Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Catalina Rivera Date

Inferring phenomenological models of biological systems

By

Catalina Rivera
Doctor of Philosophy

Physics

Ilya Nemenman, Ph.D.
Advisor

Gordon Berman, Ph.D.
Committee Member

Stefan Boettcher, Ph.D.
Committee Member

Anatoly B. Kolomeisky, Ph.D.
Committee Member

Daniel Weissman, Ph.D.
Committee Member

Accepted:

Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

Date

Inferring phenomenological models of biological systems

By

Catalina Rivera
B.A., Los Andes University, Colombia, 2011
M.Sc., Los Andes University, Colombia, 2013

Advisor: Ilya Nemenman, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Physics
2020

Abstract

Inferring phenomenological models of biological systems

By Catalina Rivera

Biological systems and processes are complex. They are governed by a large number of units interacting with each other in many different ways and on many different time scales. Thus, constructing mechanistically accurate models capable of explaining the emergent macroscopic behavior of these system and making non-trivial predictions based on such models is often infeasible. To alleviate this problem, new approaches are needed to infer functional, phenomenological models of biological systems directly from data. Here, we develop and apply tools capable of doing this, using statistical inference to automatically construct phenomenological models for different types of biological processes across multiple spatio-temporal scales. Our approach is enabled by the increases in computational power and the development of statistical inference and machine learning methods on the one hand, and the high-throughput biological experimental techniques on the other, which we have experienced over the last decades. First, we focus on inferring phenomenological models for studying systems that can be seen as First Passage Processes—where a relevant or observable biological phenomenon happens after a certain state is achieved in a series of stochastic transitions among internal states of the system. We develop a phenomenological approach where simple models get *refined* (rather than complex models getting coarse-grained), adding progressively more details until the experimental first passage time distribution is well approximated. In this way, our approach avoids the pitfalls of having to build a detailed, mechanistically accurate model first, in route to phenomenological, functional understanding. In particular, we infer models explaining Purkinje Cells (a type of a neuron) spike generation and single-enzyme turnover times distribution. We show that our inferred models allow us to uncover minimal constraints on more mechanistically accurate models of the involved phenomena. Our second set of phenomenological model inference tools revolves around developing a mathematical framework to study animal behavioral evolution. In particular, we study the behavioral evolution of six closely related species of fruit flies. We show that, by reconstructing ancestral behavioral repertoires, a very simple model describing the stochasticity in behavioral evolution lets us infer the nature of the intra- vs inter-species variability. Our approach provides a new framework to study behavioral evolution and to develop an understanding of its genetic basis. All of the phenomenological inference approaches proposed in this Dissertation show the potential of using statistical inference tools to help us achieve physics-level model-based understanding of various functions of complex biological systems.

Inferring phenomenological models of biological systems

By

Catalina Rivera
B.A., Los Andes University, Colombia, 2011
M.Sc., Los Andes University, Colombia, 2013

Advisor: Ilya Nemenman, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Physics
2020

Acknowledgments

I would like to thank my advisor Dr. Ilya Nemenman, for giving me the opportunity to come to the U. S. and study for a Ph. D. These years have been of constant learning, not only academically but also personally. I thank him for significantly increasing my motivation to learn science, for his mentorship, for being very patient with my slow rate of processing new information, and for being incredibly supportive in the most critical periods as a Ph. D. student.

I thank Dr. Gordon Berman for his guidance and for his willingness to share his knowledge in the best possible way. I thank him for patiently answering all my questions and for always having a contagious positive attitude towards life.

I also want to thank the members of my committee, Dr. Anatoly Kolomeisky for collaborating with me during the last period of my Ph. D. studies and for the interesting discussions, Dr. Daniel Weissman and Dr. Stefan Boettcher, for their engaging questions and contributions to my work.

I thank Dr. Damián G. Hernández, Dr. David Hofmann, and Dr. Baohua Zhou for not just collaborating with me, but most importantly for becoming my friends.

I thank all the people (and those who at some point were) in the Emory Physics department, especially Dr. Joe Natale, Dr. Xianxian Shao, Dr. Audrey Sederberg, Dr. Michael Martini, Ahmed Roman, Arabind Swain, and other members of the Nemenman Lab, for making it a very enjoyable learning place.

I want to thank Prof. Bernardo Recamán for sharing with me his passion for Math.

I thank all my friends in Atlanta especially Maria José Ramirez, Johanna Rodriguez, Juliana Soto, Andres Caballero, Oliver Giraldo, and Gabriela Artasanchez for becoming my family in the U.S and making these years much more enjoyable. I also thank my boyfriend Emmanuel Donate for all his support during the last period of this process, especially during Coronavirus times.

I want to thank, with all my heart, my mother Gloria Isabel Mesias and my

brother Alejandro Rivera for being my biggest source of love, support and happiness during my entire life.

Finally, I want to dedicate this Dissertation to the memory of my father Luciano Rivera because remembering him brings me peace.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Many questions in analysis of biological systems fall within two categories. On the one hand, there are those aiming to figure out *what* the system does. On the other are those investigating *how* the system does what it does. Take for example neural activity. We may ask: *What* does the neuron do? In other words, what is the time distribution characterizing the generation of a neural action potential (aka, a spike)? What kind of measurements are appropriate to describe the process (e.g., is measuring the membrane potential of a neuron useful)? What is the effect of external perturbations on the statistical properties of the process? etc. Alternatively, we could ask: *how* does the neuron do all of this? To answer this latter question, we would need to build a mechanistic model of the neuron at a molecular or cellular scale, which would reproduce the emergence of a spike through a fast depolarization of the membrane potential.

At first, it may seem that building a mechanistic model (i.e., by partitioning the system into smaller biologically meaningful subsystems, properly interacting to produce the macroscopically measured behavior) should answer all the *what* questions as well. Since once a detailed mechanistic model is built we should be able find the relevant features that matter at a macroscopical scale. In traditional physical systems,

symmetries and constraints on the interactions among the subsystems are known. This allows the successful application of the Renormalization Group (RG) theory [152, 153, 70]. RG theory constructs a microscopic model of the system and coarse grains it into an effective macroscopic model, identifying the features or model parameters that remain relevant at large length scales. However, these types of symmetries and constraints are far from obvious [100] in biological systems. There are so many components interacting with each other in many different ways and time scales that any microscopic model would involve an exponentially large number of parameters. Taking this approach to first build models that effectively describe the macroscopic behavior of the system (i.e., model *what* the system is doing) may not be very useful (or even possible) in practical terms. In particular, we may not be interested in everything that the system does, but only in some aspects of it. For example, in the aforementioned case of neural action potential generation, we may be interested in the distribution of times between action potentials, but not in the actual voltage profile of the spike. Therefore, alternative approaches to answer the *what* questions without answering the *how* as an intermediate step should also be tried. These type of approaches are likely to be particularly useful for systems that are very complex, such as at cellular or organismal scales.

Interestingly, the *what* questions are fundamentally inference questions: one needs to infer a succinct mathematical description of statistical properties of the underlying system from experimental data, which are usually noisy, undersampled, and often limited in their extent by what the experimentalists can or choose to measure. Thus working on *what* questions is particularly timely in the modern age of machine learning, when tools for statistical inference can suddenly compete with humans in their ability to make sense of data. And yet, generic machine learning tools are not expected to achieve the physics-level understanding of natural processes, and likely will always suffer from limited data [101, 133]. New methods for statistical inference,

rooted in and well-matched to the types of models we would like to use to describe specific biological systems must be developed. In this Dissertation, we develop such inference tools to build various types of phenomenological models to answer specific *what* questions for a variety of biological systems.

In Chapters 2 and 3 we develop methods for inference of phenomenological models for two different kinds of biochemical systems that can be seen as First Passage Processes (FPP). In such processes, a phenomenon of interest gets triggered when the system reaches a particular state in its stochastic trajectory over its internal states. Mathematical techniques to model FPP (but not the inference tools) have been well developed in the past decades [111, 120, 148, 74]. Many biological processes can be studied as such FPP [66, 159, 27, 18], including various systems at molecular, cellular, and even population scales. For instance, transcriptional regulation is aided by a transcription factor protein binding at a promoter site on the DNA. This can be viewed as a FPP, where a protein performs a search along the DNA, or in a 3d nuclear or cellular volume until it finds the correct site and initiates transcription [123, 76, 18, 10, 58, 94, 78, 145]. As another example, in a description of a molecular motor protein walking on a microtubule, every individual step that the protein takes can be seen as a FPP, where after a series of chemical reactions chemical energy is finally converted into mechanical work [77, 28]. Similarly, cellular division happens after a long and complex sequence of biochemical events, also making it a FPP. Indeed, division includes DNA replication, the dissolution of the nuclear membrane, the separation and transport of the chromosome to each pole of the cell, and then the division of the cell membrane to create two new cells [65]. Finally, the action potential generation process in a neuron is also a FPP: a complex set of biochemical processes results in a depolarization of the membrane potential after the interchange of ions between the exterior and interior of the cell. When the depolarization gets to a certain threshold a spike gets triggered [141, 138, 140]. While we will not study all of

these and many others systems and processes in this Dissertation, we will show that our approach to inference of phenomenological models of FPP is broadly applicable.

One of the most important questions when studying a FPP, related to the *what* questions, is: How long does the system takes to get to the configuration of interest (also known as *absorbing state*)? In principle, since the process is stochastic, the entire completion probability density is required to answer this question. Several theoretical and experimental results for many biological systems have characterized what is known as the Mean First Passage Time (MFPT), that is, the mean of the completion time probability density [94, 107, 9]. However, a single time scale is oftentimes not enough to describe the First Passage Times of these complex processes, as we will see in the next chapters. Fortunately, in the last two decades, the rapid development of experimental techniques to monitor chemical and biological systems with single molecule [53, 155, 64, 149, 22] and single cell precision [65, 96, 75, 93] has opened the door for us to understand better the statistical fluctuations in the timing of the First Passage (FP) events that can have important biological functional effects.

To answer these interesting questions, Chapters 2 and 3 introduce and apply methods of statistical inference to building models for FP time distributions in various biological systems. As explained above our approach focuses on *refining* phenomenological models [35] for these systems, because their complexity precludes us from building mechanistically accurate models first.

In Chapter 2 we formally introduce a phenomenological representation for inferring predictive models of the entire probability distribution of FP times. Our representation offers a way to effectively interpret the complexity of the underlying FPP. Furthermore, the inference methods that we use (Bayesian model selection) are especially useful for modeling experimental data sets taken under varying external conditions. Interestingly, in the well-sampled regime, this representation allows us to infer minimal constraints on the structure of the underlying kinetics towards building

mechanistic models for FPP. We first show the utility of our approach on neurophysiological data. Modeling a neuron with molecular accuracy in order to describe the stochastic dynamics of the membrane potential would require to take into account a large number of complex processes related to ion channel gating, membrane capacitance, and leakage [40, 16, 95]. By looking at the spike generation as a FPP we developed a phenomenological representation for one of the most morphologically complex types of neurons, the Purkinje cells (PCs). We provide a complete and accurate statistical description of the ISI distribution under different external conditions, without introducing any of these microscopic details.

In Chapter 3, we extend the methods to a biochemical example, studying it as a FPP to show the utility and the flexibility of our inference approach. We develop very simple phenomenological models to represent an enzymatic reaction controlled by a single molecule of an enzyme β-galactosidase under different substrate concentrations. For these experimental data, our models outperform previous models of the system proposed in the literature. Moreover, guided by the mechanistic hints provided by our phenomenological model, we use the same type of statistical inference tools to develop a different style of models, focusing on maintaining some biochemical realism. Here we automatically infer the smallest biochemically realistic network that can explain the enzymatic turnover times under different substrate concentrations. We show that no other model can provide a better fit to the data, even in principle, and that accurate predictions for the entire FP time distributions can be made using our method.

In Chapter 4, we develop another model-inference approach to study a very different biological process, namely animal behavioral evolution. For these type of processes, experimental observations are only available for the extant species. In contrast to morphological evolution [151, 119, 127] the basic genetic changes behind behavioral evolution are poorly understood [50, 23, 156, 41, 117]. The difficulty arises not only due to the lack of observations along the process (there are no fossil records

for behavior), but from building a good representation to measure behavior. The latter is really not an obvious task given the changes in genetic, neurological, and physiological factors that result in the emergence of certain behaviors as a response to environmental changes. Thus the high complexity of the process, in practice, requires the construction of data-driven models that can provide us with appropriate behavioral representations uncovering the relevant features (answering the *what*) of the behavioral dynamics along the evolutionary process. We use an unsupervised approach developed in [13] to obtain a full behavioral representation of six phylogenetically closely related species of fruit flies (*Drosophila*). We show that this behavioral representation contains enough information to distinguish between species, and, therefore, can be used as a quantitative measure to understand the evolution of behavior along the phylogenetic process. Thus under the assumption that these behavioral traits evolve through time as a result of the addition of small effects of multiple genetic loci or selective fluctuations, we can model the evolution of the flies as a diffusion process happening in the behavioral space, starting at the common ancestral behavioral repertoire. The simplest stochastic model that we could think of to account for this assumption and for the phylogenetic information available is a Generalized Linear Mixed Model (GLMM). The model not only allows us to reconstruct the ancestral behavioral repertoires (initial state of the process), but it also infers the intra- and inter-species variability that optimally explain the emergence of the extant species' behavior. Our results show that the inferred within species variability is closely related to some hidden long-lasting internal states of the flies, i.e latent states of the flies. And the inter-species variability predicts the existence of groups of behaviors that may have evolved together and that potentially could be used as behavioral meta-traits to improve genetic mapping.

All of these Chapters, while studying very different biological systems, argue that it is possible to build tools to infer phenomenological models, and then use them to

answer the *what* questions without building mechanistic models to answer the *how* questions along the way. We hope that these Chapters, collectively, will convince the reader that we are at a special point in the development of the field, where creation of methods to answer the *what* questions is now possible, and that this Dissertation makes tangible contributions in this direction.

Chapter 2

Inferring Phenomenological models of First Passage processes

(This chapter is based on: *Inferring Phenomenological models of First Passage processes.* Catalina Rivera, David Hofmann, Ilya Nemenman. Submitted to *PLOS Computational Biology.* https://arxiv.org/abs/2008.05007)

2.1 Introduction

Processes in living cells are governed by complex networks of stochastically interacting biochemical species. Understanding such processes holistically does not necessarily imply having a detailed description of the system at a microscopic, mechanistic level. Indeed, many microscopic networks can result in equivalent experimentally observable behaviors [8], so that distinguishing alternative networks may be impossible. Even if competing models are not exactly equivalent, they may approximate each other in many key measurable behaviors [54]. Thus a lot of ink has been expended on developing methods for constructing reduced, coarse-grained models of biological processes as alternatives to unidentifiable mechanistically accurate ones [121, 86, 135, 15, 60, 30, 31, 97, 59, 73, 71, 2, 63, 110, 90]. This is usually a

challenging task, requiring construction of a (possibly inaccurate) detailed mechanical model as an intermediate step. In this Chapter, we focus on an alternative approach of *refining* phenomenological models of stochastic biological processes rather than coarse-graining mechanistic ones. Our approach optimally adapts the level of complexity to match the amount and quality of the experimental observations while accurately predicting specific macroscopic properties of the processes.

A large number of biological processes – and the sole focus of this work – can be viewed as First Passage (FP), or completion processes [111, 67, 27, 18, 159]: certain molecules must interact, certain compounds must be created, or certain states must be visited, before an event of interest occurs. For such systems, one is often interested in when the final event occurs (i.e., what the FP time is), rather than in details of which molecules got created or which states were visited in the process. Thus such systems represent a fruitful field for coarse-grained modeling. Crucially, often the available experimental data are sufficiently precise to allow investigation of the whole probability distribution of the FP time, and the fact that the time is stochastic and often broadly distributed can have important functional effects [108, 67, 17, 98].

A natural approach to characterizing the FP distribution based strictly on the statistical information contained in the samples of the FP time involves progressively estimating its higher order cumulants. However, this approach suffers from a well-known problem that such cumulant expansions cannot be truncated at any order but the second, and still give rise to a proper probability distribution [146]. Here we propose a different method for systematically inferring phenomenological models of first passage distributions from empirical data. The approach does not strive for the mechanistic accuracy. Instead, following ideas from [35], we develop a family of models of FP processes, whose complexity can be grown adaptively as data requires, to fit arbitrary FP time distributions. We then choose the optimal model of the appropriate complexity within the family using Bayesian model selection [115, 72, 26, 112, 7, 87].

Our model family consists of mixtures of Gamma distributions, which we argue to have a natural interpretation in the context of FP kinetics. In the well-sampled regime, this natural interpretation allows us to infer mechanistic constraints on the underlying kinetics using fits within our model family [82]. Specifically, the element of the mixture that dominates the passage for short times, sets the minimal number of internal states that a mechanistically accurate stochastic process would need to generate the data. Furthermore, our approach provides a framework to study effects of external perturbations or experimental conditions on the first passage statistics in a systematic way. Specifically, by doing model selection simultaneously on all data sets across multiple experimental conditions, we can obtain a single phenomenological model that explains all of the available data, relating parameters of such global model to the values characterizing the perturbations.

We test the utility of our approach on neurophysiological data sets. Most neurons are too complex to be modeled mechanistically with molecular accuracy, so that any model will involve some element of phenomenology, making this a good testing ground for our approach. Indeed, spontaneous activity of neurons of different types is often modeled under the assumption that the spike trains can be described by renewal processes [137, 33, 113, 52, 80, 125, 134]. Since, in such models, all inter-spike intervals (ISIs) are independent and identically distributed, the spike generation can be specified fully by the ISI distribution, and hence can be seen as a FP process in our framework. While one usually models the ISI distribution as a Gamma distribution [14, 33], more complex constructions are often warranted [21, 139]. Ability of our method to adapt the complexity of the model as required by the quality and the quantity of the data thus promises to be useful in this context. To investigate this, we build models describing the ISI distribution in a certain type of neurons, called Purkinje cells (PCs), under a variety of experimental conditions, and with data coming from real experiments and from biophysically realistic models of the neuron. Purkinje

cells are some of the most morphologically complex neurons, and, indeed, we discover that even *phenomenological* models of their ISI distributions need to be a lot more complex than a single Gamma distribution. For example, we show that 5 or 6 terms in the mixture are needed to describe PCs of a Rhesus monkey. At the same time, even the most detailed computational model of the process is fitted well with just 4 terms, hinting at a room for improvement of biophysical models.

We conclude this article with a discussion of other applications where our method may be useful.

2.2 Results

2.2.1 The model family

The simplest possible stochastic model to represent a FP process is a two state system as shown in Fig. 2.1A. With a constant transition time τ between the initial and the absorbing state, we get an exponentially decaying completion time probability distribution $P(t) = \exp(-t/\tau)/\tau$. A natural extension is a multi-step activation process, where the system irreversibly passes through a number of intermediate states before reaching the absorbing state, see Fig. 2.1B. A simple induction shows that the completion probability distribution in this case is given by the Gamma distribution, Eq. (2.1):

$$P(t|\tau, L) = \frac{t^{L-1}}{\tau^L (L-1)!} \exp(-t/\tau), \tag{2.1}$$

where L corresponds to the number of intermediate states before FP and τ is the average transition time between the intermediate states, which we take to be the same for all states for simplicity and, as we show later, without the loss of generality. This simple model is commonly used to describe neural ISI distributions. However, often times neural spikes exhibit more complex ISI distributions [5, 37, 103, 118, 136, 61].

Motivated by these empirical findings, we built a set of models that are hierarchically organized, so that their complexity can be adapted to the quality and the quantity of empirical data by adding additional Gamma-distributed completion paths as shown in Fig. 2.2A.

The mathematical expression of our model with M different completion paths is:

$$\begin{aligned} P(t \mid \vec{\theta}, M) &= p_1 P(t|\tau_1, L_1) + p_2 P(t|\tau_2, L_2) + ... + p_M P(t|\tau_M, L_M), \\ p_1 &= \frac{1}{1 + x_2 + ... + x_M}, \quad p_2 = \frac{x_2}{1 + x_2 + ... + x_M}, \quad ... \quad , \\ p_M &= \frac{x_M}{1 + x_2 + ... + x_M}, \end{aligned} \tag{2.2}$$

where $\vec{\theta} = (\tau_1, L_1; x_2, \tau_2, L_2; \ldots; x_M, \tau_M, L_M)$ are parameters to be fitted and $P(t|\tau_i, L_i)$ are defined as in Eq. (2.1). Notice that when there is only one completion path, $M = 1$, with only one non-absorbing state $L_1 = 1$, we recover the exponential distribution function with the decay time τ_1. Figure 2.2B shows examples of FP time distributions that can emerge from models with different small values of M by changing parameter values. These distributions can approximate processes, such as neuronal bursts, which have multiple characteristic time scales.

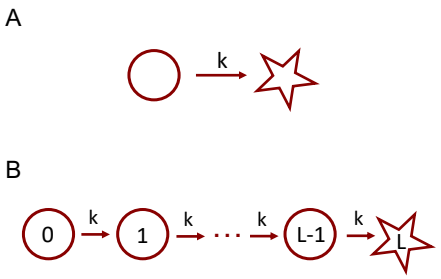

Figure 2.1: Simple FP processes. A: Exponential completion, with $k = 1/\tau$. B: Multi-step completion, with the Gamma-distributed completion time.

We will call the union of all models $P(t|\vec{\theta}, M)$, with $M = 1, \ldots, \infty$, the *multi-path*

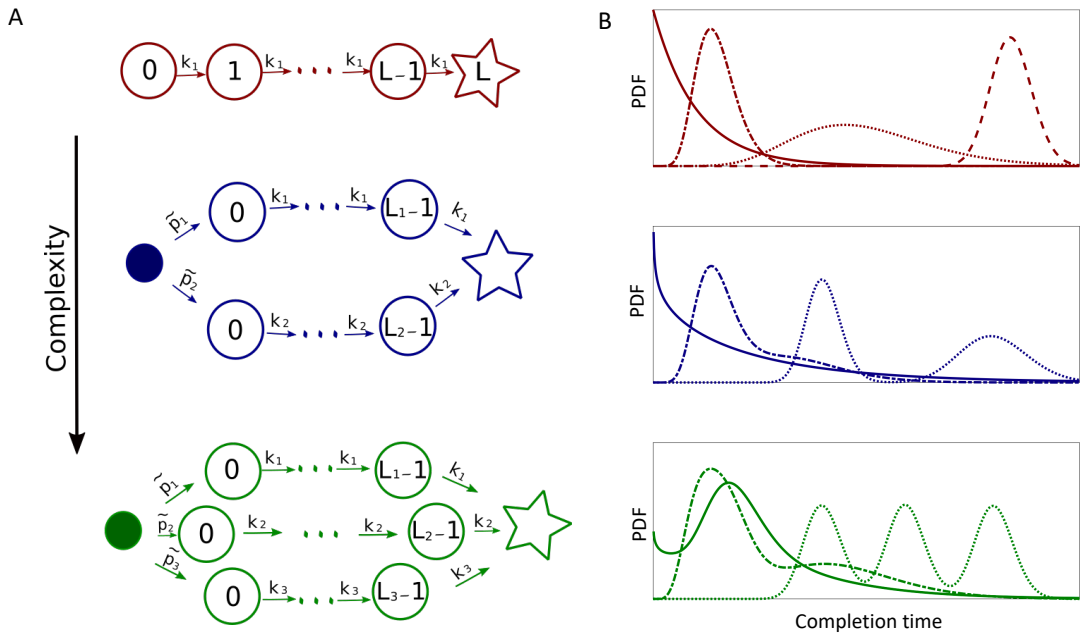

Figure 2.2: Multi-path model family. A: Kinetic schemes of the first three models in the hierarchical set. Each next model in the hierarchy is built by adding another completion path, where $k_i = 1/\tau_i$ is the transition rate between intermediate states, and p_i is the probability of completion through the path i. B: Examples of FP probability densities that can be generated with the corresponding models with different parameter values.

model family of FP distributions. We will focus on Bayesian inference of phenomenological models of FP processes within this family for the rest of this work. One would like such statistical inference to be *consistent*, so that, in the limit of infinite data, one would recover the true model if it belongs to the model family being used in the inference. For an infinite model family to allow such consistent statistical inference using Bayesian approaches, it is sufficient for the family to be *nested* and *complete* [99]. Nestedness (or hierarchy) means that models within the family can be ordered in such a way that the set of solutions of a given model is contained in the set of solutions of the next model in the hierarchy. Completeness means that every data set can be fitted arbitrarily well by some (possibly very complex) model in the hierarchy.

The multi-path model family is trivially nested: if we set $p_M = 0$, then the model

with M paths reduces to the one with $M-1$. The proof of completeness is a bit more subtle, see *Methods*. With that, we know that estimating the posterior probability of the model within the family given the observed data D, and then choosing M that maximizes the posterior probability $P(M \mid D)$, will typically result in consistent inference and in "selection" of the most probable model. Specifically, we need to evaluate

$$P(M \mid D) \propto \int P(D \mid \vec{\theta}, M) P(\vec{\theta} \mid M) d\vec{\theta}, \tag{2.3}$$

where

$$P(D \mid \vec{\theta}, M) = \prod_i P_M(t_i|\vec{\theta}), \tag{2.4}$$

and t_i is the i'th completion time in the experimental data set being fitted. Evaluating this integral and hence building the most probable phenomenological model of the data is the goal of this Chapter.

Unfortunately, as M grows in Eq. (2.3), the involved integral becomes high-dimensional, and it is very difficult to estimate reliably. One usually assumes that the integrand is strongly peaked near the *maximum likelihood* value $\vec{\theta}_0$, which maximizes $P(D \mid \vec{\theta}, M)$. A variety of approximate methods exist for the evaluation [115, 130, 49, 124, 72, 38], which make different assumptions about the structure of the integrand near its maximum likelihood argument $\vec{\theta}_0$. We observed that, for most data sets we tried, $P(D \mid \vec{\theta}, M)$ were far from Gaussian, thus prohibiting the use of the simple Laplace approximation to evaluate the integral [115, 130]. Therefore, we used importance sampling [105, 49] to evaluate Eq. (2.3), see *Methods*.

Experimental data is usually quantized in units of the experimental time resolution. To fit such data we, therefore, transform Eq. (2.2) into its discrete time version by integrating FP probabilities over a time discretization window Δt. That

is, Eq. (2.2) becomes

$$\begin{aligned} P_{\Delta t}(t \mid \vec{\theta}, M) =& p_1 \int_{t-\Delta t}^{t} P(t|\tau_1, L_1)dt \\ &+ p_2 \int_{t-\Delta t}^{t} P(t|\tau_2, L_2)dt + \cdots + p_M \int_{t-\Delta t}^{t} P(t|\tau_M, L_M)dt \\ \approx& p_1 P(t|\tau_1, L_1)\Delta t + p_2 P(t|\tau_2, L_2)\Delta t + \cdots + p_M P(t|\tau_M, L_M)\Delta t. \end{aligned} \tag{2.5}$$

The code to implement the *multi-path model family* for FPP is available at https://github.com/criver9/Inferring-FPP.git.

2.2.2 Model for interspike intervals for Purkinje cells

Purkinje Cells (PCs) are neurons present in the cerebellum of vertebrate animals, that participate in learning. They have large and intricate dendritic arbors and produce complex action potentials with a multiscale distribution of the interspike intervals (ISIs). Due to the complexity of the cells, their typical models involve many dozens of compartments, each described by a handful of biophysical parameters [36, 95, 114, 79, 46]. Crucially, the process of generating a spike can be seen as a FP process, where the neuron goes through a set of different effective states, not necessarily in a simple sequence, before crossing a certain voltage threshold (the absorbing state that results in a spike generation). Thus here we ask whether the ISI distribution for PCs, indeed, requires so many features to model well, or if, in contrast, the structural complexity of PCs does not result in a similarly high complexity of the spike generation. To answer this, we use ISIs of PCs corresponding to simple spikes (spontaneously generated by the cell) of a Rhesus monkey (*Macaca mulatta*), obtained from [61], and we search for the best phenomenological model of this distribution using our approach.

Figure 2.3 shows the best fits for each of the model in our hierarchy, $M \leq 7$, to the PC ISI distribution data. The figure and Tbl. 2.1 suggest that the simplest

phenomenological model of the process contains about $M = 5$ effective independent paths (for this data set, we cannot discriminate between models with $M = 5, 6$ based on the values of $P(M \mid D)$). Notice that, by gradually adding additional completion paths, we can approximate not only the right tail of the distribution, but also the left tail – the behavior at early times. We measure this quality of fit by showing, in Fig. 2.3, the entropy, $H_0 = -\sum_{i=1}^{N} p_i \ln p_i$ (evaluated using the NSB estimator [102]), of data being fitted, as well as the cross-entropy entropy, $H_M = -\sum_{i=1}^{N} p_i \ln P_{\Delta t}(t_i \mid \vec{\theta}, M)$, between the data and each of the best fit models with different M (this corresponds to minus the normalized value of the log-likelihood, Eq. (2.4)). To the extent that H_M approaches H_0 for larger M, the fits are quite good. And since $H_M \approx H_{M+1}$ for large M, the fits stop becoming much better, so that the Bayesian Model Selection [87] then penalizes models with large M, forcing us to settle at $M \approx 5$.

We next check how the selected model depends on the amount of data being fitted. As seen in Tbl. 2.2, increasing the number of spikes in the data set from 1000 to ~ 30000 allows us to identify finer details in the data which require more accurate models to be fitted. Thus the most likely model has $M = 2$ for a small data set, gradually increasing to $M = 5$ for full data. Since the last three-fold increase in the amount of data does not result in a further growth of the best M, we conclude that the phenomenological model likely has reached the complexity needed to explain the system, and the model with $M = 5$ is, in some sense, equivalent to the full complexity of simple spike generation of a real Purkinje cell.

This analysis illustrates two crucial points. First, a relatively simple model with $M \approx 5$ is able to explain the *experimental* ISI distribution from a complex neuron, so that much of the physiological complexity of the cell *does not* translate into a functional complexity, at least at the scale of a simple spike generation. Second, *quantitatively* fitting the data favors models with $M \geq 5$ by a factor of $\sim 10^{20}$ (see

Tbl. 2.1). In other words, PC spiking is not trivially simple, and guessing this ISI model without the automated inference procedure developed here would likely be impossible.

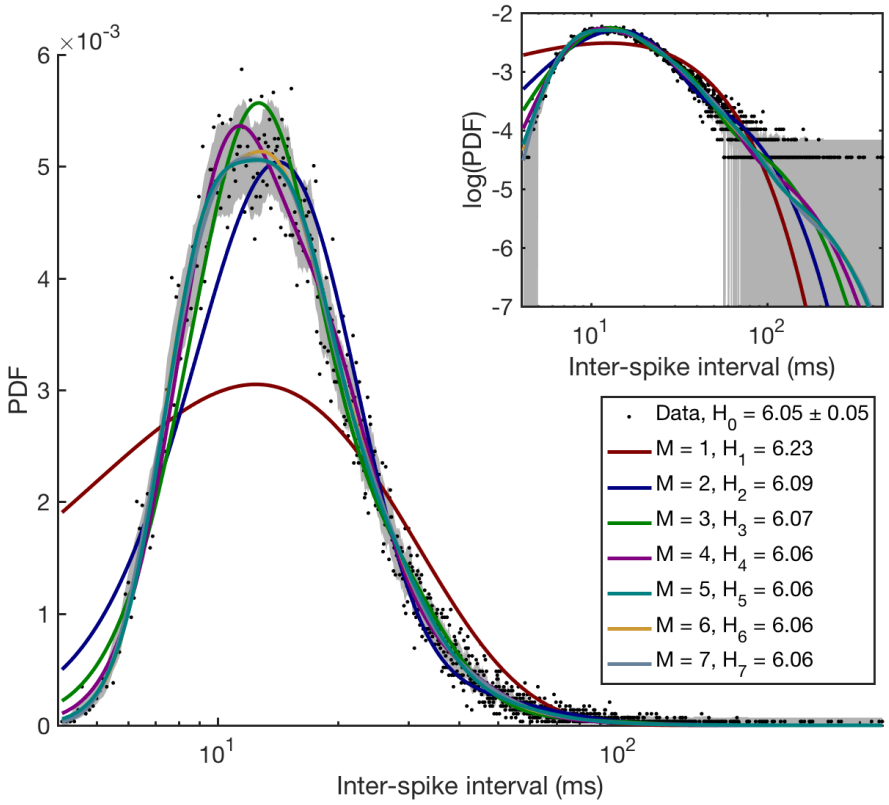

Figure 2.3: Best fit models, $M = 1 \ldots 7$, for Purkinje cells ISI distribution. Dots indicate the histogram of the real data, and the grey band denotes the standard error of every dot. Color lines show the average fit line sampled from the posterior distribution of each of the first seven models in the hierarchy, error bands (too narrow to see) on these fits where estimated using the standard deviation from the sampled curves, see *Methods*. The legend illustrates how the cross-entropy between the data and the models decreases with the model complexity towards the data entropy. Note that the horizontal axis is logarithmic. Inset: same data, but on log-log axes.

M	$\ln P(D \mid M)$
1	-180379
2	-176368
3	-175826
4	-175694
5	**-175649**
6	-175651
7	-175653

Table 2.1: Model selection results for ISI of experimental PC. Posterior likelihoods of the first seven models in the model family are shown for $N = 28966$ spikes (the full data set). The model with the highest marginal likelihood, $M = 5$, is highlighted. Note that models with $M = 6, 7$ cannot be ruled out, as they have very similar posterior likelhoods.

$\ln P(D \mid M)$						
M	1000	5000	10000	15000	20000	28966
1	-6192.08	-30901	-61792	-92841	-124040	-180379
2	**-6133.59**	-30427	-60673	-91189	-121736	-176368
3	-6133.88	-30369	-60536	-90915	-121356	-175826
4	-6142.02	**-30349**	-60486	-90858	-121262	-175694
5	-6151.81	-30350	**-60481**	**-90840**	**-121230**	**-175649**
6	———	-30357	-60485	-90844	-121232	-175651

Table 2.2: Model selection as a function of the number of samples. First row shows the size of the data set, $1000 \ldots 28966$, and the rest of the table shows the posterior probability of each model in the family for these data. As the number of samples increases, more complex models are required to explain the details of PC spiking, but the complexity eventually saturates, presumably having matched the complexity of the real cells observed at the given experimental accuracy.

2.2.3 Model for ISI of synthetic PC

One of our interests is to develop phenomenological models that are able to predict the change in the FP distributions for a system under the influence of various external perturbations. We would like to illustrate this using PCs. However, we are not aware of readily available large, precise data sets measuring the ISI distribution in PCs under external perturbations. Thus instead we focus on synthetic data, generated using a biophysically realistic, multi-compartmental model that resembles the morphologically complex structure of PCs, the Miyasho et al. model [95], which is a

modified version of the earlier De Schutter and Bower model [36]. To illustrate the complexity of the Miyasho model, we point out that it uses 1087 compartments to describe the dendritic arbor of a PC and one compartment for the soma. Additionally, the dynamics are defined by around 150 parameters that specify 12 different types of voltage-gated ion channels [95].

We used this model to simulate the behavior of the membrane potential dynamics of a PC, affected by different electric currents injected into the soma. White noise currents with standard deviation $\sigma = 3$ nA and mean values $I = 0.1, 0.5, 0.7, 1, 2, 3$ nA where injected, thus generating six different data sets, with which to explore the ISI probability distributions of the PC model. Following the procedure described earlier, we selected the simplest phenomenological model that can explain the ISI statistics of the PC model, but in this case we focus on optimizing the posterior likelihood over *all* stimulus values simultaneously. Figure 2.4 shows the best model fits for two different injected currents which produce qualitatively different ISI distributions. Fits for other current values can be found in Fig. A.2. To build the optimal model for all injected currents simultaneously, we estimate the posterior likelihood of each model in the family for $M \leq 5$ for each of the synthetic data sets, see Tbl. 2.3. Since for different currents, the ISI generated are independent of each other, the log-likelihood for the entire data set is simply the sum of log-likelihoods for each I. As always, we choose the optimal model as the one with the largest overall log-likelihood.

Table 2.3 shows that, for our data sets, $M = 4$ effective independent paths are enough to explain simultaneously the PCs behavior under six different injection currents. As can be seen in Fig. A.1, when the injected current increases the cell goes from the non-bursting to the bursting state, and the entropy of the completion time distribution decreases (see Figs. 2.4, A.2). Table 2.3 indicates that higher entropy distributions, corresponding to $I = 0.1, 0.5$ nA need $M \geq 5$ completion paths to be properly explained. Lower entropy distributions, on the other hand, not only require

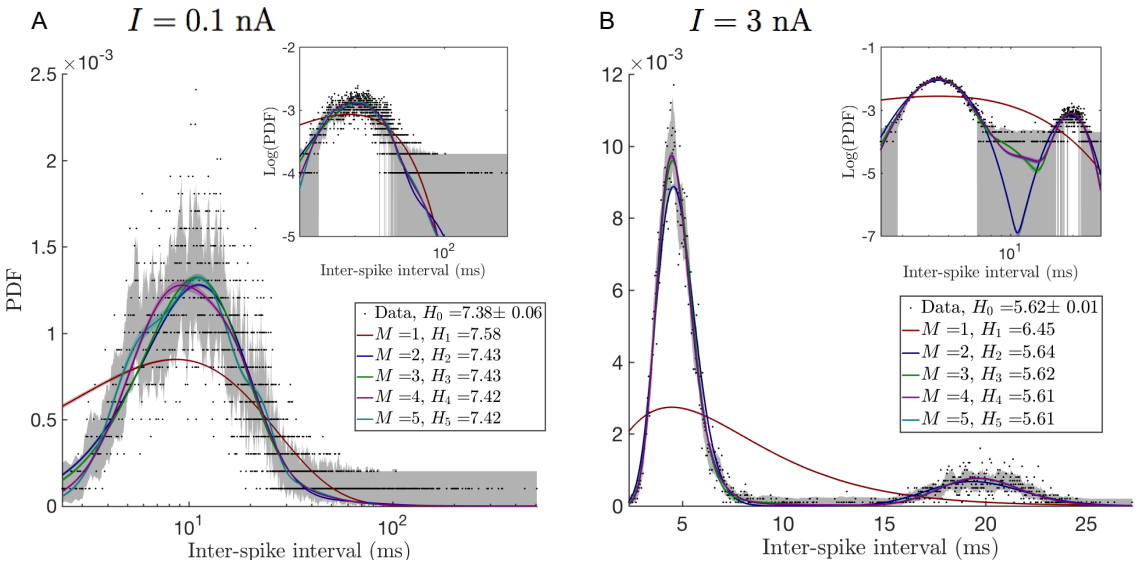

Figure 2.4: Best fits for different models in the model family for the distribution of ISIs of synthetic PCs. Color lines and color bands show the mean and standard deviation respectively of different models sampled from the posterior distribution of each of the first five models in the family (see details in the Appendix 2.4.3). The legend shows how the cross entropy decreases with the model complexity towards its minimum value of the entropy of the histogram of the observed data. According to Tbl. 2.3, 4 paths are needed to explain the ISI characteristics of synthetic PCs under different external conditions. A: injected current $I = 0.1$ nA, and B: $I = 3$ nA.

fewer paths, but also more deterministic paths, as can be observed from the coefficient of variation estimates in Fig. 2.5. This suggests, that under low external stimulus ($I < 0.5$ nA), spike generation in the cell can happen through multiple pathways. Instead, when a certain current threshold is reached ($I > 0.5$ nA), only a few of these pathways get activated. Nonetheless, more than one pathway is needed even for high currents, since, at least, two time scales are required to explain the bursting activity.

In Fig. 2.5, we explore how the properties of the model selected in Tbl. 2.3 ($M = 4$) change as a function of the injected current, I. Each independent path is described by specifying its average completion time $\overline{T}_i = \tau_i L_i$, the coefficient of variation $\mathrm{CV}_i^2 = 1/L_i$, and the probability p_i of completion along this path, and these three quantities

$\ln P(D \mid M)$							
M/I(nA)	0.1	0.5	0.7	1.0	2.0	3.0	Total
1	-75437	-71282	-66821	-66309	-64654	-64488	-408992
2	-74070	-70034	-65283	-63578	-58932	-56462	-388359
3	-74019	-70001	**-65238**	-63520	**-58773**	**-56211**	-387762
4	-74003	-69993	-65239	**-63516**	-58794	-56212	**-387753**
5	-73994	-69976	-65249	-63527*	-58805*	-56223*	-387775

Table 2.3: Model selection results for ISI of synthetic PCs. Posterior likelihood of the first five models in the family for each data set, corresponding to the six different injected currents. Last column shows that a model with 4 completion paths is optimal over the combined data. Asterisk marks those cases where the optimal parameter values fell at the boundary of the search space, usually because there were paths with near-zero flux through them (see *Methods*). Note that the numbers in the first two columns increase monotonically with M, so that the best model in the family is not found for $M \leq 5$. We chose to truncate the exploration at $M = 5$ since we are interested in the overall maximum of the log-likelihood for all I, which is reached at $M = 4$ (last column).

are plotted for each path for different values of I. There is a sharp change in these features when the PC transitions from a non-bursting to a bursting state, between $I = 0.5$ and 0.7 nA. For example, completion times and coefficient of variation for all paths drop drastically at this point. In particular, Fig. A.3 shows that the paths with the longest completion time explain very different aspects of the non-bursting and the bursting ISI distributions. For the non-bursting cases, these paths help to fit mostly the tails. Instead, for the bursting cases, these paths explain the intra-burst time interval, which happens to be a much more deterministic process, as can be seen from the behavior of the coefficients of variation, Fig. 2.5.

To test whether the phenomenological model correctly captures the time scales of the underlying biophysical processes, we predict the ISI distribution for input currents that the model was not exposed to during fitting. To achieve this we first need to determine a relationship between model parameters and the input current means, which we can then use to infer model parameters for currents different from the ones used for fitting the model. As our test case, we employed the model with $M = 4$ and tracked the dependence of its parameters on the current as shown in Fig. 2.5. A priori

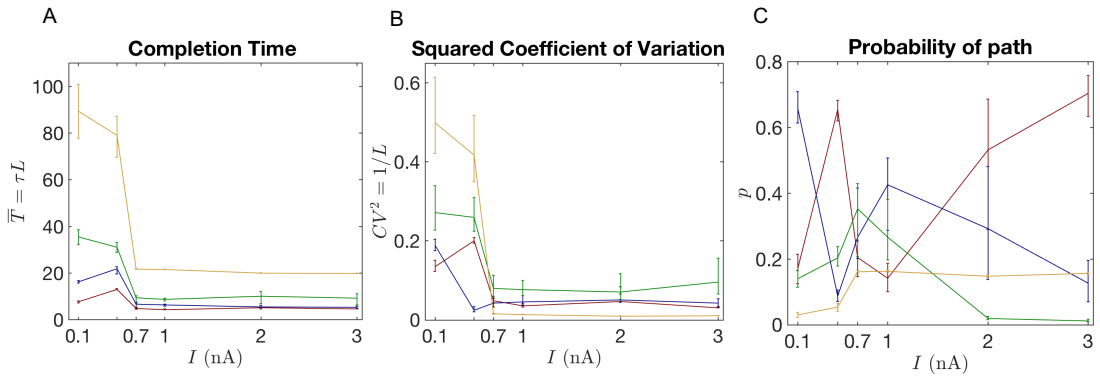

Figure 2.5: Properties of completion paths change as a function of the external parameter for the best model selected across all experiments. A: Average completion times for each of the $M = 4$ independent paths are plotted as a function of the injected current in the soma, I. Color (same in B and C) identifies paths according to how long they take to complete the process on average. B: Coefficient of variation and C: probability of taking each of the independent paths of the model as a function of I.

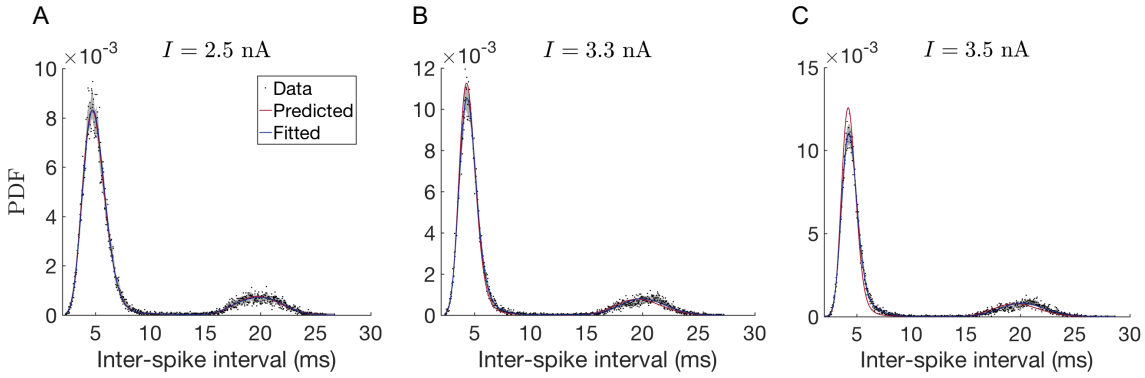

Figure 2.6: Predicted PDFs for non-measured values of the injected current. Predicted model (in red) was obtained by interpolating parameter values from Fig. 2.5. It is compared with the model (in blue) fitted directly to data. A: Prediction for $I = 2.5$ nA (interpolation). B and C: Prediction for $I = 3.3$ and $I = 3.5$ nA respectively (extrapolation).

it is unclear how to build correspondence between the four model paths for separate input currents. In our example in Fig. 2.5, we chose to establish the correspondence by ordering the paths according to their completion time, thus relating the model paths with the smallest completion time, then the second to smallest and so on. This ordering provides relationships between input currents and all model parameters, based on which we can infer parameter values for new current values using linear interpolation (for currents that fall between two fitted values) or linear extrapolation (for currents outside of the fitted range). We note that the choice to relate parameter values by completion time rather than another parameter is arbitrary. Indeed there are many possibilities to create the pathway correspondence for different current values. Besides ordering based on average completion time (confront Fig. 2.5) we also tested ordering based on the coefficient of variation or the probability path which led to no improvement over the presented case (not shown). While it is possible that other orderings can lead to better predictions we leave a more systematic exploration of this aspect for future work.

To validate our predictions, we generated new data for mean currents $I = 2.5$, 3.3, and 3.5 nA and compared predicted ISI distributions to the simulation results (see Fig. 2.6). The predicted model for $I = 2.5$ nA (where parameters interpolate between the known values at $I = 2.0$ and 3.0 nA) is almost indistinguishable from the fitted one (Fig. 2.6A). Even the extrapolation to $I = 3.3$ and $I = 3.5$ nA (Fig. 2.6B-C) show very good agreement between predicted model and simulation data.

To quantify the accuracy of these predictions, we need to calculate their quality with respect to some baseline. We chose the Jensen-Shannon Divergence (JSD)[83] as a measure of the quality of fit, and we measure it relative to two baselines. First, we quantify how an extrapolated or an interpolated prediction compares to the fit done directly on a data set; certainly the fit is expected to outperform the prediction. Second, we check how two statistically equivalent realizations of data fit each other;

this should be the ceiling, which neither the fit nor the prediction can outperform (if both are not overfitted). Both of these baselines depend on the specific data set used, and thus one needs to estimate probability distributions of the relevant JSDs, rather than their single values. However, generating data from the PC model takes hours even on a modern computer, and hence we generate only a single additional, validation, data set beyond the training and the testing sets, which we then additionally bootstrap (resample with replacement) to produce statistics of the JSDs. Specifically, Fig. 2.7 plot histograms of (i) the JSD between the test data and the bootstrapped versions of the validation data (this is the statistics that requires us to have two independent samples, test and validation, to remain unbiased), (ii) the JSD between the bootstrapped validation data and fits to these data, and (iii) the JSD between the prediction and the bootstrapped validation data. Our first observation is that all three JSD distributions are very close to each other, indicating very good fits and predictions. For $I = 2.5$ nA, the fits/predictions have smaller JSD than different realizations of data have with themselves, which is consistent with a very good fit, and suggests, as expected, that the variability across bootstrapped data sets is somewhat larger than would have been across independent samples. As I increases, and interpolation gives way to extrapolation, the prediction quality deteriorates (still remaining only a few percent worse than the fits).

Inferring mechanistic constraints

Our approach to modeling FP time probability distribution is purely phenomenological. However, the *multi-path model family* allows us additionally to constrain mechanistic, biophysical models of the underlying processes. Specifically, we can make predictions for the minimal number of intermediate states that a mechanistic model requires to explain the data. Indeed, for any FP problem, the short-time behavior of the completion probability density provides information about the length of the

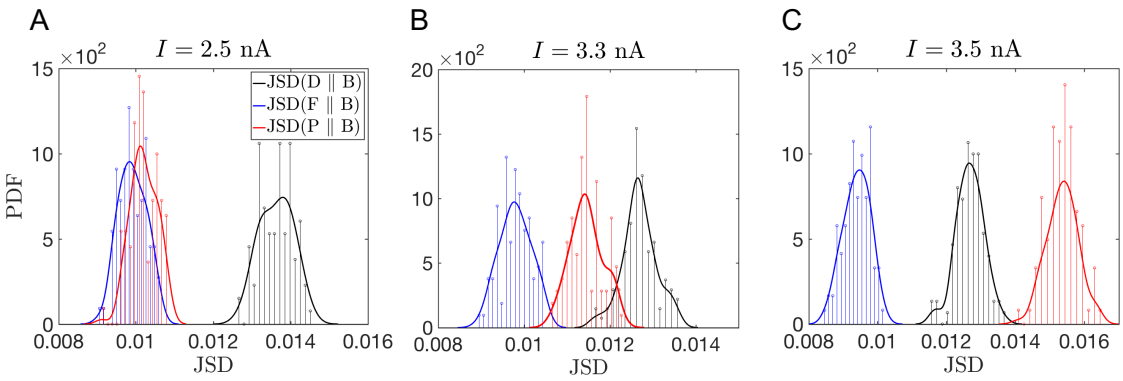

Figure 2.7: Quantifying quality of the predictions. We plot the histograms of the JSD between the test data set and the bootstrapped samples from the validation data set (in green), the JSD between the bootstrapped validation data sets and models fitted to each of these data sets (in blue), and the JSD between the bootstrapped data and the prediction based on interpolating or extrapolating the model parameters fitted to the original data (in red). To the extent that the distributions are close, predictions are good. A-C: Data for $I = 2.7, 3.3, 3.5$ nA, respectively. The first is interpolation, the other two are extrapolations.

shortest completion path [82, 142]. That is, assume that the process starts in a state i and ends at the absorbing state j of an arbitrary Markovian chemical reaction network. Then, at short times, the completion probability density can be approximated as $\rho_{ij} \propto t^m$, where m is the number of intermediate states of the shortest path connecting states i and j [142]. In principle, this means that by estimating the exponent of the power law that fits the left tail of the completion time distribution, one can put a lower limit on the number of intermediate states in a mechanistic model. Then any candidate model with a fewer number of steps can be rejected.

In practice, making use of this result is hard because it requires data with very high temporal resolution, and a very well sampled left tail. However, our *multi-path* representation allows for an extension of the approach to the case where the sampling is good, but the time resolution may not be sufficient for simpler methods. Once the most probable model in the model family is selected and fitted, we propose to

determine if the first few fastest events can be explained by a single independent path i of length L_i. We use 50 events in our analysis, which provides for a sufficient number of the events to seek a power law fit, and yet is small enough so that only the very end of the left tail is explored. Since at short time scales the Cumulative Distribution Function (CDF) of the FP time probability density is $\propto t^{L_i}$ (from Eq. (2.1)), one can insist that any mechanistic model built to describe the data will need at least L_i states, establishing a lower bound on the size of the network.

For concreteness, the short time behaviors of the CDFs obtained from the best model, $M = 4$, describing the ISIs of PCs under six different injected currents are shown in Fig. 2.8. Only for $I = 0.7$ nA the first 50 events (0.5% of sample size) can be explained by a single path with ~ 20 intermediate states, while for larger values of I, the distribution can be fitted by one or more of such paths. In all of these cases, it is thus clear that any realistic biophysical model of a PC must include, at least, ~ 20 internal states.

2.3 Discussion

In this study, we developed a mathematical structure of the (*multi-path model family*) to infer phenomenological models describing FP time distribution for biological processes. As an example of application of our approach, we show that this representation allows us to build models capable of describing the complexity of the ISI distributions of PCs by successfully explaining not only the bulk, but also the tails of the distribution. Our results show that the process of a spike generation in PCs is more complex than a simple renewal process with a Gamma-distributed completion time, which is typically used in the field. For spontaneously generated spikes, $M \geq 5$ independent Gamma-distributed paths are required. We also showed that only $M \approx 4$ paths (11 independent parameters) are needed to explain the behavior of synthetic

PCs over all injected current values $I > 0.5$ nA. This illustrates that (i) morphological complexity of PCs notwithstanding, their dynamics is not very complex at the level of the FP time distribution, and (ii) our fully phenomenological approach can, nonetheless, point out when biophysically-realistic models are inconsistent with features of experimental data. By identifying how parameters of the inferred model change with the external stimulus and extrapolating or interpolating them, we can predict the FP time distribution of the system in response to novel stimulus values. These predictions focus not just on the mean and the variance, but on the entire completion time distribution, and we have shown that the predictions are remarkably accurate, as compared to statistical fluctuations in the data themselves. Finally, we showed how our purely phenomenological approach can establish the minimum size of a mechanistically accurate biochemical network underlying the system, at least for well-sampled data sets.

The specific model family hierarchy we developed here is only one of many possible hierarchies that are both complete and nested. Like in [35], different hierarchies may be better suited for phenomenological modeling of different biological processes, and thus their relative success would reveal salient properties of the modeled processes. We hope to develop such additional hierarchies, and explore their pros and cons in future work. Additionally, here we assumed that every completion time is independent and identically distributed. This is a strong assumption, which is not always realized. Even for PCs, the burstiness of spiking suggests dependence among the successive ISIs (i.e., within a burst, a short ISI is usually followed by another short ISI). In the future, it should be possible to extend our approach to model such processes by either modeling the statistics of FP time for a sequence of events, or by extending the model family to incorporate a latent variable that controls the dependence among subsequent completion events.

Our models offer only limited understanding of the mechanistic details of the

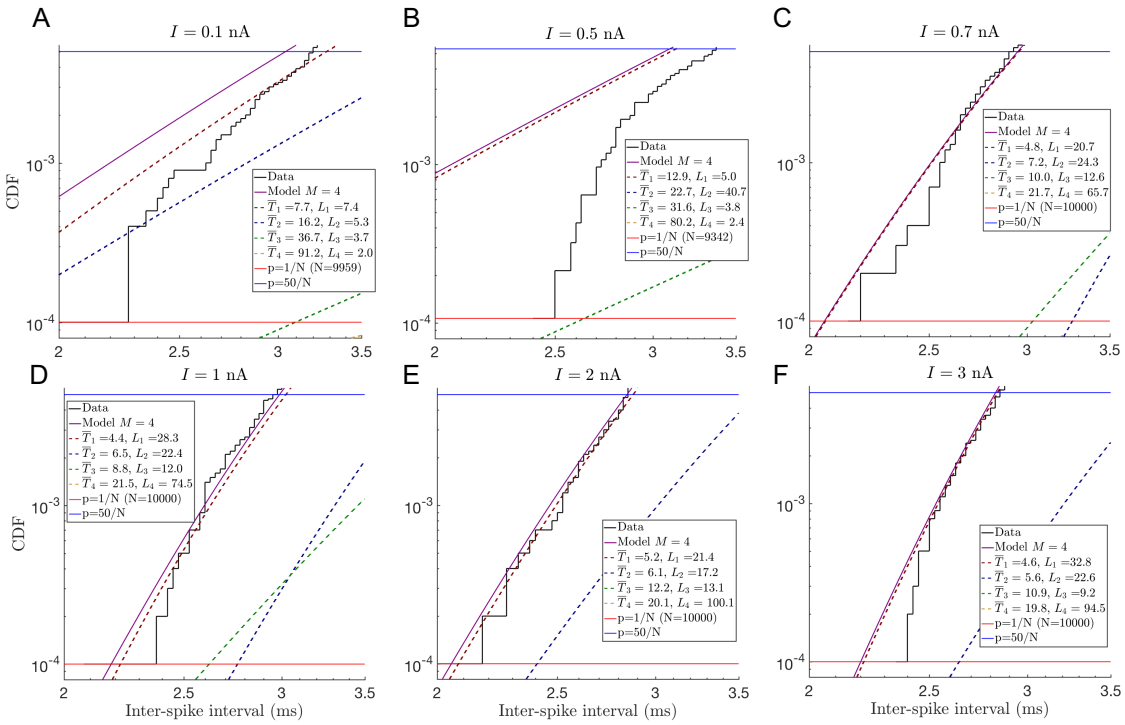

Figure 2.8: Decomposition of the Cumulative Distribution Functions (CDFs) of completion time at early times into the four completion pathways. Black line represents the CDFs from data; horizontal red and blue lines in each plot correspond to the probability of the 1st and the 50th events, respectively. Solid purple lines are CDFs from the best-fitted model with $M = 4$, and each of the dashed lines represents contributions from the constituent completion paths. A-F: $I = 0.1, 0.5, 0.7, 1.0, 2.0, 3.0$ nA, respectively. Panels C-F show that the first fifty events can be explained by one or more paths with $\sim 20 \ldots 30$ intermediate states. Therefore, any biophysically accurate reaction network explaining these data needs to have at least > 20 internal states. Notice that, even though a model with $M = 4$ is optimal over all values of I according to Tbl. 2.3, it does not explain the early time behavior in panels (A,B).

modeled biological process. Nonetheless, there are many advantages to our approach, and phenomenological modeling in general. Indeed, the complexity that biological processes have acquired over eons of evolution oftentimes makes building detailed microscopic models an extremely challenging task. And yet the functional properties of the behavior might be rather simple, with the structural complexity existing, for example, to ensure robustness of the function to various perturbations. Then focus-

ing on the phenomenological model allows us to elucidate, predict, and eventually use properties of the functional behavior even if microscopic details of the mechanisms used to produce it remain unclear. Our specific approach to phenomenological modeling is different from many others in that it does not coarse-grain a microscopic model (requiring a laborious task of building one as an intermediate step), but rather it *refines* phenomenological models, adding progressively more details until the functional behavior is well approximated. Bayesian model selection is used to find the optimal point in the refinement hierarchy. The computational advantages of taking such an adaptive, refining approach can be huge, especially when the studied complex system exhibits a simple behavior. The computational complexity of our approach is dominated by searching for optimal fits, which scales linearly with the data set size, and exponentially with the model complexity. However, the latter is rarely more than a few dozen parameters even for very complex systems, such as the PCs, at least for realistic experimental resolution and data set sizes. Thus we expect our approach to be useful for modeling any biological system for which (i) the quantity that we need to predict is the completion time, (ii) the underlying biophysics is very complex, with microscopic details not always affecting the macroscopic completion properties, and where (iii) large, high quality experimental data sets are available for different experimental conditions, requiring (iv) to predict the behavior of the system as a function of these conditions, for their yet-untested values.

2.4 Materials and Methods

2.4.1 Completeness

Here we show that the model family studied in this work, Eq. (2.5), is complete. That is, any data set describing the distribution of the completion times of the first passage process can be approximated arbitrarily well by a gamma mixture model with

sufficient complexity.

We note that experimentally measured and numerically simulated completion times are constrained by finite resolutions which essentially discretizes the time axis. Thus we can write the completion time likelihood as a multinomial

$$L(\vec{q} \mid \vec{n}) = q_1^{n_1} q_2^{n_2} ... (1 - q_1 - q_2 - ... q_{K-1})^{(N - n_1 - n_2 - ... n_{K-1})}, \tag{2.6}$$

where n_i counts how often the completion time falls into the ith out of K time interval bins $(t_i - \Delta t, t_i]$, N is the total number of completion time events, and q_i is the probability of completion in the time interval defined by bin i, given by $q_i = P_{\Delta t}(t_i \mid \vec{\theta}, M)$ (see Eq. (2.5)). Trivially, the maximum of $L(\vec{q} \mid \vec{n})$ is achieved when $q_1 = n_1/N$, $q_2 = n_2/N$, $...q_K = n_K/N$. Therefore, our aim must be to construct a model that can bring $\vec{q}$ arbitrarily close to this maximum. The rational of the proof is to have a path per time bin whose average waiting time is the center of the respective time bin and whose variance can get arbitrarily small, effectively approximating a delta function. That is, we want to construct a model such that for any $\epsilon > 0$, we have $n_i/N - \epsilon \leq P_{\Delta t}(t_i \mid \vec{\theta}, K) \leq n_i/N + \epsilon$.

To prove this we set the parameters in Eq. (2.5) to what follows. For the probability of every gamma path take $p_i = n_i/N$, with expected completion time given by $T_i = L_i \tau_i = t_i - \Delta t/2$ and variance (arbitrarily small) $\sigma_i^2 = T_i \tau_i = \frac{\Delta t^2}{4} \epsilon_i$, where $\epsilon_i = \min(\frac{\epsilon}{p_1 + p_2 + ... + p_{i-1} + p_{i+1} + ... p_K}, \frac{\epsilon}{p_i})$. Then we can show that:

$$\begin{aligned} P_{\Delta t}(t_i \mid \vec{\theta}, K) = \frac{n_1}{N} \int_{t_i - \Delta t}^{t_i} P(\tau_1, L_1) dt + \frac{n_2}{N} \int_{t_i - \Delta t}^{t_i} P(\tau_2, L_2) dt + ... + \\ \frac{n_K}{N} \int_{t_i - \Delta t}^{t_i} P(\tau_k, L_K) dt \leq \epsilon_i [\frac{n_1}{N} + ... \frac{n_{i-1}}{N} + \frac{n_{i+1}}{N} + ... \frac{n_K}{N}] + \frac{n_i}{N} \\ \leq \epsilon + \frac{n_i}{N} \end{aligned} \tag{2.7}$$

where we used Chebyshev's inequality ($\Pr(|t - T_i| \geq \alpha \sigma_i) \leq \frac{1}{\alpha^2}$, with $\alpha = 1/\sqrt{\epsilon_i}$) to

set a bound to all the integrals but the ith. For the ith integral we note that, since most of the probability mass falls in this bin, it reaches close to one and is naturally bounded by one. This concludes the upper bound on the q_i. For the lower bound we simply subtract one from both sides of the Chebyshev inequality and multiply by negative one to get $\Pr(|t-T_i| \leq \alpha\sigma_i) \geq 1-\frac{1}{\alpha^2}$. This gives a bound for the ith integral of Eq. (2.7):

$$P_{\Delta t}(t_i \mid \vec{\theta}, K) \geq \frac{n_i}{N} \int_{t_i-\Delta t}^{t_i} P(t|\tau_i, L_i)dt \geq \frac{n_i}{N}(1-\epsilon_i) \geq \frac{n_i}{N} - \epsilon, \tag{2.8}$$

showing that this model family can approximate any sufficiently smooth distribution arbitrarily well. In real applications, we may not need to have as many paths as there are bins to achieve a high approximation accuracy, so the construction above is the worst case scenario.

2.4.2 Model Selection

To choose the most likely model from the family, we evaluate and maximize the posterior probability of each model M:

$$P(M \mid D) = \frac{P(D \mid M)P(M)}{P(D)} \propto P(D \mid M), \tag{2.9}$$

where we assumed that all models in the hierarchy are a priori equally likely. The likelihood $P(D \mid M)$ is given by:

$$P(D \mid M) = \int P(D \mid \vec{\theta}, M)P(\vec{\theta} \mid M)d\vec{\theta}, \tag{2.10}$$

where the likelihood of the data set and the prior are chosen to be:

$$P(D \mid \vec{\theta}, M) = \prod_{i=1}^{K} P_{\Delta t}(t_i \mid \vec{\theta}, M)^{n_i}, \tag{2.11}$$

$$P(\vec{\theta} \mid M) = \frac{1}{(Z_x)^{M-1}} \prod_{j=1}^{M} \frac{\exp(-\frac{\tau_j}{Z_\tau})}{Z_\tau} \frac{\exp(-\frac{L_j}{Z_L})}{Z_L}. \tag{2.12}$$

Here $P_{\Delta t}(t \mid \vec{\theta}, M)$ is given by Eq. (2.5), and n_i is the number of events with completion time between $(t_i - \Delta t, t_i)$. The parameters of our prior are Z_L, Z_τ, and Z_x. The values of Z_L and Z_τ are set such that the priors are reasonably wide compare to the measured time scales and throughout our study we set them to $Z_L = 20$ and $Z_\tau = 20$ ms. Z_x sets the upper boundary for the support of the x_i and was set to $Z_x = 10^3$. Finally, we note that our choice of prior assumes no correlation among model parameters.

In most cases, the integration in Eq. (2.10) is analytically intractable. A typical approach in such a case is to use the Laplace approximation to compute the integral. However, in our considered problems the posterior distributions fall much slower than Gaussians, ruining the quality of the Laplace approximation. Thus we used importance sampling [105] instead. Specifically, we sampled from the multi-variate normal distribution $G(\vec{\theta}) = \det(2\pi\Sigma)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\vec{\theta} - \vec{\theta^*})'\Sigma^{-1}(\vec{\theta} - \vec{\theta^*})\right)$ centered at the optimal value $\vec{\theta^*}$ of the integrand $F(\vec{\theta}) := P(D \mid \vec{\theta}, M)P(\vec{\theta} \mid M)$ with the covariance matrix Σ defined by the Hessian of $F(\vec{\theta})$:

$$(\Sigma^{-1})_{ij} = (-\operatorname{Hess} \log F \mid_{\vec{\theta^*}})_{ij} \equiv -\frac{\partial^2 \log F}{\partial\theta_i \partial\theta_j}\Big|_{\vec{\theta^*}}. \tag{2.13}$$

This way we ensured that $G(\vec{\theta}) > 0$ for $F(\vec{\theta}) > 0$, at least around the domain of the local optimum at $\vec{\theta^*}$. See below for details of how we estimated the covariance

matrices. Then the importance sampling estimate of the integral in Eq. (2.10) is

$$P(D \mid M) \sim \frac{1}{N} \sum_{i=1}^{N} \frac{P(D \mid \vec{\theta}_i, M) P(\vec{\theta}_i \mid M)}{G(\vec{\theta}_i)}, \tag{2.14}$$

where $\vec{\theta}_i \sim \mathcal{N}(\vec{\theta}^*, \Sigma)$ and we used $N = 10^6$ samples to achieve the desired accuracy. Since the likelihood values exceeded numerical resolution, we instead computed the $\ln P(D \mid M)$:

$$\ln P(D \mid M) \sim \ln F(\vec{\theta}^*) + \ln\Big(\sum_{i=1}^{N} \frac{\exp(\log F(\vec{\theta}_i) - \log F(\vec{\theta}^*))}{G(\vec{\theta}_i)}\Big) - \ln N. \tag{2.15}$$

Covariance matrix estimation

Application of our importance sampling scheme requires knowing the maximum of the integrand and the Hessian around the optimum. The optimal values $\vec{\theta}^*$ were obtained using the MATLAB function `fminsearchbnd`. We used MATLAB version R2017a for our analysis. Most of the optimal values obtained for different models and data sets fell in the interior of the parameter's domain; we mark those where the optimal values fell at the boundary with an asterisk everywhere in the text.

We first explain how we computed the covariance matrix for the cases where the optimal values fell in the interior of the parameters' domain set. Using Eq. (2.11) to estimate the Hessian, we get

$$\begin{aligned} -\frac{\partial^2 \log F}{\partial\theta_k \partial\theta_j}\Big|_{\vec{\theta}^*} &= -\sum_{i}^{M} n_i \frac{\partial^2 \log(P_{\Delta t}(t_i \mid \vec{\theta}, M))}{\partial\theta_k \partial\theta_j}\Big|_{\vec{\theta}^*} - \frac{\partial^2 \log P(\vec{\theta} \mid M)}{\partial\theta_k \partial\theta_j}\Big|_{\vec{\theta}^*} \\ &= \sum_{i}^{M} n_i \Bigg[\frac{1}{P_{\Delta t}(t_i \mid \vec{\theta}, M)^2} \frac{\partial P_{\Delta t}(t_i \mid \vec{\theta}, M)}{\partial\theta_k} \frac{\partial P_{\Delta t}(t_i \mid \vec{\theta}, K)}{\partial\theta_j}\Big|_{\vec{\theta}^*} \\ &\quad - \frac{1}{P_{\Delta t}(t_i \mid \vec{\theta}, M)} \frac{\partial^2 P_{\Delta t}(t_i \mid \vec{\theta}, M)}{\partial\theta_k \partial\theta_j}\Big|_{\vec{\theta}^*} \Bigg]. \end{aligned} \tag{2.16}$$

Notice that the contribution to the Hessian coming from the prior in the previous expression cancels out. We then evaluated Eq. 2.16 numerically using Eq. (2.5).

For those cases, for which the optimal values are located at the boundary of the parameters' domain due to the presence of a trivial completion path we use the following trick. Given that the flux through a certain path j is zero, the likelihood $P(D \mid \theta, M)$ stays constant for all values of τ_j and L_j corresponding to this trivial path. However, the prior decays exponentially and therefore $F(\vec{\theta})$ also decays exponentially in the directions of τ_j and L_j. The optimal value of $F(\theta)$ can be written as $(\vec{x}_p, x_d = 0, \tau_d = 0, L_d = 0)$ with $\vec{x}_p$ is the best fit for the previous model in the family, with only $d-1$ completion paths. Then the covariance matrix is:

$$\Sigma = \left(\begin{array}{c|ccc} \Sigma_p & \multicolumn{3}{c}{\mathbf{0}} \\ \hline & \alpha_x^2 & 0 & 0 \\ \mathbf{0} & 0 & \alpha_\tau^2 & 0 \\ & 0 & 0 & \alpha_L^2 \end{array}\right), \tag{2.17}$$

where Σ_p is the covariance matrix at the best fit of the previous model in the family; α_x^2 is an upper bound on the variance along the parameter controlling the probability flux through d-th completion path estimated from the symmetric function $F_s(\theta) = F(|\theta|)$. We used $\alpha_x = 0.01$ for all the cases marked with an asterisk in Tbl. 2.3. On the other hand α_τ and α_L where estimated using the variances of the independent exponential distributions of the prior, Eq. (2.12), $Z_\tau = Z_L = 20$. We chose $\alpha_\tau^2 = \alpha_L^2 = (3\sigma_\tau)^2 = 3600$. Notice that, along these last two directions where $F(\theta)$ decays exponentially we chose the variance of the importance distribution nine times larger in these two directions to make sure that it contains most of the important domain of $F(\theta)$.

Parameter Degeneracy

The posterior distributions that we obtain often have multiple modes that correspond to parameter degeneracy, which arises by relabeling the completion paths. To account for this degeneracy in caclulating posterior likelihoods, we multiplied the likelihoods of each model with M gamma pathways by $(M-1)!$. Here we use $M-1$ instead of M because the first is different from the others: transition rate to this path is set to one and is used as a references.

Generalized Bayesian model selection

In order to find the model in the family that best fits the simultaneous description of the system under s different external conditions, we need to estimate the integral Eq. (2.10) for s independent data sets,

$$\begin{aligned} P(D_1, D_2, ..., D_s \mid M) &= \int P(D_1, D_2, ..., D_s \mid \vec{\theta}_1, \vec{\theta}_2, ..., \vec{\theta}_s, M) P(\vec{\theta}_1, \vec{\theta}_2, ..., \vec{\theta}_s \mid M) d\vec{\theta} \\ &= \prod_{j=1}^{s} \int P(D_j \mid \vec{\theta}_j, M) P(\vec{\theta}_j \mid M) d\vec{\theta}_j. \end{aligned} \tag{2.18}$$

The last equality results from each data set having its own, independent set of parameters. Taking the natural logarithm on both sides of Eq. (2.18), we obtain the following result, which we used to compute the values in Tbl. 2.3:

$$\ln P(D_1, D_2, ..., D_s \mid M) = \sum_{j=1}^{s} \ln P(D_j \mid M) \tag{2.19}$$

2.4.3 Expected values and uncertainty of fits

The fits and the error bars for curves for all of the fitted models in all Figures are the expected values and the standard deviations of the model curves over the posterior

probability distributions. That is,

$$\langle f(t \mid M)\rangle = \int f(t \mid \vec{\theta}, M) P(\vec{\theta} \mid D, M) d\vec{\theta}, \tag{2.20}$$

$$\mathrm{Var}(f(t \mid M)) = \int (f(t \mid \vec{\theta}, M) - \langle f(t \mid M)\rangle)^2 P(\vec{\theta} \mid D, M) d\vec{\theta}, \tag{2.21}$$

where $f(t \mid \vec{\theta}, M) = P_{\Delta t}(t \mid \vec{\theta}, M)$, and the posterior probability is

$$P(\vec{\theta} \mid D, M) = \frac{P(D \mid \vec{\theta}, M) P(\vec{\theta} \mid M)}{P(D \mid M)} = \frac{F(\vec{\theta})}{P(D \mid M)}. \tag{2.22}$$

As explained above, we used importance sampling to estimate the expectation values. For example, notice that Eq. (2.20) can be rewritten as

$$\langle f(t \mid M)\rangle = \frac{\int f(t \mid \vec{\theta}, M) F(\vec{\theta}) d\vec{\theta}}{\int F(\vec{\theta}) d\vec{\theta}}. \tag{2.23}$$

Using Eq. (2.15), this becomes

$$\left\langle \hat{f}(t \mid M) \right\rangle \approx \frac{\sum_{i=1}^{N} \frac{f(t|\vec{\theta}_i, M) \exp(\log F(\vec{\theta}_i) - \log F(\vec{\theta}^*))}{G(\vec{\theta}_i)}}{\sum_{i=1}^{N} \frac{\exp(\log F(\vec{\theta}_i) - \log F(\vec{\theta}^*))}{G(\vec{\theta}_i)}} \tag{2.24}$$

Similarly, for the variance, we have

$$\mathrm{Var}(f(t \mid M)) = \frac{\int (f(t \mid \vec{\theta}, M) - \langle f(t \mid M)\rangle)^2 F(\vec{\theta}) d\vec{\theta}}{\int F(\vec{\theta}) d\vec{\theta}}, \tag{2.25}$$

which results in

$$\mathrm{Var}(\hat{f}(t \mid M)) \approx \frac{\sum_{i=1}^{N} \frac{f(t|\vec{\theta}_i, M)^2 \exp(\log F(\vec{\theta}_i) - \log F(\vec{\theta}^*))}{G(\vec{\theta}_i)}}{\sum_{i=1}^{N} \frac{\exp(\log F(\vec{\theta}_i) - \log F(\vec{\theta}^*))}{G(\vec{\theta}_i)}} - \langle \hat{f}(t \mid M)\rangle^2. \tag{2.26}$$

Chapter 3

Inferring phenomenological models for biochemical reactions

3.1 Introduction

In this Chapter, we build a phenomenological model of another biochemical process that can be studied as a FPP: a single enzymatic chemical reaction.

The development of experimental techniques affording single-molecule precision characterization of biochemical transformations [53, 155, 64, 149, 22] has allowed the community to study the effects of stochasticity in these systems, generating sufficient data so that not only the mean times of the transformation, but the entire probability distributions of these times (the FPP distributions in our language) are well sampled. We focus on a reaction driven by the enzyme β-galactosidase, which transforms lactose into glucose and galactose. This is a storied model system, coming from the metabolism of the bacterium *E. coli*, and analyzed as far back as Jacob and Monod [68], with a lot of new developments recently. While the classic Michaelis-Menten model [92] provides a successful mathematical description of enzymatic kinetics for a large number of different enzymes [32, 45], relatively recent work [43] has shown this

description to be insufficient when single molecule measurements are performed, in particular, in the context of β-galactosidase. The immediate reaction by the community was to model the enzyme with a large, nearly continuum, structure of internal states, the so called multi-state interconverting model [43]. Here instead we explore if our methods for inference of the structure of completion time of FPP can build a simpler, more accurate, and more parsimonious model of the process.

We start this Chapter by using the multi-path model family from Chapter 2 to infer the phenomenological model of this enzymatic transformation. Such models are not necessarily a great match for the enzymatic catalysis happening without energy dissipation (as is the case for β-galactosidase) since each of the constitutive reactions in the multi-path model family is irreversible and, generically, does not obey detailed balance. Nonetheless, we show that a phenomenological model from this family, with only $M = 2$ paths (two-time scales), is flexible enough to explain the reaction times of a single enzyme β-galactosidase immersed at different (low and high) substrate concentrations. Crucially, in contrast to the the multi-state interconverting model [43], our model correctly fits the short-time behavior of the enzyme turnover time probability distribution as well, even though it is substantially simpler structurally. Further, guided by the analysis of the inferred models at short time scales, we introduce a different model hierarchy, which preserves the reversible nature of chemical transformations and hence is a better match for the enzymatic data. Using this model family, we build a minimal mechanistic Markovian model of the process, which successfully explains the experimental data. Finally, we show that this new modeling approach can predict the entire probability distribution characterizing the reaction times at non-tested concentrations, and not just for the concentrations, at which the model was trained. It is this ability to extrapolate that encourages us to think that our approach will be generally useful in understanding and predicting chemical kinetics processes for complex chemical systems.

3.2 Single enzyme reaction times

Going beyond deterministic chemical kinetics to understand the stochastic nature of enzymes (organic catalysts that speed up many biochemical processes inside cells) is crucial for understanding how cells perform their functions, and how these functions are dysregulated to cause diseases. At the cellular and molecular scale, we often expect stochastic fluctuations since these processes often involve chemical reactions with a small number of reactants of each species. These stochastic fluctuations are important because they can change the fundamental behavior of the system. Fortunately, the development of single-molecule imaging [155, 64, 22, 149] has allowed us to measure the distribution and fluctuations of enzymatic turnovers, which were unattainable from ensemble data, opening the door for understanding the role of stochasticity in these systems.

In particular, the enzyme β-galactosidase is important since it breaks down lactose into simpler sugars like glucose and galactose [69], and is thus a central part of metabolism in many organisms. This is a classic model system in cellular biochemistry [47, 68], which until recently was only studied on the deterministic chemical kinetics scale. In contrast, in a relatively recent pioneering work [43], a fluorogenic molecule was used as a substrate of a single β-galactosidase molecule submerged in a homogeneous substrate concentration. Upon hydrolysis this molecule generated a fluorescent burst that was detected in the confocal volume of the microscope. This allowed the researchers to generate large, high quality datasets describing the times between successive enzymatic conversion events. In their turn, knowing these times (the FP times in our language), allows us to understand the underlying stochasticity of the enzymatic reaction. In this work, we will focus on using these completion time distributions (reproduced from Ref. [43] for four different substrate concentrations in Fig. 3.1) to infer phenomenological and biophysically-realistic models of the enzyme kinetics, able to predict properties of the catalysis at arbitrary substrate concentra-

tions.

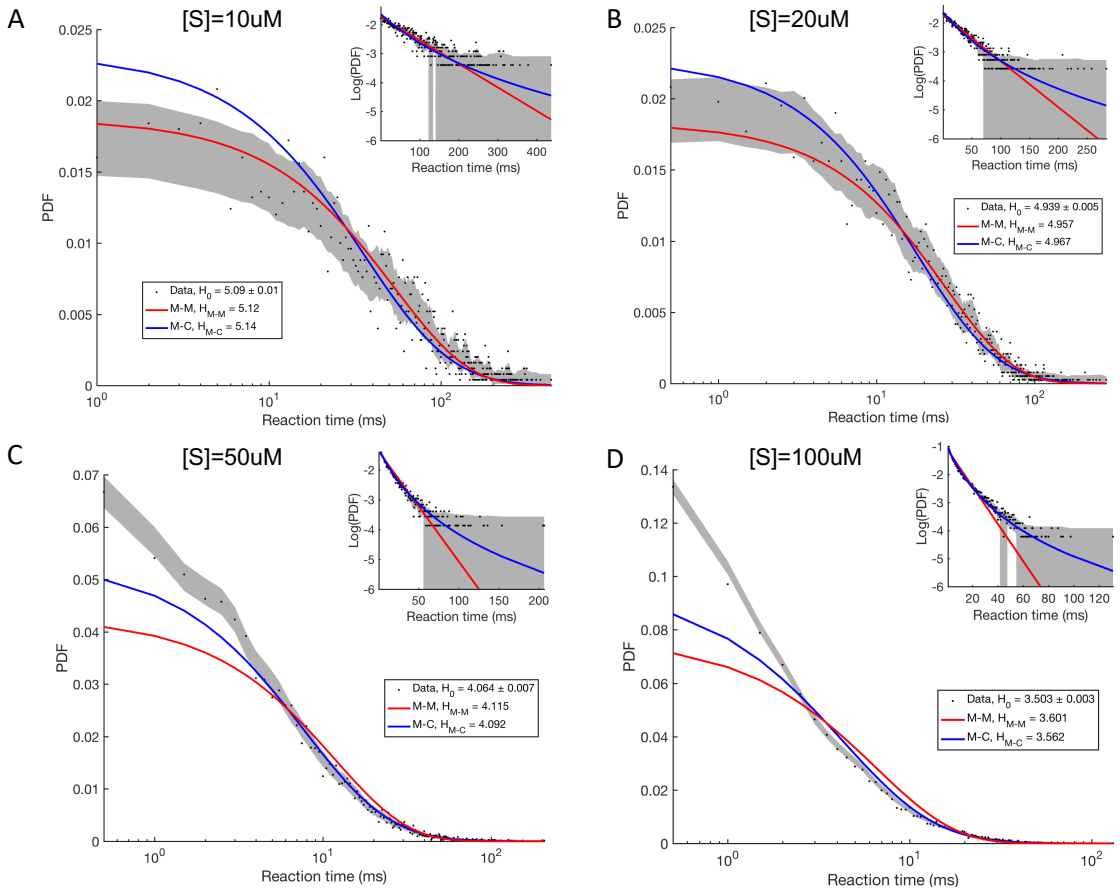

Figure 3.1: β-galoctosidase enzymatic catalysis from Ref. [43]. Dots indicate the experimental histograms for the enzymatic reaction times; grey bands are the standard error of every dot. Colored lines are two types of models of the data from Ref. [43]. Red lines represent the best fit of the classic Michaelis-Menten model, using the following parameters: $k_1 = 5 \times 10^7 \mathrm{M}^{-1}\mathrm{s}^{-1}$, $k_{-1} = 18300\ \mathrm{s}^{-1}$ and $k_2 = 730\ \mathrm{s}^{-1}$ (see Main Text for the explanation of the parameters). Blue lines are the multi-state interconverting model, with the same best fit parameters k_1, k_{-1}, and with $w(k_2) = 1/(b^a\Gamma(a))k_2^{a-1}\exp(-k_2/b)$, where $a = 4.2$ and $b = 220\ \mathrm{s}^{-1}$ (see Main Text for the parameter definitions). A-D: Catalysis for $[S] = 10, 20, 50, 100$ uM, respectively.

The simplest classic model for enzymatic catalysis is the Michaelis-Menten model, which has successfully explained data from enumerable experiments in over a hundred years of its existence [92]. For a single molecule enzymatic reaction, the chemical

kinetic diagram of the Michaelis-Menten reaction has the following structure

$$\mathrm{E} + \mathrm{S} \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} \mathrm{ES} \xrightarrow{k_2} \mathrm{P} + \mathrm{E}^{\mathrm{o}}, \quad \mathrm{E}^{\mathrm{o}} \xrightarrow{k_3} \mathrm{E}. \tag{3.1}$$

Here S is the substrate, which binds to the enzyme E with the rate k_1, forming an enzyme-substrate complex ES. The complex can either dissociate with the rate k_{-1}, or produce the product P with the rate k_2, leaving the enzyme in a modified state E^o. The enzyme then regenerates back to E with the rate k_3, and is again available to catalyze a new reaction. It is easy to show [85, 154] that, if $k_3 \gg k_2$, then the probability density of the waiting time τ for the entire enzymatic reaction to occur, $f(\tau)$, is given by:

$$f_{MM}(\tau) = \frac{k_1 k_2 [S]}{2A} [\exp{(A+B)\tau} - \exp{(B-A)\tau}], \tag{3.2}$$

where $A = \sqrt{(k_1[S] + k_{-1} + k_2)^2/4 - k_1 k_2 [S]}$ and $B = -(k_1[S] + k_{-1} + k_2)/2$.

Figure 3.1 reproduces the fits of the Michaelis-Menten model to the experimental data for the enzymatic turnover times from Ref. [43]. Clearly, the standard Michaelis-Menten model cannot explain both the short and the long time tails of the data histograms. To overcome this problem, a multi-state model involving a large number n of interconverting conformations of the enzyme was proposed [43], see Fig. 3.2 for the kinetic diagram of the proposed model. E, ES and ES^0 can exist in multiple stable molecular conformations. The main idea is to incorporate multiple time scales that can account for the non-monoexponential decay at high substrate concentrations. At low substrate concentrations (monoexponential decay), the enzyme substrate binding is rate limiting and it can be assume that the rates k_{1i} are narrowly distributed. Instead at high substrate concentrations k_{2i} becomes rate limiting and a broader distribution on these rates should account for the multiexponential decay, as long as the interconversion between ES_i is slow. Therefore, assuming that $k_{1i} = k_1$, $k_{-1i} =$

k_{-1} $\forall i$, while k_{2i} are not all equal and are sampled from the Gamma distribution, $w(k_2)$ and when the interconversion between conformations ES_i is slow compared to the enzymatic reactions (k_{2i} ($\beta_{ij}/k_{2i} \to 0$ and α_{ij}/k_{2i} small but nonzero, see details in [43]), the probability density of the enzymatic turnover time is given by:

$$f_{\mathrm{MC}}(\tau) = \int_0^\infty w(k_2) \frac{k_1 k_2 [S]}{2A} [\exp{(A+B)\tau} - \exp{(B-A)\tau}], \tag{3.3}$$

with A and B as defined in Eq. (3.2), and $w(k_2) = 1/(b^a \Gamma(a)) k_2^{a-1} \exp(-k_2/b)$. The best fits of this model with 4 parameters (k_1, k_{-1}, a and b) for different substrate concentrations are shown in Fig. 3.1. The multistate model is clearly better than the simpler Michaelis-Menten model, explaining, in particular, the long time asymptotics of the completion time distribution. However, the behavior at early times for high concentrations (Fig. 3.1 C-D) is still not well approximated. As in Chapter 2, we measure the quality of the fits by estimating the entropy of the data being fitted, $H_0 = -\sum_{i=1}^{N} p_i \ln p_i$ (using the NSB entropy estimator [102]), as well as the cross-entropy, $H = -\sum_{i=1}^{N} p_i \ln f(t_i \mid \vec{\theta}) \Delta t$, between the data and each of the proposed models (the cross-entropy corresponds to minus the normalized value of the log-likelihood). To the extent that the cross-entropy values are still far from H_0, cf. Fig. 3.4, not even the multi-state model is able to explain the data set well.

3.3 Results

3.3.1 Modeling the enzymatic turnover times using the multi path model family

We use the multi-path model family proposed in Chapter 2 to infer the best phenomenological model that can simultaneously explain the enzymatic turnover time histograms, Fig. 3.4, for all the measured substrate concentrations. Despite the

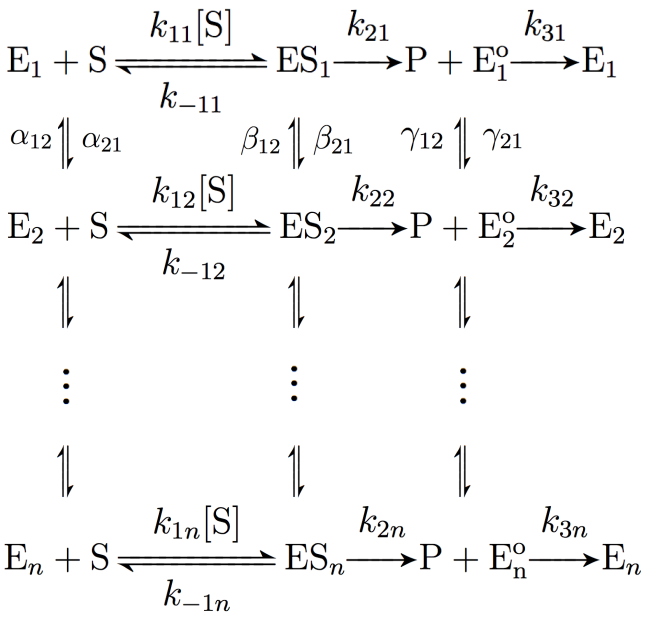

Figure 3.2: Inter-converting conformers model. E, ES and E_0 can exist in multiple molecular conformations. See details in the main text.

fact that the multi-path model family does not account for reversible reactions, like the Michaelis-Menten mechanism does, its completeness property ensures that after adding enough number of completion paths we should be able to get a perfect fit.

Figure 3.3 shows the best fits for $M \leq 3$ in our model family (see Chapter 2 for the algorithmic details and the model family description). Specifically, Tbl. 3.1 indicates that, by our Bayesian model selection criterion, a model with only two completion paths $M = 2$ (Fig. 2.2) is the most likely overall models to explain all four data sets simultaneously, by at least a factor of $\sim 10^{18}$. Notice that this model is capable of explaining both the long time, as well as the short time behavior of the distributions for all substrate concentrations, as can be seen from Fig. 3.3. Quantitatively this is supported by the fact that the cross-entropy values between the data and the model for $M = 2$ are all indistinguishable from H_0 within the statistical error. That is, not only is our $M = 2$ model better than the Michaelis-Menten or the interconverting

model, but also no other model can be a better fit for the experimental data analyzed here.

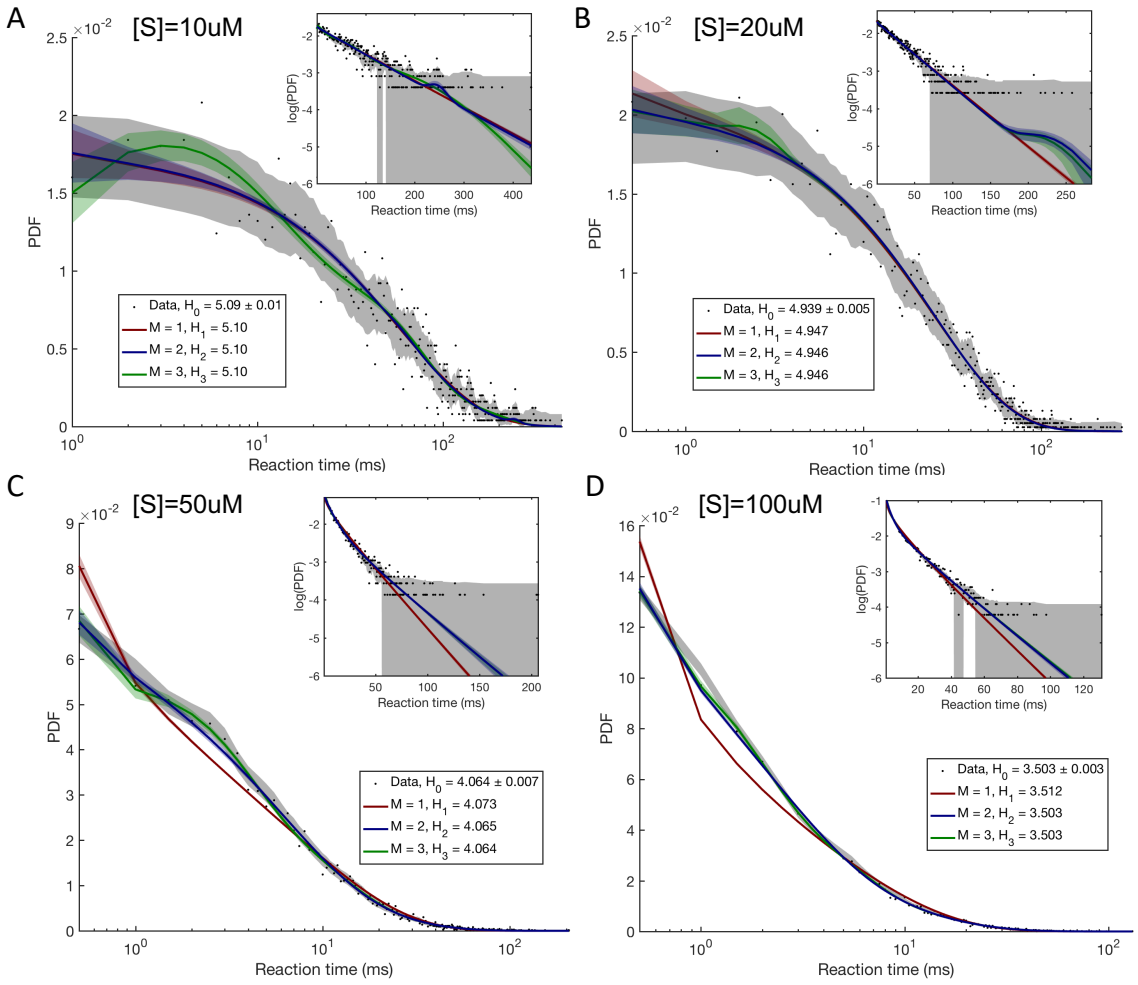

Figure 3.3: Best fit models with $M = 1, 2, 3$ for enzyme turnover time distribution. Color lines and color bands show the average fit line and the posterior standard deviation for the first three models in the family; red, blue, green correspond to $M = 1, 2, 3$, respectively. See Chapter 2 for algorithmic details for fitting the models and sampling from the posterior. The cross entropy between the data and the models (see legends) decreases towards the entropy value as the model complexity increases. A-D: $[S] = 10, 20, 50, 100$ uM, respectively.

Interestingly, by looking at how the parameters of the fitted model with $M = 2$ change as a function of the substrate concentration, we can recognize a drastic change in the properties of the process when the underlying substrate concentration

$\ln P(D \mid M)$					
M	[S]=10 uM	[S]=20 uM	[S]=50 uM	[S]=100 uM	Total
1	**-12784**	**-19029**	-29834	-58226	-119872
2	-12794	-19038	**-29787**	**-58086**	**-119705**
3	-12799	-19052	-29795	-58099	-119745
4	-12801	-19046	-29806	-58103	-119755

Table 3.1: Model selection results using multi-path model family for enzymatic turnover times. Posterior likelihoods for the first three models of the family for each substrate concentration are shown. Last column assesses the total fit. It shows that the data favors model with $M = 2$ by a factor of at least $\sim 10^{18}$.

transitions between the low ($[S] = 10, 20$ uM) and the high ($[S] = 50, 100$ uM) values, see Fig. 3.4. At low concentrations, there is effectively a single dominant path (Fig. 3.4C in blue) explaining the process. This path is very simple, with only a single transition step (Fig. 3.4B). Its mean completion time is concentration-dependent (Fig. 3.4A). We think this path is likely representing the enzyme-substrate binding process, which is difussion-limited and, at low concentrations, becomes also rate limiting. As the substrate concentration increases, the second path is required to explain the long time turnover statistics (cf. Figs. 3.4C, B.1). This suggests that the second time scale, probably related to internal conformational changes of the enzyme-substrate complex, becomes comparable to the enzyme-substrate search and binding time scale.

Ideally, similar to Chapter 2, we would like to use the parameter dependence on S from Fig. 3.4 to see whether our model with $M = 2$ is able to predict the reaction time distributions for substrate concentrations not included during the fitting procedure. However, since we are dealing with experimental data set, we have no access to data sets for non-measured concentrations to test our model. Instead, we chose to leave one data set (for $[S] = 20$ uM) out, and predict its reaction time distribution by linearly interpolating the corresponding parameter values from the fitted values at $[S] = 10$ and $[S] = 50$ uM, see Fig. 3.5A. Unfortunately, such interpolation does not result in an accurate prediction, cf. Fig. 3.5C. Similarly, the interpolation does not

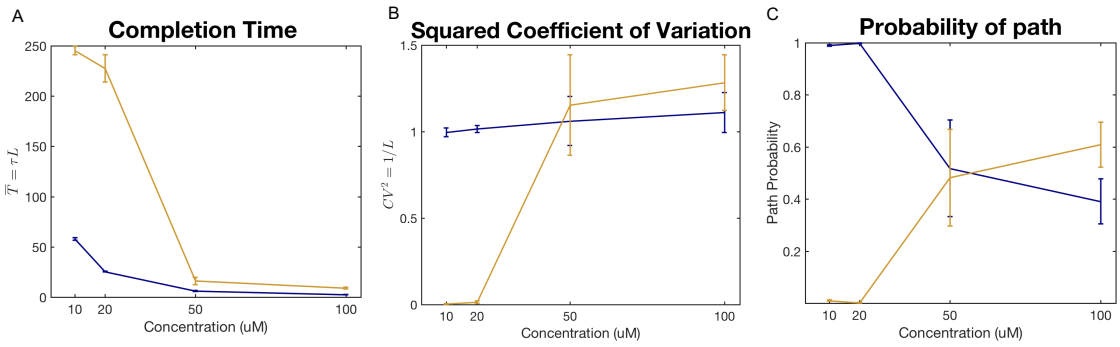

Figure 3.4: Properties of completion paths as a function of the substrate concentration for the best model, $M = 2$, across all experiments. A: Average completion times for each of the $M = 2$ independent paths are plotted as a function of the concentration $[S]$. Color (same in B and C) identifies paths according to how long they take to complete the process on average (yellow: slow path, blue: fast path). B: Coefficient of variation and C: probability of taking each of path as a function of $[S]$.

work well if we leave $[S] = 50$ uM data aside and try to predict it, cf. Fig. 3.5B, D. Notice that this does not contradict the fact that, for each of the high concentration values, ($[S] = 50$ and $[S] = 100$ uM) a model with $M = 2$ is the best possible model. In fact, as pointed out above, this model fits the data over all concentration values as well as is statistically possible. What this suggests, however, is that the multi-path model family is not a good match for data that comes from enzymatic catalysis, likely because the underlying mechanisms include reversible reactions.

We can repeat the analysis of Section 2.2.3 to establish the minimum network size required by any mechanistic model to explain the enzymatic turnover times.

Figure 3.6 shows the short time behavior of the cumulative distribution function (CDF) obtained from the best model, $M = 2$. For $[S] = 10, 20$ uM, at least the first 20% of the events are explained by a single path with only a single transition step. In contrast, for $[S] = 50, 100$ uM, the left tail of the distribution can only be explain using both completion paths with ≤ 1 transition steps. This suggests that even if a simple Michaelis-Menten mechanism does not explain the observations, it is

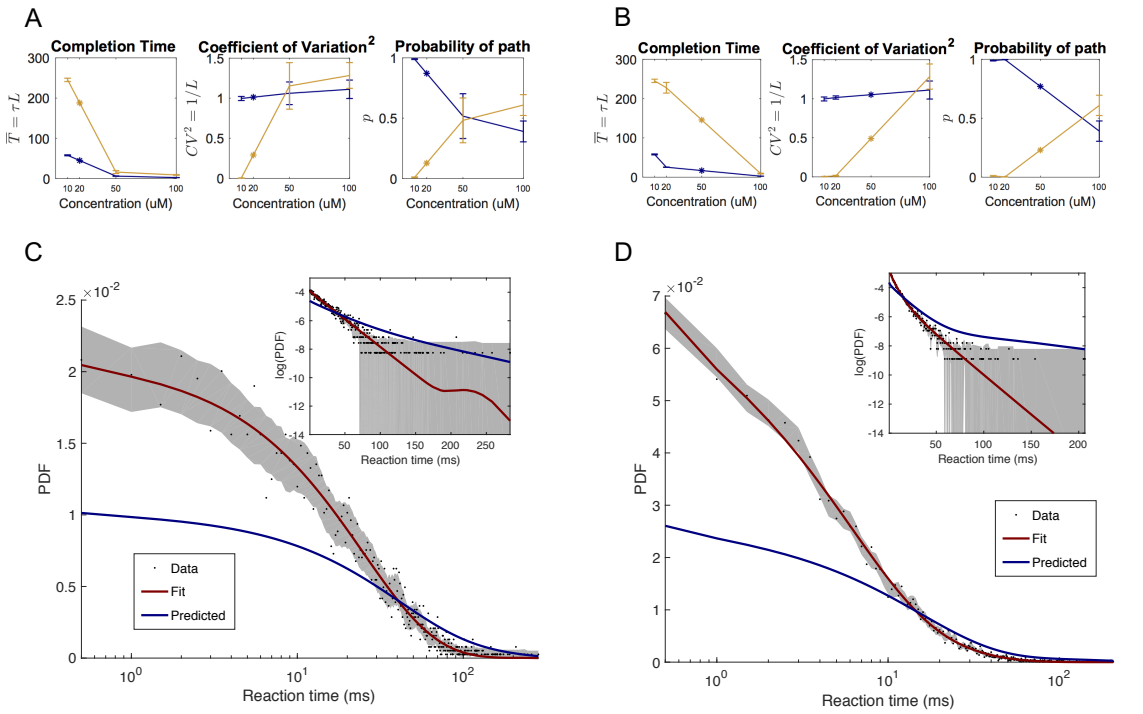

Figure 3.5: Predicted PDFs for reaction times using multi-path model family. A and B: Linear interpolation of parameter values for $[S] = 20$ uM and $[S] = 50$ uM, respectively, using nearby concentration values. C and D: Predicted models (in blue) obtained using interpolated parameter values from A and B respectively. Both predictions are off from the corresponding fitted model (in red), which fits the data nearly perfectly.

still possible that another *biochemically realistic* small network does. To explore this we created a different hierarchy of models, the *Biochemically realistic model family*, which we study in the next section.

3.3.2 Modeling the enzymatic turnover times using the Biophysically-realistic model family

In order to build a *biochemically realistic model family*, we must account for the reversibility of internal transitions, which happen before reaching the absorbing state. For example, for the Michaelis-Menten enzyme, the substrate can bind the enzyme

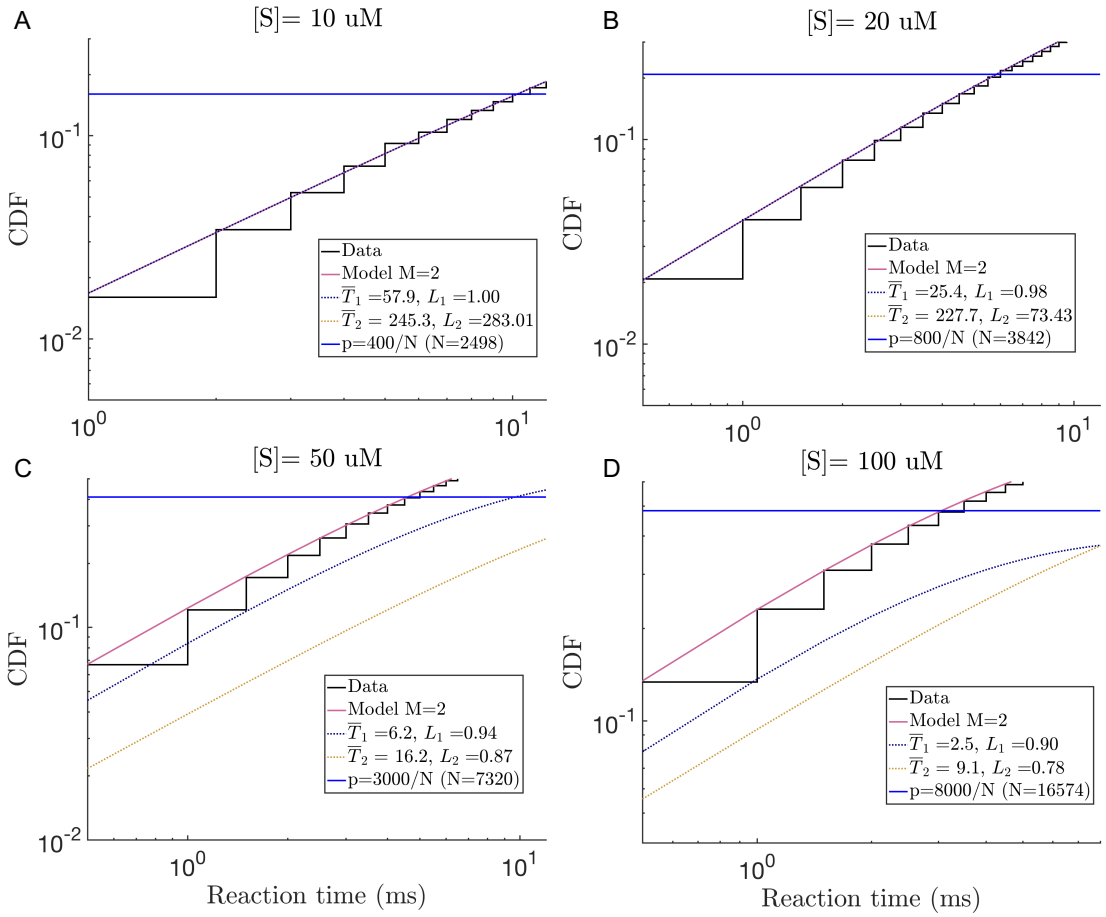

Figure 3.6: Decomposition of the CDFs at short time scales into two paths: Solid pink lines are the CDFs corresponding to the best-fit model with $M = 2$, and the dashed lines are the contributions from individual completion paths. Blue lines represent the completion probability of the first $\sim 20\%$ of the events for A: [S]= 10 uM and B: [S]= 20 uM. These events can be explained by a single path (blue dashed line) with a single completion step. For C: [S]=50 uM and D: [S]=100 uM, the left tail of the distribution ($\sim 50\%$) of events can be explained combining to paths with effectively $<= 1$ transition step. Time resolution is not enough to discriminate between paths in cases.

and get unbound before the actual catalysis begins. Since our goal is to build a minimal biochemically realistic network capable of explaining—and predicting—experimental observations, we start the model family with a single transition step model, see Fig. 3.7. This is the simplest network describing a FPP as explained in the previous chapter (Section 2.2.1). Next, we add an intrinsic (filled) node and two

edges forming a second non-reversible (two-step) completion path. Intrinsic nodes represent states where the enzyme-substrate molecules are bound and they will have a special property that we will describe later. Then, progressively, we add directional edges until the graph excluding the absorbing state becomes fully connected (note that a reversible reaction is accounted for by two directional edges). We then add an extrinsic (empty) node forming a new single-step completion path. Again, gradually we add directional edges until the graph excluding the absorbing state becomes fully connected. Then, we add another intrinsic state and the entire procedure is iterated. We refer to the union of all such models as the *Biochemically realistic model family*, cf. Fig. 3.7.

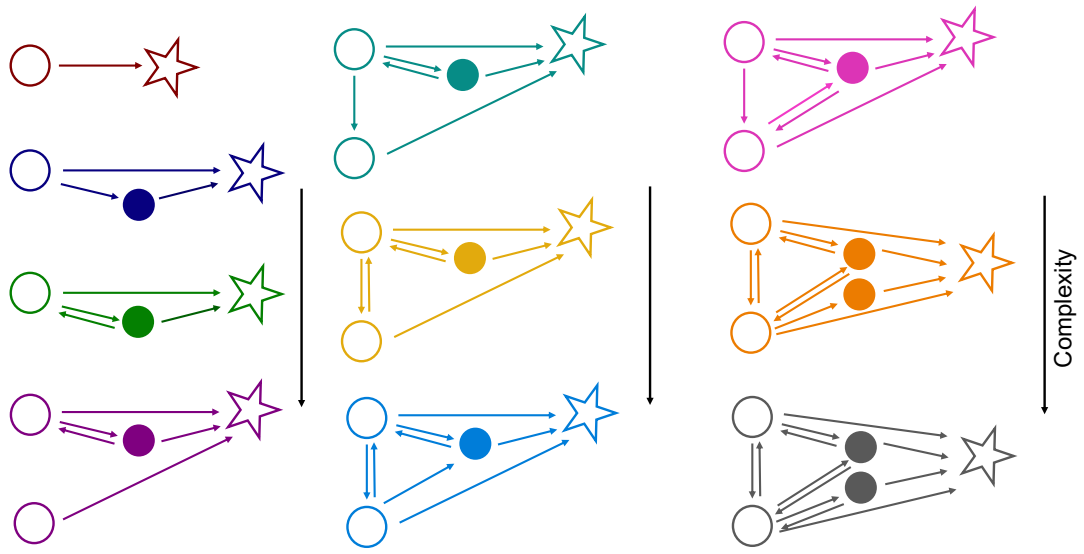

Figure 3.7: Biochemically-realistic model family. Kinetic schemes of the first 10 models in this hierarchy. Stars represent the absorbing state. Color-filled circles represent internal enzyme-substrate bounded conformations.

The networks in Fig. 3.7 represent the interacting structure of a Markovian process [144]. Let $\vec{p} = [p_1, p_2, ..., p_L]$ be a column vector representing the probabilities of each state in the kinetic diagram, where state 1 represents the absorbing state. This

probability distribution evolves according to the master equation:

$$\frac{d\vec{p}}{dt} = A\vec{p}, \tag{3.4}$$

where A represents the matrix describing the transition rates. For example A_{ij} represents the transition rate from state j to state i (note that this transition rate is time independent). For the relatively small dimensions that we are dealing with here, this system of first order differential equations can be easily solved after specifying the initial value $\vec{p}_0$ of the probability distribution. Notice that $p_1(t)$ represents the probability that the system is at the absorbing state at time t, which is the CDF of the FP time distribution, which we are trying to model. To fit this distribution to the data, we use a discretized version of the FP time distribution, as in Chapter 2:

$$P_{\Delta t}(t \mid \vec{\theta}, M_B) = p_1(t \mid \vec{\theta}, M_B) - p_1(t - \Delta t \mid \vec{\theta}, M_B), \tag{3.5}$$

where $\vec{\theta}$ represent the model parameters and M_B stands for the specific kinetic diagram in Fig. 3.7. Adding an extrinsic node to the kinetic diagram adds a parameter to be fitted: namely, the value of the probability of being in that state at time $t = 0$. All intrinsic nodes are assumed to have zero probability initially (the enzyme is not bound to the substrate), so no new parameters are associated with them. Finally, every directional arrow has a single kinetic rate associated with it, which needs to be fitted. In other words, at any stage of model building, going to the next model in the hierarchy adds at most two parameters to the previous model: we either add a single edge and zero nodes (1 parameter), an intrinsic node and two edges (two parameters) or an extrinsic node and one edge (2 parameters one of them accounting for the initial condition of the new node). Thus the statistical complexity does not increase abruptly in each of these steps, so that a good balance between the under- and the over-fitting is likely to be found for each particular data set.

It is easy to show that the biochemically-realistic model family satisfies the properties to make the statistical inference *consistent*. By simple inspection of Fig. 3.7, we observe that the model family is *nested*, since each network is a subnetwork of the subsequent one. Furthermore, the model family is also *complete*, assuming that the process being fitted can be described as a discrete state space Markov chain. In fact, any possible network with one absorbing state will be a subnetwork of at least one model in the hierarchy. This follows from the fact that for any number of nodes, excluding the absorbing state, the fully connected network will be eventually incorporated in the hierarchy.

Consequently, we can implement Bayesian model selection (see Section 2.4.2) to estimate the most probable model M_B within the *Biochemically-realistic model family*. The results for the β-galoctosidase data sets are shown in Tbl. 3.2. Quantitatively, the best model capable of representing all data sets simultaneously is the $M_B = 4$ (bottom left corner in Fig. 3.7). The best fits for each of the first five models in the hierarchy are shown in Fig. 3.8. Notice that the cross-entropy values between the data and the model ($M_B = 4$) are all indistinguishable from the entropy, H_0, within statistical error, (see Section 2.2.2 for details). Thus a simple network with only 4 nodes (Fig. 3.9A) can explain all experimental data sets better or equal than any other model. In this sense, this model is as good as the multi-path model family inferred in the previous section. Notice that the multi-path model had (5 parameters) and only one more parameter is required to build a biochemically realistic model.

Interestingly, this model strongly suggests the need to incorporate a path with an intermediate state representing the enzyme-substrate internal conformation. Furthermore, it suggests that the data can be fit by the model where there is not one, but two different configurations at the start of the process, so that enzyme may exist in two different functional states.

By looking at the changes of the fitted parameters for the $M_B = 4$ model as a

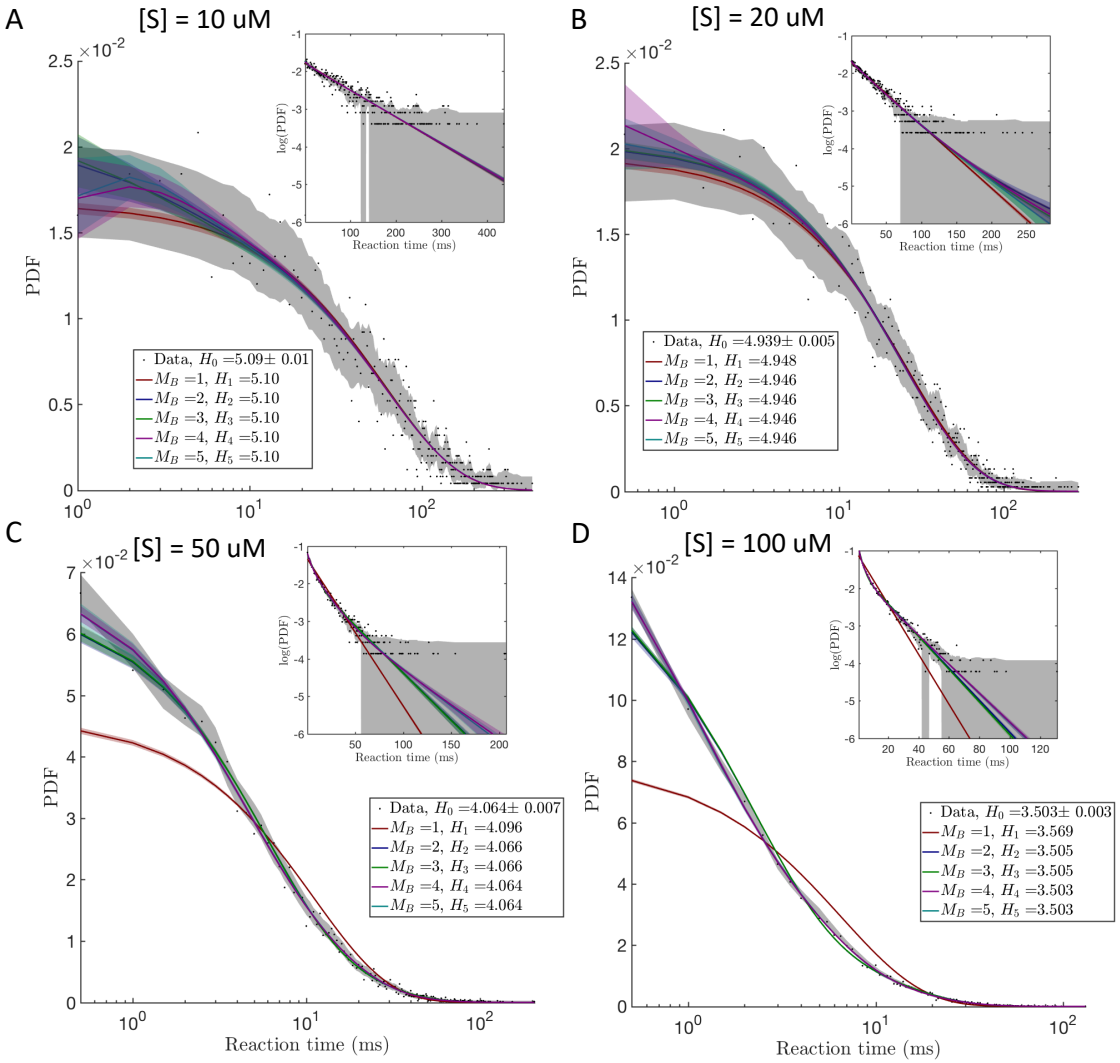

Figure 3.8: Best fit models of the Biochemically-realistic model family for the enzyme turnover time distribution. Color lines and color bands show the average fit line and the posterior standard deviation for the first five models in the family. The cross entropy between the data and the models (see legends) decreases towards the data entropy value H_0 as the model complexity increases. A-D: Substrate concentrations $[S] = 10, 20, 50, 100$ uM, respectively.

$\ln(P(M \mid \vec{x}))$					
c	[S]=10 uM	[S]=20 uM	[S]=50 uM	[S]=100 uM	Total
1	**-12754**	-19016	-29990	-59159	-120919
2	-12759	-19020	-29775	-58098	-119651
3	-12759	-19022	-29777	-58098	-119656
4	-12756	**-19015**	**-29774**	**-58069**	**-119614**
5	-12762	-19016	-29776	-58072	-119622

Table 3.2: Model selection results for the Biochemically-realistic model family to the enzymatic turnover data. Posterior likelihoods for the first five models of the hierarchy are shown. Last column assesses the total fit over all substrate concentrations. Data favors the model with $M_B = 4$ by a factor of $\sim 10^4$.

function of the substrate concentration, (Fig. 3.9B), we observe that one of the two paths with the single-step transition is concentration-dependent (see k_{21} in Fig. 3.9B). This is consistent with what we found using the inferred model from the multi-path model family. In addition, we see that the transition rate k_{31} does not depend on the substrate concentration, within the admittedly large error bars of the inference process. This is expected since it suggests that, once the substrate binds the enzyme, the concentration of substrate molecules in the solvent does not affect the velocity of the catalysis.

Lastly, using the procedure similar to Section 3.3.1, we make predictions for the completion time PDFs at concentrations $[S] = 20, 50$ uM. Here we linearly interpolate each of the k_{ij} parameter using the nearby concentration values in Fig. 3.9B. The predictions are shown in Fig. 3.10. These predictions are better than those obtained using the multi-path model family. The cross-entropy values for the predicted distributions are $H = 4.957$ and $H = 4.100$ for $[S] = 20$uM and $[S] = 50$ respectively. To the extent that these values are close to the corresponding entropy values of the data (cf. Fig. 3.8) the predictions are reasonably good. This suggests that unlike the multi-path model family, the biochemically-realistic model family offers not only a minimal biochemically plausible description, but also, as has been argued before in [35], models that capture better the biophysical constraints are able to make better

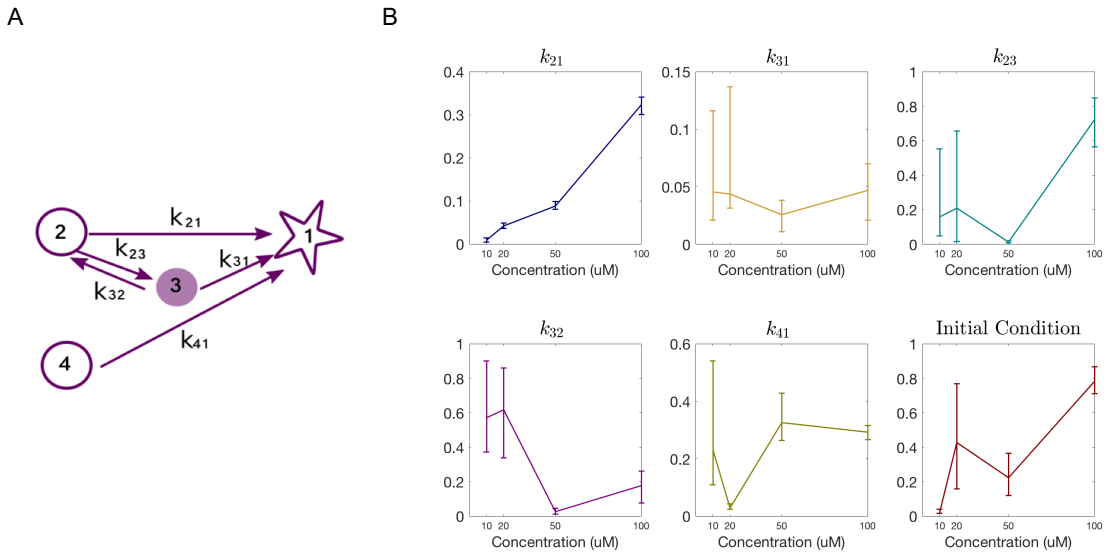

Figure 3.9: Properties of the best fit model from the biochemically-realistic model family for the β-galactosidase turnover times. A: The kinetic scheme of the best fit model, $M_B = 4$. Model parameters associated with each transition are labeled. B: The model parameters as a function of the substrate concentration $[S]$.

predictions.

3.4 Discussion

In this Chapter, we showed how a phenomenological inference approach can successfully explain the behavior of single enzyme reactions. We first used the multi-path model family developed in Chapter 2, to infer the best model that can explain the β-galactosidase enzyme turnover times under four different substrate concentrations. We show that a simple model with $M = 2$ completion paths can accurately explain the experimental data. In fact, our model achieves better fits than those proposed in the literature [43], where the right tail of the distribution could not be explained. Even more, we show that no other model can better fit the data than ours.

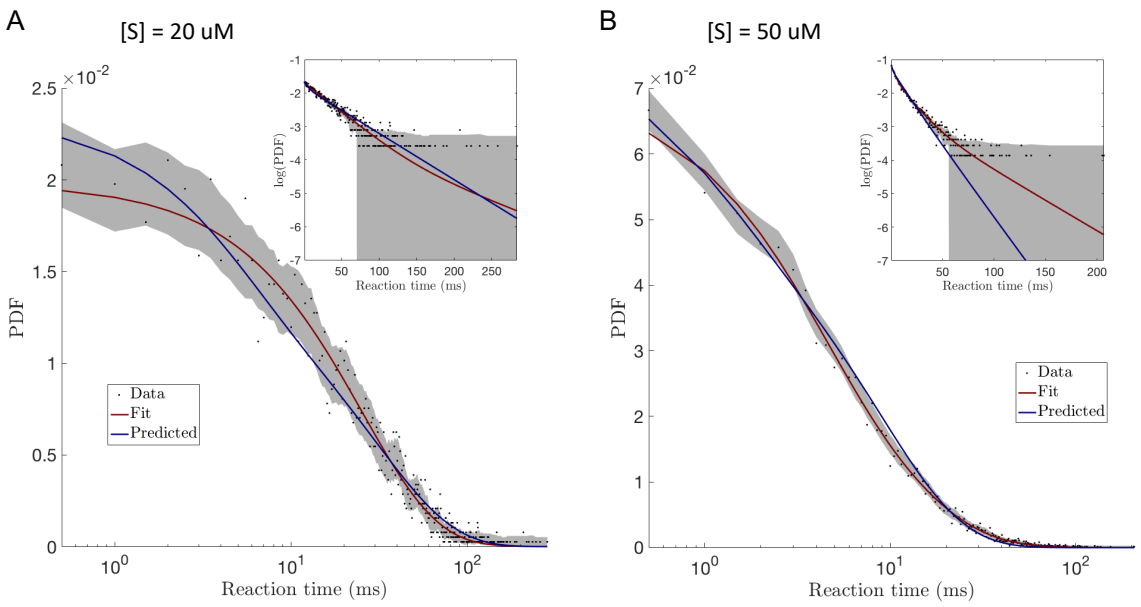

Figure 3.10: Predicted completion time PDFs for enzyme turnover times using the biochemically-realistic model family. Predicted models (in blue) were obtained by interpolating the corresponding parameter values from Fig. 3.9B. Best fits done directly to the data are shown in red for comparison. A: and B: predictions for $[S] = 20, 50$ uM respectively.

Nevertheless, predictions from the multi-path model family for the entire FP time distribution for substrate concentrations not included in the fitting procedure were poor. We attribute this to the fact that despite offering a simple model with a nearly perfect fit, this model family does not account for the reversible transitions present in biochemical reactions, does not account for the fact that only a handful of parameters need to be concentration dependent, and hence, generally, is not the best match for the data. Based on the study of the early time behavior of the FP time distributions we could rule out the possibility that a large network would be required to explain the experimental data set. Therefore, in the second part of this Chapter we formally introduced a different mathematical structure, the biophysically-realistic model family, specially designed to find the minimal biochemically realistic network that could explain the experimental observations. We showed that a simple Markov

network of four nodes can explain all experimental data sets better than the inter-converting conformational model and in fact better or equal than any other model. Moreover, predictions for the FP time distributions were considerably improved. Our model suggests that the enzyme should exist at least in two different functional states or conformations, determining different starting conditions of the process. On the other hand, only one conformation for the enzyme-substrate configuration is required to accurately explain the observations.

These results shows that, in principle many hierarchies can produce parsimonious models capable of explaining the experimental data. However, those that can better represent the biophysical constraints of the system are able to make better predictions.

Chapter 4

A framework for studying behavioral evolution by reconstructing ancestral repertoires

(This chapter is based on: *A framework for studying behavioral evolution by reconstructing ancestral repertoires.* Damián G. Hernández *, Catalina Rivera*, Jessica Cande, Baohua Zhou, David L. Stern,Gordon J. Berman. Submitted to *eLife*.)

*These authors contributing equally to this work.

4.1 Introduction

Behavior is one of the most rapidly evolving phenotypes, with notable differences even between closely-related species [84, 88]. Variable behaviors and rapid behavioral evolution likely allows species to adapt rapidly to new or varying environments [4, 150]. Despite the importance of animal behavior, progress in revealing the genetic basis

of behavioral evolution has been slow [50, 156, 41, 117]. In contrast, recent decades have seen significant progress in understanding the genetic causes of morphological evolution [151, 119, 81, 127].

While there are many potential reasons for the discrepancy between studies of behavioral and morphological evolution, including the lack of a fossil record for behavior, a key difficulty is identifying which aspects of an animal's development and physiology are responsible for the observed changes in animals' actions. Changes in behavior along a phylogeny could emerge from alterations in the developmental patterning of neural circuitry (e.g., brain networks, descending commands, central pattern generators), hormonal regulation that affects the expression of behaviors, or the gross morphology of an animal's body or limbs [6]. Each of these possibilities could result in behavioral effects at different, yet overlapping, scales – from muscle twitches to stereotyped behaviors to longer-lived states like foraging or courtship or aging that may control the relative frequency of a given behavior – making it difficult to identify the precise aspects of behavior that are changing.

To address these difficulties, the standard approach in the study of behavioral evolution has been to identify focal behaviors that exhibit robust differences between species, such as courtship behavior in fruit flies [23, 25, 39] or burrow formation in deermice [147, 62]. With these robust differences in phenotype, it is possible to perform analyses that isolate regions of the genome that correlate with quantitative changes in the performance of the focal behavior. However, there tend to be multiple such regions identified, each containing many genes. Given the large number of putative genes involved, combined with the possibility of epistatic interactions between loci, identification of the contributions of individual genes to behavioral evolution has moved slowly.

An alternative approach to focusing on single behaviors is to examine the full repertoire of movements that an animal performs. By identifying sets of behaviors

that evolve together, it may be possible to identify regulators of these suites of behaviors. This approach has been made possible by recent progress in unsupervised identification of animal behaviors across length and time scales [11, 20]. In this study, we introduce a quantitative framework for studying the evolutionary dynamics of large suites of behavior. We have focused initially on fruit flies, which provide a convenient model for this problem because they exhibit a wide range of complex behaviors and unsupervised approaches can be used to map all of the animal movements captured in video recordings [13, 24, 12].

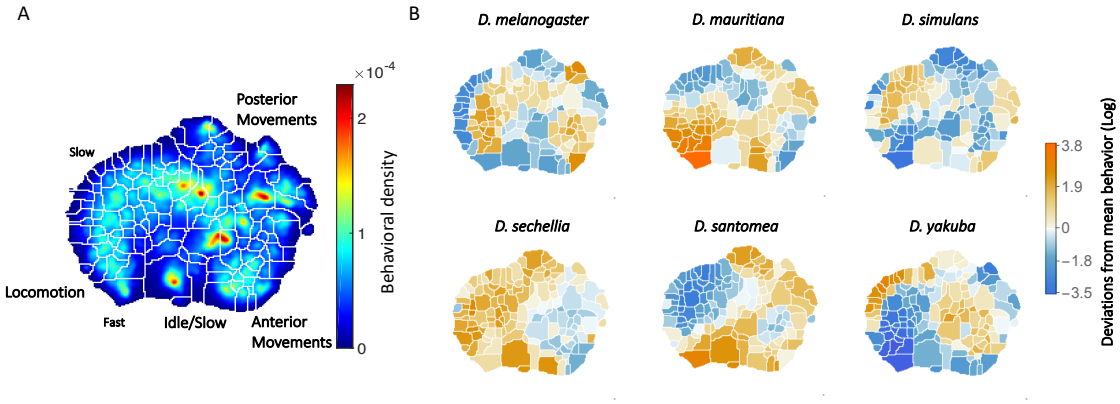

Figure 4.1: Behavioral repertoires of *Drosophila*. A: The behavioral space probability density function, obtained using the unsupervised approach described in [13] on the entire data set of 561 individuals across all species. Coarse grained behaviors corresponding to the different types of movements exhibited in the map are shown as well. B: The relative performance of each of the 134 stereotyped behaviors for each of the six species. Each region here represents a behavior, and the color scale indicates the logarithm of the fraction of time that species performs the specified behavior divided by the average across all species.

We recorded movies of isolated male flies from six species in a nearly stimulus-free environment. Because we did not record flies experiencing social and other environmental cues, we did not observe many charismatic natural behaviors, such as courtship and aggression. Nevertheless, we found that the behaviors they performed, includ-

ing walking and grooming, contain species-specific information. We thus hypothesized that our quantitative representations of behaviors could be studied in an evolutionary context. To infer the evolutionary trajectories of behavioral evolution, we estimated ancestral behavioral repertoires with a Generalized Linear Mixed Model (GLMM) approach [55], which builds upon Felsenstein's approach to reconstructing ancestral states [44, 56]. Using these results, we develop a framework that allows us to model the behavioral traits that co-vary both within a species and along the phylogeny. We find that within-species variance is related primarily to long-lasting internal states of the animal, what might be called a fly's "mood," and that inter-species variance can capture how disparate behaviors may evolve together. This latter finding points towards the presence of higher-order behavioral traits that may be amenable to further evolutionary and genetic analysis.

4.2 Experiments and behavioral quantification

We captured video recordings of all behaviors performed by single flies isolated in a largely featureless environment for multiple individuals from six species of the *Drosophila melanogaster* species subgroup: *D. mauritiana*, *D. melanogaster*, *D. santomea*, *D. sechellia*, *D. simulans*, and *D. yakuba* [24]. Although the animals could not jump or fly in these chambers and were not expected to exhibit social or feeding behaviors, the flies displayed a variety of complex behaviors, including locomotion and grooming. Each of these behaviors involves multiple body parts that move at varying time scales. The species studied here were chosen because their phylogenetic relationships are well understood [42, 104, 29, 116] (summarized in the tree seen in Fig. 4.3), and genetic tools are available for most of these species [126]. Since a single strain represents a genomic "snapshot" of each species, we assayed individuals from multiple strains from each species to attempt to capture species-specific differences,

and not variation specific to particular strains (see Materials and Methods). In total, we collected data from 561 flies, each measured for an hour at a sampling rate of 100 Hz.

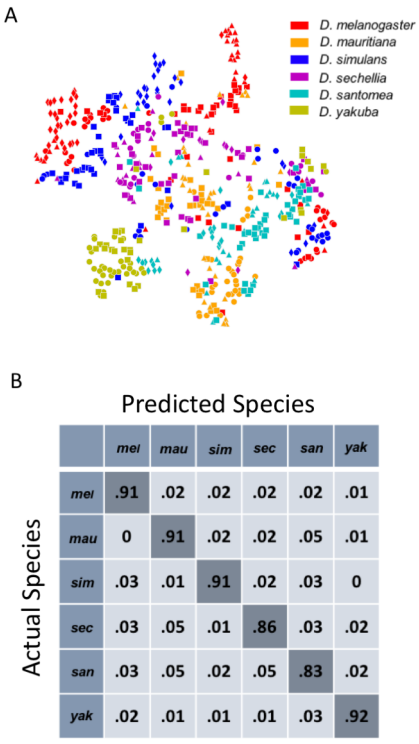

Figure 4.2: Classification of fly species based on behavioral repertoires. A: A t-SNE embedding of the behavioral repertoires shows that behavioral repertoires contain some species-specific information. Each dot represents one individual fly, with different colors representing different species and different symbols with the same color representing different strains within the same species. The distance matrix (561 by 561) used to create the embedding is the Jensen-Shannon divergence between the behavioral densities of individual flies. B: Confusion matrix for the logistic regression with each row normalized. All the values are averaged from 100 different trials. The standard error is less than 0.01 for the diagonal elements and less than 0.005 for each of the off-diagonal elements.

While previous studies have identified differences in specific behaviors, such as courtship behavior, between these species [23, 39, 157, 3], here we assayed the full repertoire of behaviors the flies performed in the arena, with the aim of identifying combinations of behaviors that may be evolving together. To measure this repertoire,

we used a previously-described behavior mapping method [13, 24] that starts from raw video images and attempts to find each animal's stereotyped movements in an unsupervised manner. The output of this method is a two-dimensional probability density function (PDF) that contains many peaks and valleys (Fig. 4.1A), where each peak corresponds to a different stereotyped behavior (e.g., right wing grooming, proboscis extension, running, etc).

Briefly, to create the density plots, raw video images were rotationally and translationally aligned to create an egocentric frame for the fly. The transformed images were decomposed using Principal Components Analysis into a low-dimensional set of time series. For each of these postural mode time series, a Morlet wavelet transform was applied, obtaining a local spectrogram between 1 Hz and 50 Hz (the Nyquist frequency). After normalization, each point in time was mapped using t-SNE [143] into a two dimensional plane. Finally, convolving these points with a two-dimensional gaussian and applying the watershed transform [91], produced 134 different regions, each of these containing a single local maximum of probability density that corresponds to a particular stereotypical behavior. Thus, by integrating the density of the region for a particular fly, we can associate to each of them a 134-dimensional real-valued vector that represents the probability of the fly performing a certain stereotyped behavior at a given time. We will refer to this quantity as the animal's *behavioral vector* $\vec{P}$.

The behavioral map averaged across all six species is shown in Fig. 4.1A and displays a pattern of movements similar to those we found in previous work, where locomotion, idle/slow, anterior/posterior movements, etc. are segregated into different regions [13, 24]. Averaging across all individuals of each species, we found the mean behavioral vector for each species (Fig. 4.1B) and observed that each species performs certain behaviors with different probabilities. For example, *D. mauritiana* individuals spend more time performing fast locomotion than all other species on average, and *D. yakuba* individuals spend much of their time performing an almost

species-unique type of slow locomotion, but little time running quickly.

These average probability maps provide some insight into potential species differences, but to identify species-specific behaviors, we also need to account for variation in the probability that individuals of each species perform each behavior. One way to address this problem is to ask whether an individual's species identity can be predicted solely from its multi-dimensional behavioral vector. To explore this question, we first used t-SNE to project all 561 individuals into a 2 dimensional plane (Fig. 4.2A), using the Jensen-Shannon divergence as the distance metric between individual behavioral vectors. In this plot, different colors represent different species, and different symbols with the same color represent different strains within the same species. Although there is not a clear segregation of all species in this plane, the distribution of species is far from random, with individuals from the same species tending to group near to each other.

To quantify this observation, we applied a multinomial logistic regression classifier that performed a six-way classification based solely on the high-dimensional behavioral vectors. After training, the classifier correctly classified $89 \pm .2\%$ of vectors (using a randomly-selected test set of 30% of the entire data set). Moreover, the confusion matrix (Fig. 4.2B) revealed no systematic misclassifications bias amongst the species. Note that we have used a relatively simple classifier compared to modern deep learning methods [51], so these results likely represent a lower bound on the distinguishability of the behavioral vectors. Thus, behavioral vectors appear to contain considerable species-specific information. We therefore proceeded to explore how these behavioral vectors may have evolved along the phylogeny.

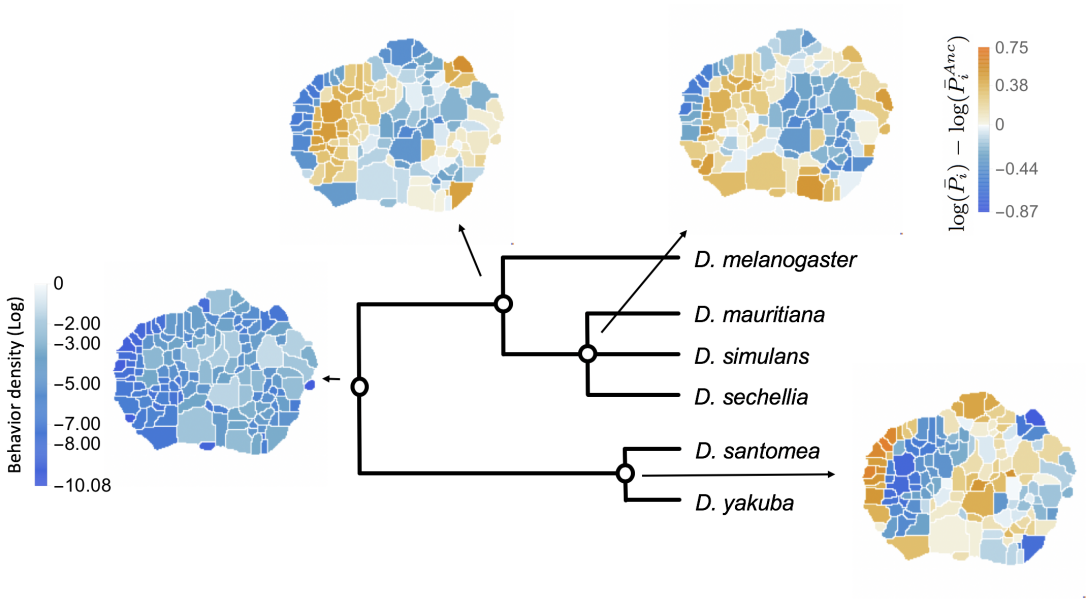

Figure 4.3: Reconstructed behavioral repertoires using the GLMM. Inferred probabilities of the behavioral traits for the ancestral states are plotted in logarithmic scale. Except for the ancestral root, other ancestral states are plotted with respect to the closest ancestor. Here, all but the root ancestor are plotted with respect to their closest ancestral state. Therefore, for each behavioral trait, i, we show: $\log(\bar{P}_i) - \log(\bar{P}_i^{Anc})$, where $\bar{P}i$ and $\bar{P}_i^{Anc}$ correspond to the inferred mean behavioral trait for the given ancestor and its closest ancestor, respectively.

4.3 Reconstructing Ancestral Behavioral Repertoires

Multiple methods have been proposed for reconstructing ancestral states solely from data collected from extant species [44, 158]. These methods generally fall into two camps: parsimony reconstruction, which attempts to reconstruct evolutionary history with the fewest number of evolutionary changes [34], and diffusion-processes, which model evolution as a random walk on a multi-dimensional landscape [57]. Given the high-dimensional behavioral vectors that we are attempting to model, a diffusion process is more likely to capture the inter-trait correlations that we would like to understand. Thus, we focus on a diffusion-based model here.

Given a phylogeny for a collection of species, we modeled how species-specific complexes of behaviors might have emerged. Specifically, we assumed that each behavior is a quantitative trait, that is, each behavioral difference results from the additive effects of many genetic loci, each of small effect. We do not, however, assume that all behaviors evolve independently of each other. Thus, we are interested in predicting (1) how behaviors co-vary and (2) whether intra- and inter-species variation can be separated to identify independently evolving sets or linear combinations of behaviors.

We assumed that the flies' behaviors evolved via a diffusion process, where initially the process starts at the common ancestor behavioral representation and eventually each individual's trajectory performs a random walk with Gaussian noise along the known phylogenetic tree. Note that this is a less stringent assumption than neutrality, as multiple traits under selection may evolve in a correlated manner. More precisely, we fit a Multi-response Generalized Linear Mixed Model (GLMM) to the data, using the approach described in [55]:

$$\vec{l} = \vec{\mu} + \vec{\rho} + \vec{e} \tag{4.1}$$

where $\vec{l} = (l_1, ..., l_{K=134})$ denotes the logarithm of the *behavioral vector* $\vec{P}$ for each individual, $\vec{\mu}$ is the mean behavior of the common ancestor (treated as the fixed effects of this model), and $\vec{\rho}$ and $\vec{e}$ are the random effects corresponding to the phylogenetic and individual variability, respectively. We assume that these random effects are generated from the multi-dimensional normal distributions $\mathcal{N}(\vec{0}, A \otimes V^{(a)})$ (phylogenetic) and $\mathcal{N}(\vec{0}, I \otimes V^{(e)})$ (individual). Here, the matrix A represents the information contained in the phylogenetic tree, with A_{ij} being proportional to the length of the path from the most recent common ancestor of species i and j to the main ancestor. This matrix is normalized so that the diagonal elements are all equal to 1. I is the identity matrix, and $V^{(a)}$ and $V^{(e)}$ are the covariance matrices that govern the process. We fit μ, $V^{(a)}$, and $V^{(e)}$ using Markov Chain Monte Carlo (MCMC) simulation (see Materials and Methods). We checked that the MCMC converged using the Gelman-Rubin

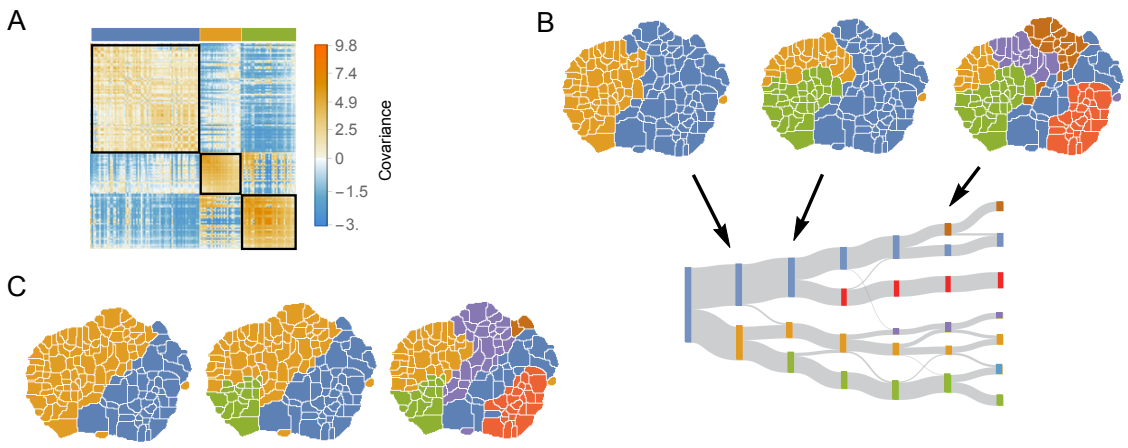

Figure 4.4: The structure of variability between flies of the same species relates to long timescale transitions in behavior. A: The intra-species behavioral covariance matrix ($V^{(e)}$), with columns and rows ordered via an information-based clustering algorithm [122]. The black squares represent behaviors that are grouped together in the three cluster solution. B: Behavioral map representation of the clustering solutions. The two, three, and six cluster solutions are shown on top (colors on the three cluster solution match those above the plot in A). The clusters are all spatially contiguous and break down hierarchically (see Fig. C.3 for more examples). C: Clustering structure of the behavioral space obtained finding the optimally predictive groups of behaviors (see text for details). Note how these clusterings are nearly the same as the clusterings in B, despite having been derived from an entirely independent measure.

diagnostic (see Materials and Methods, Fig. C.1). In addition to the inferred behavioral states corresponding to the common ancestor, $\bar{P}^{Anc}$, we also reconstructed the mean behavioral representations for the intermediate ancestors (Fig. 4.3). Further validation of our results corresponding to the current species behavior is shown in Fig. C.2.

4.4 Individual variability and long timescale correlations

While it is not possible to directly test the accuracy of our ancestral state reconstructions, the inferred covariance matrices generate predictions about genetic correlations

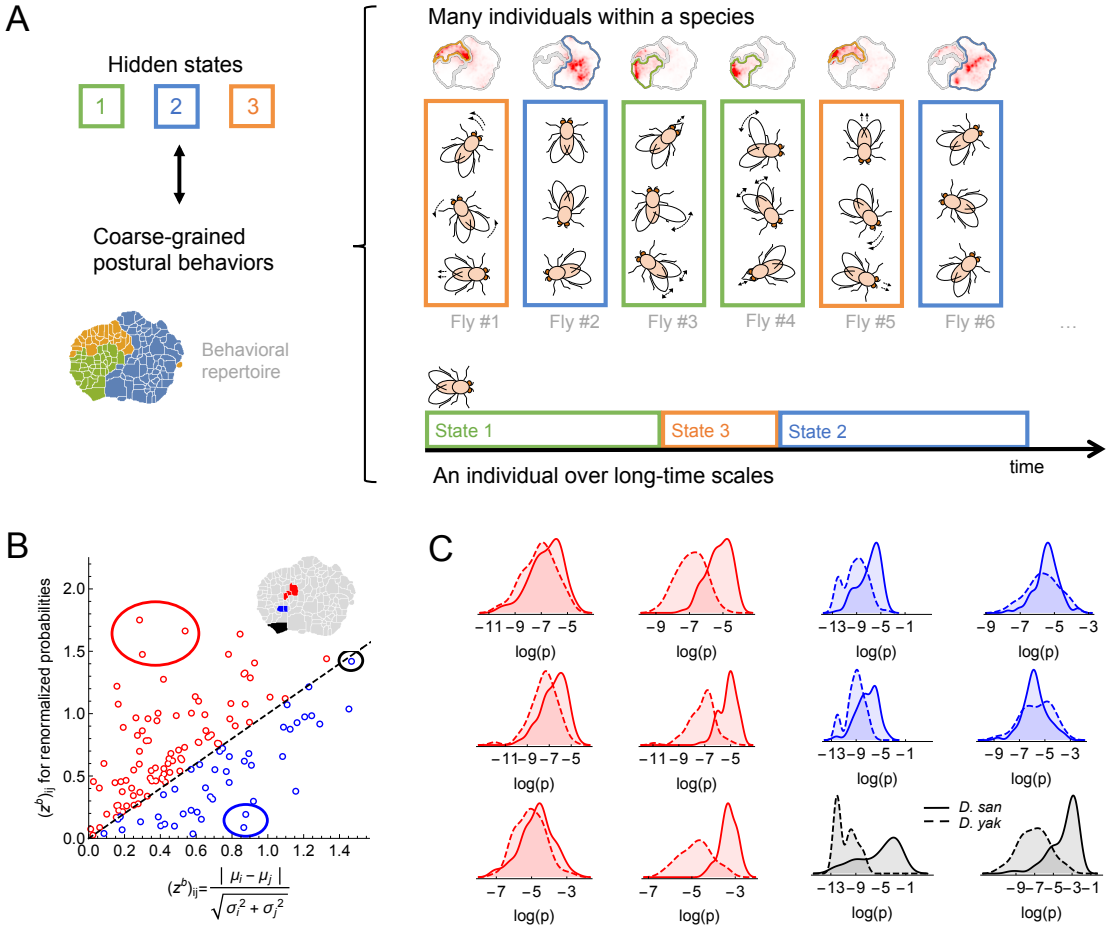

Figure 4.5: Variability within a species, long timescale transitions, and hidden states modulating behavior. A: A cartoon of the hypothesized relation between individual variability within a species and long timescale transitions through hidden states. B: Accounting for the long timescale dynamics - by adjusting for the amount of time spent in each coarse-grained region (here, the six cluster solution at the top right of Fig. 4.4C) - affects the measured behavioral distributions between *D. santomea* and *D. yakuba*. Shown is the comparison of the Mahalanobis distance ($(z^b)_{ij}$) between behavioral distributions before (x-axis) and after (y-axis) adjusting. C: Kernel density estimates of the distributions for the circled behaviors in B) on the left before (left) and after (right) adjustments. Solid lines represent *D. santomea* and dashed lines represent *D. yakuba*.

that are, in principle, testable. We therefore focus on our fitted covariance matrix, $V^{(e)} \in \Re^{134\times134}$, which accounts for within-species random effects.

We first note that $V^{(e)}$ exhibits a modular structure (Fig. 4.4A). After rearranging

the behavior order via an information-based clustering procedure [122], we see that a block diagonal pattern emerges, with positive correlations lying within the blocks and negative correlations lying off the diagonal. This clustering approach minimizes the functional $\mathcal{F} = \langle d \rangle + \beta I(C; b)$, where $\langle d \rangle$ is the average within-cluster distance between behaviors (defined here as $d_{ij} = \frac{1}{2}[1 - V_{ij}^{(e)}/\sqrt{V_{ii}^{(e)} V_{jj}^{(e)}}]$), $I(C; b)$ is the mutual information between cluster assignment and behavior number, and β modulates the relative importance of the two terms (see Materials and Methods). This modular structure emerges when applying other clustering methods as well (Fig. C.3). Quantifying the matrix's modularity, we find that $\langle d \rangle \approx 0.30$ and 0.22 for the 3 and 6-cluster solutions respectively. These values are significantly smaller than the average distances obtained using random cluster assignments ($\langle d \rangle = 0.46 \pm 0.03$ and 0.45 ± 0.04 for 3 and 6 clusters respectively, see Fig. C.5). Strikingly, these clusters are spatially contiguous in the behavioral map – implying that similar behaviors explain much of the intra-species variance [12]. Moreover, new clusters emerge in a hierarchical fashion, where coarse-grained behaviors sub-divide into new clusters (Fig. 4.4B), a feature that is not guaranteed by the information-based clustering algorithm.

This hierarchical structure of the behavioral space is reminiscent of the hierarchical temporal structure of behavior that was hypothesized originally by ethologists [131] and was observed to optimally explain the long timescale structure of *Drosophila melanogaster* behavioral transitions [12]. To explore this connection further, we found coarse-grainings of the behavioral space that are optimally predictive of the future behaviors that the flies perform via the Deterministic Information Bottleneck (DIB) [128]. Similar to the previously described information-based clustering method, this approach minimizes a functional, $\mathcal{J}_\tau = -I(b(t); Z(t+\tau)) + \gamma \mathcal{H}(Z)$, where $b(t)$ is a fly's behavior at time t, $Z(t+\tau)$ is the coarse-grained behavior visited at time $t+\tau$, $\tau = 50$, $I(b(t); Z(t+\tau))$ is the mutual information between these quantities, γ is a positive constant, and $\mathcal{H}(Z)$ is the entropy of the coarse-grained representation (see

Material and Methods). As γ is increased, progressively coarser representations are found.

Applying this method to the data pooled across all six species (Figs. 4.4C, C.4), we again found the same type of hierarchical division in the behavioral space that was observed for freely moving *D. melanogaster* [12]. Moreover, we found that the structure of the space using this approach closely mirrors the structure found via clustering $V^{(e)}$ (Fig. 4.4C). We quantify the similarity between both clustering partitions by calculating the Weighted Similarity Index (WSI), a modification of the Rand Index [109] (Materials and Methods). The WSI between the information-based clustering method and the predictive information bottleneck for three clusters is $WSI = 0.73$ and $WSI = 0.87$ for six clusters. For random clusterings, we would expect to observe 0.51 ± 0.02 and 0.70 ± 0.01 for 3 and 6 clusters, respectively, indicating a non-random overlap between these two partitions. Fig. C.3, shows that this result is independent of the clustering method and the number of clusters.

The overlap between these two coarse-grainings indicates that most individual variability in the behaviors we observe results from non-stationarity in behavioral measurements, rather than from individual-specific variation. That is, much of the intraspecific variation appears to reflect flies recorded when they were experiencing different hidden behavioral states (i.e. "moods"), rather than reflecting fixed (environmental or genetic) differences between flies. This variation may have arisen because, although we controlled many variables (e.g., fly age, circadian cycle, temperature, and humidity), it is not possible to control for all internal factors (e.g., hunger, arousal, etc.) that affect an animal's behavioral patterns [1]. The temporal coarse-graining of the behavioral space that we found via the DIB, gives insight into these non-stationarities, as they are optimally-predictive of the fly's future behaviors. Given the contiguous nature of these regions, this result means that flies tended to stay within specific regions of the behavioral space much longer than one would

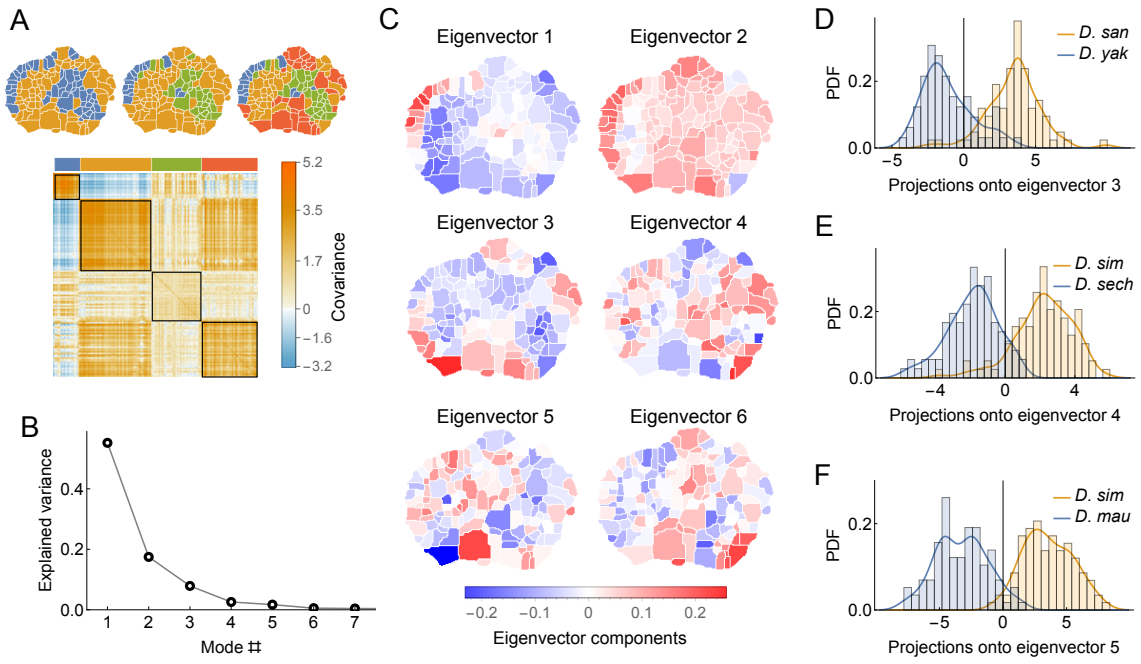

Figure 4.6: Phylogenetic variability and behavioral meta-traits. A: (top) Clustering the phylogenetic covariance matrix (using the same information-based clustering method from Fig. 4.4), we observe that the clusters are no longer spatially contiguous. (bottom) The phylogenetic covariance matrix reordered according to four clusters (colors corresponding to the four-cluster map above). B: Fraction of variance explained by the largest eigenvalues of the phylogenetic covariance matrix. C: The eigenvectors corresponding to the largest six eigenvalues. D: Distributions of the projections of individual density vectors from *D. santomea* and *D. yakuba* onto eigenvector 3. E: Same as in D but using projections of individuals from *D. sechellia* and *D. simulans* onto eigenvector 4. F: Same as in D but using projections of individuals from *D. simulans* and *D. mauritiana* onto eigenvector 5.

assume from a Markov model.

This observation implies that variation in behavior observed among individuals, especially in non-manipulated settings, is likely to often reflect a large component of hidden behavior states (Fig. 4.5A). Thus, it may be possible to improve upon behavioral measurements in many settings by controlling for the variability associated with these hidden states. For example, just because one fly performs less anterior grooming than another may reflect that the animal is in a different long timescale behavioral state, rather than that the animal has a genetically encoded preference for

reduced grooming.

A potential method for accounting for these artifacts is to normalize each individual's behavioral density such that the amount of time that the animal spends in each of the coarse-grained regions is equalized. In other words, the amount of time spent anterior grooming, locomoting, etc. are set to be the same for all animals, thus accounting for the variability associated our inferred hidden states. Mathematically, if P_i is the probability of observing behavior i, and C_i is the clustering assignment of this behavior, we can define a normalized probability, $\hat{P}_i$, via

$$\hat{P}_i = \frac{\bar{P}^{(C_i)}}{P_i^{(C_i)}} P_i, \tag{4.2}$$

where $P_i^{(C)} = \sum_{k \in C} P_k$ is the total density in cluster C for an individual fly and $\bar{P}^{(C)}$ is the average across all animals.

We found that applying this normalization to our data often results in substantial changes in the inferred distributions of behavioral densities. For example, Fig. 4.5B displays how the difference in behavioral density between *D. santomea* and *D. yakuba* (as measured by the Mahalanobis distance between the distributions) alters as a result of normalization. For some behaviors, the signal increases (red points), and in some cases, it reverses (blue points). Thus, it is important to take these non-stationary effects into account when estimating how often single behaviors are performed in studies of behavioral evolution. To measure these non-stationary effects, many behaviors must be measured, not just a focal behavior.

4.5 Identifying phylogenetically linked behaviors

One of the advantages of our approach is that we separate variations in behavior corresponding to evolutionary patterns, the phylogenetic variability, from variations among individuals of the same species. By studying the properties of the phylogenetic

covariance matrix ($V^{(a)}$), we can identify behaviors that may be evolving together.

We first characterized the coarse-grained structure within $V^{(a)}$ through the information-based clustering described in the previous section [122]. As seen in Fig. 4.6A, these clusters are not spatially contiguous in the behavioral space. This pattern contrasts to the spatial contiguity we observed for the individual covariance matrix (Fig. 4.4B). For example, the two-cluster solution (Fig. 4.6A, left) separates the behavioral space into side legs movements (middle) and certain locomotion gaits (far left) from the rest of behaviors. Similarly, non-localized structure is also observed when the matrix is clustered into a larger numbers of clusters as well.

One possible interpretation of these discontinuous clusters is that at the neural level, each of these groups of movements may reflect a motor response to shared upstream commands [24]. For example, different types of locomotion might be controlled through the same descending neural circuitry, but due to evolutionary changes, the same commands could lead to different behavioral outputs, as has been observed in fly courtship patterns [39]. Thus, examination of phylogenetically course-grained regions such as these may provide a more biologically realistic view of suites of evolving behaviors than does focus on single behaviors.

To quantify these patterns as traits, we decomposed $V^{(a)}$ via an eigendecomposition. As seen in Fig. 4.6B, almost all of the variance within the matrix can be explained with only the first six eigenmodes. These eigenvectors (Fig. 4.6C) share similar non-local structure to the clusterings described above. By projecting individual behavioral vectors onto these eigenvectors, the resulting dot products represent a meta-trait that is a linear combination of phylogenetically linked behaviors.

These evolving meta-traits may be suitable targets for further neurobiological or genetic studies. Three examples of these distributions are shown in in Fig. 4.6D for several pairs of closely related species. These three examples were not chosen at random, but instead because they showed significant differentiation between species.

The aim of this analysis is not to show that all meta-traits would differ between all pairs of species, which strikes us as unlikely, but rather that it is possible to identify synthetic meta-traits that could be further interrogated with experimental methods.

4.6 Discussion

We have developed a quantitative framework to study the evolution of behavioral repertoires, using fruit flies (*Drosophila*) as a model system. We started with observations of 561 individuals from six extant species behaving in an unremarkable environment. This assay did not include social behaviors, such as courtship and aggression, nor many foraging behaviors. Thus, at first glance, it might seem like we had excluded most species-specific behaviors from the analysis. Nonetheless, we found that other complex behaviors, like walking, running, and grooming, exhibit species-specific features that can be used to reliably assign individuals to the correct species. Thus, the motor patterns of behaviors that are not normally investigated for their species-specific features are clearly evolving between even closely related species. It is not clear if these differences reflect natural selection or genetic drift on the details of these motor patterns. But, all of these behaviors would seem to be critical to individual survival, so it is possible that these behaviors have evolved, at least in part, in response to natural selection. It is clear, however, that the underlying neural circuitry controlling these behaviors must have evolved.

Inspired by these observations, we estimated patterns of behavioral evolution in the context of a well-understood phylogeny. We fit a Generalized Mixed Linear Model to our behavioral measurements and the given phylogeny to reconstruct ancestral behavioral repertoires and the intra- and inter-species covariance matrices. We found that the patterns of intra-species variability are similar to long timescale behavioral dynamics. This suggests that much of the intraspecific variability that emerged by

sampling flies under well-controlled conditions reflects variability in the hidden behavioral states of individual flies. This variability is a clear confound for evolutionary and experimental studies of behavior and we therefore propose a method to control for these internal states and improve the accuracy of behavioral phenotyping. We showed that controlling for these internal states can dramatically alter estimates of the "heritable" elements of behavior.

Given our estimates for how suites of behaviors evolved, we examined whether the inter-species covariance matrix could be used to identify behavioral meta-traits that might be subjected to further evolutionary and experimental analysis. We identified multiple suites of behaviors that differed between closely related species, providing a starting point for further analysis of how the mechanismns underlying these suites of behaviors have evolved.

The analysis framework introduced here represents the first attempt to analyze full behavioral repertoires to gain insight into evolution. In principle, this approach could be applied to any data set where a large number of behaviors have been sampled in many species. We envision several areas where future improvements may yield more detailed, comprehensive, and biologically meaningful results. First, we recorded behavior from only six species of flies. Adding additional species would place more constraints on the evolutionary dynamics, likely resulting in less variance in the ancestral state estimations and potentially adding more structure to the relatively low rank covariance matrices. Additionally, further work is required to determine the balance between sampling within and between strains and species that optimizes estimates of evolutionary dynamics.

Second, our framework assumes that all evolutionary changes in behavior resemble a diffusion process. Although this assumption is a reasonable initial hypothesis [44], it may be possible to test this assumption. For example, deeper sampling of additional species may allow identification of specific behaviors on particular lineages where

neutrality can be rejected [129].

In addition, all of our analyses involved measuring the fraction of behaviors performed during the recording time, ignoring the temporal structure and sequences of movements. While we show here that much of this information can be related to the structure of the intra-species covariance matrix, the order in which behaviors occured may also provide important biological information. It should be possible to incorporate temporal structure directly into the regression. Deciding exactly which quantities to measure and how they should be incorporated, however, are complex questions that are outside the scope of this initial study.

Lastly, capturing the full range of animal behaviors for a large number of animals presents a number of technological challenges, which is why we focused on measuring behavior in a highly simplified environment. However, a more complete understanding of the structure of behavior will require more sophisticated ways to capture behavioral dynamics in more naturalistic settings and during complex social arrangements. While modern deep learning methods have made tracking animals in more realistic settings increasingly plausible [106, 89], there are still considerable hurdles to translating this information into a form that can be subjected to the kind of analysis we propose here.

Despite these limitations, this work represents a new way to quantitatively characterize the evolution of complex behaviors, which may provide new phenotypes that can be subjected to experimental analysis. In the absence of a behavioral fossil record, reconstructing ancestral behaviors requires an inferential approach like the one we present here. In addition, more complex models could be built to test assumptions underlying this initial, diffusion-based, model. Finally, a strength of our approach is that it makes falsifiable predictions about how behaviors are linked mechanistically, providing predictions that can be tested experimentally to provide further insight in the genetic and neurobiological structure of behavior.

4.7 Materials and Methods

4.7.1 Data collection

All imaging of fly behavior followed the procedures described in [24], but without any red light stimulation. In total, we collected data from 561 individual from 18 strains and six species. These included three strains of *D. mauritiana* (*mau29*: 29 flies, *mau317*: 35 flies, *mau318*: 32 flies), four strains of *D. melanogaster* (*Canton-S*: 31 flies, *Oregon-R*: 33 flies, *mel54*: 34 flies, *mel56*: 31 flies), three strains of *D. santomea* (*san00*: 29 flies, *san1482*: 33 flies, *STO OBAT*: 22 flies), three strains of *D. sechellia* (*sech28*: 32 flies, *sech340*: 25 flies, *sech349*: 33 flies), three strains of *D. simulans* (*sim5*: 33 flies, *sim199*: 30 flies, *Oxnard*: 34 flies), and two strains of *D. yakuba* (*yak01*: 34 flies, *CYO2*: 31 flies).

4.7.2 Generalized Linear Mixed Model

We fit our GLMM (Eq. 4.1) using the software introduced in [55]. The covariance matrices $V^{(e)}$ and $V^{(a)} \in \Re^{K\times K}$, $K = 134$ and the mean vector $\vec{\mu} \in \Re^{K\times 1}$, were inferred from the posterior distribution via MCMC sampling. Prior distributions for the covariance matrices were given by Inverse Wishart Distributions (conjugate priors for the multi-Gaussian model) with K degrees of freedom and $\frac{1}{K+1}\frac{I+J}{2}$ as scale matrix, with J and I the unit and identity matrices respectively. Tree branch length were estimated from [116].

4.7.3 Gelman-Rubin convergence diagnostic

This test evaluates MCMC convergence by analyzing the difference between several Markov chains. Convergence is evaluated by comparing the estimated between-chains and within-chain variances for each parameter of the model. Large differences between these variances indicate non-convergence [48]. Let θ be the model parameter of in-

terest and $\{\theta_m\}_{t=1}^{N}$ be the mth simulated chain, $m = 1, 2, ..., M$. Denote, $\hat{\theta}_m$ and $\hat{\sigma}_m^2$ be the sample posterior mean and variance of the mth chain. If $\hat{\theta} = \frac{1}{M}\sum_{m=1}^{M}\hat{\theta}_m$ is the overall posterior mean estimator, the between-chains and within-chain variances are given by:

$$B = \frac{N}{M-1}\sum_{m=1}^{M}(\hat{\theta}_m - \hat{\theta})^2,\ W = \frac{1}{M}\sum_{m=1}^{M}\hat{\sigma}_m^2. \tag{4.3}$$

In reference [48], it is shown that the following weighted average of W and B is an unbiased estimator of the marginal posterior variance of θ: $\hat{V} = \frac{N-1}{N}W + \frac{M+1}{NM}B$.

The ratio $\hat{V}/W$ should get close to one as the M chains converge to the target distribution with $N \to \infty$. In reference [19] this ratio known as the potential scale reduction factor (PSRF) was corrected to account for the the sampling variability using $R_c = \sqrt{\frac{d+3}{d+1}\frac{\hat{V}}{W}}$, where d is the degrees of freedom estimate of a t-distribution. Values of PSRF for all model parameters such that $R_c < 1.1$ are used in [19] as a criteria for convergence of the MCMC chains. Here, we used 20 independent chains, each with a different initialization.

4.7.4 Information-based clustering

Minimizes the distance between elements within clusters while compressing the original representation. The method minimizes the functional $\mathcal{F} = \langle d \rangle + TI(C;i)$, where $I(C;i) = \sum_{i=1}^{N}\sum_{C=1}^{N_c} P(C;i)\log[\frac{P(C|i)}{P(C)}]$ is the mutual information between the original behavioral variable i and the representation C. $\langle d \rangle = \sum_{C=1}^{N_c} P(C)d(C)$, and $d(C)$ is the average distance of elements chosen out of a single cluster:

$$d(C) = \sum_{i_1}^{N}\sum_{i_2}^{N} P(i_1 \mid C)P(i_2 \mid C)d(i_1, i_2), \tag{4.4}$$

with $d(i_1, i_2)$ being the distance measure between a pair of elements and $P(i \mid C)$ being the probability to find element i in cluster C.

Given $\mid C \mid = N_c$, T and a random initial condition for $P(C \mid i)$, a solution is ob-

tained by iterating the following self-consistent equations until the criteria $\frac{\mathcal{F}_t - \mathcal{F}_{t+1}}{\mathcal{F}_t} < 10^{-5}$ is satisfied. We chose $40,000$ different random values of $T \in [0.1, 1000]$, N_c between 2 and 20, and performed the optimization in each case until the convergence criterion was met. We defined the Pareto front as the set of solutions $P(C \mid i)$ such that no other solution presents a smaller $\langle d \rangle$ and a smaller $I(C;i)$. Finally, for each number of clusters we selected the solution with the lowest $\langle d \rangle$.

For each number of clusters, the significance of the optimal value found for $\langle d \rangle$ is shown by comparing it to the average distance corresponding to random cluster assignments. These assignments are made in such a way that the amount of elements per cluster is conserved by randomly shuffling the vector that assigns each behavior to a particular cluster. The values presented in the main text correspond to the mean and standard deviation of $\langle d \rangle$ over 50 different random trials.

4.7.5 Deterministic Information Bottleneck

Here we use the Deterministic Information Bottleneck (DIB) method to find coarse-grainings of the behavioral space that optimally predict future states [128]. Inspired by the Information Bottleneck [132], DIB replaces the compression measure $I(X,Z)$ with the entropy $H(Z)$, thus emphasizing constraints on the representation. DIB minimizes the functional:

$$\mathcal{L}_\alpha = H(Z) - \alpha H(Z \mid X) - \beta I(Z;Y), \tag{4.5}$$

with respect to $p(z \in Z | x \in X)$ and takes the limit as $\alpha \to 0$.

To apply DIB to the behavioral dynamics, we count time in units of the transitions between states, providing a discrete time series of behaviors, $b(n)$ that can take on $N = 134$ different integer values at each discrete time n. Here, we relate the joint distributions of $b(n)$ (X in Eqn. 4.5) and $b(n+\tau)$ (Y) through a coarse-grained

clustering of the behavioral states (Z). We chose $10,000$ different pairs of random values for β between 0.1 and 10^4 and N_c between 2 and 30 clusters. Given N_c, β and a random initial condition for $p(t \mid x)$, we find a solution by iterating the self-consistent equations [128] until the convergence criteria $\mid \mathcal{L}^t - \mathcal{L}^{t+1} \mid < 10^{-6}$ is satisfied. If any cluster has its probability become zero at any iteration, then that cluster is dropped for all future iterations, thus N_c is the maximum number of clusters that can be returned. Of these $10,000$ solutions, we keep all solutions that are on the Pareto front (i.e., no other solution has both a higher $I(Y;T)$ and a smaller $H(T)$). The displayed clusters are the solutions on the Pareto front with the largest $I(Y;Z)$ for a given number of clusters.

4.7.6 Weighted Similarity Index

We quantify the similarity between clustering partitions by calculating the Weighted Similarity Index (WSI), a modification of the Rand Index [109] such that behaviors contribute the index according to their overall probability. Specifically,

$$\text{WSI} = \sum_{i,j \in S_a} \text{W}_{ij} + \sum_{k,l \in S_b} \text{W}_{kl}, \; \text{W}_{ij} \;\; = \frac{P_i P_k}{\sum_{kl} P_k P_l}, \tag{4.6}$$

where $S_a(S_b)$ is the set of pairs of behaviors that belong to the same (different) cluster in the two partitions and P_k is the probability of observing behavior k.

Chapter 5

Conclusions

In this Dissertation, we have made progress towards solving a major challenge in modern theoretical biophysics: namely, developing tools for inferring phenomenological models from data. We then used these statistical inference tools to answer some of the *what* type of questions for different types of biological processes, from enzymatic catalysis, to behavioral evolution. The inferred models, besides being able to accurately explain experimental observations, often allow us to make predictions about the systems, and sometimes even get insight into the underlying biological and physical mechanisms.

In Chapters 2 and 3 of this Dissertation, we build on the idea initially proposed in Ref. [35], where phenomenological models of living systems were built without having to construct mechanistically accurate model first, to be coarse-grained later. The complexity of biological systems would make the route requiring a mechanistic model as an intermediate step a very challenging and in many cases a hopeless task. In contrast, our approach of *refining* phenomenological models until their complexity reaches what is needed to explain the data does not require to build a mechanistic model as a stepping stone. Instead, we start from the simplest possible model and progressively add details to it, until the complexity is sufficient to explain the studied

data set. Which details are to be added, and in which order, is controlled by the hierarchical structure, also known as a *model family*, which our approach adopts. The multi-path model family, which we propose to study systems that can be seen as FPP, successfully explain the completion probability distributions of two very different types of biochemical processes (the spike generation in Purkinje Cells and the single enzyme chemical reaction). Despite our models offering little understanding of the mechanistic details of the processes, we show that no other models can have a better fit than ours. Furthermore, they can provide meaningful insights and minimal constraints on the mechanistic properties of the systems. The inferred model that best describes the interspike interval (ISI) distributions of PC under different injected currents makes good predictions of the entire ISI distribution at the values of the current not used in training. Furthermore, the inferred model points out that any biophysically realistic description of the spike generation process would require at least 20 different states to explain the experimental data.

In contrast, for the single enzyme reaction, predictions for the FP time distributions not included in the training set are very poor, even if fits were great, as seen in Chapter 3. However, the inferred model suggested that a small biochemical network should be able to explain the enzymatic turnover times under different substrate concentrations. Following on this observation, in Chapter 3, we proposed a different model family, the biochemically-realistic model family, specially designed to infer the smallest biochemically-realistic network to explain single enzyme reactions. The inferred model within this hierarchy shows that a network with only four different states can fit the data as good as the model inferred using the multi-path model family. Even more, good predictions for the entire FP time distributions can be made using this hierarchy. Therefore, in this Chapter, we show that in principle there could be many hierarchies that can explain the data in a parsimonious way. However, the model that encapsulates mechanistic biophysical constraints to recover the coarse-grained

behavior makes better predictions. This signals that using generic machine learning methods to infer physical models may not be very useful, and such methods need to be tailored to the systems one wants to describe. On the other hand, the generic multi-path model family gave us a lot of valuable information about the systems, and we are very confident that this model family will have its use as a starting point for description of many FPPs in biology.

In Chapter 4 we develop what we think is the first attempt to infer phenomenological models of animal behavioral evolution. We show that under a simple and *realistic* assumption that individuals of phylogenetically closely related species, evolve using a difussion-like process in the behavioral space, we could infer the ancestral behavioral repertories and the intra and inter-species behavioral variability. These results suggest that a lot of the intra-species variability is related mostly to long-lasting internal states of the animal, while the variability between individuals of different species can capture different groups of behavior that could have evolved together. Despite we do not have experimental evidence to test the reconstructed behavioral repertories, we think that our approach provides a new framework for identifying co-evolving behaviors and may provide new opportunities to facilitate the study of the genetic bases of behavioral evolution.

We conclude this Dissertation with the following notes. While our attempts above represent some of the first, and sometimes disparate, attempts to study the *what* in biological systems using automatic inference tools, we believe that they illustrate a few important points. First, we think that our efforts show that the developed computational tools will be useful in understand many diverse biological systems, further advancing the field of phenomenological modeling. Second, our approach illustrates the power of merging the physics-level model-based understanding with modern statistical inference approaches, aligning with the booming trend of using machine learning methods to learn new physics from experimental data, but doing

this in a model-driven, interpretable way. Third, all of our developed methods require a certain level of mathematical and computational sophistication. This signals a lot of opportunities for future work to make these computational tools more user-friendly, so that, at some point, they can be easily used by experimentalists themselves to analyze other biological systems for which the appropriate experimental observations have been done.

Appendix A

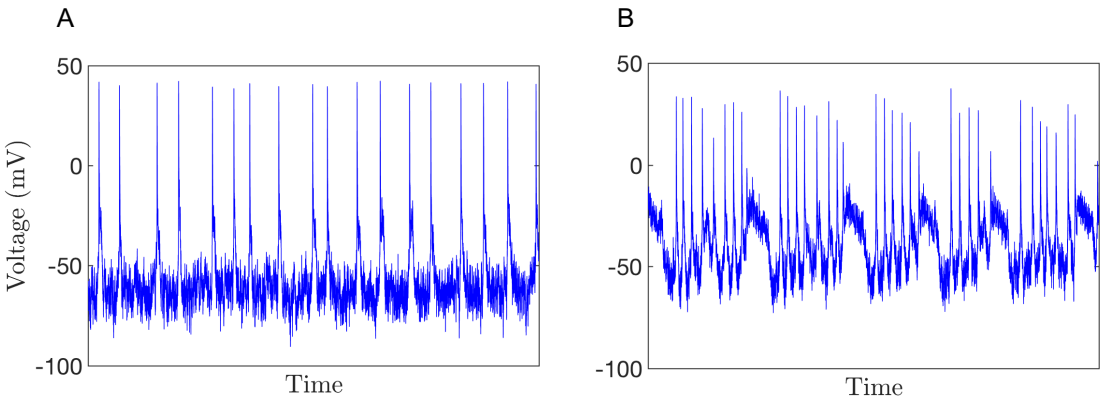

Figure A.1: Simulated PC membrane potential using the multi-compartmental model proposed in [95] for A: low ($I = 0.5$ nA) and B: high ($I = 3$ nA) values of the injected current.

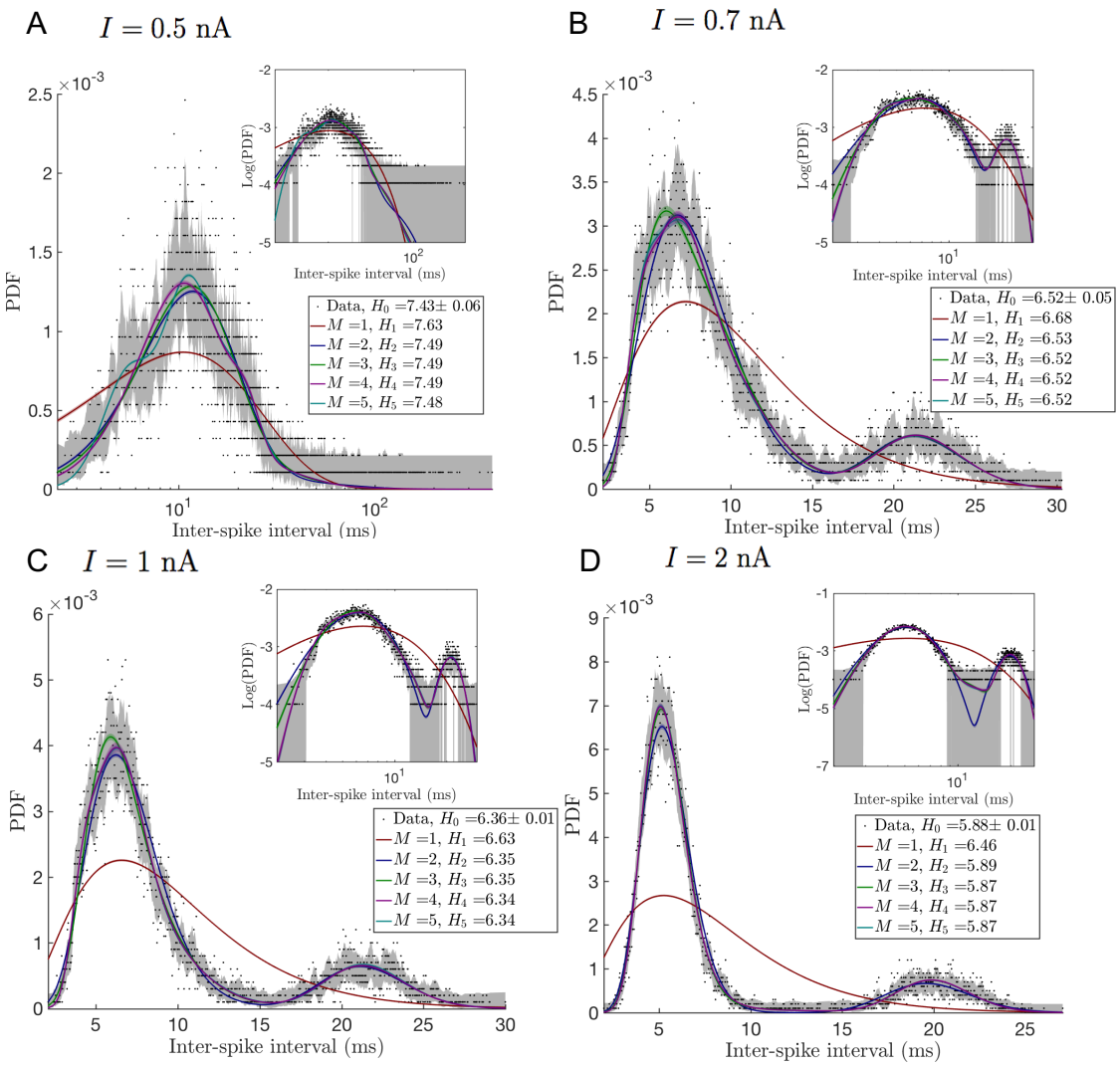

Figure A.2: Best fit models from the multi-path model family for the ISIs distributions of PCs. Color lines and bands (the latter often too narrow to be seen) show the mean and the standard deviation of different models sampled from the posterior distribution of each of the first five models in the family. The legends illustrate the decrease of the cross entropy with the model complexity towards its minimum value of the entropy of the histogram of the observed data. According to Tbl. 2.3, 4 paths are needed to explain the ISI characteristics of synthetic under different external conditions. (A, B, C, D) injected currents $I = 0.5, 0.7, 1.0, 2.0$ nA, respectively.

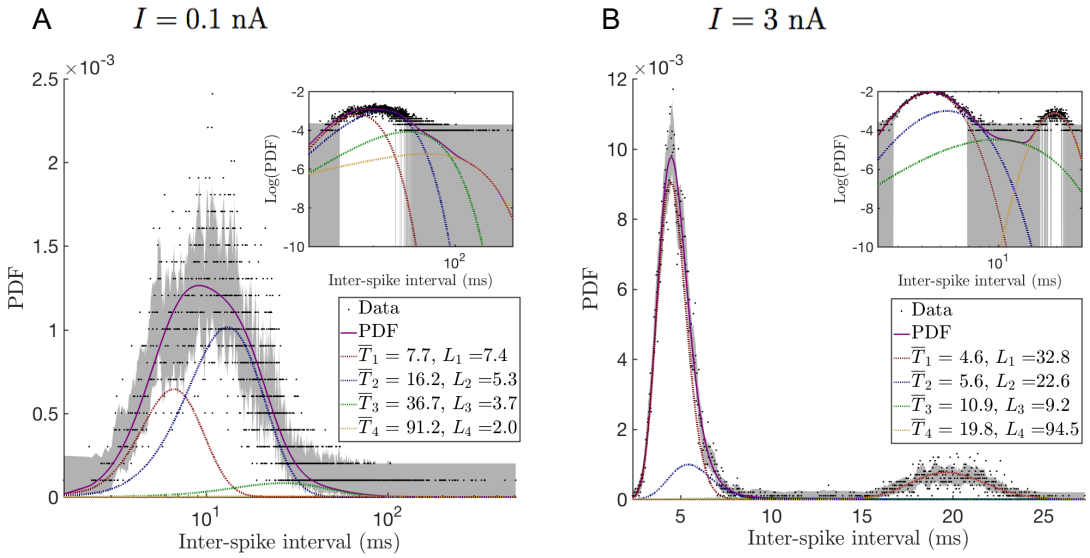

Figure A.3: Decomposition of the completion time PDF into contributions from different paths for (A) $I = 0.1$ nA and (B) $I = 3.0$ nA. Insets show the same data in log-log units. In (A), the two pathways with the shortest completion time explain the bulk of the distribution while the pathway with the longest average completion time approximate the left tail of the distribution. In (B), pathways with shortest/longest completion time contribute mostly to the intra/inter burst time scales.

Appendix B

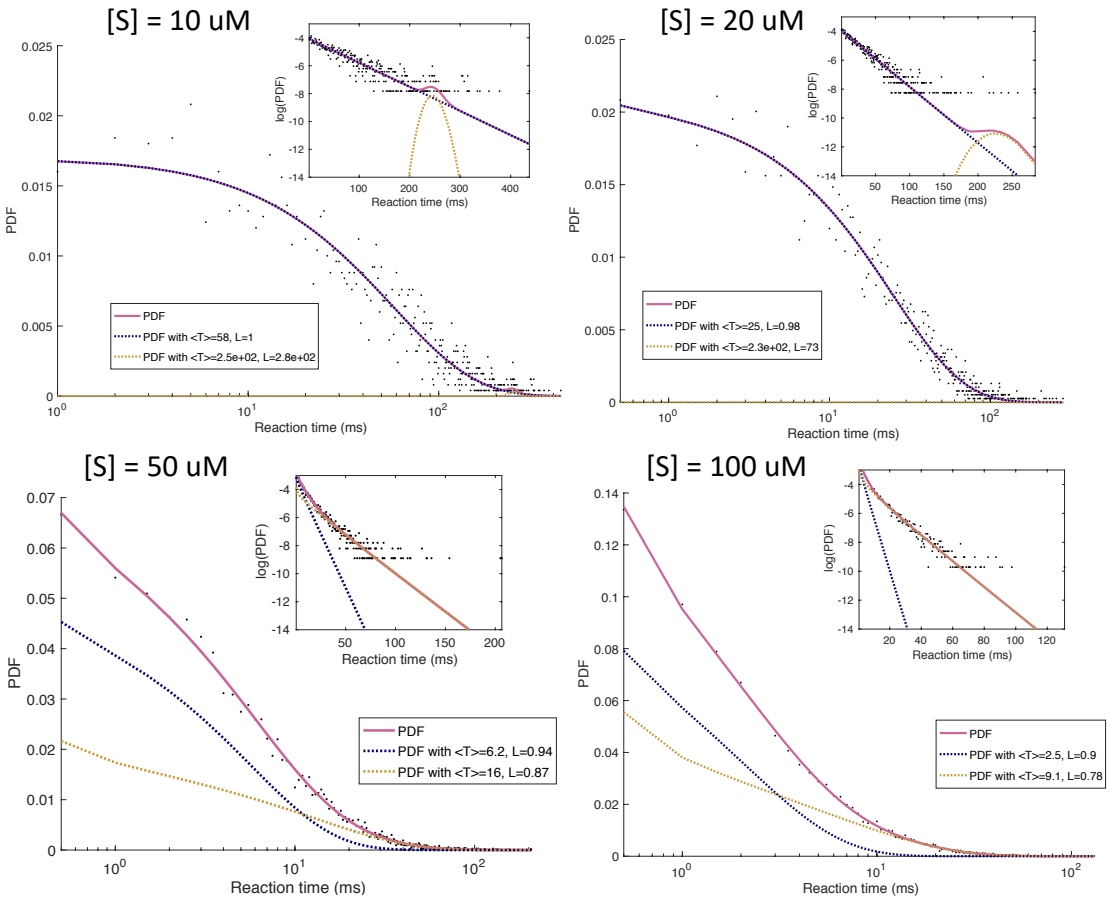

Figure B.1: Best fitted model with $M = 2$ decomposed into completion paths. The yellow path for low substrate concentrations is not needed (Tbl. 3.1), but for high concentrations it explains the tail of the distributions

Appendix C

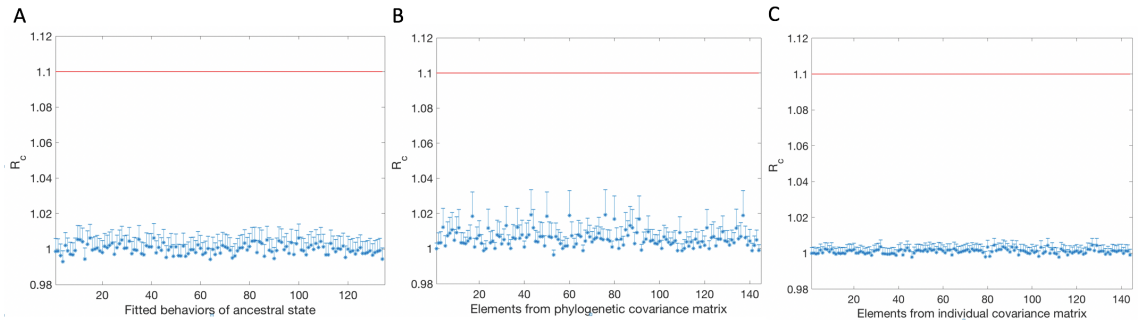

Figure C.1: Gelman Rubin diagnostic for model parameters inferred using MCMC. A: PSRF for the 134 ancestral behaviors inferred in the GLMM. 20 MCMC chains with different initial conditions were used. B: PSRF for the phylogenetic covariance matrix elements corresponding to the 10% most common behaviors performed by the measured flies. C: PSRF for the individual covariance matrix elements corresponding to the 10% most common behaviors performed by the measured flies. The PSRF values for all of these inferred parameters indicate that the MCMC chains are converging.

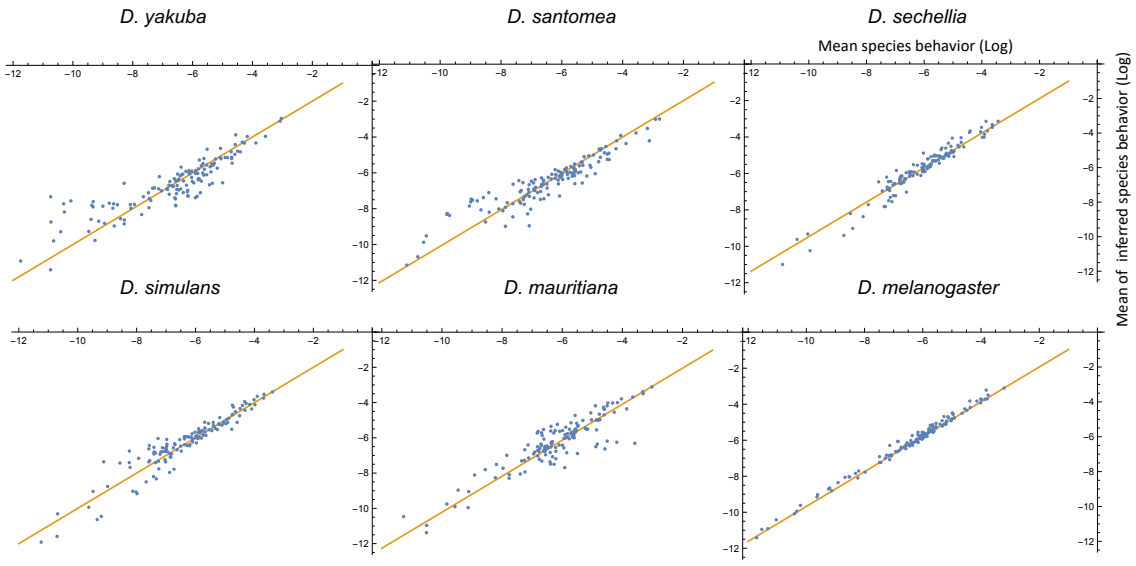

Figure C.2: Comparison between measured and inferred behaviors (in log scale) for each of the extant species. The mean of the measured behavioral repertoires for all the individuals of a particular species is taken in the log scale. Each measured behavioral mean gets compared to the mean obtained from the components of the MCMC samples corresponding to that particular species and behavioral mode (i.e., the inferred behavioral repertories from the GLMM). The biggest differences occur mostly in the low probability behaviors.

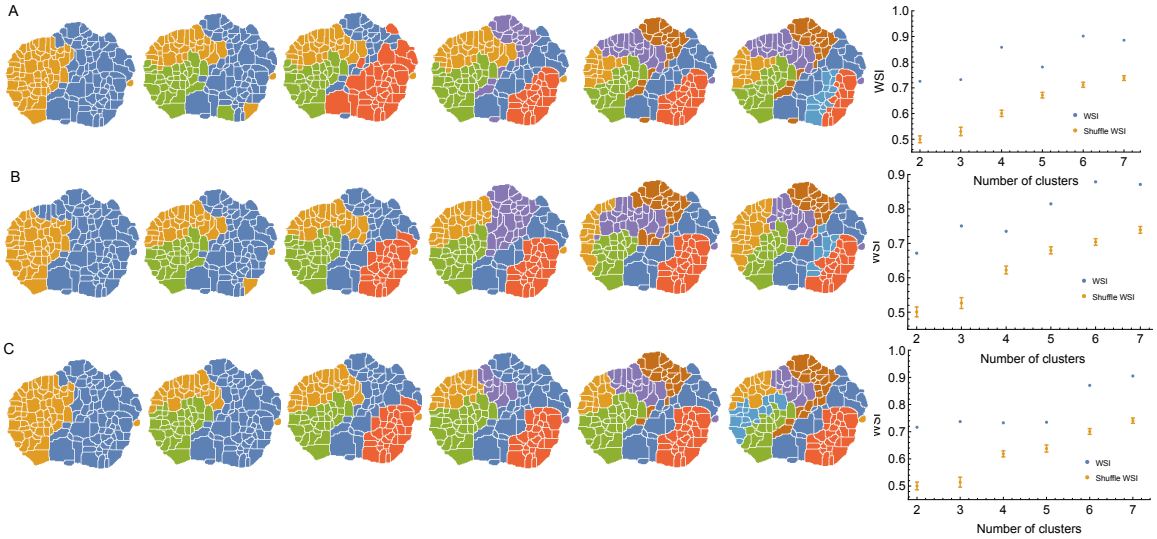

Figure C.3: Behaviors clustered according to information of the individual covariance matrix using three different clustering methods. A: Results using k-medoids clustering method with distance matrix $d_{ij} = (1-\rho_{ij})/2$ for 2,3,..7 clusters. To the right, the WSI between the clusters obtained using k-medoids and those obtained using predictive information bottleneck method. Clearly, the similarity between these two orthogonal measurements is significant, as can be shown when compared to the WSI calculated by randomly shuffling the labels of the k-medoids clustering corresponding to each number of clusters. B: Same as in A but we used Spectral clustering instead of k-medoids. The similarity index between Spectral clustering and predictive information bottleneck is also statistically significant. C: Same as in A but we used Information based clustering instead of k-medoids. The similarity index between Information based clustering and predictive information bottleneck is statistically significant as well.

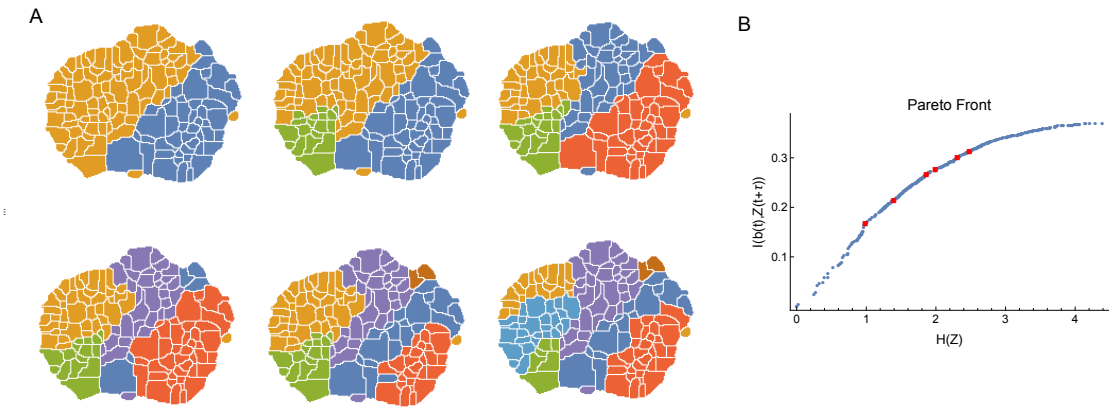

Figure C.4: Coarse-grained behavioral representations that are optimally predictive of the future behavior states via DIB. A: Behavioral representation with 2,3,...,7 clusters using $\tau = 50$ in Eq. 4.5. B: Optimal trade-off curve (Pareto Front) between complexity of coarse grained description against predictive power. For each number of clusters, representations in A correspond to points (in red) in this curve with the highest predictive information

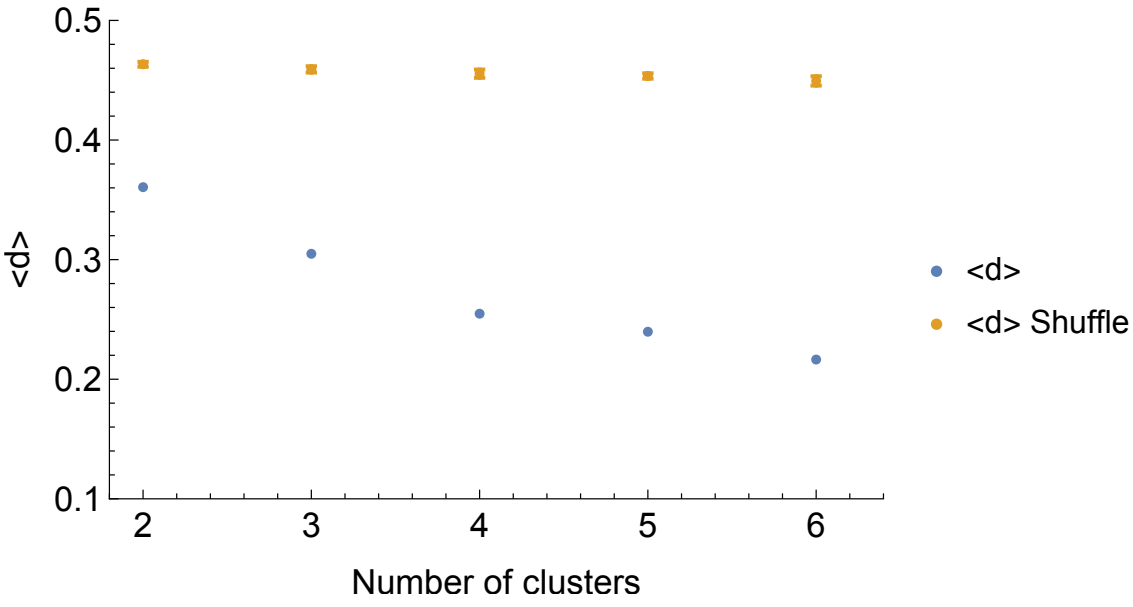

Figure C.5: Modularity measure of the intra-species behavioral covariance matrix using information based clustering. $< d >$ corresponds to the average distance among elements of the same clusters, (see Materials and Methods for definition). We show for different number of clusters that matrix modularity is significantly smaller (in blue) than expected by random assignation of behaviors to clusters (in orange).

Bibliography

[1] D. J. Anderson. Circuit modules linking internal states and social behaviour in flies and mice. *Nature Reviews Neuroscience*, 17(11):692, 2016.

[2] J. Anderson, Y.-C. Chang, and A. Papachristodoulou. Model decomposition and reduction tools for large-scale networks in systems biology. *Automatica*, 47(6):1165–1174, 2011.

[3] T. O. Auer and R. Benton. Sexual circuitry in Drosophila. *Current opinion in neurobiology*, 38:18–26, 2016.

[4] F. Baier and H. E. Hoekstra. The genetics of morphological and behavioural island traits in deer mice. *Proceedings of the Royal Society B*, 286(1914):20191697, 2019.

[5] W. Bair, C. Koch, W. Newsome, and K. Britten. Power spectrum analysis of bursting cells in area mt in the behaving monkey. *Journal of Neuroscience*, 14(5):2870–2892, 1994.

[6] B. S. Baker, B. J. Taylor, and J. Hall. Are complex behaviors specified by dedicated regulatory genes? Reasoning from Drosophila. *Cell*, 105(1):13–24, Apr. 2001.

[7] V. Balasubramanian. Statistical inference, occam's razor, and statistical me-

chanics on the space of probability distributions. *Neural computation*, 9(2):349–368, 1997.

[8] G. Bel, B. Munsky, and I. Nemenman. The simplicity of completion time distributions for common complex biochemical processes. *Physical biology*, 7(1):016003, 2009.

[9] O. Benichou, T. Guérin, and R. Voituriez. Mean first-passage times in confined media: from markovian to non-markovian processes. *Journal of Physics A: Mathematical and Theoretical*, 48(16):163001, 2015.

[10] O. Bénichou, C. Loverdo, M. Moreau, and R. Voituriez. Intermittent search strategies. *Reviews of Modern Physics*, 83(1):81, 2011.

[11] G. J. Berman. Measuring behavior across scales. *BMC Biology*, 16(1):23, Feb. 2018.

[12] G. J. Berman, W. Bialek, and J. W. Shaevitz. Predictability and hierarchy in drosophila behavior. *Proceedings of the National Academy of Sciences*, 113(42):11943–11948, 2016.

[13] G. J. Berman, D. M. Choi, W. Bialek, and J. W. Shaevitz. Mapping the stereotyped behaviour of freely moving fruit flies. *Journal of The Royal Society Interface*, 11(99):20140672, 2014.

[14] P. Bishop, W. Levick, and W. Williams. Statistical analysis of the dark discharge of lateral geniculate neurones. *The Journal of physiology*, 170(3):598–612, 1964.

[15] N. M. Borisov, A. S. Chistopolsky, J. R. Faeder, and B. N. Kholodenko. Domain-oriented reduction of rule-based network models. *IET systems biology*, 2(5):342–351, 2008.

[16] J. M. Bower and D. Beeman. *The book of GENESIS: exploring realistic neural models with the General Neural Simulation System*. Springer Science & Business Media, 2012.

[17] P. C. Bressloff. *Stochastic processes in cell biology*, volume 41. Springer, 2014.

[18] P. C. Bressloff and J. M. Newby. Stochastic models of intracellular transport. *Reviews of Modern Physics*, 85(1):135, 2013.

[19] S. P. Brooks and A. Gelman. General methods for monitoring convergence of iterative simulations. *Journal of computational and graphical statistics*, 7(4):434–455, 1998.

[20] A. E. X. Brown and B. de Bivort. Ethology as a physical science. *Nature Physics*, 20:410, Apr. 2018.

[21] B. D. Burns and A. Webb. The spontaneous activity of neurones in the cat's cerebral cortex. *Proc. R. Soc. Lond. B*, 194(1115):211–223, 1976.

[22] C. Bustamante, Z. Bryant, and S. B. Smith. Ten years of tension: single-molecule dna mechanics. *Nature*, 421(6921):423–427, 2003.

[23] J. Cande, P. Andolfatto, B. Prud'homme, D. L. Stern, and N. Gompel. Evolution of multiple additive loci caused divergence between drosophila yakuba and d. santomea in wing rowing during male courtship. *PLoS One*, 7(8):1–10, 08 2012.

[24] J. Cande, S. Namiki, J. Qiu, W. Korff, G. M. Card, J. W. Shaevitz, D. L. Stern, and G. J. Berman. Optogenetic dissection of descending behavioral control in drosophila. *eLife*, 7:e34275, 2018.

[25] J. Cande, D. L. Stern, T. Morita, B. Prud'homme, and N. Gompel. Looking

under the lamp post: neither fruitless nor doublesex has evolved to generate divergent male courtship in drosophila. *Cell reports*, 8(2):363–370, 2014.

[26] H. Chipman, E. I. George, R. E. McCulloch, M. Clyde, D. P. Foster, and R. A. Stine. The practical implementation of bayesian model selection. *Lecture Notes-Monograph Series*, pages 65–134, 2001.

[27] T. Chou and M. R. D'Orsogna. First passage problems in biology. In *First-passage phenomena and their applications*, pages 306–345. World Scientific, 2014.

[28] T. Chou, K. Mallick, and R. Zia. Non-equilibrium statistical mechanics: from a paradigmatic model to biological transport. *Reports on progress in physics*, 74(11):116601, 2011.

[29] S. Chyb and N. Gompel. *Atlas of Drosophila Morphology: Wild-type and classical mutants*. Academic Press, 2013.

[30] L. A. Chylek, L. A. Harris, J. R. Faeder, and W. S. Hlavacek. Modeling for (physical) biologists: an introduction to the rule-based approach. *Physical biology*, 12(4):045007, 2015.

[31] H. Conzelmann, D. Fey, and E. D. Gilles. Exact model reduction of combinatorial reaction networks. *BMC systems biology*, 2(1):78, 2008.

[32] P. F. Cook and W. W. Cleland. *Enzyme kinetics and mechanism*. Garland Science, 2007.

[33] M. Correia and J. Landolt. A point process analysis of the spontaneous activity of anterior semicircular canal units in the anesthetized pigeon. *Biological cybernetics*, 27(4):199–213, 1977.

[34] C. W. Cunningham, K. E. Omland, and T. H. Oakley. Reconstructing ancestral character states: a critical reappraisal. *Trends in Ecology & Evolution*, 13(9):361–366, 1998.

[35] B. C. Daniels and I. Nemenman. Automated adaptive inference of phenomenological dynamical models. *Nature communications*, 6:8133, 2015.

[36] E. De Schutter and J. Bower. An active membrane model of the cerebellar Purkinje cell I. Simulation of current clamps in slice. *Neurophysiology*, 71, 1994.

[37] B. DeBusk, E. DeBruyn, R. Snider, J. Kabara, and A. Bonds. Stimulus-dependent modulation of spike burst length in cat striate cortical cells. *Journal of Neurophysiology*, 78(1):199–213, 1997.

[38] J. Ding, V. Tarokh, and Y. Yang. Model selection techniques: An overview. *IEEE Signal Processing Magazine*, 35(6):16–34, 2018.

[39] Y. Ding, J. L. Lillvis, J. Cande, G. J. Berman, B. J. Arthur, M. Xu, B. J. Dickson, and D. L. Stern. Neural evolution of context-dependent fly song. *Current Biology*, 29:1089–1099, 2019.

[40] E. D'Angelo, S. Solinas, J. Garrido, C. Casellato, A. Pedrocchi, J. Mapelli, D. Gandolfi, and F. Prestori. Realistic modeling of neurons and networks: towards brain simulation. *Functional neurology*, 28(3):153, 2013.

[41] C. Ellison, C. Wiley, and K. Shaw. The genetics of speciation: genes of small effect underlie sexual isolation in the hawaiian cricket laupala. *Journal of Evolutionary Biology*, 24(5):1110–1119, 2011.

[42] *Drosophila* 12 Genomes Consortium. Evolution of genes and genomes on the Drosophila phylogeny. *Nature*, 450(7167):203 – 218, 2007.

[43] B. P. English, W. Min, A. van Oijen, K. Taek Lee, G. Luo, H. Sun, B. J. Cherayil, S. C. Kou, and S. Xie. Ever fluctuating single enzyme molecules: Michaelis-Menten equation revisited. *Nat. Chem. Biol*, 17, 2005.

[44] J. Felsenstein. Phylogenies and the Comparative Method. *The American Naturalist*, 125(1):1–15, 1985.

[45] A. Fersht et al. *Structure and mechanism in protein science: a guide to enzyme catalysis and protein folding*. Macmillan, 1999.

[46] M. D. Forrest, M. J. Wall, D. A. Press, and J. Feng. The sodium-potassium pump controls the intrinsic firing of the cerebellar purkinje neuron. *PloS one*, 7(12):e51169, 2012.

[47] A. V. Fowler and I. Zabin. The amino acid sequence of beta-galactosidase of escherichia coli. *Proceedings of the National Academy of Sciences*, 74(4):1507–1510, 1977.

[48] A. Gelman and D. B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, pages 457–472, 1992.

[49] J. Geweke. Bayesian inference in econometric models using monte carlo integration. *Econometrica: Journal of the Econometric Society*, pages 1317–1339, 1989.

[50] J. M. Gleason and M. G. Ritchie. Do quantitative trait loci (qtl) for a courtship song difference between drosophila simulans and d. sechellia coincide with candidate genes and intraspecific qtl? *Genetics*, 166(3):1303–1311, 2004.

[51] I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT press, Cambridge, MA, 2016.

[52] R. Grossman and L. Viernstein. Discharge patterns of neurons in cochlear nucleus. *Science*, 134(3472):99–101, 1961.

[53] D. Grünwald and R. H. Singer. In vivo imaging of labelled endogenous β-actin mrna during nucleocytoplasmic transport. *Nature*, 467(7315):604–607, 2010.

[54] R. N. Gutenkunst, J. J. Waterfall, F. P. Casey, K. S. Brown, C. R. Myers, and J. P. Sethna. Universally sloppy parameter sensitivities in systems biology models. *PLoS Comput Biol*, 3(10):e189, 2007.

[55] J. Hadfield. Mcmc methods for multi–response generalised linear mixed models: The mcmcglmm r package. *Journal of Statistical Software*, 33:1–22, 2010.

[56] J. D. Hadfield and S. Nakagawa. General quantitative genetic methods for comparative biology: phylogenies, taxonomies and multi-trait models for continuous and categorical characters. *Journal of evolutionary biology*, 23(3):494–508, 2010.

[57] J. D. Hadfield and S. Nakagawa. General quantitative genetic methods for comparative biology: phylogenies, taxonomies and multi-trait models for continuous and categorical characters. *Journal of evolutionary biology*, 23(3):494–508, 2010.

[58] S. E. Halford and J. F. Marko. How do site-specific dna-binding proteins find their targets? *Nucleic acids research*, 32(10):3040–3052, 2004.

[59] E. L. Haseltine and J. B. Rawlings. Approximate simulation of coupled fast and slow reactions for stochastic chemical kinetics. *The Journal of chemical physics*, 117(15):6959–6969, 2002.

[60] W. S. Hlavacek, J. R. Faeder, M. L. Blinov, R. G. Posner, M. Hucka, and W. Fontana. Rules for modeling signal-transduction systems. *Sci. STKE*, 2006(344):re6–re6, 2006.

[61] S. Hong, M. Negrello, M. Junker, A. Smilgin, P. Thier, and E. De Schutter. Multiplexed coding by cerebellar purkinje neurons. *Elife*, 5, 2016.

[62] C. K. Hu and H. E. Hoekstra. Peromyscus burrowing: A model system for behavioral evolution. *Seminars in cell & developmental biology*, 61:107–114, 2016.

[63] H. Huang, M. Fairweather, J. Griffiths, A. Tomlin, and R. Brad. A systematic lumping approach for the reduction of comprehensive kinetic models. *Proceedings of the Combustion Institute*, 30(1):1309–1316, 2005.

[64] A. Ishijima and T. Yanagida. Single molecule nanobioscience. *Trends in biochemical sciences*, 26(7):438–444, 2001.

[65] S. Iyer-Biswas, C. S. Wright, J. T. Henry, K. Lo, S. Burov, Y. Lin, G. E. Crooks, S. Crosson, A. R. Dinner, and N. F. Scherer. Scaling laws governing stochastic growth and division of single bacterial cells. *Proceedings of the National Academy of Sciences*, 111(45):15912–15917, 2014.

[66] S. Iyer-Biswas and A. Zilman. First passage processes in cellular biology. *arXiv preprint arXiv:1503.00291*, 2015.

[67] S. Iyer-Biswas and A. Zilman. First-passage processes in cellular biology. *Advances in Chemical Physics*, 160:261–306, 2016.

[68] F. Jacob and J. Monod. Genetic regulatory mechanisms in the synthesis of proteins. *Journal of molecular biology*, 3(3):318–356, 1961.

[69] D. H. Juers, B. W. Matthews, and R. E. Huber. Lacz β-galactosidase: structure and function of an enzyme of historical and molecular biological importance. *Protein Science*, 21(12):1792–1807, 2012.

[70] L. P. Kadanoff. Scaling laws for ising models near t c. *Physics Physique Fizika*, 2(6):263, 1966.

[71] H.-W. Kang, T. G. Kurtz, et al. Separation of time-scales and model reduction for stochastic reaction networks. *The Annals of Applied Probability*, 23(2):529–583, 2013.

[72] R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the american statistical association*, 90(430):773–795, 1995.

[73] J. K. Kim and E. D. Sontag. Reduction of multiscale stochastic biochemical reaction networks using exact moment derivation. *PLoS computational biology*, 13(6):e1005571, 2017.

[74] S. K. Kim. Mean first passage time for a random walker and its application to chemical kinetics. *The Journal of Chemical Physics*, 28(6):1057–1067, 1958.

[75] D. J. Kiviet, P. Nghe, N. Walker, S. Boulineau, V. Sunderlikova, and S. J. Tans. Stochasticity of metabolism and growth at the single-cell level. *Nature*, 514(7522):376–379, 2014.

[76] A. B. Kolomeisky. Physics of protein–dna interactions: mechanisms of facilitated target search. *Physical Chemistry Chemical Physics*, 13(6):2088–2095, 2011.

[77] A. B. Kolomeisky and M. E. Fisher. Molecular motors: a theorist's perspective. *Annu. Rev. Phys. Chem.*, 58:675–695, 2007.

[78] A. B. Kolomeisky and A. Veksler. How to accelerate protein search on dna: Location and dissociation. *The Journal of chemical physics*, 136(12):03B615, 2012.

[79] I. B. Kulagina, S. M. Korogod, G. Horcholle-Bossavit, C. Batini, and S. Tyc-Dumont. The electro-dynamics of the dendritic space in purkinje cells of the cerebellum. *Archives italiennes de biologie*, 145(3):211–233, 2007.

[80] Y. Lamarre, M. Filion, and J. Cordeau. Neuronal discharges of the ventrolateral nucleus of the thalamus during sleep and wakefulness in the cat i. spontaneous activity. *Experimental brain research*, 12(5):480–498, 1971.

[81] M. Levine and E. H. Davidson. Gene regulatory networks for development. *Proceedings of the National Academy of Sciences*, 102(14):4936–4942, 2005.

[82] X. Li and A. B. Kolomeisky. Mechanisms and topology determination of complex chemical and biological network systems from first-passage theoretical approach. *The Journal of chemical physics*, 139(14):10B606_1, 2013.

[83] J. Lin. Divergence measures based on the Shannon entropy. 37(1):145–151, 1991.

[84] K. Z. Lorenz. The Evolution of Behavior. *Scientific American*, 199(6):67–78, 1958.

[85] H. P. Lu, L. Xun, and S. Xie. Single-molecule enzymatic dynamics. *Science*, 282(5395):1877–1882, 1998.

[86] B. B. Machta, R. Chachra, M. K. Transtrum, and J. P. Sethna. Parameter space compression underlies emergent theories and predictive models. *Science*, 342(6158):604–607, 2013.

[87] D. J. MacKay and D. J. Mac Kay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[88] E. P. Martins and E. L. P. Martins. *Phylogenies and the comparative method in animal behavior*. Oxford University Press, 1996.

[89] M. W. Mathis and A. Mathis. Deep learning tools for the measurement of animal behavior in neuroscience. *Current Opinion in Neurobiology*, 60:1–11, 2020.

[90] M. Maurya, S. Bornheimer, V. Venkatasubramanian, and S. Subramaniam. Reduced-order modelling of biochemical networks: application to the gtpase-cycle signalling module. *IEE Proceedings-Systems Biology*, 152(4):229–242, 2005.

[91] F. Meyer. Topographic distance and watershed lines. *Signal processing*, 38(1):113–125, 1994.

[92] L. Michaelis and M. Menten. Die kinetik der invertinwirkung biochem z 49: 333–369. *Find this article online*, 1913.

[93] M. Mir, Z. Wang, Z. Shen, M. Bednarz, R. Bashir, I. Golding, S. G. Prasanth, and G. Popescu. Optical measurement of cycle-dependent cell growth. *Proceedings of the National Academy of Sciences*, 108(32):13124–13129, 2011.

[94] L. Mirny, M. Slutsky, Z. Wunderlich, A. Tafvizi, J. Leith, and A. Kosmrlj. How a protein searches for its site on dna: the mechanism of facilitated diffusion. *Journal of Physics A: Mathematical and Theoretical*, 42(43):434013, 2009.

[95] T. Miyasho, H. Takagi, H. Suzuki, S. Watanabe, M. Inoue, Y. Kudo, and H. Miyakawa. Low-threshold potassium channels and a low-threshold calcium channel regulate ca2+ spike firing in the dendrites of cerebellar purkinje neurons: a modeling study. *Brain research*, 891(1-2):106–115, 2001.

[96] J. R. Moffitt, J. B. Lee, and P. Cluzel. The single-cell chemostat: an agarose-based, microfluidic device for high-throughput, single-cell studies of bacteria and bacterial communities. *Lab on a Chip*, 12(8):1487–1494, 2012.

[97] B. Munsky and M. Khammash. The finite state projection algorithm for the solution of the chemical master equation. *The Journal of chemical physics*, 124(4):044104, 2006.

[98] B. Munsky, G. Neuert, and A. Van Oudenaarden. Using gene expression noise to understand gene regulation. *Science*, 336(6078):183–187, 2012.

[99] I. Nemenman. Fluctuation-dissipation theorem and models of learning. *Neural Comput*, 17, 2005.

[100] I. Nemenman. Renormalizing complex models: It is hard without landau. *J. Club Condens. Matter Phys*, 2017.

[101] I. Nemenman. Ai theorist? not yet. https://www.condmatjclub.org/uploads/2020/05/JCCM_May_2020_02.pdf, 2020.

[102] I. Nemenman, F. Shafee, and W. Bialek. Entropy and inference, revisited. In *Advances in neural information processing systems*, pages 471–478, 2002.

[103] L. G. Nowak, R. Azouz, M. V. Sanchez-Vives, C. M. Gray, and D. A. McCormick. Electrophysiological classes of cat primary visual cortical neurons in vivo as revealed by quantitative analyses. *Journal of neurophysiology*, 89(3):1541–1566, 2003.

[104] D. J. Obbard, J. Maclennan, K.-W. Kim, A. Rambaut, P. M. O'Grady, and F. M. Jiggins. Estimating divergence dates and substitution rates in the Drosophila phylogeny. *Molecular Biology and Evolution*, 29(11):3459–3473, Nov. 2012.

[105] A. B. Owen. *Monte Carlo theory, methods and examples.* 2013.

[106] T. D. Pereira, D. E. Aldarondo, L. Willmore, M. Kislin, S. S. H. Wang,

M. Murthy, and J. W. Shaevitz. Fast animal pose estimation using deep neural networks. *Nature Methods*, 16(1):117–125, 2018.

[107] S. Pillay, M. J. Ward, A. Peirce, and T. Kolokolnikov. An asymptotic analysis of the mean first passage time for narrow escape problems: Part i: Two-dimensional domains. *Multiscale Modeling and Simulation*, 8(3):803–835, 2010.

[108] A. Raj and A. van Oudenaarden. Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell*, 135(2):216–226, 2008.

[109] W. M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.

[110] S. Rao, A. Van der Schaft, K. Van Eunen, B. M. Bakker, and B. Jayawardhana. A model reduction method for biochemical reaction networks. *BMC systems biology*, 8(1):52, 2014.

[111] S. Redner. *A guide to first-passage processes*. Cambridge University Press, 2001.

[112] J. Rissanen. Hypothesis selection and testing by the mdl principle. *The Computer Journal*, 42(4):260–269, 1999.

[113] R. Rodieck, N.-S. Kiang, and G. Gerstein. Some quantitative methods for the study of spontaneous activity of single neurons. *Biophysical Journal*, 2(4):351–368, 1962.

[114] F. Santamaria, D. Jaeger, E. De Schutter, and J. M. Bower. Modulatory effects of parallel fiber and molecular layer interneuron synaptic activity on purkinje cell responses to ascending segment input: a modeling study. *Journal of computational neuroscience*, 13(3):217–235, 2002.

[115] G. Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

[116] A. S. Seetharam and G. W. Stuart. Whole genome phylogeny for 21 Drosophila species using predicted 2b-RAD fragments. *PeerJ*, 1:e226, 2013.

[117] K. L. Shaw and S. C. Lesnick. Genomic linkage of male song and female acoustic preference qtl underlying a rapid species radiation. *Proceedings of the National Academy of Sciences*, 106(24):9737–9742, 2009.

[118] J. Y. Shih, C. A. Atencio, and C. E. Schreiner. Improved stimulus representation by short interspike intervals in primary auditory cortex. *Journal of neurophysiology*, 105(4):1908–1917, 2011.

[119] N. Shubin, C. Tabin, and S. B. Carroll. Deep homology and the origins of evolutionary novelty. *Nature*, 457(7231):818–823, 2009.

[120] A. J. Siegert. On the first passage time probability problem. *Physical Review*, 81(4):617, 1951.

[121] N. Sinitsyn, N. Hengartner, and I. Nemenman. Adiabatic coarse-graining and simulations of stochastic biochemical networks. *Proceedings of the National Academy of Sciences*, 106(26):10546–10551, 2009.

[122] N. Slonim, G. S. Atwal, G. Tkačik, and W. Bialek. Information-based clustering. *Proceedings of the National Academy of Sciences*, 102(51):18297–18302, 2005.

[123] M. Slutsky and L. A. Mirny. Kinetics of protein-dna interaction: facilitated target location in sequence-dependent potential. *Biophysical journal*, 87(6):4021–4035, 2004.

[124] A. F. Smith and G. O. Roberts. Bayesian computation via the gibbs sampler

and related markov chain monte carlo methods. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(1):3–23, 1993.

[125] M. Steriade, P. Wyzinski, and V. Apostol. Differential synaptic reactivity of simple and complex pyramidal tract neurons at various levels of vigilance. *Experimental brain research*, 17(1):87–110, 1973.

[126] D. L. Stern, J. Crocker, Y. Ding, N. Frankel, G. Kappes, E. Kim, R. Kuzmickas, A. Lemire, J. D. Mast, and S. Picard. Genetic and Transgenic Reagents for Drosophila simulans, D. mauritiana, D. yakuba, D. santomea, and D. virilis. *G3*, 7(4):1339–1347, 2017.

[127] D. L. Stern and N. Frankel. The structure and evolution of cis-regulatory regions: the *shavenbaby* story. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1632):20130028, 2013.

[128] D. Strouse and D. J. Schwab. The deterministic information bottleneck. *Neural Computation*, 29(6):1611–1630, 2017.

[129] F. Tajima. Simple methods for testing the molecular evolutionary clock hypothesis. *Genetics*, 135(2):599–607, 1993.

[130] L. Tierney and J. B. Kadane. Accurate approximations for posterior moments and marginal densities. *Journal of the american statistical association*, 81(393):82–86, 1986.

[131] N. Tinbergen. *The Study of Instinct*. Oxford University Press, Oxford, U. K., 1951.

[132] N. Tishby, F. C. Pereira, and W. Bialek. The information bottleneck method. In *Proceedings of the 37th Annual Allerton Conference on Communication, Control*

and Computing, pages 368–377. University of Illinois Press, Urbana-Champaign, IL, 1999.

[133] E. TMLS. Learning new physics with machine learning. https://www.youtube.com/watch?v=DRh1OlGlRxo&t=1927s, 2020.

[134] D. Tolhurst, J. A. Movshon, and I. Thompson. The dependence of response amplitude and variance of cat visual cortical neurones on stimulus contrast. *Experimental brain research*, 41(3-4):414–419, 1981.

[135] M. K. Transtrum, B. B. Machta, K. S. Brown, B. C. Daniels, C. R. Myers, and J. P. Sethna. Perspective: Sloppiness and emergent theories in physics, biology, and beyond. *The Journal of chemical physics*, 143(1):07B201_1, 2015.

[136] Y. Tsubo, Y. Isomura, and T. Fukai. Power-law inter-spike interval distributions infer a conditional maximization of entropy in cortical neurons. *PLoS Comput Biol*, 8(4):e1002461, 2012.

[137] H. C. Tuckwell. *Introduction to theoretical neurobiology: volume 2, nonlinear and stochastic theories*. Cambridge University Press, 1988.

[138] H. C. Tuckwell and D. K. Cope. Accuracy of neuronal interspike times calculated from a diffusion approximation. *Journal of Theoretical Biology*, 83(3):377–387, 1980.

[139] H. C. Tuckwell and W. Richter. Neuronal interspike time distributions and the estimation of neurophysiological and neuroanatomical parameters. *Journal of theoretical biology*, 71(2):167–183, 1978.

[140] H. C. Tuckwell, R. Rodriguez, and F. Y. Wan. Determination of firing times for the stochastic fitzhugh-nagumo neuronal model. *Neural computation*, 15(1):143–159, 2003.

[141] H. C. Tuckwell and F. Y. Wan. Time to first spike in stochastic hodgkin–huxley systems. *Physica A: Statistical Mechanics and its Applications*, 351(2-4):427–438, 2005.

[142] A. Valleriani, X. Li, and A. B. Kolomeisky. Unveiling the hidden structure of complex stochastic biochemical networks. *The Journal of chemical physics*, 140(6):02B608_1, 2014.

[143] L. van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

[144] N. G. Van Kampen. *Stochastic processes in physics and chemistry*, volume 1. Elsevier, 1992.

[145] A. Veksler and A. B. Kolomeisky. Speed-selectivity paradox in the protein search for targets on dna: is it real or not? *The journal of physical chemistry B*, 117(42):12695–12701, 2013.

[146] D. L. Wallace. Asymptotic approximations to distributions. *The Annals of Mathematical Statistics*, 29(3):635–654, 1958.

[147] J. N. Webert, B. K. Peterson, and H. E. Hoekstra. Discrete genetic modules are responsible for complex burrow evolution in peromyscus mice. *Nature*, 493:402–405, 01 2013.

[148] G. H. Weiss. First passage time problems in chemical physics. 1967.

[149] S. Weiss. Fluorescence spectroscopy of single biomolecules. *Science*, 283(5408):1676–1683, 1999.

[150] M. J. West-Eberhard. *Developmental plasticity and evolution*. Oxford University Press, Oxford, U.K., 2003.

[151] T. M. Williams and S. B. Carroll. Genetic and molecular insights into the development and evolution of sexual dimorphism. *Nature Reviews Genetics*, 10(11):797–804, 2009.

[152] K. G. Wilson. Renormalization group and critical phenomena. i. renormalization group and the kadanoff scaling picture. *Physical review B*, 4(9):3174, 1971.

[153] K. G. Wilson. Renormalization group and critical phenomena. ii. phase-space cell analysis of critical behavior. *Physical Review B*, 4(9):3184, 1971.

[154] S. Xie. Single-molecule approach to enzymology. *Single Molecules*, 2(4):229–236, 2001.

[155] S. Xie and J. K. Trautman. Optical studies of single molecules at room temperature. *Annual review of physical chemistry*, 49(1):441–480, 1998.

[156] D. Yamamoto and Y. Ishikawa. Genetic and neural bases for species-specific behavior in drosophila species. *Journal of Neurogenetics*, 27(3):130–142, 2013.

[157] D. Yamamoto and Y. Ishikawa. Genetic and Neural Bases for Species-Specific Behavior in Drosophila Species. *Journal of Neurogenetics*, 27(3):130–142, 2013.

[158] Z. Yang et al. *Computational molecular evolution*, volume 284. Oxford University Press Oxford, Oxford, U.K., 2006.

[159] Y. Zhang and O. K. Dudko. First-passage processes in the genome. *Annual review of biophysics*, 45:117–134, 2016.