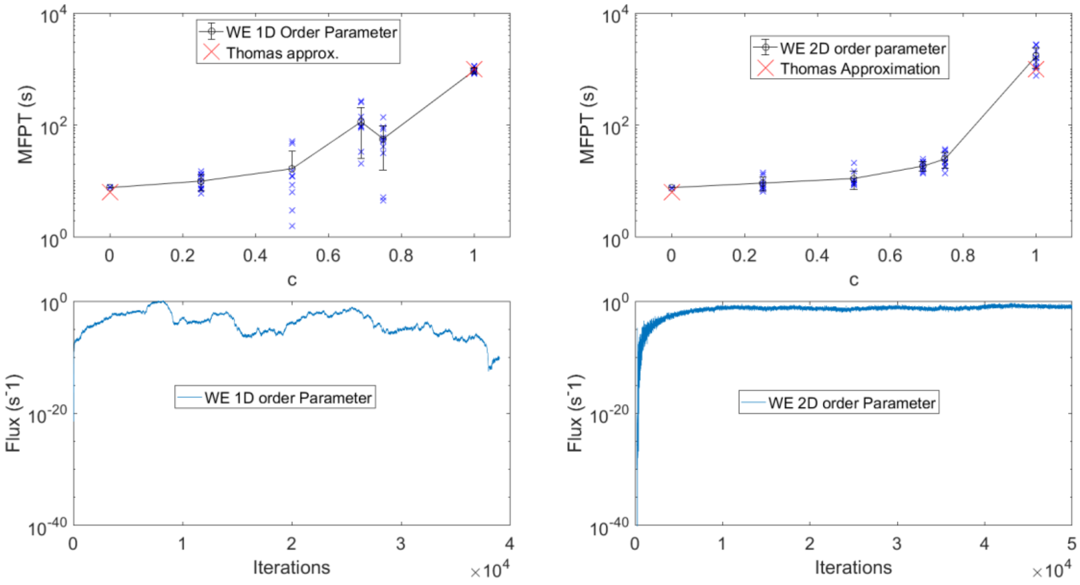


Figure 4.5 2D order parameter improves estimation by separating the independent processes

Ten simulations were averaged to produce the data points and the error bars (standard deviation). (A) Using a WE 1D order parameter, The mean first passage time (in seconds) is plotted as a function of the contact displacement for the interface case to reach target state $Z^* = 25$. The Thomas approximation is plotted as validation when there is only one process present. (B) Using a WE 2D order parameter, The mean first passage time (in seconds) is plotted as a function of the contact displacement for the interface case to reach target state $Z^* = 25$. (C) Using a WE 1D order parameter, the flux, the probability entering the target state per unit time, is plotted as a function of the simulation iteration at $Z^* = 9$. (D) Using a WE 2D order parameter, the flux, the probability entering the target state per unit time, is plotted as a function of the simulation iteration at $Z^* = 45$ and $c = 0.75$. The following parameters were used for the 1D order parameter M_{targ} = 200, $\tau_{WE} = 1 \times 10^{-8}$, and $\langle WE_{iter} \rangle = 70{,}000$. A bin edge was assigned at every 0.3 Z^* up to 15 and 0.1 Z^* above. The following parameters were used for the 2D order parameter M_{targ} = 32, $\tau_{WE} = 1 \times 10^{-8}$, and $\langle WE_{iter} \rangle = 40{,}000$. A bin edge was assigned at every 1 Z^*. The 2D OU model parameters were: $\tau_{fast} = 8.18 \times 10^{-5}$, $\tau_{slow} = 5.22 \times 10^{-7}$, $\sigma_{\text{fast}} = \sigma_{\text{slow}} = 4.27$ nm.

4.4 Discussion

The current study has applied WE to discern the behavior of mean first passage time as a function of the separation between two membranes with the condition of only thermal undulations. Based on the growth trend, the mean first passage time exceeds the biological expected timescale for critical binding in T-cell membranes. The single membrane case was used as validation for the method since a theoretical approximation existed for the 1D OU process [159]. WE was able to match the theoretical values even at large displacements where the event was extremely rare. A challenge arose when WE was applied to the interface case at large displacements. The more the slow process dominated, the more unstable flux became and there would be several large outliers, despite having more sampling than in the single membrane case. The issue was shown to be amended by including a 2D order parameter, one for each component of the 2D OU process. The fluxes and MFPTs were considerably more stable and narrowly distributed. It is likely due to the increased consistency of crossings due to explicit binning of both variables and also the emphasis on the positive values of Z^* which contribute more to the flux.

This case study is a prime example of where automation of the binning of collective variables would be beneficial to obtain more accurate estimations of the MFPT. Just by increasing from one dimension to two dimensions made it much harder to obtain accurate estimates since the flux became much more unstable. Strategies for higher dimensional binning have been suggested previously. For example, Zhang et. al [160] suggested a successive binning strategy in higher dimensions, where an initial binning is decided in one dimension and more bins in the second coordinate is added based on the configurations in the first coordinate. One can propose a strategy where influential transition paths are maximized when designing the order parameter. It is difficult to discern the optimal order since there can be multiple degrees of freedom which are relevant. If there are multiple important degrees of freedom, a possible strategy may to be systematically build the order parameter by starting out with only degree of freedom as the order parameter. When a crossing of significant probability flux occurs, one could then backtrack to subdivide along the other important degrees of freedom. To prevent adding too many bins, a probability cutoff could be used to determine if the weight of the trajectory was high enough to consider subdividing based on the current flux. To test the effectiveness, a simple diagnostic can be based on the stability of the flux after adding more bins in the new dimensions. Of course, this strategy would still require knowledge of the relevant progress coordinates to the transition to work effectively in improving the quality of the flux.

5. Adpative Binning for High Dimensional Order Parameter in Weighted Sampling

5.1 Introduction

In the context of transition path sampling [38], the order parameter is a collective set of variables which can distinguish the initial and final states of a metastable system. It is an approximation of the reaction coordinate which is the true variable that represents the mechanism driving the transition between the final and initial state. The quality of approximation of the order parameter can be evaluated based on its committor function, which is the probability for a state to make to the final state before the initial state. First, the separatix, the saddle point where the free energy is highest is identified. According to transition path theory, this is where the committor function should be at 0.5 since the transition state lies at the separatix. For a landscape with two metastable states, an order parameter that adequately represents the reaction coordinate will possess a narrowly peaked probability distribution for the committor function at the separatix centered on 0.5. The choice of order parameter is a challenge as complex systems can have an immense degrees of freedom where the mechanism is not fully understood. Another issue with the order parameter is when the number of collective variables grow, the number of possible bins grows exponentially. Hence a method is needed to keep the number of bins at a computationally manageable size while also preserving adequate sampling of the transitions.

A method that has been developed previously called the string method [161] can project a high dimensional space into a smaller one that connects the initial and final states and thus keeps the number of bins manageable. It has been tested on primarily metastable systems, but it is unknown whether it is viable for systems where the final state is transient. In this chapter, an algorithm based on forward transitions is suggested to control the number of bins as the dimensions grow with increasing collective variables.

5.2 An Algorithm Sketch

A strategy to reduce the number of bins while adequately sampling the transitions is to incrementally define new progress coordinates as simulation data is gathered on the forward transitions. A forward transition is defined when a trajectory is closer to the target state than it was previously. By gathering statistics on these trajectories, we are able to sample on the important areas of the state space.

The first step in defining the bins is to decide on the initial progress coordinate. This step is no different from the original weighted ensemble algorithm. The bin edges should be placed in the transition region and the choice of coordinate should easily distinguish between the initial and final state.

To define subsequent bin edges for additional coordinates, we can make use of simulation data of the forward transitions. At each iteration of the algorithm, the trajectories are stored if it travels along a forward path. After gathering enough statistics within each bin such that the running mean of the new coordinate is steady, new edges on the next coordinate can be defined around the mean value of the coordinate in each bin. The spacing of these bins can be kept similar to the first coordinate. The advantage defining coordinates in a sequential manner is that it is more than efficient than using a grid based bins since that will lead to exponential growth when scaling to more variables.

A simple diagnostic to judge the effectiveness of the algorithm is to track the flux with the initial coordinate and observe the steadiness of the flux. If flux is unsteady, additional bins can be added based on the already gathered simulation data. With the new coordinate, if the new bins were effective then the flux should be steadier around its mean value. Note that this approach is mainly focused on deciding the edges of the bins, it is still up to the user to know or decide what the actual coordinates should be.

Bibliography

[1] H. H. McAdams and A. Arkin, “It’s a noisy business! Genetic regulation at the nanomolar scale,” *Trends Genet. TIG*, vol. 15, no. 2, pp. 65–69, Feb. 1999.

[2] N. Komin and A. Skupin, “How to address cellular heterogeneity by distribution biology,” *Curr. Opin. Syst. Biol.*, vol. 3, pp. 154–160, Jun. 2017.

[3] S. S. Andrews, “Serial rebinding of ligands to clustered receptors as exemplified by bacterial chemotaxis,” *Phys. Biol.*, vol. 2, no. 2, pp. 111–122, Jun. 2005.

[4] J. B. Gurdon and P. Y. Bourillot, “Morphogen gradient interpretation,” *Nature*, vol. 413, no. 6858, pp. 797–803, Oct. 2001.

[5] “Division accuracy in a stochastic model of Min oscillations in Escherichia coli | PNAS.” [Online]. Available: http://www.pnas.org/content/103/2/347. [Accessed: 08-Oct-2018].

[6] A. Grosse-Wilde *et al.*, “Stemness of the hybrid Epithelial/Mesenchymal State in Breast Cancer and Its Association with Poor Survival,” *PLOS ONE*, vol. 10, no. 5, p. e0126522, May 2015.

[7] Z. Tan *et al.*, “Aneuploidy underlies a multicellular phenotypic switch,” *Proc. Natl. Acad. Sci.*, vol. 110, no. 30, pp. 12367–12372, Jul. 2013.

[8] M. Mojtahedi *et al.*, “Cell Fate Decision as High-Dimensional Critical State Transition,” *PLOS Biol.*, vol. 14, no. 12, p. e2000640, Dec. 2016.

[9] A. Mogilner, R. Wollman, and W. F. Marshall, “Quantitative Modeling in Cell Biology: What Is It Good for?,” *Dev. Cell*, vol. 11, no. 3, pp. 279–287, Sep. 2006.

[10] A. D. Lander, “The edges of understanding,” *BMC Biol.*, vol. 8, no. 1, p. 40, Apr. 2010.

[11] E. N. Olson, “Gene regulatory networks in the evolution and development of the heart,” *Science*, vol. 313, no. 5795, pp. 1922–1927, Sep. 2006.

[12] C. Espinosa-Soto, P. Padilla-Longoria, and E. R. Alvarez-Buylla, “A Gene Regulatory Network Model for Cell-Fate Determination during Arabidopsis thaliana Flower Development That Is Robust and Recovers Experimental Gene Expression Profiles,” *Plant Cell*, vol. 16, no. 11, pp. 2923–2939, Nov. 2004.

[13] E. H. Davidson, “Emerging properties of animal gene regulatory networks,” *Nature*, vol. 468, no. 7326, pp. 911–920, Dec. 2010.

[14] S. A. Kauffman, “Control Circuits for Determination and Transdetermination,” *Science*, vol. 181, no. 4097, pp. 310–318, Jul. 1973.

[15] N. G. van Kampen, *Stochastic processes in physics and chemistry*. Amsterdam; Boston; London: Elsevier, 2007.

[16] V. Likhoshvai and A. Ratushny, “Generalized hill function method for modeling molecular processes,” *J. Bioinform. Comput. Biol.*, vol. 5, no. 2B, pp. 521–531, Apr. 2007.

[17] E. Abranches *et al.*, “Stochastic NANOG fluctuations allow mouse embryonic stem cells to explore pluripotency,” *Development*, vol. 141, no. 14, pp. 2770–2779, Jul. 2014.

[18] A. Arkin, J. Ross, and H. McAdams, “Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected Escherichia coli cells,” *Genetics*, vol. 149, no. 000075232100002, pp. 1633–1648, Dec. 1997.

[19] A. Raj and A. van Oudenaarden, “Nature, Nurture, or Chance: Stochastic Gene Expression and Its Consequences,” *Cell*, vol. 135, no. 2, pp. 216–226, Oct. 2008.

[20] J. F. Allard, O. Dushek, D. Coombs, and P. A. van der Merwe, “Mechanical Modulation of Receptor-Ligand Interactions at Cell-Cell Interfaces,” *Biophys. J.*, vol. 102, no. 6, pp. 1265–1273, Mar. 2012.

[21] V. T. Chang *et al.*, “Initiation of T cell signaling by CD45 segregation at ‘close contacts,’” *Nat. Immunol.*, vol. 17, no. 5, pp. 574–582, Mar. 2016.

[22] G. Furlan, T. Minowa, N. Hanagata, C. Kataoka-Hamai, and Y. Kaizuka, “Phosphatase CD45 Both Positively and Negatively Regulates T Cell Receptor Phosphorylation in Reconstituted Membrane Protein Clusters,” *J. Biol. Chem.*, vol. 289, no. 41, pp. 28514–28525, Oct. 2014.

[23] S. J. Davis and P. A. van der Merwe, “The kinetic-segregation model: TCR triggering and beyond,” *Nat. Immunol.*, vol. 7, no. 8, pp. 803–809, Aug. 2006.

[24] Y. Kaizuka and J. T. Groves, “Hydrodynamic damping of membrane thermal fluctuations near surfaces imaged by fluorescence interference microscopy,” *Phys. Rev. Lett.*, vol. 96, no. 11, p. 118101, Mar. 2006.

[25] L. C.-L. Lin, J. T. Groves, and F. L. H. Brown, “Analysis of shape, fluctuations, and dynamics in intermembrane junctions,” *Biophys. J.*, vol. 91, no. 10, pp. 3600–3606, Nov. 2006.

[26] A. Vázquez-Quesada and M. Ellero, “Analytical solution for the lubrication force between two spheres in a bi-viscous fluid,” *Phys. Fluids*, vol. 28, no. 7, p. 073101, Jul. 2016.

[27] J.-E. Dietrich and T. Hiiragi, “Stochastic patterning in the mouse pre-implantation embryo,” *Dev. Camb. Engl.*, vol. 134, no. 23, pp. 4219–4231, Dec. 2007.

[28] T. Kalmar *et al.*, “Regulated Fluctuations in Nanog Expression Mediate Cell Fate Decisions in Embryonic Stem Cells,” *PLoS Biol.*, vol. 7, no. 7, p. e1000149, Jul. 2009.

[29] M. J. Tse, B. K. Chu, C. P. Gallivan, and E. L. Read, “Rare-event sampling of epigenetic landscapes and phenotype transitions,” *PLOS Comput. Biol.*, vol. 14, no. 8, p. e1006336, Aug. 2018.

[30] G. A. Huber and S. Kim, “Weighted-ensemble Brownian dynamics simulations for protein association reactions,” *Biophys. J.*, vol. 70, no. 1, pp. 97–110, Jan. 1996.

[31] M. C. Zwier, J. W. Kaus, and L. T. Chong, “Efficient Explicit-Solvent Molecular Dynamics Simulations of Molecular Association Kinetics: Methane/Methane, Na+/Cl−, Methane/Benzene, and K+/18-Crown-6 Ether,” *J. Chem. Theory Comput.*, vol. 7, no. 4, pp. 1189–1197, Apr. 2011.

[32] J. L. Adelman *et al.*, “Simulations of the Alternating Access Mechanism of the Sodium Symporter Mhp1,” *Biophys. J.*, vol. 101, no. 10, pp. 2399–2407, Nov. 2011.

[33] M. C. Zwier, A. J. Pratt, J. L. Adelman, J. W. Kaus, D. M. Zuckerman, and L. T. Chong, “Efficient Atomistic Simulation of Pathways and Calculation of Rate Constants for a Protein–Peptide Binding Process: Application to the MDM2 Protein and an Intrinsically Disordered p53 Peptide,” *J. Phys. Chem. Lett.*, vol. 7, no. 17, pp. 3440–3445, Sep. 2016.

[34] D. Bhatt and D. M. Zuckerman, “Heterogeneous path ensembles for conformational transitions in semi–atomistic models of adenylate kinase,” *J. Chem. Theory Comput.*, vol. 6, no. 11, pp. 3527–3539, Oct. 2010.

[35] R. M. Donovan, A. J. Sedgewick, J. R. Faeder, and D. M. Zuckerman, “Efficient Stochastic Simulation of Chemical Kinetics Networks using a Weighted Ensemble of Trajectories,” *J. Chem. Phys.*, vol. 139, no. 11, p. 115105, Sep. 2013.

[36] “Unbiased Rare Event Sampling in Spatial Stochastic Systems Biology Models Using a Weighted Ensemble of Trajectories.” [Online]. Available: http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004611. [Accessed: 19-Mar-2018].

[37] A. K. Faradjian and R. Elber, “Computing time scales from reaction coordinates by milestoning,” *J. Chem. Phys.*, vol. 120, no. 23, pp. 10880–10889, May 2004.

[38] P. G. Bolhuis, D. Chandler, C. Dellago, and P. L. Geissler, “TRANSITION PATH SAMPLING: Throwing Ropes Over Rough Mountain Passes, in the Dark,” *Annu. Rev. Phys. Chem.*, vol. 53, no. 1, pp. 291–318, 2002.

[39] R. J. Allen, C. Valeriani, and P. Rein ten Wolde, “Forward flux sampling for rare event simulations,” *J. Phys. Condens. Matter*, vol. 21, no. 46, p. 463102, Nov. 2009.

[40] D. Bhatt and I. Bahar, “An adaptive weighted ensemble procedure for efficient computation of free energies and first passage rates,” *J. Chem. Phys.*, vol. 137, no. 10, p. 104101, Sep. 2012.

[41] J. L. Adelman and M. Grabe, “Simulating rare events using a weighted ensemble-based string method,” *J. Chem. Phys.*, vol. 138, no. 4, p. 044105, 2013.

[42] D. Bhatt, B. W. Zhang, and D. M. Zuckerman, “Steady-state simulations using weighted ensemble path sampling,” *J. Chem. Phys.*, vol. 133, no. 1, Jul. 2010.

[43] A. Dickson and C. L. Brooks, “WExplore: Hierarchical Exploration of High-Dimensional Spaces Using the Weighted Ensemble Algorithm,” *J. Phys. Chem. B*, vol. 118, no. 13, pp. 3532–3542, Apr. 2014.

[44] E. Suárez *et al.*, “Simultaneous Computation of Dynamical and Equilibrium Information Using a Weighted Ensemble of Trajectories,” *J. Chem. Theory Comput.*, vol. 10, no. 7, pp. 2658–2667, Jul. 2014.

[45] S.-H. Ahn, J. W. Grate, and E. F. Darve, “Efficiently sampling conformations and pathways using the concurrent adaptive sampling (CAS) algorithm,” *J. Chem. Phys.*, vol. 147, no. 7, p. 074115, Aug. 2017.

[46] M. C. Zwier *et al.*, “WESTPA: An Interoperable, Highly Scalable Software Package for Weighted Ensemble Simulation and Analysis,” *J. Chem. Theory Comput.*, vol. 11, no. 2, pp. 800–809, Feb. 2015.

[47] E. Suárez, A. J. Pratt, L. T. Chong, and D. M. Zuckerman, “Estimating first-passage time distributions from weighted ensemble simulations and non-Markovian analyses,” *Protein Sci. Publ. Protein Soc.*, vol. 25, no. 1, pp. 67–78, Jan. 2016.

[48] D. M. Zuckerman and L. T. Chong, “Weighted Ensemble Simulation: Review of Methodology, Applications, and Software,” *Annu. Rev. Biophys.*, vol. 46, no. 1, pp. 43–57, 2017.

[49] W. Xiong and J. E. Ferrell, “A positive-feedback-based bistable ‘memory module’ that governs a cell fate decision,” *Nature*, vol. 426, no. 6965, pp. 460–465, Nov. 2003.

[50] J. Zhou and S. Huang, “Understanding gene circuits at cell-fate branch points for rational cell reprogramming,” *Trends Genet. TIG*, vol. 27, no. 21146896, pp. 55–62, Feb. 2011.

[51] M. Lu, M. K. Jolly, R. Gomoto, B. Huang, J. Onuchic, and E. Ben-Jacob, “Tristability in Cancer-Associated MicroRNA-TF Chimera Toggle Switch,” *J. Phys. Chem. B*, vol. 117, no. 42, pp. 13164–13174, Oct. 2013.

[52] H. Feng and J. Wang, “A new mechanism of stem cell differentiation through slow binding/unbinding of regulators to genes,” *Sci. Rep.*, vol. 2, no. 22870379, p. 550, Jan. 2012.

[53] B. Zhang and P. Wolynes, “Stem cell differentiation as a many-body problem,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 111, no. 24946805, pp. 10185–90, Jul. 2014.

[54] P. Wang, C. Song, H. Zhang, Z. Wu, X.-J. Tian, and J. Xing, “Epigenetic state network approach for describing cell phenotypic transitions,” *Interface Focus*, vol. 4, no. 3, p. 20130068, Jun. 2014.

[55] T. Hong, J. Xing, L. Li, and J. Tyson, “A mathematical model for the reciprocal differentiation of T helper 17 cells and induced regulatory T cells,” *PLoS Comput. Biol.*, vol. 7, no. 21829337, p. e1002122, Jul. 2011.

[56] T. Graf and T. Enver, “Forcing cells to change lineages,” *Nature*, vol. 462, no. 19956253, pp. 587–94, Dec. 2009.

[57] S. Huang, “The molecular and mathematical basis of Waddington’s epigenetic landscape: a framework for post-Darwinian biology?,” *BioEssays News Rev. Mol. Cell. Dev. Biol.*, vol. 34, no. 22102361, pp. 149–57, Feb. 2012.

[58] A. H. Lang, H. Li, J. J. Collins, and P. Mehta, “Epigenetic Landscapes Explain Partially Reprogrammed Cells and Identify Key Reprogramming Genes,” *PLoS Comput. Biol.*, vol. 10, no. 8, p. e1003734, Aug. 2014.

[59] M. B. Elowitz, “Stochastic Gene Expression in a Single Cell,” *Science*, vol. 297, no. 5584, pp. 1183–1186, Aug. 2002.

[60] E. M. Ozbudak, M. Thattai, I. Kurtser, A. D. Grossman, and A. van Oudenaarden, “Regulation of noise in the expression of a single gene,” *Nat. Genet.*, vol. 31, no. 1, pp. 69–73, May 2002.

[61] I. Golding, J. Paulsson, S. M. Zawilski, and E. C. Cox, “Real-Time Kinetics of Gene Activity in Individual Bacteria,” *Cell*, vol. 123, no. 6, pp. 1025–1036, Dec. 2005.

[62] N. Balaban, J. Merrin, R. Chait, L. Kowalik, and S. Leibler, “Bacterial persistence as a phenotypic switch,” *Science*, vol. 305, no. 15308767, pp. 1622–5, Sep. 2004.

[63] M. Acar, J. T. Mettetal, and A. van Oudenaarden, “Stochastic switching as a survival strategy in fluctuating environments,” *Nat. Genet.*, vol. 40, no. 4, pp. 471–475, Apr. 2008.

[64] S. Sharma *et al.*, “A chromatin-mediated reversible drug-tolerant state in cancer cell subpopulations,” *Cell*, vol. 141, no. 20371346, pp. 69–80, Apr. 2010.

[65] H. H. Chang, M. Hemberg, M. Barahona, D. E. Ingber, and S. Huang, “Transcriptome-wide noise controls lineage choice in mammalian progenitor cells,” *Nature*, vol. 453, no. 7194, pp. 544–547, May 2008.

[66] L. Yuan *et al.*, “A role of stochastic phenotype switching in generating mosaic endothelial cell heterogeneity,” *Nat. Commun.*, vol. 7, p. 10160, 2016.

[67] C. Waddington and H. Kacser, *The Strategy of the Genes*. Routledge, 1957.

[68] J. Wang, K. Zhang, L. Xu, and E. Wang, “Quantifying the Waddington landscape and biological paths for development and differentiation,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 108, no. 21536909, pp. 8257–62, May 2011.

[69] G. Karlebach and R. Shamir, “Modelling and analysis of gene regulatory networks,” *Nat. Rev. Mol. Cell Biol.*, vol. 9, no. 10, pp. 770–780, Oct. 2008.

[70] T. Kepler and T. Elston, “Stochasticity in transcriptional regulation: Origins, consequences, and mathematical representations,” *Biophys. J.*, vol. 81, no. 000172407800010, pp. 3116–3136, Dec. 2000.

[71] D. GILLESPIE, “EXACT STOCHASTIC SIMULATION OF COUPLED CHEMICAL-REACTIONS,” *J. Phys. Chem.*, vol. 81, no. A1977EE49800008, pp. 2340–2361, Dec. 1976.

[72] B. Munsky and M. Khammash, “The finite state projection algorithm for the solution of the chemical master equation,” *J. Chem. Phys.*, vol. 124, no. 16460146, p. 044104, Jan. 2006.

[73] Y. Cao and J. Liang, “Optimal enumeration of state space of finitely buffered stochastic molecular networks and exact computation of steady state landscape probability,” *BMC Syst. Biol.*, vol. 2, no. 1, p. 30, 2008.

[74] V. Wolf, R. Goel, M. Mateescu, and T. A. Henzinger, “Solving the chemical master equation using sliding windows,” *BMC Syst. Biol.*, vol. 4, no. 1, p. 42, 2010.

[75] C. D. Pahlajani, P. J. Atzberger, and M. Khammash, “Stochastic reduction method for biological chemical kinetics using time-scale separation,” *J. Theor. Biol.*, vol. 272, no. 1, pp. 96–112, Mar. 2011.

[76] R. B. Sidje and H. D. Vo, “Solving the chemical master equation by a fast adaptive finite state projection based on the stochastic simulation algorithm,” *Math. Biosci.*, vol. 269, pp. 10–16, Nov. 2015.

[77] S. Huang, Y.-P. Guo, G. May, and T. Enver, “Bifurcation dynamics in lineage-commitment in bipotent progenitor cells,” *Dev. Biol.*, vol. 305, no. 2, pp. 695–713, May 2007.

[78] R. Ma, J. Wang, Z. Hou, and H. Liu, “Small-number effects: a third stable state in a genetic bistable toggle switch,” *Phys. Rev. Lett.*, vol. 109, no. 23368390, p. 248107, Dec. 2012.

[79] Y. Cao, H.-M. Lu, and J. Liang, “Probability landscape of heritable and robust epigenetic state of lysogeny in phage lambda,” *Proc. Natl. Acad. Sci.*, vol. 107, no. 43, pp. 18445–18450, Oct. 2010.

[80] B. Munsky, Z. Fox, and G. Neuert, “Integrating single-molecule experiments and discrete stochastic models to understand heterogeneous gene transcription dynamics,” *Methods*, vol. 85, pp. 12–21, Sep. 2015.

[81] C. Schütte, A. Fischer, W. Huisinga, and P. Deuflhard, “A Direct Approach to Conformational Dynamics Based on Hybrid Monte Carlo,” *J. Comput. Phys.*, vol. 151, no. 1, pp. 146–168, May 1999.

[82] W. C. Swope, J. W. Pitera, and F. Suits, “Describing Protein Folding Kinetics by Molecular Dynamics Simulations. 1. Theory †,” *J. Phys. Chem. B*, vol. 108, no. 21, pp. 6571–6581, May 2004.

[83] D. S. Chekmarev, T. Ishida, and R. M. Levy, “Long-Time Conformational Transitions of Alanine Dipeptide in Aqueous Solution: Continuous and Discrete-State Kinetic Models,” *J. Phys. Chem. B*, vol. 108, no. 50, pp. 19487–19495, Dec. 2004.

[84] G. R. Bowman, X. Huang, and V. S. Pande, “Network models for molecular kinetics and their initial applications to human health,” *Cell Res.*, vol. 20, no. 6, pp. 622–630, Jun. 2010.

[85] P. Érdi and J. Tóth, *Mathematical models of chemical reactions: theory and applications of deterministic and stochastic models*. Princeton, N.J: Princeton University Press, 1989.

[86] B. Munsky and M. Khammash, “The finite state projection algorithm for the solution of the chemical master equation,” *J. Chem. Phys.*, vol. 124, no. 4, p. 044104, 2006.

[87] K. R. Sanft, S. Wu, M. Roh, J. Fu, R. K. Lim, and L. R. Petzold, “StochKit2: software for discrete stochastic simulation of biochemical systems with events,” *Bioinformatics*, vol. 27, no. 17, pp. 2457–2458, Sep. 2011.

[88] *MATLAB Release 2015a.* Natick, Massachusetts, United States: The MathWorks, Inc.

[89] M. K. Scherer *et al.*, “PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models,” *J. Chem. Theory Comput.*, vol. 11, no. 11, pp. 5525–5542, Nov. 2015.

[90] J.-H. Prinz *et al.*, “Markov models of molecular kinetics: Generation and validation,” *J. Chem. Phys.*, vol. 134, no. 17, p. 174105, 2011.

[91] J. D. Chodera and F. Noé, “Markov state models of biomolecular conformational dynamics,” *Curr. Opin. Struct. Biol.*, vol. 25, pp. 135–144, Apr. 2014.

[92] P. Deuflhard, W. Huisinga, A. Fischer, and C. Schütte, “Identification of almost invariant aggregates in reversible nearly uncoupled Markov chains,” *Linear Algebra Its Appl.*, vol. 315, no. 1–3, pp. 39–59, Aug. 2000.

[93] S. Röblitz and M. Weber, “Fuzzy spectral clustering by PCCA+: application to Markov state models and data classification,” *Adv. Data Anal. Classif.*, vol. 7, no. 2, pp. 147–179, Jun. 2013.

[94] P. Deuflhard, W. Huisinga, A. Fischer, and C. Schütte, “Identification of almost invariant aggregates in reversible nearly uncoupled Markov chains,” *Linear Algebra Its Appl.*, vol. 315, no. 1–3, pp. 39–59, Aug. 2000.

[95] P. Deuflhard and M. Weber, “Robust Perron cluster analysis in conformation dynamics,” *Linear Algebra Its Appl.*, vol. 398, pp. 161–184, Mar. 2005.

[96] N.-V. Buchete and G. Hummer, “Coarse Master Equations for Peptide Folding Dynamics †,” *J. Phys. Chem. B*, vol. 112, no. 19, pp. 6057–6069, May 2008.

[97] K. A. Beauchamp, G. R. Bowman, T. J. Lane, L. Maibaum, I. S. Haque, and V. S. Pande, “MSMBuilder2: Modeling Conformational Dynamics on the Picosecond to Millisecond Scale,” *J. Chem. Theory Comput.*, vol. 7, no. 10, pp. 3412–3419, Oct. 2011.

[98] S. Kube and M. Weber, “A coarse graining method for the identification of transition rates between molecular conformations,” *J. Chem. Phys.*, vol. 126, no. 2, p. 024103, 2007.

[99] W. E. and E. Vanden-Eijnden, “Towards a Theory of Transition Paths,” *J. Stat. Phys.*, vol. 123, no. 3, pp. 503–523, May 2006.

[100] P. Metzner, C. Schütte, and E. Vanden-Eijnden, “Transition Path Theory for Markov Jump Processes,” *Multiscale Model. Simul.*, vol. 7, no. 3, pp. 1192–1219, Jan. 2009.

[101] F. Noe, C. Schutte, E. Vanden-Eijnden, L. Reich, and T. R. Weikl, “Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations,” *Proc. Natl. Acad. Sci.*, vol. 106, no. 45, pp. 19011–19016, Nov. 2009.

[102] D. Schultz, A. M. Walczak, J. N. Onuchic, and P. G. Wolynes, “Extinction and resurrection in gene networks,” *Proc. Natl. Acad. Sci.*, vol. 105, no. 49, pp. 19165–19170, Dec. 2008.

[103] H. Feng and J. Wang, “A new mechanism of stem cell differentiation through slow binding/unbinding of regulators to genes,” *Sci. Rep.*, vol. 2, Aug. 2012.

[104] M. J. Morelli, S. Tănase-Nicola, R. J. Allen, and P. R. ten Wolde, “Reaction Coordinates for the Flipping of Genetic Switches,” *Biophys. J.*, vol. 94, no. 9, pp. 3413–3423, May 2008.

[105] S. Huang, “Reprogramming cell fates: reconciling rarity with robustness,” *BioEssays News Rev. Mol. Cell. Dev. Biol.*, vol. 31, no. 19319911, pp. 546–60, May 2009.

[106] S. Huang, “Hybrid T-Helper Cells: Stabilizing the Moderate Center in a Polarized System,” *PLoS Biol.*, vol. 11, no. 8, p. e1001632, Aug. 2013.

[107] V. S. Pande, K. Beauchamp, and G. R. Bowman, “Everything you wanted to know about Markov State Models but were afraid to ask,” *Methods*, vol. 52, no. 1, pp. 99–105, Sep. 2010.

[108] Gardner, Timothy, S. Charles, R. Cantor, and James J. Collins, “Construction of a genetic toggle switch in Escherichia coli,” *Nature*, vol. 403, no. 6767, pp. 339–342, 2000.

[109] A. Lipshtat, A. Loinger, N. Balaban, and O. Biham, “Genetic toggle switch without cooperative binding,” *Phys. Rev. Lett.*, vol. 96, no. 16712399, p. 188101, May 2006.

[110] D. Schultz, A. M. Walczak, J. N. Onuchic, and P. G. Wolynes, “Extinction and resurrection in gene networks,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 105, no. 49, pp. 19165–19170, Dec. 2008.

[111] T. J. Lane, G. R. Bowman, K. Beauchamp, V. A. Voelz, and V. S. Pande, “Markov State Model Reveals Folding and Functional Dynamics in Ultra-Long MD Trajectories,” *J. Am. Chem. Soc.*, vol. 133, no. 45, pp. 18413–18419, Nov. 2011.

[112] M. J. Tse, B. K. Chu, M. Roy, and E. L. Read, “DNA-Binding Kinetics Determines the Mechanism of Noise-Induced Switching in Gene Networks,” *Biophys. J.*, vol. 109, no. 8, pp. 1746–1757, Oct. 2015.

[113] W. E. and E. Vanden-Eijnden, “Towards a Theory of Transition Paths,” *J. Stat. Phys.*, vol. 123, no. 3, pp. 503–523, May 2006.

[114] P. Metzner, C. Schütte, and E. Vanden-Eijnden, “Transition Path Theory for Markov Jump Processes,” *Multiscale Model. Simul.*, vol. 7, no. 3, pp. 1192–1219, Jan. 2009.

[115] A. Berezhkovskii, G. Hummer, and A. Szabo, “Reactive flux and folding pathways in network models of coarse-grained protein dynamics,” *J. Chem. Phys.*, vol. 130, no. 20, p. 205102, 2009.

[116] A. M. Walczak, J. N. Onuchic, and P. G. Wolynes, “Absolute rate theories of epigenetic stability,” *Proc. Natl. Acad. Sci.*, vol. 102, no. 52, pp. 18926–18931, Dec. 2005.

[117] J. Wang, K. Zhang, and E. Wang, “Kinetic paths, time scale, and underlying landscapes: A path integral framework to study global natures of nonequilibrium systems and networks,” *J. Chem. Phys.*, vol. 133, no. 12, p. 125103, 2010.

[118] I. Chambers *et al.*, “Nanog safeguards pluripotency and mediates germline development,” *Nature*, vol. 450, no. 7173, pp. 1230–1234, Dec. 2007.

[119] J. D. Chodera, N. Singhal, V. S. Pande, K. A. Dill, and W. C. Swope, “Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics,” *J. Chem. Phys.*, vol. 126, no. 15, p. 155101, 2007.

[120] G. R. Bowman, K. A. Beauchamp, G. Boxer, and V. S. Pande, “Progress and challenges in the automated construction of Markov state models for full protein systems,” *J. Chem. Phys.*, vol. 131, no. 12, p. 124101, 2009.

[121] G. Pérez-Hernández, F. Paul, T. Giorgino, G. De Fabritiis, and F. Noé, “Identification of slow molecular order parameters for Markov model construction,” *J. Chem. Phys.*, vol. 139, no. 1, p. 015102, 2013.

[122] Burrage, Kevin, Hegland, M., Macnamara, Shev, & Sidje, Roger, “A Krylov-based finite state projection algorithm for solving the chemical master equation arising in the discrete modelling of biological systems,” in *Proceedings of the Markov 150th Anniversary Conference*, 2006.

[123] Y. Cao, A. Terebus, and J. Liang, “Accurate Chemical Master Equation Solution Using Multi-Finite Buffers,” *Multiscale Model. Simul.*, vol. 14, no. 2, pp. 923–963, Jan. 2016.

[124] M. Hegland, C. Burden, L. Santoso, S. MacNamara, and H. Booth, “A solver for the stochastic master equation applied to gene regulatory networks,” *J. Comput. Appl. Math.*, vol. 205, no. 2, pp. 708–724, Aug. 2007.

[125] E. L. Haseltine and J. B. Rawlings, “Approximate simulation of coupled fast and slow reactions for stochastic chemical kinetics,” *J. Chem. Phys.*, vol. 117, no. 15, p. 6959, 2002.

[126] S. Peleš, B. Munsky, and M. Khammash, “Reduction and solution of the chemical master equation using time scale separation and finite state projection,” *J. Chem. Phys.*, vol. 125, no. 20, p. 204104, 2006.

[127] Y. Kuroda, A. Suenaga, Y. Sato, S. Kosuda, and M. Taiji, “All-atom molecular dynamics analysis of multi-peptide systems reproduces peptide solubility in line with experimental observations,” *Sci. Rep.*, vol. 6, p. 19479, Jan. 2016.

[128] G. Jayachandran, V. Vishal, and V. S. Pande, “Using massively parallel simulation and Markovian models to study protein folding: Examining the dynamics of the villin headpiece,” *J. Chem. Phys.*, vol. 124, no. 16, p. 164902, 2006.

[129] N. Singhal, C. D. Snow, and V. S. Pande, “Using path sampling to build better Markovian state models: Predicting the folding rate and mechanism of a tryptophan zipper beta hairpin,” *J. Chem. Phys.*, vol. 121, no. 1, p. 415, 2004.

[130] J. J. Tapia, J. R. Faeder, and B. Munsky, “Adaptive coarse-graining for transient and quasi-equilibrium analyses of stochastic gene regulation,” 2012, pp. 5361–5366.

[131] B. W. Zhang, D. Jasnow, and D. M. Zuckerman, “The ‘weighted ensemble’ path sampling method is statistically exact for a broad class of stochastic processes and binning procedures,” *J. Chem. Phys.*, vol. 132, no. 5, p. 054107, 2010.

[132] Weber, Marcus, and K. Fackeldey, “G-pcca: Spectral clustering for non-reversible markov chains,” *ZIB Rep.*, vol. 15, Jul. 2015.

[133] D. Schultz, J. N. Onuchic, and P. G. Wolynes, “Understanding stochastic simulations of the smallest genetic networks,” *J. Chem. Phys.*, vol. 126, no. 24, p. 245102, 2007.

[134] N. J. Burroughs, Z. Lazic, and P. A. van der Merwe, “Ligand Detection and Discrimination by Spatial Relocalization: A Kinase-Phosphatase Segregation Model of TCR Activation,” *Biophys. J.*, vol. 91, no. 5, pp. 1619–1629, Sep. 2006.

[135] T. R. Weikl and R. Lipowsky, “Pattern Formation during T-Cell Adhesion,” *Biophys. J.*, vol. 87, no. 6, pp. 3665–3678, Dec. 2004.

[136] “Synaptic pattern formation during cellular recognition | PNAS.” [Online]. Available: http://www.pnas.org/content/98/12/6548. [Accessed: 12-Aug-2018].

[137] E. M. Schmid *et al.*, “Size-dependent protein segregation at membrane interfaces,” *Nat. Phys.*, vol. 12, no. 7, pp. 704–711, Jul. 2016.

[138] J. Newby and J. Allard, “First-Passage Time to Clear the Way for Receptor-Ligand Binding in a Crowded Environment,” *Phys. Rev. Lett.*, vol. 116, no. 12, Mar. 2016.

[139] N. Le Novère and T. S. Shimizu, “STOCHSIM: modelling of stochastic biomolecular processes,” *Bioinforma. Oxf. Engl.*, vol. 17, no. 6, pp. 575–576, Jun. 2001.

[140] J. S. van Zon and P. R. ten Wolde, “Green’s-function reaction dynamics: a particle-based approach for simulating biochemical networks in time and space,” *J. Chem. Phys.*, vol. 123, no. 23, p. 234910, Dec. 2005.

[141] J. Hattne, D. Fange, and J. Elf, “Stochastic reaction-diffusion simulation with MesoRD,” *Bioinforma. Oxf. Engl.*, vol. 21, no. 12, pp. 2923–2924, Jun. 2005.

[142] S. S. Andrews, N. J. Addy, R. Brent, and A. P. Arkin, “Detailed Simulations of Cell Biology with Smoldyn 2.1,” *PLoS Comput. Biol.*, vol. 6, no. 3, p. e1000705, Mar. 2010.

[143] C. W. Gardiner, K. J. McNeil, D. F. Walls, and I. S. Matheson, “Correlations in stochastic theories of chemical reactions,” *J. Stat. Phys.*, vol. 14, no. 4, pp. 307–331, Apr. 1976.

[144] “Zur kinetischen Theorie der Brownschen Molekularbewegung und der Suspensionen - von Smoluchowski - 1906 - Annalen der Physik - Wiley Online Library.” [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/andp.19063261405. [Accessed: 12-Aug-2018].

[145] R. M. Donovan *et al.*, “Unbiased Rare Event Sampling in Spatial Stochastic Systems Biology Models Using a Weighted Ensemble of Trajectories,” *PLOS Comput. Biol.*, vol. 12, no. 2, p. e1004611, Feb. 2016.

[146] P. Deuflhard, M. Dellnitz, O. Junge, and C. Schütte, “Computation of Essential Molecular Dynamics by Subdivision Techniques,” in *Computational Molecular Dynamics: Challenges, Methods, Ideas*, Springer, Berlin, Heidelberg, 1999, pp. 98–115.

[147] C. Schütte and W. Huisinga, “Biomolecular conformations can be identified as metastable sets of molecular dynamics,” in *Handbook of Numerical Analysis*, vol. 10, Elsevier, 2003, pp. 699–744.

[148] R. J. Allen, P. B. Warren, and P. R. Ten Wolde, “Sampling rare switching events in biochemical networks,” *Phys. Rev. Lett.*, vol. 94, no. 1, p. 018104, Jan. 2005.

[149] R. J. Allen, D. Frenkel, and P. R. ten Wolde, “Simulating rare events in equilibrium or nonequilibrium stochastic systems,” *J. Chem. Phys.*, vol. 124, no. 2, p. 024102, Jan. 2006.

[150] C. W. Cairo *et al.*, “Dynamic Regulation of CD45 Lateral Mobility by the Spectrin-Ankyrin Cytoskeleton of T Cells,” *J. Biol. Chem.*, vol. 285, no. 15, pp. 11392–11401, Apr. 2010.

[151] V. Rajani, G. Carrero, D. E. Golan, G. de Vries, and C. W. Cairo, “Analysis of Molecular Diffusion by First-Passage Time Variance Identifies the Size of Confinement Zones,” *Biophys. J.*, vol. 100, no. 6, pp. 1463–1472, Mar. 2011.

[152] T. Schlick, *Innovations in Biomolecular Modeling and Simulations*. Royal Society of Chemistry, 2012.

[153] O. Yakovian *et al.*, “Gp41 dynamically interacts with the TCR in the immune synapse and promotes early T cell activation,” *Sci. Rep.*, vol. 8, no. 1, p. 9747, Jun. 2018.

[154] A. Carlson and L. Mahadevan, “Elastohydrodynamics and Kinetics of Protein Patterning in the Immunological Synapse,” *PLoS Comput. Biol.*, vol. 11, no. 12, p. e1004481, Dec. 2015.

[155] M. Mani, A. Gopinath, and L. Mahadevan, “How Things Get Stuck: Kinetics, Elastohydrodynamics, and Soft Adhesion,” *Phys. Rev. Lett.*, vol. 108, no. 22, p. 226104, May 2012.

[156] W. Helfrich, “Steric Interaction of Fluid Membranes in Multilayer Systems,” *Z. Für Naturforschung A*, vol. 33, no. 3, pp. 305–315, 2014.

[157] G. Barnocky and R. H. Davis, “The lubrication force between spherical drops, bubbles and rigid particles in a viscous fluid,” *Int. J. Multiph. Flow*, vol. 15, no. 4, pp. 627–638, Jul. 1989.

[158] Kai Liu, Brian Chu, Jay Newby, Elizabeth Read, John Lowengrub, and Jun Allard, “Hydrodynamics of transient cell-cell contact: The role of membrane permeability and active protrusion length,” *Rev.*

[159] M. U. Thomas, “Some mean first-passage time approximations for the Ornstein-Uhlenbeck process,” *J. Appl. Probab.*, vol. 12, no. 3, pp. 600–604, Sep. 1975.

[160] B. W. Zhang, D. Jasnow, and D. M. Zuckerman, “Efficient and verified simulation of a path ensemble for conformational change in a united-residue model of calmodulin,” *Proc. Natl. Acad. Sci.*, vol. 104, no. 46, pp. 18043–18048, Nov. 2007.

[161] “String method in collective variables: minimum free energy paths and isocommittor surfaces. - PubMed - NCBI.” [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/16848576. [Accessed: 02-Nov-2018].