BAYESIAN NONPARAMETRICS FOR BIOPHYSICS

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Meysam Tavakoli

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

May 2020

Purdue University

West Lafayette, Indiana

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF DISSERTATION APPROVAL

Dr. Andrew D. Gavrin, Chair

Indiana University-Purdue University Indianapolis

School of Science, Department of Physics

Dr. Steve Pressé

Arizona State University

School of Science, Department of Physics

Dr. Horia Petrache

Indiana University-Purdue University Indianapolis

School of Science, Department of Physics

Dr. Stephen R. Wassall

Indiana University-Purdue University Indianapolis

School of Science, Department of Physics

Dr. Andrew Mugler

Purdue University

School of Science, Department of Physics

Approved by:

Dr. Andrew D. Gavrin

Head of the Physics Department

To my wife, Azin, who provided me with the greatest support and love, also, to the sweetest thing in my life, Hamta.

ACKNOWLEDGMENTS

I would like to express my sincere appreciation to many people who have played important roles during my PhD career.

First and foremost, I would like to thank my advisor, Dr. Steve Pressé, for teaching, and supporting me over the past five years. It has been a real pleasure to work with you during these years. You have consistently encouraged me to preform my best with frequent discussions and showed a lot of patience during my studies. I will never forget how patient you were at the beginning of our work, and how I learned modeling and inference from you. In a word, you shaped my future career. I enjoyed working with you and I will always be grateful to you. My last word for you is "you are one of my BEST friends".

I am also grateful to chair of my committee, Dr. Andrew D. Gavrin, head of Physics Department at IUPUI, with whom I had the privilege of working for the last two years. I also thank Dr. Gavrin for his valuable feedback on my thesis and more importantly my future career. I enjoyed working with you both during my thesis and my teaching duties.

I am also grateful to my other committee members, Dr. Horia Petrache and Dr. Stephen R. Wassall from Department of Physics at IUPUI, and Dr. Andrew Mugler from Purdue University at West Lafayette for their helpful suggestions and valuable feedback on my thesis as well as their encouragement during my studies.

I would like to thank Dr. Ricardo Decca for his consistent assistance with administrative forms. As the graduate program coordinator, he had to deal with official forms and documentations not only during my application process, but also with my graduation forms.

I would like to thank everyone in our friendly Department of Physics including all faculties whom I took valuable courses with during my PhD, the helpful administra-

tive staff, and all the fellow graduate students for their support and warm interactions over these years.

I would like to specially thank the following people from Dr. Steve Pressé's inference group from Arizona State University for their help and participation in scientific discussions: Dr. Ioannis Sgouralis, Postdoctoral Research Fellow, and Dr. Zeliha Kilic, Postdoctoral Research Fellow. Moreover, I express my sincere appreciation to my valuable friend, Sina Jazani, graduate student, who enormously helped and supported me regarding all aspects of the projects in this thesis.

I would like to thank the IUPUI Physics Department for financially supporting me for the last five years.

Finally, I am very grateful to my family and my friends for their constant support.

TABLE OF CONTENTS

Page

Page

LIST OF FIGURES

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

ABSTRACT

Tavakoli, Meysam Ph.D., Purdue University, May 2020. Bayesian Nonparametrics for Biophysics. Major Professor: Andrew D. Gavrin.

The main goal of data analysis is to summarize huge amount of data (as our observation) with a few numbers that come up us with some sort of intuition into the process that generated the data. Regardless of the method we use to analyze the data, the process of analysis includes (1) create the mathematical formulation for the problem, (2) data collection, (3) create a probability model for the data, (4) estimate the parameters of the model, and (5) summarize the results in a proper way-a process that is called "statistical inference".

Recently it has been suggested that using the concept of Bayesian approach and more specifically Bayesian nonparametrics (BNPs) is showed to have a deep influence in the area of data analysis [1], and in this field, they have just begun to be extracted [2–4]. However, to our best knowledge, there is no single resource yet available that explain it, both its concepts, and implementation, as would be needed to bring the capacity of BNPs to relieve on data analysis and accelerate its unavoidable extensive acceptance.

Therefore, in this dissertation, we provide a description of the concepts and implementation of an important, and computational tool that extracts BNPs in this area specifically its application in the field of biophysics. Here, the goal is using BNPs to understand the rules of life (in vivo) at the scale at which life occurs (single molecule) from the fastest possible acquirable data (single photons).

In chapter 1, we introduce a brief introduction to Data Analysis in biophysics. Here, our overview is aimed for anyone, from student to established researcher, who plans to understand what can be accomplished with statistical methods to modeling

and where the field of data analysis in biophysics is headed. For someone just getting started, we present a special on the logic, strengths and shortcomings of data analysis frameworks with a focus on very recent approaches.

In chapter 2, we provide an overview on data analysis in single molecule biophysics. We discuss about data analysis tools and model selection problem and mainly Bayesian approach. We also discuss about BNPs and their distinctive characteristics that make them ideal mathematical tools in modeling of complex biomolecules as they offer meaningful and clear physical interpretation and let full posterior probabilities over molecular-level models to be deduced with minimum subjective choices.

In chapter 3, we work on spectroscopic approaches and fluorescence time traces. These traces are employed to report on dynamical features of biomolecules. The fundamental unit of information came from these time traces is the single photon. Individual photons have information from the biomolecule, from which they are emitted, to the detector on timescales as fast as microseconds. Therefore, from confocal microscope viewpoint it is theoretically feasible to monitor biomolecular dynamics at such timescales. In practice, however, signals are stochastic and in order to derive dynamical information through traditional meanssuch as fluorescence correlation spectroscopy (FCS) and related methodsfluorescence time trace signals are gathered and temporally auto-correlated over many minutes. So far, it has been unfeasible to analyze dynamical attributes of biomolecules on timescales near data acquisition as this requests that we estimate the biomolecule numbers emitting photons and their locations within the confocal volume. The mathematical structure of this problem causes that we leave the normal ("parametric") Bayesian paradigm. Here, we utilize novel mathematical tools, BNPs, that allow us to extract in a principled fashion the same information normally concluded from FCS but from the direct analysis of significantly smaller datasets starting from individual single photon arrivals. Here, we specifically are looking for diffusion coefficient of the molecules. Diffusion coefficient allows molecules to find each other in a cell and at the cellular level, determination of the diffusion coefficient can provide us valuable insights about how molecules in-

teract with their environment. We discuss the concepts of this method in assisting significantly reduce phototoxic damage on the sample and the ability to monitor the dynamics of biomolecules, even down to the single molecule level, at such timescales.

In chapter 4, we present a new approach to infer lifetime. In general, fluorescence Lifetime Imaging (FLIM) is an approach which provides us information on the number of species and their associated lifetimes. Current lifetime data analysis methods rely on either time correlated single photon counting (TCSPC) or phasor analysis. These methods require large numbers of photons to converge to the appropriate lifetimes and do not determine how many species are responsible for those lifetimes. Here, we propose a new method to analyze lifetime data based on BNPs that precisely takes into account several experimental complexities. Using BNPs, we can not only identify the most probable number of species but also their lifetimes with at least an order magnitudes less data than competing methods (TCSPC or phasors). To evaluate our method, we test it with both simulated and experimental data for one, two, three and four species with both stationary and moving molecules. Also, we compare our species estimate and lifetime determination with both TCSPC and phasor analysis for different numbers of photons used in the analysis.

In conclusion, the basis of every spectroscopic method is the detection of photons. Photon arrivals encode complex dynamical and chemical information and methods to analyze such arrivals have the capability to reveal dynamical and chemical processes on fast timescales. Here, we turn our attention to fluorescence lifetime imaging and single spot fluorescence confocal microscopy where individual photon arrivals report on dynamics and chemistry down to the single molecule level. The reason this could not previously be achieved is because of the uncertainty in the number of chemical species and numbers of molecules contributing for the signal (i.e., responsible for contributing photons). That is, to learn dynamical or kinetic parameters (like diffusion coefficients or lifetime) we need to be able to interpret which photon is reporting on what process. For this reason, we abandon the parametric Bayesian paradigm and use the nonparametric paradigm that allows us to flexibly explore and learn numbers

of molecules and chemical reaction space. We demonstrate the power of BNPs over traditional methods in single spot confocal and FLIM analysis in fluorescence lifetime imaging.

1. INTRODUCTION TO DATA ANALYSIS IN BIOPHYSICS

Some parts of this section are coming from introduction part of below paper which also appears in arXiv: https://arxiv.org/pdf/1606.00403.pdf, and published in Advances in Chemical Physics.
Meysam Tavakoli, J. Nicholas Taylor, Chun-Biu Li, Tamiki Komatsuzaki, Steve Pressé*. (2017). Single Molecule Data Analysis: An Introduction. In Advances in Chemical Physics (eds S.A. Rice and A.R. Dinner).
DOI:10.1002/9781119324560.ch4

The traditional route to model-building in chemistry and physics – a strategy by which implausible hypotheses are eliminated to arrive at a quantitative framework – has been successfully applied to the realm of biological physics [5]. For instance, polymer models have predicted how DNA's extension depends on externally applied force [6] while thermodynamic models – with deep origin-of-life implications – explain how lipid vesicles trapping long RNA molecules grow at the expense of neighboring vesicles trapping shorter RNA segments in buffer [7].

Another modeling route is the atomistic – molecular dynamics (MD) – approach in which one investigates complex systems by monitoring the evolution of their many degrees of freedom. Novel algorithms along with machine architectures for high-speed MD simulations of biological macromolecules have now even allowed small proteins to be folded into their native state [8].

Both routes have important *pros* but they also have *cons*. For instance, potentials in MD are constructed to reproduce behaviors in regimes for which they are parametrized and cannot rigorously treat chemical reactions at the heart of biology.

In general, the goal of experimental science is to relate data (i.e. observations) into biological information (e.g., rate constants, lifetimes, and diffusion coefficients). Unfortunately, the experimental methods are utilized in biological physics or similar areas rarely and directly extract useful biological information [9]. Instead, the researchers have to "analyse" the data in such a way to extract biological information from the data. Furthermore, it is not always clear how simple physics-based models – while intuitive – should be adapted to treat complex biological data [10–12]. For example, diffusion in complex environments – such as telomeres inside mammalian cell nuclei [13]; bacterial chromosomal loci [14] and mRNA inside the bacterial cytoplasm [15]; and viruses inside infected cells [16] – has often been termed "anomalous". This is because "normal" diffusion models often fail in living systems where there is molecular crowding [17–21], biomolecular interactions and binding [10, 22–26] and active transport [27–31].

In the complement of above explanation, to the average of experimental physicists and chemists, the area of data analysis is so complex that one simply refers data analysis to an expert or uses software packages to analyse the data [32]. The main issue with this idea is the fact that data analysis experts and use of computer software often have no real concept of the science behind the experiments [33, 34]. The important point here is to reveal that people can often effect a major transformation of the amount of experimental information which can be acquired from a set of experiments by direct incorporation of "scientific concepts" into the analysis of the data [9, 35].

Drastic revolutions in the natural sciences have often been triggered by new observations and new observations, in turn, have presented important modeling challenges [36]. These challenges now include the heterogeneity of data collected at room temperature or in living systems [10, 37, 38] and the noise that rattles nanoscale systems [39–43].

Necessity is the mother of invention and these new challenges have motivated statistical, data-driven, approaches to model-building in biophysics that explicitly deal with various sources of uncertainty [44–52].

Statistical data-driven analysis methods are the focus of this dissertation. Statistical approaches have been invaluable in generating detailed mechanistic insight into realms previously inaccessible at every step along biology's central dogma [53, 54]. They have also unveiled basic molecular mechanisms from noisy single molecule data that have given rise to detailed energy landscape [55–61] and kinetic scheme [41,62–66] models. Furthermore, one's choice of analysis methods can deeply alter the interpretation of experiment [67, 68].

We focus this review on parametric as well as more recent information theoretic and non-parametric statistical approaches to biophysical data analysis with an emphasis on single molecule applications. We review simpler parametric approaches starting from an assumed model with unknown parameters. We later expand our discussion to include information theoretic and non-parametric approaches that have broadened our perspective beyond a strict "parametric" requirement that the model be fully specified from the onset [10, 11, 47, 69, 70].

These more general methods have, under some assumptions, relaxed important requirements to: know the fluorophore photophysics *a priori* to count single molecules from superresolution imaging [45]; prespecify the number of states in the analysis of single molecule fluorescence resonance energy transfer (smFRET) time traces [47]; or the number of diffusion components contributing to fluorescence correlation spectroscopy curves [10, 71, 72]; or the number of diffusion coefficients sampled by cytoplasmic proteins from single protein tracking trajectories [46]. In fact, these efforts bring us closer towards a non-parametric treatment of the data.

As so many model selection [73, 74] and statistical methods developed to tackle problems in physics, and later biophysics, have been motivated by Shannon's information [75, 76], we take an information theoretic approach whenever helpful [77, 78]. Beyond model selection, topics we will discuss also include parameter estimation, image deconvolution, outliers, change-point detection, clustering and state identification.

In this review, we do not focus on how methods are implemented algorithmically. Rather, we cite the appropriate literature as needed. Neither do we discuss the experimental methods from which the data are drawn [79]. Since our focus is on an introduction to data analysis for single molecule, there are also many topics in data analysis that we do not discuss at all or in much detail (p-values, type I and II errors, point estimates, hypothesis testing, likelihood ratio tests, credible intervals, bootstrapping, Kalman filtering, single particle tracking, localization, feature modeling, aspects of density estimation, etc...).

Our review [79] is intended for anyone, from student to established researcher, who wants to understand what can be accomplished with statistical approaches to modeling and where the field of data analysis in biophysics is headed. For someone just getting started, we place a special emphasis on the logic, strengths and shortcomings of different data analysis frameworks with a focus on very recent approaches.

2. SINGLE MOLECULE DATA ANALYSIS: AN INTRODUCTION

This paper appears in arXiv: https://arxiv.org/pdf/1606.00403.pdf, and published in Advances in Chemical Physics.
Meysam Tavakoli, J. Nicholas Taylor, Chun-Biu Li, Tamiki Komatsuzaki, Steve Pressé*. (2017). Single Molecule Data Analysis: An Introduction. In Advances in Chemical Physics (eds S.A. Rice and A.R. Dinner).
DOI:10.1002/9781119324560.ch4
Contribution: MT conceived, designed, and wrote the theory for sections "Bayesian Parametric Approaches", "Model Selection", and "Introduction to Bayesian Nonparametrics". SP conceived, designed, and wrote the theory for sections "Final Thoughts on Data Analysis", and "Information Theory", and "Information Theory as a Data Analysis Tool". JNT, CBL, and TK conceived, designed, and wrote the theory for section "Information Theory: State Identification and Clustering". MT and SP expanded and revised the write up. SP oversaw all aspects of the review.
This paper appears in arXiv: https://arxiv.org/pdf/1606.00403.pdf.

2.1 Author Summary

This chapter considers statistical data-driven analysis methods, and focuses on parametric as well as more recent information theoretic and nonparametric statistical approaches to biophysical data analysis with an emphasis on single-molecule applications. It then reviews simpler parametric approaches starting from an assumed model with unknown parameters. Model selection criteria are widely used in biophysical data analysis from image deconvolution to single-molecule step detection and continue to

be developed by statisticians. The goal of successful model selection criteria is to pick models whose complexity is penalized, in a principled fashion, to avoid overfitting and that convincingly fit the data provided (the training set). The chapter summarizes both information theoretic as well as Bayesian model selection criteria. Finally, the chapter discusses efforts to use information theory in experimental design and ends with some considerations on the broader applicability of information theory.

2.2 Frequentist and Bayesian Parametric Approaches: A Brief Review

2.2.1 Frequentist inference

Conceptually, the simplest data-driven approach is *parametric* and *frequentist*. By "parametric", we mean a model M is pre-specified and its parameters, $\boldsymbol{\theta} = \{\theta_1, \theta_2, \cdots, \theta_K\}$, are unknown and to be determined from the data, $\mathbf{D} = \{D_1, D_2, \cdots, D_N\}$.

By "frequentist", we mean that model parameters are determined exclusively from frequencies of repeated experiments by contrast to being informed by prior information, which we will turn to shortly in our discussion of Bayesian methods.

Model parameters can, in principle, be determined by binning and subsequently fitting histograms, such as histograms of photon arrivals. Fitting histograms is avoided in data analysis in both frequentist and Bayesian methods.

In particular, to avoid selecting an arbitrary histogram bin size and having to collect enough data to build a reliable histogram, model parameters are determined by maximizing the probability, $p(\mathbf{D}|M)$, of observing the sequence of outcomes under the assumptions of a model whose parameter values, $\boldsymbol{\theta}$, have yet to be determined. This probability, $p(\mathbf{D}|M) = p(\mathbf{D}|\boldsymbol{\theta})$, is termed the likelihood which is a central object in frequentist inference.

For multiple independent data points, the likelihood is the product over each independent observation

$$p(\mathbf{D}|\boldsymbol{\theta}) = \prod_i p(D_i|\boldsymbol{\theta}). \tag{2.1}$$

As an example of maximum likelihood estimation, suppose our goal is to estimate a molecular motor's turnover rate r from a single measurement of the number of stepping events, n, in some time interval ΔT. We begin by pre-specifying a model: the probability of observing n events is Poisson distributed. Under these assumptions, our likelihood is

$$p(D = n|\theta = r) = \frac{(r\Delta T)^n}{n!} e^{-r\Delta T}. \tag{2.2}$$

Maximizing this likelihood with respect to r yields the estimator $\hat{r} = n/\Delta T$. That is, it returns the most likely turnover rate under the assumptions of the model.

We can also write likelihoods to explicitly account for correlations in time in a time series (a sequence of data points ordered in time) even for continuous time. For instance, the likelihood – for a series of events occurring at times $\mathbf{D} = \mathbf{t} = \{t_1, t_2, \cdots, t_N\}$ in continuous time with possible time correlations – is

$$p(\mathbf{D} = \mathbf{t}|\boldsymbol{\theta}) = p(t_N|t_{N-1}, \cdots, t_1, \boldsymbol{\theta}) \cdots p(t_2|t_1, \boldsymbol{\theta}) p(t_1|\boldsymbol{\theta}) = \prod_{i=2}^{N} [p(t_i|\{t_j\}_{j<i}, \boldsymbol{\theta})] p(t_1|\boldsymbol{\theta}). \tag{2.3}$$

Returning to our molecular motor example, we can also investigate how sharply peaked our likelihood is around $r = \hat{r}$ to give us an estimate for the variability around $\hat{r}$. A lower bound on the variance – with $var(r) \equiv E(r^2) - (E(r))^2$ where "$\equiv$" denotes a definition, not an equality, and E denotes an expectation – evaluated at $\hat{r}$ is given by the inverse of the expectation of the likelihood's curvature

$$var(\hat{r}) \geq \frac{1}{-E(\partial_r^2 \log p(n|r))}. \tag{2.4}$$

Intuitively, this shows that the variance is inversely proportional to the likelihood's sharpness at its maximum. This inequality is called the Cramér-Rao bound and the denominator of the right hand side is called the Fisher information [80]. A formal proof of this bound follows from the Cauchy-Schwarz inequality [80]. However, informally, the equality of the above bound can be understood as follows. Consider our likelihood as a product of independent observations

$$p(\mathbf{D}|r) = \prod_i p(D_i|r) \equiv e^{N \log f(\mathbf{D}|r)} \tag{2.5}$$

where we have defined $f(\mathbf{D}|r)$ through the expression above and $f(\mathbf{D}|r)$ is a function that scales like N^0. We then expand the likelihood around its maximum $r = r*$

$$p(\mathbf{D}|r) = e^{N \log f(\mathbf{D}|r)} = e^{N \log f(\mathbf{D}|r*) + N \frac{(r-r*)}{2!}^2 \partial_r^2 \log f(\mathbf{D}|r*) + R} \tag{2.6}$$

where R is the remainder. For large enough N, a quadratic expansion of the likelihood, Eq. (2.6), is a sufficiently good approximation to the exact $p(\mathbf{D}|r)$. By this same reasoning, the $var(r)$ that we compute using the approximate $p(\mathbf{D}|r)$ is a good approximation to the exact $var(r)$. Only when the quadratic expansion is exact – and R is zero – do we recover the lower Cramér-Rao bound upon computing the variance. In fact, only in this limit do r^* and $\hat{r}$ coincide.

Maximum likelihood estimation: Applications to hidden and aggregated Markov models

While our molecular motor example is simple and conceptual, this frequentist approach, that we now discuss in more detail, has been used to learn kinetic rates of protein folding [81] or protein conformational changes [82].

As a more realistic example, we imagine a noisy two state trajectory with high and low signal. Fig. (2.1a) is a cartoon of an experimental single molecule force spectroscopy setup that generates the types of time traces – transitions of an RNA hairpin between zipped, unzipped and an intermediate state – shown in Fig. (2.1b) [68]. The noise level around the signal is high enough that the peaks of the intensity histograms (shown in grey at the extreme right of Fig. (2.1b)) overlap. Thus, even in the idealized case where there is no drift in the time trace over time, it is not possible to draw straight lines through the time trace of Fig. (2.1b) in order to establish in what state the system finds itself in at any point in time. In fact, looking for crossings of horizontal lines as an indication of a state change would grossly overcount the number of transitions between states.

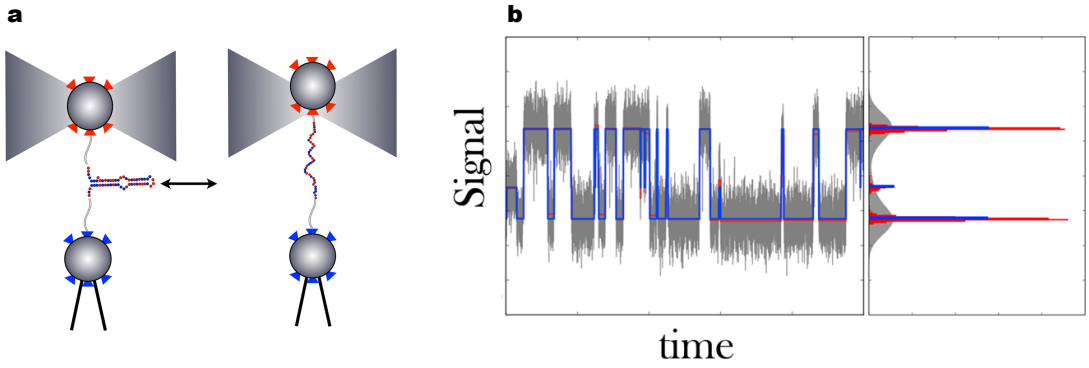

Fig. 2.1. **Single molecule experiments often generate time traces. The goal is to infer models of single molecule behavior from these time traces. a)** A cartoon of a single molecule force spectroscopy setup probing transitions between zipped and unzipped states of an RNA hairpin [68]. Change-point algorithms, that we later discuss, were used in **b)** to determine when the signal suddenly changes (red line). The signal indicates the changes in the conformation of the RNA hairpin obscured by noise. Clustering algorithms, also discussed later, were then used to regroup the "denoised" intensity levels (red line) into distinct states (blue line).

Markov models:

Time series analysis is an important topic in single molecule inference as is treating noise explicitly on a pathwise basis in single particle tracking data [83, 84] or in two-level time traces [41]. Here we will not deal directly with single particle tracking problems [85] nor the extensive literature on Kalman filtering in time series analysis [86].

Rather, we immediately focus on Hidden Markov models, HMMs, commonly used in time series analysis that deal directly with the types of time traces shown in Fig. (2.1b). Before we tackle noisy time traces, we discuss idealized traces with no noise (such as thermal noise that rattles molecules, effective noise from unresolved motion on fast timescales as well as measurement noise). Our goal here is to extract transition rates between states observable in our noiseless time trace without histogramming data.

Markov models start by assuming that the system can occupy a total of K states and that transitions between these states are fully described by a transition matrix, $\mathbf{A}$, whose matrix elements, a_{ij}, coincide with the transition probability of state s_i to state s_j, $a_{ij} = p(s_j|s_i)$. Given idealized (noiseless) time traces for now, our goal is to find the model parameters, $\boldsymbol{\theta}$, which consist of all transition matrix elements as well as all initial state probabilities. That is, $p(s_j|s_i)$ for all pairs of states i and j and $p(s_i)$ for all i.

To obtain these parameters, we define a likelihood, for each trajectory, of having observed a definite state sequence $\mathbf{D} = \{s_1, s_2, ..., s_N\}$ – where $\{s_1, s_2, ..., s_N\}$ here are numbers that serve as labels for states – in discrete time

$$\mathrm{L}(\boldsymbol{\theta}|\mathbf{D}) \equiv p(s_1, s_2, ..., s_N|\boldsymbol{\theta}) = \prod_{i=2}^{N} \left[p(s_i|s_{i-1})\right] p(s_1). \tag{2.7}$$

We have written the transition probability from time t to $t + \delta t$ as $p(s_{t+\delta t}|s_t)$ where s_t is the state the system finds itself in at time t. We subsequently maximize these likelihoods with respect to all unknown parameters, $\boldsymbol{\theta}$. We add that we need more than one trajectory to estimate the initial state probabilities.

Finally, while the likelihood is the probability of the data given the model, the likelihood is typically maximized with respect to its parameters and its parameters are treated as its variables. For this reason we write $L(\boldsymbol{\theta}|\mathbf{D})$ not $L(\mathbf{D}|\boldsymbol{\theta})$.

Hidden Markov models:

While Eq. (2.7) is used for theoretical illustrations [49, 87, 88], it must be augmented to treat noise for real data analysis applications. These resulting models, HMMs [41, 89, 90], have been broadly used in single molecule analysis including smFRET studies [58, 91–95] and force spectroscopy [82].

In HMMs, the state of the signal in time, termed the state of the "latent" or hidden variable, is provided indirectly through a sequence of observations $\mathbf{D} = \mathbf{y} = \{y_1, y_2, ..., y_N\}$. Often this relation is captured by the probability of making the observation y_i given that the system is in state s_i, $p(y_i|s_i)$.

We can use a distribution over observations of the form $p(y_i|s_i)$ under the assumption that the noise is uncorrelated in time and that the observable only depends on the state of the underlying system at that time. A Gaussian form for $p(y_i|s_i)$ – following from the central limit theorem or as an approximation to the Poisson distribution – is common [41, 91].

The HMM model parameters, $\boldsymbol{\theta}$, now include, as before, all transition matrix elements as well as all initial state probabilities. In addition, $\boldsymbol{\theta}$ also includes the parameters used to describe $p(y_i|s_i)$. For example, for a Gaussian distribution over

observations, where μ_k and σ_k designate the mean and variance of the signal for the system in state k at time point i, we have

$$p(y_i|s_i = k) \propto e^{-\frac{(y_i-\mu_k)^2}{2\sigma_k^2}} \tag{2.8}$$

where the additional parameters include the means and variances for each state. For this reason, to be clear, we may have chosen to make our probability over observations depend explicitly on $\boldsymbol{\theta}$, $p(y_i|s_i, \boldsymbol{\theta})$.

In discrete time, where i denotes the time index, the likelihood used for HMMs is

$$\mathrm{L}(\boldsymbol{\theta}|\mathbf{D}) = p(\mathbf{y}|\boldsymbol{\theta}) = \sum_{\mathbf{s}} p(\mathbf{y}, \mathbf{s}|\boldsymbol{\theta}) = \sum_{\mathbf{s}} \prod_{i=2}^{N} \left[p(y_i|\mathbf{s}_i)p(\mathbf{s}_i|\mathbf{s}_{i-1})\right] p(y_1|\mathbf{s}_1)p(\mathbf{s}_1) \tag{2.9}$$

where $\mathbf{s} = \{\mathbf{s_1}, \cdots, \mathbf{s_N}\}$. Unlike in Eq. (2.7), in Eq. (2.9), the $\mathbf{s}_i$ are bolded as they are variables not numbers. We sum over these state variables since we are not, in general, interested in knowing the full distribution $p(\mathbf{y}, \mathbf{s}|\boldsymbol{\theta})$. Rather we are only interested in obtaining the marginal likelihood $p(\mathbf{y}|\boldsymbol{\theta})$ describing the probability of the observation given the model irrespective of what state the system occupied at each point in time.

Put differently, while the system only occupies a definite state at any given time point, we are unaware of what state the system is in. And, for this reason, we must sum over all possibilities.

An alternative way to represent the HMM is to say

$$\begin{aligned} s_1 &\sim p(\mathbf{s_1}) \\ s_i|s_{i-1} &\sim p(\mathbf{s_i}|\mathbf{s_{i-1}}) \\ y_i|s_i, \theta &\sim p(y_i|\mathbf{s}_i, \boldsymbol{\theta}). \end{aligned} \tag{2.10}$$

That is s_1 – a number, i.e. a realization of $\mathbf{s_1}$ – is sampled from $p(\mathbf{s_1})$. Then for any $i > 1$, $s_i|s_{i-1}$ – a realization of $\mathbf{s_i}$ conditioned on s_{i-1} – is sampled from the conditional $p(\mathbf{s_i}|\mathbf{s_{i-1}})$ while its observation $y_i|s_i, \theta$ is sampled from $p(y_i|\mathbf{s}_i, \boldsymbol{\theta})$.

The goal is now to maximize the likelihood, Eq. (2.9), over each parameter $\boldsymbol{\theta}$. There is a broad literature describing multiple strategies available to numerically evaluate and maximize the likelihood functions generated from HMMs (as well as AMMs described in the next section) [96] including the Viterbi algorithm [91,97], and, most often used, forward-backward algorithms and expectation maximization [89,98].

Aggregated Markov models:

Aggregated Markov models (AMMs) [99] can be thought of as a special case of HMMs in which many states of the latent variable have identical output.

AMMs were popularized in biophysics in the analysis of single ion-channel patch clamp experiments [99–102] since, often, two or more distinct inter-converting molecular states of an ion channel may not be experimentally distinguishable. For example, both states may carry current.

Microscopic states that cannot be distinguished experimentally form an "aggregate of states". In its simplest formulation, AMMs describe transitions between two aggregates (such as the open and closed aggregates of states). Each aggregate is composed of multiple, possible interconverting, microscopic states that cannot be directly observed. Instead, each aggregate of states belongs to an "observability class". For instance, one can say that a particular microscopic state belongs to the "open observability class" for an ion channel or the "dark observability class" for a fluorophore.

AMMs are relevant beyond ion channels. In smFRET, a low FRET state (the "low fluorescence observability class") could arise from photophysical properties of the fluorophores or an internal state of the labeled protein [92]. In fact, most re-

cently, AMMs have been used to address the single molecule counting problem using superresolution imaging data [45].

For simplicity, consider a rate matrix, $\mathbf{Q}$, containing only two observability classes, 1 and 2,

$$\mathbf{Q} = \begin{bmatrix} \mathbf{Q}_{11} & \mathbf{Q}_{12} \\ \mathbf{Q}_{21} & \mathbf{Q}_{22} \end{bmatrix}. \tag{2.11}$$

The submatrices $\mathbf{Q}_{ij}$ are populated by matrix elements, indexed $k\ell$ say, describing the transition rates from state k in observability class i to state ℓ in observability class j.

The logic from this point forward is identical to the logic of the previous section on HMMs. We must write down a likelihood and subsequently maximize this likelihood with respect to the model parameters. Ignoring noise, the likelihood of observing the sequence of observability classes $\mathbf{D} = \{a_1, a_2, ..., a_N\}$ in continuous time is [103]

$$\mathrm{L}(\boldsymbol{\theta}|\mathbf{D}) = \mathbf{1}^T \cdot \prod_{j=1}^{N-1} \mathbf{G}_{a_j a_{j+1}}(t_j) \cdot \boldsymbol{\pi}_{a_1} \tag{2.12}$$

where the i^{th} element of the column vector, $\boldsymbol{\pi}_{a_1}$, denotes the initial probability of being in state i from the a_1 observability class and where

$$\mathbf{G}_{ab}(t_j) = \mathbf{Q}_{ab} e^{\mathbf{Q}_{aa} t_j}. \tag{2.13}$$

In other words, $k\ell^{th}$ element of $\mathbf{G}_{ab}(t_j)$ is the probability that you enter from the k^{th} state of observability class a, dwell there for time t_j and subsequently transition to the ℓ^{th} state of observability class b. The row vector, $\mathbf{1}^T$, in Eq. (2.12) is used as a mathematical device to sum over all final microscopic states of the observability class, a_N, observed at the last time point. We do so because we only know in which final

observability class we are at the N^{th} measurement, not which microscopic state of the system we are in.

The parameters, $\boldsymbol{\theta}$, here include transitions between all microscopic states across all observability classes as well as initial probabilities for each state within each observability class. Since the number of parameters exceeds the number of observability classes in AMMs, AMMs often yield underdetermined problems [104].

The AMM treatment above can be generalized to include noise or treated in discrete time [105,106]. Both AMMs and HMMs can also be generalized to include the possibility of missed transitions [45,107]. Missed transitions arise in real applications when a system in some state (in HMMs) or observability class (in AMMs), say k, undergoes rapid transitions – for example rapid as compared to t_d, the camera's data acquisition time – to another state, say ℓ. Then the real transition probability in state k must account for all possible missed transitions to ℓ and recoveries back to k that could have occurred within t_d. We account for these missed transitions be resuming over all possible events that could have occurred within the interval t_d. The technical details are described in Refs. [45, 107].

2.2.2 Bayesian inference

Frequentist inference yields model parameter estimates – like $\hat{r}$ we saw earlier which are called "point estimates" – and error bounds. Just as we've treated the data in the previous section as random variables – that is, realizations of an experiment – and model parameters as fixed quantities to be determined, Bayesian analysis treats both the data as well as model parameters as random variables [108]. For the same amount of data that is used in frequentist inference, Bayesian methods return parameter distributions whose usefulness is contingent on the choice of likelihood and prior, which we describe shortly.

Bayesian methods are now widely used across biophysical data analysis [46, 47, 109–115]. For instance, they have been used to infer models describing how mRNA-protein complexes transition between active transport and Brownian motion [110].

Of central importance in Bayesian analysis is the posterior, $p(\mathbf{M}|\mathbf{D})$: the conditional probability over models $\mathbf{M}$ given observations, $\mathbf{D}$, i.e. the probability of the model *after* observations have been made. Since there may be many (choices of) models, we have bolded the model variable, $\mathbf{M}$. By contrast, the probability over $\mathbf{M}$ *before* observations are made, $p(\mathbf{M})$, is called a prior.

We construct the posterior from Bayes' rule (or theorem) using the likelihood and the prior as inputs. That is, we set

$$\begin{aligned} p(\mathbf{D}, \mathbf{M}) &= p(\mathbf{M}, \mathbf{D}) \\ p(\mathbf{M}|\mathbf{D})p(\mathbf{D}) &= p(\mathbf{D}|\mathbf{M})p(\mathbf{M}) \\ p(\mathbf{M}|\mathbf{D}) &= \frac{p(\mathbf{D}|\mathbf{M})p(\mathbf{M})}{p(\mathbf{D})} \end{aligned} \tag{2.14}$$

where $p(\mathbf{D})$ is obtained by normalization from

$$p(\mathbf{D}) = \int d\mathbf{M} p(\mathbf{D}, \mathbf{M}) = \int d\mathbf{M} p(\mathbf{D}|\mathbf{M})p(\mathbf{M}) \tag{2.15}$$

and $p(\mathbf{D}, \mathbf{M})$ is called the joint probability of the model and the data. Typically, in parametric Bayesian inference, when there is a single model, the integration in Eq. (2.15) is meant as an integration over the model's parameters. However there are cases where many parametric models $\mathbf{M}$ are considered and the integration is interpreted as a sum over models (if the models are discrete) and a subsequent integration over their associated parameters (if the parameters are continuous).

Furthermore, we can marginalize (integrate over) posteriors to describe the posterior probability of a particular model, say M_ℓ, from the broader set of models $\mathbf{M}$ irrespective of its associated parameter values ($\boldsymbol{\theta}_\ell$)

$$p(M_\ell|\mathbf{D}) \propto \int d\boldsymbol{\theta}_\ell p(\mathbf{D}|M_\ell, \boldsymbol{\theta}_\ell) p(\boldsymbol{\theta}_\ell|M_\ell) p(M_\ell). \tag{2.16}$$

Here we make it a point to distinguish a model from its parameters, while earlier $\mathbf{M}$ re-grouped both models and their parameters. Furthermore, to be clear, we note that the following notations are equivalent

$$\int d\boldsymbol{\theta}_\ell \leftrightarrow \int d^K\theta \leftrightarrow \int \prod_{k=1}^{K} d^k\theta \tag{2.17}$$

where K designates the total number of parameters, $\boldsymbol{\theta}$. The quantity $p(M_\ell|\mathbf{D})$ can then be used to compare different models head-to-head. For instance, in single particle tracking, we may be interested in computing the posterior probability that a particle's mean square displacement arises from one of many models of transport (Brownian motion versus directed motion) irrespective of any value assigned to parameters such as the diffusion coefficient [111].

Priors

As the number of observations, N, grows, the likelihood determines the shape of the posterior and the choice of likelihood becomes critical as we will illustrate shortly in Fig. (2.2). By the central limit theorem, for sufficiently independent observations, the likelihood function's breadth will narrow with respect to its mean as $N^{-1/2}$. Provided abundant data, more attention should be focused on selecting an appropriate likelihood function than selecting a prior.

However, if provided with insufficient data, our choice of prior may deeply influence the posterior. This is perhaps best illustrated with the extreme example of the canonical distribution in classical statistical physics where posterior distributions – over Avogadro's number of particle positions and velocities – are constructed from just one data point (total average energy with vanishingly small error) [49, 116–118]. That is, the error bar is below the resolution limit of the experiment on a macroscopic system.

While the situation is not quite as extreme in biophysics, data may still be quite limited. For instance, single particle (protein) tracks may be short because protein labels photobleach or particles move in and out of focus [46] or the kinetics into and out of intermediate states may be difficult to quantify in single molecule force spectroscopy for rarely visited conformational states [68].

A good choice of prior is therefore also important. There are two types of priors: informative and uninformative [108, 119].

Uninformative priors

The simplest uninformative prior – inspired from Laplace's principle of insufficient reason when the set of hypotheses are complete and mutually exclusive, such as with dice rolls – is the flat, uniform, distribution. Under the assumption that $p(\mathbf{M})$ is constant, or flat, over some range, the posterior and likelihood are directly related

$$p(\mathbf{M}|\mathbf{D}) \propto p(\mathbf{D}|\mathbf{M})p(\mathbf{M}) \propto p(\mathbf{D}|\mathbf{M}). \tag{2.18}$$

That is, their dependence on $\mathbf{M}$ is identical. However, a flat prior over a model parameter, say θ, is not quite as uninformative as it may appear [108] as, a coordinate transformation to the alternate variable, say e^{θ}, reveals that we suddenly know

more about the random variable e^{θ} than we did about θ since its distribution is no longer flat. Conversely, if e^{θ} is uniform on the interval [0, 1], then θ becomes more concentrated at the upper boundary, 1.

The problem stems from the fact that, under coordinate transformation, if the variable θ's range is from $[0, 1]$, then e^{θ}'s range is from $[1, e]$. To resolve this problem, we can use the Jeffreys prior [120–122] which is invariant under reparametrization of a continuous variable.

The Jeffreys prior, as well as other uninformative priors, are widely used tools across the biophysical literature [123–126]. As we will discuss shortly – as well as in detail in the last section – the Shannon entropy itself can be thought of as an uninformative prior (technically the logarithm of a prior) over probability distributions [49,127–129]. This prior is used in the analysis of data originating from a number of techniques including fluorescence correlation spectroscopy (FCS) [10, 71, 72], Electron spin resonance (ESR) [130], fluorescence resonance energy transfer (FRET) and bulk fluorescence [131–133].

Informative priors

One choice of informative prior is suggested by Bayes' theorem that is used to update priors to posteriors. Briefly, we see that when additional independent data are incorporated into a posterior, the new posterior $p(\mathbf{M}|D_2, D_1)$ is obtained from the old posterior, $p(\mathbf{M}|D_1)$, and the likelihood as follows

$$p(\mathbf{M}|D_2, D_1) \propto p(D_2|\mathbf{M})p(\mathbf{M}|D_1). \tag{2.19}$$

In this way, the old posterior, $p(\mathbf{M}|D_1)$, plays the role of the prior for the new posterior, $p(\mathbf{M}|D_2, D_1)$. If we ask – on the basis of mathematical simplicity – that

all future posteriors adopt the same mathematical form, then our choice of prior is settled: this prior – called a conjugate prior – must yield a posterior of the same mathematical form as the prior when multiplied by its corresponding likelihood. The likelihood, in turn, is dictated by the choice of experiment.

Priors, say $p(\mathbf{M}|\gamma)$, may depend on additional parameters, γ, called hyperparameters distinct from the model parameters $\boldsymbol{\theta}$. These hyperparameters, in turn, can also be distributed, $p(\gamma|\eta)$, thereby establishing a parameter hierarchy. For instance, an observable (say the FRET intensity) can depend on the state of a protein which depends on transition rates to that state (a model parameter) which, in turn, depends on prior parameters determining how transition rates are assumed to be *a priori* distributed (hyperparameter). We will see examples of such hierarchies in the context of later discussions on infinite Hidden Markov Models.

As a final note, before turning to an example of conjugacy, to avoid committing to specific arbitrary values for hyperparameters we may assume they are distributed according to a (hyper)prior and integrate over the hyperparameters in order to obtain $p(\mathbf{M}|\mathbf{D})$ from $p(\mathbf{M}|\mathbf{D}, \gamma, \eta, ...)$.

Now, we illustrate the concept of conjugacy by returning to our earlier molecular motor example. The prior conjugate to the Poisson distribution with parameter λ – Eq. (2.2) where λ is $r\Delta T$ – is the Gamma distribution

$$Gamma(\alpha, \beta) = p(\lambda = r\Delta T|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda} \tag{2.20}$$

which contains two hyperparameters, α and β. After a single observation – of n_1 events in time ΔT – the posterior is

$$p(\lambda|N, \alpha, \beta) = Gamma(n_1 + \alpha, 1 + \beta) \tag{2.21}$$

while, after N independent measurements, with $\mathbf{D} = \{n_1, \cdots n_N\}$, we have

$$p(\lambda|\mathbf{D}, \alpha, \beta) = Gamma\left(\sum_{i=1}^{N} n_i + \alpha, N + \beta\right). \tag{2.22}$$

Fig. (2.2) illustrates how the posterior is dominated by the likelihood provided sufficient data and how an arbitrary choice for the hyperparameters becomes less important for large enough N.

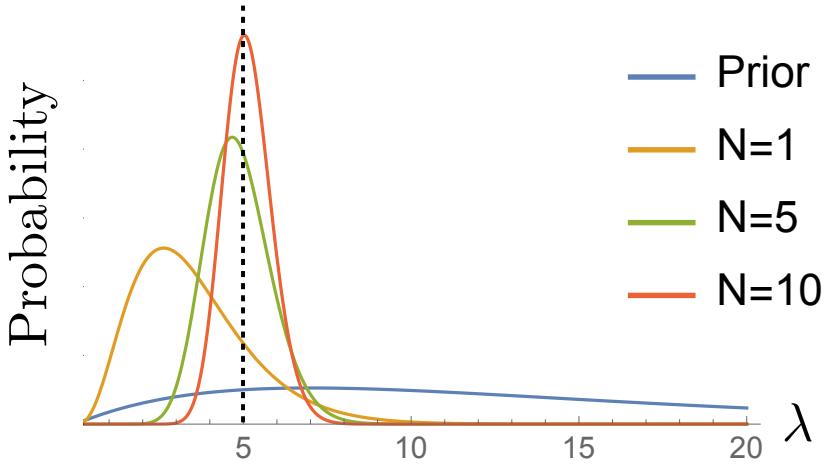

Fig. 2.2. **The posterior probability sharpens as more data are accumulated.** Here we sampled data according to a Poisson distribution with $\lambda = 5$ (designated by the dotted line). Our samples were $\mathbf{D} = \{2, 8, 5, 3, 5, 2, 5, 10, 6, 4\}$. We plotted the prior (Eq. (2.20) with $\alpha = 2$, $\beta = 1/7$) and the resulting posterior after collecting $N = 1$, then $N = 5$ and $N = 10$ points.

Single molecule photobleaching provides yet another illustrative example [134]. Here we consider the probability that a molecule has an inactive fluorophore (one that never turns on) which, in itself, is a problem towards achieving quantitative superresolution imaging [135, 136]. We define θ as the probability that a fluorophore is active (and detected). We, correspondingly, let $1 - \theta$ be the probability that the

fluorophore never turns on. The probability that y of n total molecules in a complex turns on is then binomially distributed

$$p(y|\theta) = \frac{n!}{(n-y)!\,y!}\theta^y(1-\theta)^{n-y}. \tag{2.23}$$

Over multiple measurements (multiple complexes each having n total molecules), $\mathbf{y}$, we obtain the following likelihood

$$p(\mathbf{y}|\theta) \propto \prod_i \frac{n!}{(n-y_i)!\,y_i!}\theta^{y_i}(1-\theta)^{n-y_i}. \tag{2.24}$$

One choice for $p(\theta)$ is the Beta distribution, a conjugate prior to the binomial,

$$p(\theta) = \frac{(a+b-1)!}{(a-1)!\,(b-1)!}\theta^{a-1}(1-\theta)^{b-1}. \tag{2.25}$$

By construction (i.e. by conjugacy), our posterior now takes the form of the Beta distribution

$$p(\theta|\mathbf{y}) \propto \theta^{\sum_i y_i + a - 1}(1-\theta)^{\sum_i (n-y_i)+b-1}. \tag{2.26}$$

Given these data, the estimated mean, $\hat{\theta}$, obtained from the posterior is now:

$$\hat{\theta} = \frac{\sum_i y_i + a}{\sum_i n + a + b} = \frac{\sum_i y_i}{\sum_i n + a + b} + \frac{a}{\sum_i n + a + b} \tag{2.27}$$

which is, perhaps unsurprisingly, a weighted sum over the prior expectation and the actual data.

Conjugate priors do have obvious mathematical appeal and yield analytically tractable forms for posteriors but they are more restrictive. Numerical methods to sample posteriors – including Gibbs sampling and related Markov chain Monte Carlo methods [137,138] – continue to be used [134] and developed [139,140] for biophysical

problems and have somewhat reduced the historical analytical advantage of conjugate priors. However the advantage conferred by the tractability of conjugate priors has turned out to be major advantage for more complex inference problems – such as those involving Dirichlet processes – that we will discuss later.

2.3 Information Theory as a Data Analysis Tool

2.3.1 Information theory: Introduction to key quantities

In 1948 Shannon [75] formulated a quantitative measure of uncertainty of a distribution, $p(x)$, later called the Shannon entropy

$$H(x) = -\sum_{i=1}^{K} p(x_i) \log\ p(x_i) \tag{2.28}$$

where x_i, the observable, takes on K discrete numerical values $\{x_1, x_2, \cdots, x_K\}$ such as the intensity levels observed from a single molecule time trace. Often, $-H(x)$, is called the Shannon information.

The Shannon entropy is an exact, non-perturbative, formula whose mathematical form, Eq. (2.28), has also been argued using the large sample limit of the multinomial distribution and Poisson distribution (as we will show later) [49]. However, Shannon made no such approximations and derived $H(x)$ from a simple set of axioms that a reasonable measure of uncertainty must satisfy [75].

The Shannon entropy behaves as we expect an uncertainty to behave. That is, informally, when all probabilities are uniform, $p(x_i) = 1/K$ for any x_i, then $H(x)$ is at its maximum, $H(x) = \log\ K$. In fact, as K increases, so does the uncertainty, again as we would expect. Conversely, when all probabilities, save one, are zero, then $H(x)$ is at its minimum, $H(x) = 0$. On a more technical note, Shannon also stipulated that the uncertainty of a probability distribution must satisfy the "composition property"

which quantifies how uncertainties should add if outcomes, indexed i, are arbitrarily regrouped [49, 75].

Later formalizations due to Shore and Johnson (SJ) [141], have independently arrived at precisely the same form as $H(x)$ (or any function monotonic with $H(x)$). SJ's work is closer in spirit to Bayesian methods [127, 128, 142–146]. We refer the reader to the last section of this review for a simplified version of SJ's derivation.

While we have so far dealt with distributions depending only on a single variable, the Shannon entropy can also deal with joint probability distributions as follows

$$H(x,y) = -\sum_{i=1}^{K_x}\sum_{j=1}^{K_y} p(x_i,y_j)\log\ p(x_i,y_j). \tag{2.29}$$

If both observables are statistically independent – that is, if $p(x_i,y_j) = p(x_i)p(y_j)$ – then $H(x,y)$ is the sum of the Shannon entropy of each observable, i.e. $H(x,y) = -\sum_{i,j} p(x_i)p(y_j)\log\ p(x_i)p(y_j) = -\sum_i p(x_i)\log\ p(x_i) - \sum_j p(y_j)\log\ p(y_j) = H(x) + H(y)$. This property is called "additivity".

On the other hand, if the two observables are statistically dependent – that is, if $p(x_i,y_j) \neq p(x_i)p(y_j)$ – then we can decompose the Shannon entropy as follows

$$H(x,y) = -\sum_{i,j} p(x_i,y_j)\log\left(p(x_i|y_j)p(y_j)\right) = H(y) + H(x|y) \tag{2.30}$$

where $p(x_i|y_j) = p(x_i,y_j)/p(y_j)$ is the conditional probability and in which we have $H(x|y) \equiv -\sum_{i,j} p(x_i,y_j)\log\ p(x_i|y_j)$. $H(x|y)$ is called the conditional entropy that measures the uncertainty in knowing the outcome of x if the value of y is known. In fact, this interpretation follows from Eq. (2.30): the total uncertainty in predicting the outcomes of both x and y, $H(x,y)$, follows from the uncertainty to predict y, given by $H(y)$, and the uncertainty to predict x after y is known, $H(x|y)$.

Fig. (2.3) illustrates the relationship between these entropies and, in particular, provides a conceptual picture for the relation $H(x,y) = H(y) + H(x|y) = H(x) + H(y|x)$.

The Venn diagram provides us with another important quantity, the mutual information $I(x,y)$, which corresponds to the intersecting area between $H(x)$ and $H(y)$. From Fig. (2.3), we can read out the following form of $I(x,y)$ and its relation to other entropies

$$\begin{aligned} I(x,y) &= H(x) + H(y) - H(x,y) \\ &= H(y) - H(y|x) = H(x) - H(x|y) \\ &= \sum_{i,j} p(x_i, y_j) \log \left(\frac{p(x_i, y_j)}{p(x_i)p(y_j)} \right). \end{aligned} \tag{2.31}$$

The second line of Eq. (2.31) provides an intuitive meaning for $I(x,y)$ as the amount of uncertainty reduction that knowledge of either observable provides about the other. In other words, it is interpreted as the information shared by the two observables x and y. From the last line of Eq. (2.31), we see that $I(x,y) = 0$ if and only if x and y are statistically independent, $p(x_i, y_j) = p(x_i)p(y_j)$, for all x_i and y_j.

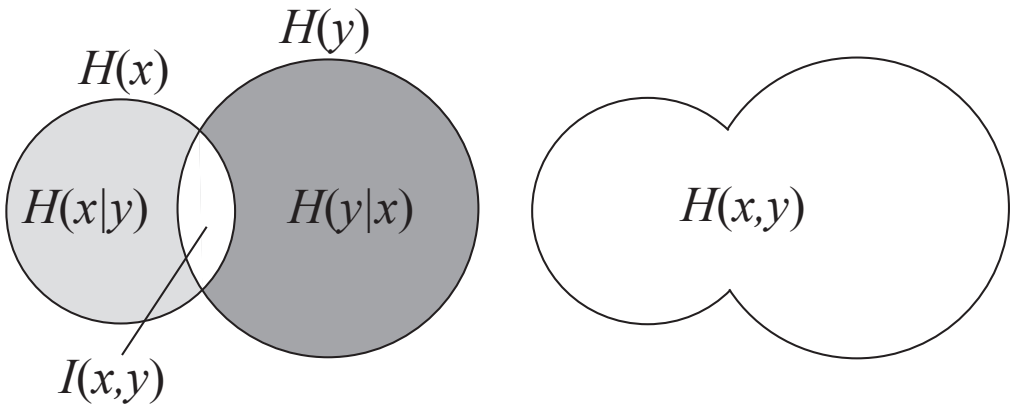

Fig. 2.3. **Venn diagram depicting different information quantities and their relationship.** The value of each entropy is represented by the enclosed area of different regions. $H(x)$ and $H(y)$ are both complete circles.

Now that we have defined the mutual information, we define the Kullback-Leibler (KL) divergence (or relative entropy) – a generalization of the mutual information – defined as [147, 148]

$$D_{\mathrm{KL}}[p(x)\|p(y)] = \sum_{i,j} p(x_i) \log \frac{p(x_i)}{p(y_j)}. \tag{2.32}$$

In Eq. (2.32), the probability distributions are distributions over single variables for simplicity only. The KL divergence vanishes if and only if $p(x) = p(y)$ but otherwise $D_{\mathrm{KL}}[p(x)\|p(y)] \geq 0$. We will see this quantity appear in our model selection section interpreted as a measure of dissimilarity in information content between $p(x)$ and $p(y)$. It is also interpreted as a pseudo-distance between $p(x)$ and $p(y)$ [149, 150] though it is not generally symmetric with respect to its arguments, $D_{\mathrm{KL}}[p(x)\|p(y)] \neq D_{\mathrm{KL}}[p(y)\|p(x)]$.

Finally, we note that the Shannon entropy defined as a measure of uncertainty is different from the entropy discussed in thermodynamics and statistical mechanics.

Confounding these concepts has lead to important misconceptions [151] and we discuss important differences between the Shannon and thermodynamic entropy in the last section of this review.

2.3.2 Information theory in model inference

Feynman often remarked that Young's double-slit experiment and its conceptual implications captured the essence of all of quantum mechanics [152]. The following thought experiment captures key aspects of information theory [142, 153].

An information theorist once noted that a third of kangaroos have blue eyes (B) and a quarter are left-handed (L). Thus three quarters are right-handed (R) and two-thirds are not blue-eyed (N).

The information theorist was then asked to compute the joint probability that kangaroos simultaneously be blue-eyed and right-handed (p_{BR}). But, this is an underdetermined problem. In fact, we have four unknown probabilities (p_{BR}, p_{BL}, p_{NR}, p_{NL}) but only three constraints (i.e. constraints on blue eyes, left-handedness but also a constraint on the normalization of probabilities). Thus, from these limited constraints and some algebra, we find that p_{BR} can take on any value from 1/12 to 1/3. Even for this simple example, there are an infinite number of acceptable models (i.e. probabilities) lying within this range.

To resolve this apparent ambiguity, the information theorist recommended that zero correlations between eye color and handedness be assumed *a priori* since none are otherwise provided by the data. The only model that now satisfies this condition, and falls within the previous range, is $p_{BR} = p_B p_R = (1/3) \times (3/4) = 1/4$.

Interestingly, this solution could equally well have been obtained by minimizing the Shannon information introduced in the previous section – or equivalently maximizing the Shannon entropy ($H = -\sum_i p_i \log p_i$) – under the three constraints of the problem

imposed using Lagrange multipliers where the index i labels each discrete outcome. This short illustration highlights many important ideas relevant to biophysical data analysis that we will use later and touches upon this critical point: information theory provides a principled recipe for incorporating data – even absent data – in model inference.

For our kangaroo example, information theory provides a recipe by which all data – both present and absent – contributed to the model-building process. In fact, by saying nothing about the structure of correlations between eye-color and handedness – what we are calling "absent data" – we say something: all but one value for p_{BR} would have imposed correlations between the model variables.

In our kangaroo example, unmeasured correlations were set to zero as a prior assumption to obtain $p_{BR} = 1/4$. This assumption on absent data is built into the process of model inference by maximizing the Shannon entropy (a process called MaxEnt). The details of this reasoning follow from the SJ axioms discussed in the last section of the review. But informally, for now, we say that MaxEnt only inserts correlations that are contained in the data (through the constraints) and assumes no other.

While the kangaroo example is conceptual, here is an example relevant to biophysics: master equations – the evolution equations describing the dynamics of state occupation probabilities – also rigorously follow from maximizing the Shannon entropy over single molecule trajectory probabilities by assuming structure of absent data. While the mathematics are detailed elsewhere [87] and reviewed in a broader context in Ref. [49], the key idea is simple. When we write a master equation with rates (transition probabilities per unit time) from state to state, we presuppose that the rates themselves – that is, conditional probabilities of hopping from one state to another – are time-independent. In order for MaxEnt to arrive at time-independent

rates, it must therefore have information on future, as of yet, unobserved transitions which should exhibit no time dependence.

Once these basic constraints on the absent (future) data are incorporated in MaxEnt, then master equations follow as a natural consequence [87]. The master equations still only follow if the data provided on transition probabilities has no spatial dependence and thus the system is, at least locally, well-stirred. Otherwise, we may need a more detailed model with, for example, spatially dependent forces [11].

As we will discuss, the structure imposed on absent data [10] is related to the concept of priors in Bayesian analysis [154] and model complexity penalties [73, 74] which we will review in later sections.

2.3.3 Maximum Entropy and Bayesian inference

Maximum entropy (MaxEnt) is a recipe to infer probability distributions from the data. The probability distributions inferred coincide with the maximum of an objective function.

Historically, in its simplest realization, Jaynes [49, 116, 117, 155] used Shannon's entropy to infer the most probable distribution of equilibrium classical degrees of freedom (positions and momenta). He did so by asking which distribution, $\{p_i\}$, maximized H given constraints (imposed using Lagrange multipliers) on normalization $\sum_i p_i = 1$ and the average energy. Mathematically, Jaynes maximized the following objective function with respect to the $\{p_i\}$ and the Lagrange multipliers

$$-\sum_i p_i \log p_i - \sum_j \lambda_j \left(\sum_i a_{ij} p_i - \bar{a}_j \right) \tag{2.33}$$

where the λ_j are the Lagrange multipliers for the j^{th} constraint, and $\bar{a}_j$ are the measured average of the quantity a_j which, for outcome i, takes on the value a_{ij}. For

example, the average dice roll would be constrained as: $(\sum_{i=1}^{6} ip_i - 2.7)$ assuming the average roll happens to be 2.7. Furthermore, if the j^{th} constraint is normalization, then it would be imposed by setting $a_{ij} = \bar{a}_j = 1$ for that constraint.

In fact, going back to our kangaroo example, we saw that acceptable values for p_{BR} that satisfied all 3 constraints were 1/12 to 1/3. We could have assumed that all values in this range were equally acceptable. However, by enforcing no correlations where none were warranted by the data, we arrived at 1/4 from MaxEnt.

To quantify just how good or bad 1/4 is, we need a posterior distribution over models, $\{p_i\}$, for the given data. In other words, we need to reconcile MaxEnt and Bayesian inference by relating the entropy to a Bayesian prior.

To do so, we first note that maximizing the constrained Shannon entropy, Eq. (2.33), is analogous to maximizing a posterior over the $\{p_i\}$: H is a logarithm of a prior over the $\{p_i\}$ while the constraints are the logarithm of the likelihood. The same restrictions that apply to selecting a likelihood in frequentist and Bayesian analysis hold for selecting its logarithm (i.e. the constraints in MaxEnt). Thus, just as Bayesian inference generates distributions over parameters, $\{p_i\}$, MaxEnt returns point estimates (the maxima, $\{p_i^*\}$, of the constrained Shannon entropy).

As we will see, the constraints that we imposed on Eq. (2.33) are highly unusual and equivalent to delta-function likelihoods infinitely sharply peaked at their mean value.

To quantitatively relate MaxEnt to Bayesian inference, we take a frequentist route [128, 142] and consider frequencies of the outcomes of an experiment by counting the number of events collected in the i^{th} bin, n_i, assuming such independent events occur with probability μ_i

$$P(\mathbf{n}|\boldsymbol{\mu}) = \prod_i \frac{\mu_i^{n_i} e^{-\mu_i}}{n_i!} \sim e^{\sum_i (n_i - \mu_i)} e^{-\sum_i n_i \log(n_i/\mu_i)} \tag{2.34}$$

where, in the last step, we have invoked Stirling's approximation valid when all n_i are large. We now define $\mathcal{N}$ as the total number of events, $\sum_i n_i$, and define probabilities $p_i \equiv n_i/\mathcal{N}$ and $q_i \equiv \mu_i/\mathcal{N}$ [49]. Then

$$P(\mathbf{p}|\mathbf{q}) \sim e^{\mathcal{N}\sum_i (p_i - q_i)} e^{-\mathcal{N}\sum_i p_i \log(p_i/q_i)}. \tag{2.35}$$

By imposing normalization on both p_i and q_i, we have

$$\begin{aligned} \mathcal{P}(\mathbf{p}\,|\mathbf{q}) &\equiv P\left(\mathbf{p}\,\middle|\mathbf{q}, \sum_i p_i = \sum_i q_i = 1\right) \\ &= \frac{e^{-\mathcal{N}\sum_i p_i \log(p_i/q_i)} \delta_{\sum_i p_i, 1} \delta_{\sum_i q_i, 1}}{Z} \\ &= \frac{e^{\mathcal{N}H} \delta_{\sum_i p_i, 1} \delta_{\sum_i q_i, 1}}{Z} \end{aligned} \tag{2.36}$$

where $Z = Z(\mathbf{q})$ is a normalization factor. That is, it is an integral of the numerator of Eq. (2.36) over each p_i from 0 to 1. In addition, $H = -\sum_i p_i \log(p_i/q_i)$ and $\delta_{x,y}$, is the Kronecker delta (i.e. is zero unless x=y in which case it is one).

For our simple kangaroo example, we can now compute the posterior probability over the model, $\mathbf{p}$, given constraints from the data, $\mathbf{D}$ by multiplying the prior, Eq. (2.36), with hard (Kronecker delta) constraints as our likelihood. This yields

$$P(\mathbf{p}|\mathbf{D}, \mathbf{q}) = \frac{\delta_{p_1+p_2,1/3}\delta_{p_3+p_4,1/4} \times \mathcal{P}(\mathbf{p}\,|\mathbf{q})}{\mathcal{Z}} \tag{2.37}$$

and $\mathcal{Z}$ is again a normalization and we have conveniently re-indexed the probabilities with numbers rather than letters. Here $P(\mathbf{p}|\mathbf{D}, \mathbf{q})$ is a posterior, $\mathcal{P}(\mathbf{p}|\mathbf{q})$ is a prior (which we have found depends on the Shannon entropy) and $\mathbf{q}$ are hyperparameters. Intuitively, we can understand $\mathbf{q}$ as being the values to which $\mathbf{p}$ defaults

when we maximize the entropy in the absence of constraints. That is, maximizing $-\sum_i p_i \log(p_i/q_i)$ returns $p_i \propto q_i$.

In fact, $P(\mathbf{p}|\mathbf{D},\mathbf{q})$ describes the probability over all allowed models. Furthermore, given identical constraints from the data – i.e. the same likelihood – and given that the normalization, $\mathcal{Z}$, does not depend on $\mathbf{p}$, the ratio of posterior probabilities then only depends on the Shannon entropy of both models

$$\frac{P(\mathbf{p}|\mathbf{D},\mathbf{q})}{P(\mathbf{p}'|\mathbf{D},\mathbf{q})} = e^{\mathcal{N}(-H(\mathbf{p}'|\mathbf{q})+H(\mathbf{p}|\mathbf{q}))}. \tag{2.38}$$

The factor of $\mathcal{N}$ quantifies the strength of our prior assumptions in much the same way that the hyperparameters α and β of Eq. (2.20) set properties of the prior. In other words, informally $\mathcal{N}$ tells us how many "data points" our prior knowledge is worth.

Now, we can evaluate, for the kangaroo example, a ratio of posteriors for the optimal MaxEnt model ($p_1 = 1/4, p_2 = 1/12, p_3 = 1/2, p_4 = 1/6$) and a variant

$$\frac{P(p_1 = 1/4, p_2 = 1/12, p_3 = 1/2, p_4 = 1/6|D)}{P(p_1 = 1/4 - \epsilon, p_2 = 1/12 + \epsilon, p_3 = 1/2 + \epsilon, p_4 = 1/6 - \epsilon|D)} = e^{\mathcal{N}(0.19)} \tag{2.39}$$

where we have assumed equal (uniform) q_i's and $\epsilon = 1/8$.

While the maximum of the posterior, Eq. (2.37), is independent of $\mathcal{N}$ for the kangaroo example – because of the artificiality of delta-function constraints – the shape of the posterior and thus the credible interval (the Bayesian analogue of the frequentist confidence interval) – certainly depend on $\mathcal{N}$ [127, 144].

The MaxEnt recipe is thus equivalent to maximizing a posterior over a probability distribution. The prior in the MaxEnt prescription is set to the entropy for fundamental reasons described in the last section of the review which also details the repercussions of rejecting the principle of MaxEnt.

MaxEnt does not assume a parametric form for the probability distribution. Also, the model parameters – that is, each individual p_i – can be very large for probability distributions discretized on a very fine grid.

2.3.4 Applications of MaxEnt: Deconvolution methods

Often, to determine how many exponential components contribute to a decay process, a decay signal is first fit to a single exponential and the resulting goodness-of-fit is quantified. If the fit is deemed unsatisfactory, then an additional decay component is introduced and new parameters (two exponential decay constants and the relative weights for each exponential in this mixture model) are determined. As described, this fitting procedure cannot be terminated in a principled way. That is, an increasingly large number of exponentials will always improve the fit [145].

MaxEnt deconvolution methods are specifically tailored to tackle this routine problem of data analysis. To give a concrete example, imagine a decay signal, $s(t)$, which is related to the distribution of decay rates, $p(r)$, through the following relation

$$s(t) = \int_0^\infty dr e^{-rt} p(r). \tag{2.40}$$

MaxEnt deconvolution solves the inverse problem of determining $p(r)$ from $s(t)$. That is, MaxEnt readily infers probability distributions such as unknown weights, the $p(r)$, which appear in mixture models [127]. These weights could include, for example, probabilities for exponentials, as in Eq. (2.40), or probabilities that cytoplasmic proteins sample different diffusion coefficients [10, 72, 156]. The fitting procedure is ultimately terminated because MaxEnt insists on simple (minimum information or maximum entropy) models consistent with observations.

More generally, for discrete data, D_i, we can write the discrete analog of Eq. (2.40)

$$D_i = \sum_j G_{ij} p_j + \epsilon_i \tag{2.41}$$

where G_{ij} is the ij^{th} matrix element of a general transformation matrix, $\mathbf{G}$, and p_j is the model. Contrary to the noiseless Eq. (2.40) here, we have added noise, ϵ_i, to Eq. (2.41).

All experimental details are captured in the matrix $\mathbf{G}$. Here are examples of this matrix:

$$\mathbf{G}_{Fluor} \cdot \mathbf{p} = \int_0^\infty dr e^{-rt} p(r) \tag{2.42}$$

$$\mathbf{G}_{FRAP} \cdot \mathbf{p} = -\int_0^\infty dD e^{-\frac{(x-x_0)^2}{2Dt}} p(D) \tag{2.43}$$

$$\mathbf{G}_{FCS} \cdot \mathbf{p} = -\int_0^\infty d\tau_D \frac{1}{n} \frac{1}{(1+\tau/\tau_D)^{3/2}} p(\tau_D) \tag{2.44}$$

where the first can be used to determine decay rate distributions [131–133]; the second is relevant to fluorescence recovery after photobleaching (FRAP) with an undetermined distribution over diffusion coefficients, D (assuming isotropic diffusion in one dimension); the third is relevant to fluorescence correlation spectroscopy (FCS) with an undetermined distribution over diffusion times, τ_D, through a confocal volume [10, 72], assuming a symmetric Gaussian confocal volume with n diffusing particles.

If a Gaussian noise model is justified – where ϵ_i is sampled from a Gaussian distribution with zero mean and standard deviation σ_i, i.e. $\epsilon_i \sim N(0, \sigma_i)$ – then, one could propose to find p_j by minimizing the following log-likelihood (equivalent

to maximizing the likelihood for a product of Gaussians) under the assumption of independent Gaussian observations

$$\chi^2 \equiv \sum_i \left(\frac{D_i - \sum_j G_{ij} p_j}{\sigma_i} \right)^2 . \tag{2.45}$$

This ill-fated optimization of Eq. (2.45), described in the first paragraph of this section, overfits the model. Additionally, depending on our choice of discretization for the index j of Eq. (2.41), we may select to have many more weights, p_j, than we have data points, D_i, and, in this circumstance, a unique minimum of the χ^2 may not even exist. For this reason, we use the entropy prior and write down our posterior

$$\begin{aligned} P\left(\mathbf{p} \middle| \mathbf{D}, \mathbf{q}, \sum_i p_i = \sum_i q_i = 1 \right) &\propto P\left(\mathbf{D}|\mathbf{p}\right) \times P\left(\mathbf{p} \middle| \mathbf{q}, \sum_i p_i = \sum_i q_i = 1 \right) \\ &\propto e^{-\chi^2/2} \times e^{-\mathcal{N} \sum_i p_i \log(p_i/q_i)} \delta_{\sum_i p_i, 1} \delta_{\sum_i q_i, 1} \end{aligned} \tag{2.46}$$

where the proportionality above indicates that we have not explicitly accounted for the normalization, $P(\mathbf{D}|\mathbf{q})$. If we are only interested in a point estimate for our model – i.e. the one that that maximizes the posterior given by Eq. (2.46) – then the objective function we need to maximize is [49]

$$-\mathcal{N} \sum_i p_i \log\left(p_i/q_i\right) - \frac{\chi^2}{2} + \lambda_0 \left(\sum_i p_i - 1 \right) + \lambda_1 \left(\sum_i q_i - 1 \right) \tag{2.47}$$

where we have used Lagrange multipliers (λ_0, λ_1) to replace the delta-function constraints. The variation of the MaxEnt objective function, Eq. (2.47), is now understood to be over each p_i as well as λ_0 and λ_1. Furthermore, if we are only interested in

the maximum of Eq. (2.47), we are free to multiply Eq. (2.47) through by constants or add constants as well. In doing so, we obtain a more familiar MaxEnt form

$$-\sum_i p_i \log\left(p_i/q_i\right) - \phi\frac{(\chi^2 - N)}{2} + \tilde{\lambda}_0\left(\sum_i p_i - 1\right) + \tilde{\lambda}_1\left(\sum_i q_i - 1\right) \tag{2.48}$$

where N – the number of independent observations – and ϕ are constants. $\mathcal{N}$ or its inverse, ϕ, is a hyperparameter that we must in principle set *a priori*.

One way to determine ϕ is to treat ϕ as a Lagrange multiplier enforcing the constraint that χ^2 be equal to its frequentist expectation

$$\chi^2 \sim N. \tag{2.49}$$

That is, summed over a large and typical number of data points – where, typically, $\left(D_i - \sum_j G_{ij}p_j\right)^2 \sim \sigma_i^2$ – we have $\chi^2 \sim N$.

Skilling and Gull [128] have argued that this frequentist line of reasoning to determine ϕ, and thus the posterior, undermines the meticulous effort that has been put into deriving Shannon's entropy from SJ's self-consistent reasoning arguments (that we discuss in the last section). Instead, they proposed [128] a method based on empirical Bayes [108] motivating the choice of hyperparameter from the data.

If we take Eq. (2.49) for now, we now find a recipe for arriving at the optimal model, $\mathbf{p}^*$,

$$\mathbf{p}^* = \max_{\mathbf{p},\phi,\tilde{\lambda}_0,\tilde{\lambda}_1}\left(-\sum_i p_i \log\left(p_i/q_i\right) - \phi\frac{(\chi^2 - N)}{2} + \tilde{\lambda}_0\left(\sum_i p_i - 1\right) + \tilde{\lambda}_1\left(\sum_i q_i - 1\right)\right). \tag{2.50}$$

Often, we select a uniform distribution (flat $\mathbf{q}$) though different choices are discussed in the literature [145].

Finally, for FCS, Eq. (2.44) is just a starting point that, for simplicity, ignores confocal volume asymmetry and triplet corrections. More sophisticated FCS deconvolution methods can account for these [10] and also explicitly account for correlated noise and ballistic motion of actively transported particles [71]. And – in part because FCS is so versatile [157, 158] and can even be used *in vivo* in a minimally invasive manner [24, 72, 159–168] – FCS data has been analyzed using multiple deconvolution methods that have provided models for the dynamics of the human islet amyloid polypeptide (hIAPP) on the plasma membrane [169] as well as the dynamics of signaling proteins in zebrafish embryos [156].

MaxEnt deconvolution: An application to FCS

Here we briefly present an application where MaxEnt was used to infer the behavior of transcription factors *in vivo* [10] and used to learn about crowding, binding effects and photophysical artifacts contributing to FCS.

Briefly in FCS, labeled proteins are monitored as they traverse an illuminated confocal volume [170]. The diffusion time, τ_D, across this volume of width w is obtained from the fluorescence time intensity correlation function, $G(\tau)$, according to [168, 170]

$$G(\tau) = \frac{1}{n}\left(1 + \frac{\tau}{\tau_D}\right)^{-1}\left(1 + \frac{1}{Q^2}\frac{\tau}{\tau_D}\right)^{-1/2} \tag{2.51}$$

where n is the average number of particles in the confocal volume and Q characterizes the confocal volume's asymmetry. The diffusion constant, D, is related to τ_D by $\tau_D = w^2/4D$ [for simplicity, Eq. (2.51) ignores triplet corrections [171]] where w designates the width of the confocal volume.

In complex environments, $G(\tau)$ often cannot be fit with a single diffusion component (Eq. (2.51)). That is, τ is no longer simply proportional to a mean square

displacement in the confocal volume, $\langle \delta r^2 \rangle$. Instead, $G(\tau)$'s are constructed for anomalous diffusion models – where $\langle \delta r^2 \rangle \propto \tau^\alpha$ and α is different from one – as follows [24, 162–165]

$$G(\tau) = \frac{1}{n} \left(1 + \left(\frac{\tau}{\tilde{\tau}_D} \right)^\alpha \right)^{-1} \left(1 + \frac{1}{\tilde{Q}^2} \left(\frac{\tau}{\tilde{\tau}_D} \right)^\alpha \right)^{-1/2} \tag{2.52}$$

where we have introduced an effective diffusion time, $\tilde{\tau}_D$, and asymmetry parameter, $\tilde{Q}$.

Circumstances under which anomalous diffusion models – where $\langle \delta r^2 \rangle \propto \tau^\alpha$ strictly holds – over many decades in time are exceptional, not generic. For example, fractional Brownian motion (FBM) – that can give rise to anomalous diffusion [172] – may arise when proteins diffuse through closely packed fractal-like heterochromatin structures [173] though it is unclear to what degree the structure of heterochromatin actually is fractal. As another example, continuous time random walks (CTRW), in turn, yield anomalous diffusion by imposing power law particle waiting time or jump size distributions to describe a single particle's trajectory [174–177] though these power laws have only rarely been observed experimentally [174, 175] and, what is more, FCS does not collect data on single particle trajectories.

A method of analysis should deal with the data as it is provided. That is, for FCS, a model should preferentially be inferred directly from $G(\tau)$ rather than conjecturing behaviors for single particle trajectories – that are not observed – that may give rise to a $G(\tau)$.

MaxEnt starts with the data at hand and provides an alternative solution to fitting data using anomalous diffusion models [10]. Rather than imposing a parametric form (Eq. (2.52)) on the data, MaxEnt has been used to harness the entire $G(\tau)$ to extract information on crowding effects, photophysical label artifacts, cluster formation as

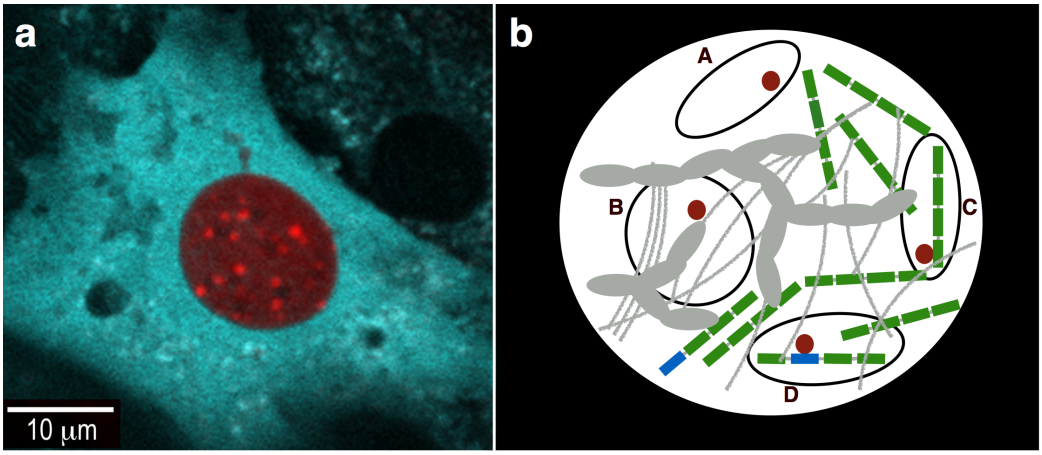

Fig. 2.4. **FCS may be used to model the dynamics of labeled particles at many cellular locations (regions of interest (ROIs)), both in the cytosol and in the nucleus**. **a)** Merged image of a cerulean-CTA fluorescent protein (FP) used to image the cytosol and mCherry red FP used to tag BZip protein domains. In Ref. [10], we analyzed FCS data on tagged BZips diffusing in the nucleus and the cytosol. We analyzed diffusion in ROIs far from heterochromatin by avoiding red FP congregation areas (bright red spots). MaxEnt analysis revealed details of the fluorophore photophysics, crowding and binding effects that could otherwise be fit using anomalous models. **b)** A cartoon of the cell nucleus illustrating various microenvironments in which BZip (red dots) diffuses (A: free region; B: crowded region; C: non-specific DNA binding region; D: high affinity binding region).

well as affinity site binding *in vivo*, a topic that has been of recent interest [178–186]; see Fig. (2.4).

To extract information on basic processes that could be contributing to the $G(\tau)$, we start with the observation that the *in vivo* confocal volume is composed of many "microenvironments"; see Fig. (2.4b). In each microenvironment, the diffusion coeffi-

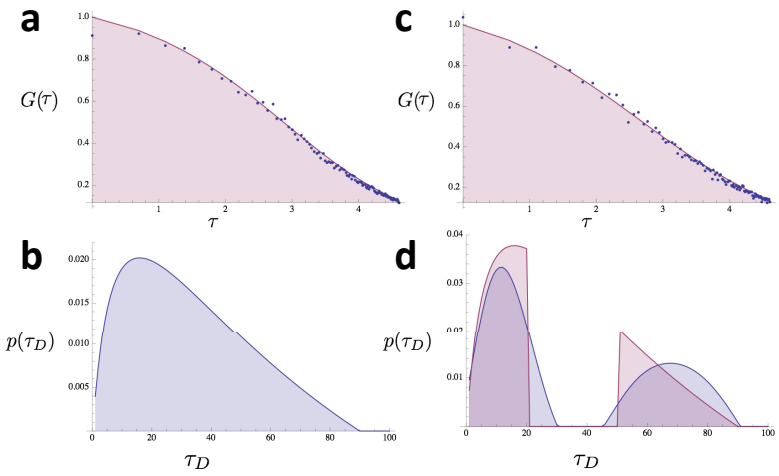

Fig. 2.5. **Protein binding sites of different affinities yield a $G(\tau)$ that is well fit by an anomalous diffusion model**. A theoretical $G(\tau)$ (containing 150 points) was created from an anomalous diffusion model, Eq. (2.52) with $\alpha = 0.9$, to which we added 5% white noise (a, blue dots, logarithmic in time). Using MaxEnt, we infer a $p(\tau_D)$ from this $G(\tau)$ (b) and, as a sanity check, use it to reconstruct a $G(\tau)$ (a, solid curve). In the main body, we discuss how protein binding sites of different affinities could give rise to such a $p(\tau_D)$. Part of $p(\tau_D)$ is then excised, yielding a new $p(\tau_D)$ (d, pink curve). Conceptually, this is equivalent to mutating a binding site which eliminates some τ_D's. We created a $G(\tau)$ from this theoretical distribution with 8% white noise (c, blue dots, logarithmic in time). We then extracted a $p(\tau_D)$ from this (d, blue curve) and we reconstructed a $G(t)$ from this $p(\tau_D)$ distribution as a check (c, solid curve). Time is in arbitrary units. See text and Ref. [10] for more details.

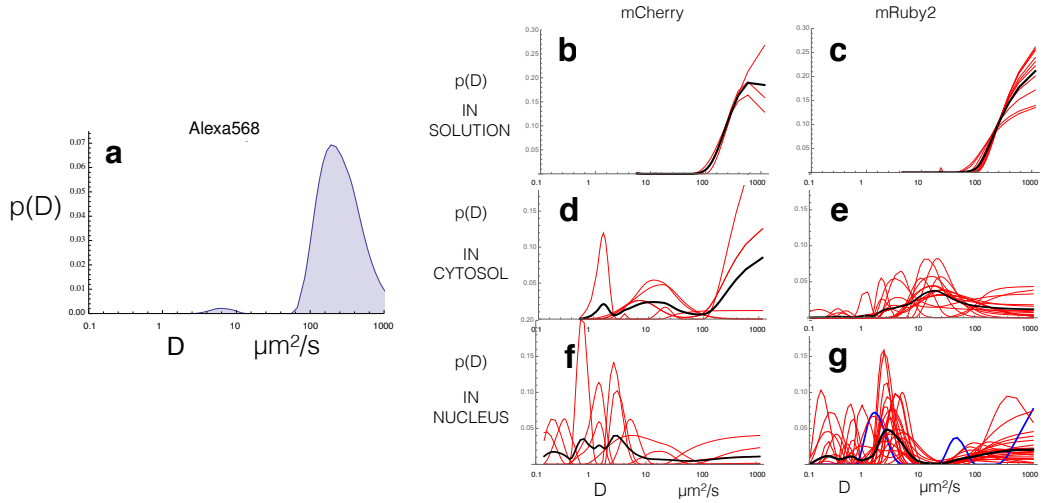

Fig. 2.6. **Probability distributions of diffusion coefficients can be inferred from FCS curves. a)** $p(D)$ for freely diffusing Alexa568 shows no "superdiffusive plateau" (defined in the text) that arises from dye flickering. Rather, it shows its main peak at $360\mu\text{m}^2/\text{s}$ very near the reported value of 363 $\mu\text{m}^2/\text{s}$ [187]. We attributed the smaller peak centered at ~ 5 $\mu\text{m}^2/\text{s}$ to dye aggregation [10]. **b) + c)** We analyzed $p(D)$'s obtained from FCS data acquired on mCherry and mRuby2 diffusing freely in solution, and **d) + e)** mCherry or mRuby2 tagged BZip protein domains in the cytosol and **f) + g)** the nucleus far from heterochromatin [187]. Black curves are averages of the red curves [total number of data sets: b:3,c:9,d:5,e:16,f:7, and g:21]. The additional blue curve in (g) shows the analysis of the best data set (i.e. the most monotonic $G(\tau)$). See text and Ref. [10] for more details.

cient differs and diffusion may thus be described using a multi-component (mixture) normal diffusion model [72, 159–162]

$$G(\tau) = \frac{1}{n} \sum_{\tau_D} p(\tau_D) \left(1 + \frac{\tau}{\tau_D}\right)^{-1} \left(1 + \frac{1}{Q^2} \frac{\tau}{\tau_D}\right)^{-1/2}. \tag{2.53}$$

MaxEnt is then used to infer the full diffusion coefficient distribution, $p(D)$, or, equivalently $p(\tau_D)$, directly from the data. In particular, Fig. (2.5a)-(2.5b) illustrates how mixture models and anomalous diffusion models may both fit the data equally well. As a benchmark, a synthetic $G(\tau)$ (blue dots, Fig. (2.5a)) is generated with an anomalous diffusion model [Eq. (2.52) with $\alpha = 0.9$ with added 5% white noise]. Fig. (2.5b) shows the resulting $p(\tau_D)$ extracted from the noisy synthetic data. From this $p(\tau_D)$, we re-create a $G(\tau)$ (solid line, Fig. (2.5a)) and verified that it closely matches the original $G(\tau)$ (blue dots, Fig. (2.5a)).

An important advantage with the mixture model, is that it provides a $p(\tau_D)$ that may be microscopically interpretable. For instance, suppose a binding site is removed either by mutating/removing a particular DNA binding site or cooperative binding partner. Based on the model [discussed in detail in Ref. [10]], we expect the resulting $p(\tau_D)$ – or, equivalently, $p(D)$ – to show a gap at some τ_D. The hypothetical $p(\tau_D)$, expected after removal of a binding site, is shown with an exaggerated excision (pink curve, Fig. (2.5d)). A corresponding noisy $G(\tau)$ is generated from this $p(\tau_D)$ (blue dots, Fig. (2.5c)). Now we ask: had we been presented with such a $G(\tau)$, would we have been able to tell that a site had been mutated? The inferred $p(\tau_D)$ (blue curve, Fig. (2.5d)) shows a clear excision directly indicating that a mutated site would have been detectable.

To illustrate here that MaxEnt also works on real data, we re-analyzed *in vitro* data on the diffusion of the small dye Alexa568 (Fig. (2.6a)) as well as previously published FCS data [24] on the diffusion of the BZip domain of a transcription factor

(TF) [CCAAT/enhancer-binding C/EBPα] tagged with red fluorescent proteins (FPs) [either mCherry or mRuby2]. We analyzed data on BZip's diffusion both in solution and a living mouse cells' cytoplasm and nucleus (away from heterochromatin) that appeared to show anomalous diffusion. The results are summarized in Fig. (2.6b)-(2.6g).

Briefly, in solution (Fig. (2.6b)-(2.6c)) we identify the effects of protein flickering on the $p(D)$. Flickering is a fast, reversible photoswitching arising from FP chromophore core instabilities [188]. Fast flickering [faster than the tagged protein's τ_D] registers as fast-moving (high diffusion) components. Since many particles flicker, $p(D)$ shows substantial density at high D values. As expected mRuby2 and mCherry flickering appears in $p(D)$ as a "superdiffusive plateau" at the highest values of D in Fig. (2.6b)-(2.6c). The plateau's lower bound coincides with the diffusion coefficient expected in the absence of flickering [10]. As a control, Alexa568 – which is well-behaved in FCS studies [187, 189] – shows no plateau; see Fig. (2.6a).

In the cytosol, we found label-dependent molecular crowding effects on protein diffusion. Beyond the superdiffusive plateau, the cytosolic $p(D)$ shows peaks for mCherry-BZip and mRuby2-BZip at $\sim 20 - 40$ μm^2/s; see Fig. (2.6d)-(2.6e). This peak's location is consistent with results from FCS and FRAP experiments [187, 190, 191] and is attributed to crowding since our labeled proteins are thought to have few cytosolic interactions [187, 192]; see Ref. [10] for details.

Finally, in the nucleus, while our data sets are well fit by anomalous diffusion models [24], our method instead finds evidence of BZip's DNA site-binding. For BZip, we expect: 1) high affinity binding to specific DNA elements as well as lower-affinity non-specific DNA binding [10, 24, 193]; 2) interactions with other chromatin binding proteins [24]; and 3) association with proteins [194, 195] such as BZip's interaction with HP1α (which binds to histones) [24, 195]. The $p(D)$ of Fig. (2.6f)-(2.6g) shows

the less prominent but expected crowding peak ($\sim$ 10 μm^2/s) and photobleaching plateau as well as features arising from interactions. For instance, for mCherry-BZip we find slow diffusion coefficients with peaks centered at about 0.2 μm^2/s, 0.8 μm^2/s and 5 μm^2/s identifying possible nuclear interactions. The blue curve in Fig. (2.6g) displays "the best data set" for mRuby2-BZip nuclear diffusion [i.e. the most monotonic $G(\tau)$ (which is what $G(\tau)$ ought to be in the absence of noise) as measured by the Spearman rank coefficient]. This $p(D)$ shows three clear peaks corresponding to diffusion coefficients of $\sim$ 1000 (flickering), $\sim$ 80 (crowding) and $\sim$ 2 μm^2/s (binding interaction with $K \approx 500nM^{-1}$ assuming $[S] = 0.1mM$).

In summary, this section highlights the important mechanistic details that can be drawn from MaxEnt deconvolution techniques.

While MaxEnt is focused on inferring probability distributions of an *a priori* unspecified form, often we do have specific parametric forms for models in mind when we analyze data. Selecting between different "nested model" – models obtained as a special case of a more complex model by either eliminating or setting conditions on the complex model's parameters – is the focus of the next section.

2.4 Model Selection

2.4.1 Brief overview of model selection

A handful of highly complex models may fit any given data set very well. By contrast, a combinatorially larger number of simpler models — with fewer and more flexible parameters – provide a looser fit to the data. While highly complex models may provide excellent fits to a single data set, they are, correspondingly, over-committed to that particular data set.

The goal of successful model selection criteria is to pick models: 1) whose complexity is penalized, in a principled fashion, to avoid overfitting; and 2) that convincingly fit the data provided (the training set).

Model selection criteria are widely used in biophysical data analysis from image deconvolution [153, 196–199] to single molecule step detection [48, 70, 200, 201] and continue to be developed by statisticians [202].

Here we summarize both Information theoretic [203–207] as well as Bayesian [46, 47, 169, 208–211] model selection criteria.

Information theoretic and Bayesian model selection

In information theory, $h(\mathbf{x}|\boldsymbol{\theta}) = -\log p(\mathbf{x}|\boldsymbol{\theta})$ is interpreted as the information contained in the likelihood for data points $\mathbf{x}$ given parameters $\boldsymbol{\theta}$ [212]. Minimizing this information over $\boldsymbol{\theta}$ is equivalent to maximizing the likelihood for parametric models. For problems where the number of parameters (K) is unknown, preference is always given to more complex models. To avoid this problem, a cost function, L, associated to each additional variable is introduced [213], $-\log p(\mathbf{x}|\boldsymbol{\theta}) + L(\boldsymbol{\theta})$. Put differently, in the language of Shannon's coding theory that we will discuss later, if $-\log p(\mathbf{x}|\boldsymbol{\theta})$ measures the message length, then the goal of model selection is to find a model of minimal description length (MDL) [214–216] or, informally, of maximum compression [75].

Information theoretic model selection criteria – such as the Akaike Information Criterion (AIC) [217–220] – start with the assumption that the data may be very complex but that an approximate, candidate, model may minimize the difference in information content between the true (hypothetical) model and the candidate model. As we will see in detail later, models that overfit the data are avoided by parametrizing the candidate model on a training data set and comparing the information between

an estimate of the true model and candidate models on a different (validation) data test set. In this way, the AIC is about prediction of a model for additional data points provided beyond those data points used in the training step.

Since the data may be very complex, as the number of data points provided grows, the complexity (number of parameters) of the model selected by the AIC grows concomitantly. Complex models are not always a disadvantage. For instance, they may be essential if we try to approximate an arbitrary non-linear function using a high-order polynomial candidate model or for models altogether too complex to represent using simple parametric forms [219].

Bayesian model selection criteria – such as the Bayesian (or Schwartz) information criterion (BIC) [73] – instead select the model that maximizes a marginal posterior [66, 221, 222]. In the marginalization step, we have integrated over all irrelevant or unknown model parameters. This marginalization step is, as we will see, critical in avoiding overfitting. This is because, by marginalizing over variables, our final marginal posterior is a sum over models including models that fit the data poorly.

Unlike the AIC, the BIC assumes that there exists a true model and it searches for this model [220, 223]. Since this model's complexity is fixed – does not depend on the number of data points N – the BIC avoids growing the dimensionality of the model with N by penalizing the number of parameters of the model according to a function of N, $\log N$. This penalizing function is derived, it is not imposed by hand. By contrast to the AIC, the BIC "postdicts" the model since, in using the BIC, we assume that we already have access to all observed data [224].

As we will see later, for slightly non-linear models, the BIC may outperform the AIC which overfits small features while for highly non-linear models – where small features are important – the AIC may outperform the BIC [219]. The performance of the AIC and BIC are illustrated for a simple example in Fig. (2.7).

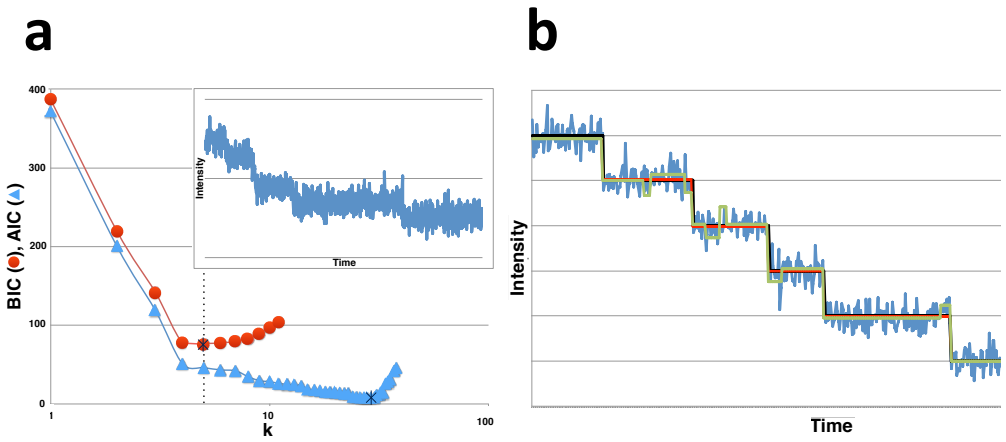

Fig. 2.7. **The AIC and BIC are often both applied to step-finding. a)** We generated 1000 data points with a background noise level, $\sigma_b = 20$. On top of the background, we added 6 dwells (5 change points) with noise around the signal having a standard deviation of $\sigma_s = 5$ (see inset). At this high noise level, and for this particular application, the BIC outperforms the AIC and the minimum of the BIC is at the theoretical value of 5 (dotted line). All noise is Gaussian and de-correlated. **b)** For our choice of parameters, the AIC (green) finds a model that overfits the true model (black) while the BIC (red) does not. However, as we increase the number of steps (while keeping the total number of data points fixed), the AIC does eventually outperform the BIC. This is to be expected. The AIC assumes the model could be unbounded in complexity and therefore does not penalize additional steps as much. The BIC, by contrast, assumes that there exists a true model of finite complexity. We acknowledge K. Tsekouras for generating this figure.

2.4.2 Information theoretic model selection: The AIC

In this section, we sketch a derivation of the AIC [203, 212, 225, 226]. While this section is theoretical and can be skipped upon a first reading, it does highlight: i) the method's limitations and applicability [227]; ii) that the penalty term follows from a principled derivation (and is not arbitrarily tunable); iii) how it conceptually differs from the BIC.

Briefly, finding the real or true model that generated the data is not achievable, and our goal is to seek a good candidate model. The AIC [74, 228], as one of the important selection criteria, is based on estimation of the KL divergence [148] between the (unknown and unknowable) true distribution that generated the data, $f(\mathbf{x})$, and a candidate distribution $p(\mathbf{x}|\boldsymbol{\theta}_0)$ parametrized by $\boldsymbol{\theta}_0$

$$D_{\mathrm{KL}}[f\|p] = \int d\mathbf{x} f(\mathbf{x}) \log\left(\frac{f(\mathbf{x})}{p(\mathbf{x}|\boldsymbol{\theta}_0)}\right) = \int d\mathbf{x} f(\mathbf{x}) \log f(\mathbf{x}) - \int d\mathbf{x} f(\mathbf{x}) \log p(\mathbf{x}|\boldsymbol{\theta}_0) \tag{2.54}$$

where $\boldsymbol{\theta}_0$ is the best possible estimate for $\boldsymbol{\theta}$ (obtainable in the limit of infinite data). The KL divergence is always positive or – in the event only realizable with synthetic data that both true model and candidate models coincide – zero. The proof of this is well known and follows from Jensen's inequality [229].

To achieve our goal – and select a model, p, that minimizes $D_{\mathrm{KL}}[f\|p]$ – we must first make some approximations as both f and $\boldsymbol{\theta}_0$ are unknown.

Given a training data set, $\mathbf{x}$, we may replace the hypothetical $\boldsymbol{\theta}_0$ with its estimate $\hat{\boldsymbol{\theta}}(\mathbf{x})$. However, using the same data set to evaluate both $\hat{\boldsymbol{\theta}}(\mathbf{x})$ and $p(\mathbf{x}|\hat{\boldsymbol{\theta}}(\mathbf{x}))$ biases our KL toward more complex models. To see this, we note that the numerical values for the likelihood satisfy $p(\mathbf{x}|\hat{\boldsymbol{\theta}}(\mathbf{x})) > p(\mathbf{x}|\hat{\boldsymbol{\theta}}(\mathbf{y}))$ since the numerical value for the likelihood is clearly worse (smaller) for a data set ($\mathbf{y}$) different from the one that was used to parametrize the model. Thus, the KL is always smaller for $p(\mathbf{x}|\hat{\boldsymbol{\theta}}(\mathbf{x}))$ than for $p(\mathbf{x}|\hat{\boldsymbol{\theta}}(\mathbf{y}))$ and becomes increasingly smaller as we add more and more parameters. That is, as we grow the dimensionality of $\hat{\boldsymbol{\theta}}(\mathbf{x})$.

To avoid this bias, we (conceptually) estimate $\boldsymbol{\theta}_0$ instead on a training set, $\mathbf{y}$, different from the validation set, $\mathbf{x}$, and estimate the KL, $\widehat{D}_{\mathrm{KL}}[f\|p]$, as follows [202]

$$\begin{aligned}\widehat{D}_{\mathrm{KL}}[f\|p] &= \text{const} - T \\ T &= E_{\mathbf{y}} E_{\mathbf{x}}[\log p(\mathbf{x}|\hat{\boldsymbol{\theta}}(\mathbf{y}))]\end{aligned} \tag{2.55}$$

where the constant, const, only depends on the hypothetical true model, see Eq. (2.54), and is, therefore, independent of our choice of candidate model. Furthermore, the expectation with respect to a distribution over some variable $\mathbf{z}$ is understood as

$$E_{\mathbf{z}}[g(\mathbf{z})] = \frac{1}{N}\sum_{i=1}^{N} g(z_i) \tag{2.56}$$

where the samples, z_i, are drawn from that distribution. Furthermore, for clarity, if f were known then

$$T \to \int d\mathbf{y} f(\mathbf{y}) \int d\mathbf{x} f(\mathbf{x}) \log p(\mathbf{x}|\hat{\boldsymbol{\theta}}(\mathbf{y})). \tag{2.57}$$

Our goal is now to estimate and, subsequently, maximize T, with respect to candidate models, in order to minimize $\widehat{D}_{\mathrm{KL}}$.

To evaluate T, we must compute a double expectation value. In the large data set limit, where $\hat{\boldsymbol{\theta}}(\mathbf{y})$ is presumably near $\boldsymbol{\theta}_0$, we expand $\hat{\boldsymbol{\theta}}(\mathbf{y})$ around $\boldsymbol{\theta}_0$

$$\begin{aligned} T &\sim E_{\mathbf{y}} E_{\mathbf{x}}\left[\log p(\mathbf{x}|\boldsymbol{\theta}_0) + \frac{1}{2}(\hat{\boldsymbol{\theta}}(\mathbf{y}) - \boldsymbol{\theta}_0)\cdot\left(\nabla_{\boldsymbol{\theta}}\nabla_{\boldsymbol{\theta}}\log p(\mathbf{x}|\hat{\boldsymbol{\theta}}(\mathbf{y}))\right)_{\hat{\boldsymbol{\theta}}(\mathbf{y})=\boldsymbol{\theta}_0}\cdot(\hat{\boldsymbol{\theta}}(\mathbf{y}) - \boldsymbol{\theta}_0)\right] \\ &= E_{\mathbf{y}} E_{\mathbf{x}}[\log p(\mathbf{x}|\boldsymbol{\theta}_0)] - \frac{1}{2}E_{\mathbf{y}}[(\hat{\boldsymbol{\theta}}(\mathbf{y}) - \boldsymbol{\theta}_0)\cdot\mathbf{I}(\boldsymbol{\theta}_0)\cdot(\hat{\boldsymbol{\theta}}(\mathbf{y}) - \boldsymbol{\theta}_0)] \\ &= E_{\mathbf{x}}[\log p(\mathbf{x}|\boldsymbol{\theta}_0)] - \frac{1}{2}E_{\mathbf{y}}[(\hat{\boldsymbol{\theta}}(\mathbf{y}) - \boldsymbol{\theta}_0)\cdot\mathbf{I}(\boldsymbol{\theta}_0)\cdot(\hat{\boldsymbol{\theta}}(\mathbf{y}) - \boldsymbol{\theta}_0)] \end{aligned} \tag{2.58}$$

where the expectation (not the value itself) of the first order term (which we have not written) vanishes and $\mathbf{I}(\boldsymbol{\theta}_0)$ is the Fisher information matrix. We will eventually want to evaluate the expectation of the second order term by further simplification (keeping in mind that any errors incurred will, by construction, be ever higher order).

However, for now, our focus is on the leading order term of Eq. (2.58), $E_{\mathbf{x}}[\log p(\mathbf{x}|\boldsymbol{\theta}_0)]$, which we expand around $\hat{\boldsymbol{\theta}}(\mathbf{x})$. The resulting T of Eq. (2.58) becomes

$$\begin{aligned} T \sim E_{\mathbf{x}}[\log p(\mathbf{x}|\hat{\boldsymbol{\theta}}(\mathbf{x})] - \frac{1}{2}E_{\mathbf{x}}[(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}(\mathbf{x})) \cdot \mathbf{I}(\hat{\boldsymbol{\theta}}(\mathbf{x})) \cdot (\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}(\mathbf{x}))] \\ - \frac{1}{2}E_{\mathbf{y}}[(\hat{\boldsymbol{\theta}}(\mathbf{y}) - \boldsymbol{\theta}_0) \cdot \mathbf{I}(\boldsymbol{\theta}_0) \cdot (\hat{\boldsymbol{\theta}}(\mathbf{y}) - \boldsymbol{\theta}_0)]. \end{aligned} \tag{2.59}$$

By construction, the first order term of Eq. (2.59) vanished. Furthermore, to leading order, both quadratic terms are identical such that

$$T \sim E_{\mathbf{x}}[\log p(\mathbf{x}|\hat{\boldsymbol{\theta}}(\mathbf{x})] - E_{\mathbf{y}}[(\hat{\boldsymbol{\theta}}(\mathbf{y}) - \boldsymbol{\theta}_0) \cdot \mathbf{I}(\boldsymbol{\theta}_0) \cdot (\hat{\boldsymbol{\theta}}(\mathbf{y}) - \boldsymbol{\theta}_0)]. \tag{2.60}$$

Evaluated near its maximum, the expectation of $(\hat{\boldsymbol{\theta}}(\mathbf{y}) - \boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}(\mathbf{y}) - \boldsymbol{\theta}_0)$ is, again to leading order, the inverse of the Fisher information matrix [202]. For a $K \times K$ information matrix, we therefore have

$$T \sim E_{\mathbf{x}}[\log p(\mathbf{x}|\hat{\boldsymbol{\theta}}(\mathbf{x}))] - K = E_{\mathbf{x}}[\log p(\mathbf{x}|\hat{\boldsymbol{\theta}}(\mathbf{x})) - K]. \tag{2.61}$$

The model that maximizes T is then equivalent to the model minimizing the AIC [202]

$$AIC \equiv -2 \log p(\mathbf{x}|\hat{\boldsymbol{\theta}}(\mathbf{x})) + 2K. \tag{2.62}$$

In other words, as we increase the number of parameters in our model and the numerical value for the likelihood becomes larger (and the AIC decreases), our complexity cost (2 for each parameter for a total of $2K$) also rises (and the AIC increases). Thus,

our goal is to find a model with a K that minimizes the AIC where, ideally, K is different from 0 or ∞.

The "penalty term" of Eq. (2.62), $+2K$, does not depend on N. This is very different from the BIC, as we will now see, that selects an absolute model without comparison to any reference true model and whose penalty explicitly depends on N.

2.4.3 Bayesian model selection

Parameter marginalization: Illustration on outliers

Parameter marginalization is essential to understanding the BIC. We therefore take a brief detour to discuss this topic in the context of outliers [153].

Suppose, for simplicity, that we are provided a signal with a fixed standard deviation as shown in the inset of Fig. (2.7a). The likelihood of observing a sequence of N independent Gaussian data points, $\mathbf{D} = \mathbf{x}$, drawn from a distribution with unknown mean, μ, and variance, σ^2, is

$$p(\mathbf{x}|\mu,\sigma) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}. \tag{2.63}$$

The posterior is then

$$p(\mu,\sigma|\mathbf{x}) = \frac{p(\mathbf{x}|\mu,\sigma)p(\mu,\sigma)}{p(\mathbf{x})} \tag{2.64}$$

where $p(\mu,\sigma)$ is the prior distribution over μ and σ and the normalization $p(\mathbf{x}) = \int d\mu d\sigma p(\mathbf{x}|\mu,\sigma)p(\mu,\sigma)$.

If we know – or can reliably estimate – the standard deviation, then we may fix σ to that value, say σ_0, and subsequently maximize the posterior to obtain an optimal μ. The form of this posterior, $p(\mu|\mathbf{x},\sigma_0) \equiv p(\mu|\mathbf{x})$, is quite sensitive to outliers because each data point is treated with the same known uncertainty [153]. For example,

the distribution over μ – blue curve in Fig. (2.8) – is heavily influenced by the two apparent outliers near 5-6.

If we have outliers, it may be more reasonable to assume that we are merely cognizant of a lower bound on σ [153]. In this case, a marginal posterior over μ is obtained by integrating σ starting from the lower bound σ_0

$$p(\mu|\mathbf{x}) = \frac{1}{p(\mathbf{x})} \cdot \int_{\sigma_0}^{\infty} d\sigma \; p(\mathbf{x}|\mu, \sigma) p(\mu, \sigma). \tag{2.65}$$

As N grows, the likelihood eventually dominates over the prior and determines the shape of the posterior. This posterior over μ – the orange curve of Fig. (2.8) – is obtained using our previous likelihood (Eq. (2.63)) with $p(\mu, \sigma) = p(\mu)p(\sigma)$ with a flat prior on μ and, as a matter of later convenience, $p(\sigma) = \sigma_0/\sigma^2$. As we no longer commit to a fixed σ, the orange curve is much broader than the blue curve. However it is still susceptible to the outliers near 5-6 since all points are treated as having the same uncertainty though only a range for that uncertainty is now specified.

Relaxing the constraint that all points must have the same uncertainty, the marginal posterior distribution over μ becomes

$$p(\mu|\mathbf{x}) = \frac{1}{p(\mathbf{x})} \cdot \int_{\sigma_0}^{\infty} \prod_i d\sigma_i \; p(x_i|\mu, \sigma_i) p(\mu, \sigma_i). \tag{2.66}$$

This marginal posterior – shown by the green line of Fig. (2.8) – is far less committal than were the previous posteriors we considered. That is, this posterior can assume a multimodal form with the highest maximum centered in the region with the largest number of data points. Unlike our two previous marginal posteriors, the location of this posterior's highest maximum is not as deeply influenced by the outliers. While the highest posterior maximum provides an estimate for μ largely insensitive to outliers, the additional maxima may help identify a possible second candidate μ. This might

be helpful to single molecule force spectroscopy say – where the noise properties depend on the state of the system – and apparent occasional outliers may suggest the presence of an additional force state.

This integration over parameters whose value we do not know – σ in the example above – is key. Through integration, we allow (sum over) a broad range of fits to the data. This naturally reduces the complexity of our model because it contributes parameter values to our marginal posterior that yield both good and bad fits to the data.

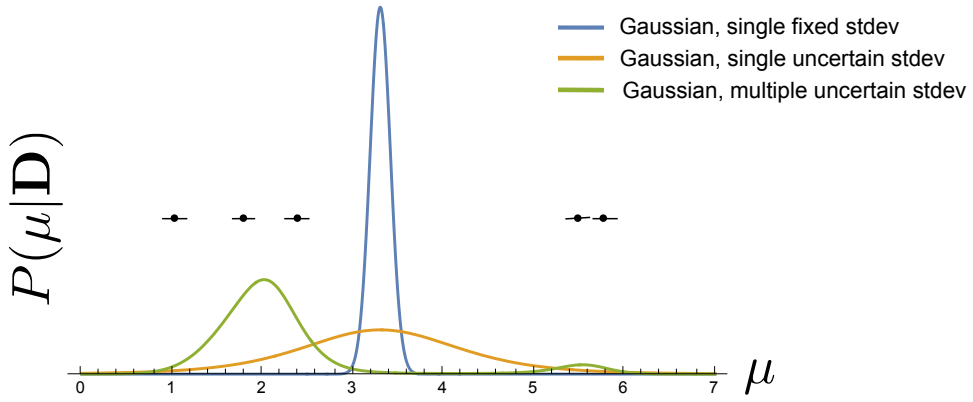

Fig. 2.8. **Noise models can be adapted to treat outliers.** We are given a sequence of data points, $\mathbf{D} = \{1, 1.8, 2.4, 5.5, 5.8\} \pm 0.25$. We want to find the posterior over μ. Blue: We assume the standard deviation is fixed at 0.25 and use a Gaussian likelihood with a single variance for all points. Orange: We assume that the standard deviation's lower bound is 0.25, see Eq. (2.65), but that we still have a single variance for all points. Green: We still assume the standard deviation's lower bound is 0.25 but that all points are assumed to have independent standard deviations, see Eq. (2.66).

The BIC obtained as a sum over models

The BIC seeks a model that maximizes the posterior marginalized over irrelevant or unknown model parameters $\boldsymbol{\theta}$. To compute this posterior, we define a likelihood, $p(\mathbf{D} = \mathbf{x}|\boldsymbol{\theta})$, describing N independent – or, at worst, weakly correlated – identical observations where $p(\mathbf{D}|\boldsymbol{\theta}) \equiv e^{N \log f(\mathbf{D}|\boldsymbol{\theta})}$. To be clear, if the data are completely independent, then f is understood as the likelihood per observation.

We write down a marginal posterior

$$p(K|\mathbf{D}) \propto \int d^K\theta \; e^{N \log f(\mathbf{D}|\boldsymbol{\theta})} \tag{2.67}$$

where K is the total number of parameters.

Since we are interested in a general model selection criterion that does not care about the particularities of the application, we consider the large N limit and, thus, ignore the prior altogether as well.

To approximate the integral in Eq. (2.67), we invoke Laplace's method as before and expand $\log f(\mathbf{D}|\boldsymbol{\theta})$ around its maximum, $\boldsymbol{\theta}^*$, to second order. In other words, we write

$$\begin{aligned} p(K|\mathbf{D}) &\sim e^{N \log f(\mathbf{D}|\boldsymbol{\theta}^*)} \int d^K\theta \; e^{-\frac{N}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\cdot\nabla_{\boldsymbol{\theta}}\nabla_{\boldsymbol{\theta}} \log f(\mathbf{D}|\boldsymbol{\theta}^*)\cdot(\boldsymbol{\theta}-\boldsymbol{\theta}^*)} \\ &= e^{N \log f(\mathbf{D}|\boldsymbol{\theta}^*)} \frac{(2\pi/N)^{K/2}}{\sqrt{det \; \nabla_{\boldsymbol{\theta}}\nabla_{\boldsymbol{\theta}} \log f(\mathbf{D}|\boldsymbol{\theta}^*)}}. \end{aligned} \tag{2.68}$$

The BIC then follows directly from Eq. (2.68)

$$BIC \equiv -2 \log p(K|\mathbf{D}) = -2 \log p(\mathbf{D}|\boldsymbol{\theta}^*) + K \log N + \mathcal{O}(N^0). \tag{2.69}$$

By contrast to the AIC given by Eq. (2.62), the BIC has a penalty that scales as $\log N$. In a later section, we will relate this penalty to the predictive information

provided by the model (Eq. (2.111)). Finally, we add that while the prior over $\boldsymbol{\theta}$ is not treated explicitly, this treatment is still distinctly Bayesian. This is because the model parameters, over which we marginalize, are treated as continuous variables rather than fixed numbers.

We now turn to an illustration of model selection drawn from change-point analysis.

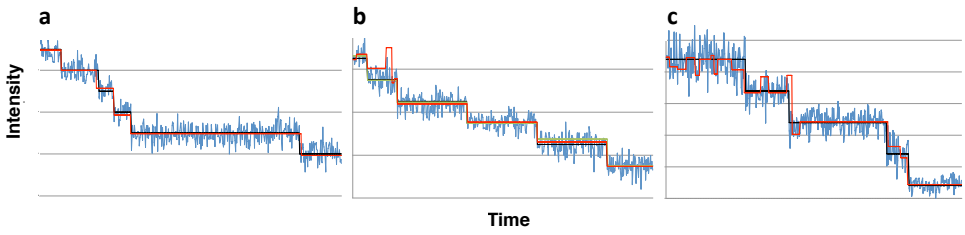

Fig. 2.9. **BIC finds correct steps when the noise statistics are well characterized. a)** Our control. We generated synthetic steps (black line) and added noise (white, decorrelated) with the same standard deviation for each data point. We used a greedy algorithm [70] to identify and compare models according to Eq. (2.73) and identify the correct step locations (red line) from the noisy time trace (blue). **b)** Here we use a different, incorrect, likelihood that does not adequately represent the process that we used to generate the synthetic data. That is, we correctly assumed that the noise was white and decorrelated but also, incorrectly, assumed that we knew and fixed σ (and therefore did not integrate over σ in Eq. (2.71)). We underestimated σ by 12%. Naturally, we overfit (red) the true signal (black). Green shows the step-finding algorithm re-run using the correct noise magnitude. **c)** Here we use the BIC from Eq. (2.73) whose likelihood assumes no noise correlation. However, we generated a signal (black) to which we added correlated noise [by first assigning white noise, ϵ_t, to each data point and then computing a new correlated noise, $\tilde{\epsilon}_t$, at time t from $\tilde{\epsilon}_t = 0.7\epsilon_t + 0.1\epsilon_{t-1} + 0.1\epsilon_{t-2} + 0.1\epsilon_{t-3}$]. As expected, the model that the BIC now selects (red) interprets as signal some of the correlated noise from the synthetic data. We acknowledge K. Tsekouras for generating this figure.

Illustration of the BIC: Change-point algorithms for a Gaussian process

Change-points algorithms locate points in the data where the statistics for a process generating the data change. There is a broad literature, including reviews [230–232], on change-point algorithms relying on the AIC [233–236], the BIC [200, 234, 235, 237–240], generalizations of the AIC [224, 227] and BIC [48, 241], wavelet transforms [242–244] and related techniques [245–249].

Here we illustrate how model selection – and the BIC in particular – are applied to a change-point detection problem for a Gaussian process like the one shown in Fig. (2.7b) with fixed but unknown standard deviation – which is the same for all data points – and with a discretely changing mean.

We begin by writing down the likelihood

$$p(\mathbf{x}|K,\sigma,\boldsymbol{\mu},\mathbf{j}) = \prod_{i=0}^{K-1}\prod_{\ell=j_i}^{j_{i+1}-1}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x_\ell-\mu_i)^2}{2\sigma^2}} \tag{2.70}$$

where K denotes the number of change-points – points where the mean of the signal changes – occurring at locations $\mathbf{j} = \{j_0, \cdots, j_K\}$. To be precise, since the standard deviation is also a parameter to be determined, we have $K+1$ total parameters here.

The model maximizing this likelihood places a change-point at every step ($x_\ell = \mu_i$ for every ℓ). That is, $p(\mathbf{x}|K,\sigma,\boldsymbol{\mu},\mathbf{j})$ peaks when $K = N$ and, expectedly, overfits the data.

To avoid overfitting and – since we are still largely ignorant of the correct values for σ and $\boldsymbol{\mu}$ – we integrate over all allowed values for σ and $\boldsymbol{\mu}$. This yields the following marginal posterior

$$\begin{aligned} p(K, \mathbf{j}|\mathbf{x}) &\propto \int d\sigma d^K\mu \; p(\mathbf{x}|K, \sigma, \boldsymbol{\mu}, \mathbf{j}) \\ &= \frac{\sqrt{2\pi}^{-(N-K)}}{n_0^{1/2} \cdots n_K^{1/2}} \cdot \frac{1}{2} \cdot \left(\frac{S}{2}\right)^{-\frac{(N-K-1)}{2}} \cdot \left(\frac{N-K-3}{2}\right)! \end{aligned} \tag{2.71}$$

where $S \equiv n_0\hat{\sigma}_0^2 + \cdots + n_K\hat{\sigma}_K^2$ and

$$\hat{\sigma}_i^2 \equiv \frac{1}{n_i} \sum_{\ell=j_i}^{j_{i+1}-1} x_\ell^2 - \frac{1}{n_i^2} \left(\sum_{\ell=j_i}^{j_{i+1}-1} x_\ell\right)^2 \tag{2.72}$$

where n_i counts the number of points contained in the i^{th} step.

Eq. (2.71) reveals that $p(K, \mathbf{j}|\mathbf{x})$ may no longer be peaked at $K = N$. This is expected since, conceptually, by summing over σ, grossly underfitting models (models with large σ) now contribute to our marginal posterior.

Taking the further simplifying assumptions that: 1) $n_i \sim N/K$; 2) all n_i are large; and 3) $\hat{\sigma}_0^2 \sim \hat{\sigma}_1^2 \cdots \sim \hat{\sigma}_K^2$ are all equal to an expected standard deviation $\hat{\sigma}^2$, we recover a form for a Gaussian process BIC seen in the literature [70]

$$BIC = -2 \log p(K, \mathbf{j}|\mathbf{x}) = N \log \hat{\sigma}^2 + K \log N + \mathcal{O}(N^0) + \text{const} \tag{2.73}$$

where the constants, const, capture all terms independent of model parameters (that may depend on N).

Fig. (2.9) shows the detection of change-points in synthetic data and illustrates just how sensitive the BIC is to the correct choice of likelihood. To address this sensitivity, BIC's have, for example, been tailored to detect change-points with time

correlated noise as would be expected from methods such as single molecule force spectroscopy [200].

Shortcomings of the AIC and BIC

We cannot compare data sets of different lengths: The objective functions for both the AIC and BIC depend on N. For this reason, we cannot directly compare numerical values for AICs and BICs obtained for data sets of different lengths [250]. This problem often arises when comparing data sets of originally the same length but with a different number of outliers removed.

Correctly characterizing the likelihood is critical: We illustrate, in Fig. (2.9b)-(2.9c), how a mischaracterization of the likelihood – and, ultimately, the process that generates the noise – can yield incorrect models.

The curvature of the likelihood function may vanish: The curvature of the likelihood arises in both the AIC and BIC. In the AIC, it appears through the Fisher information (see the second term in Eq. (2.58)) while in the BIC it arises from Laplace's method [227, 251, 252]. For singular problems, those where the likelihood's curvature vanishes, the AIC and BIC diverge. In concrete terms, this signifies that the model selection criterion becomes broadly insensitive to the model's dimensionality. A vanishing curvature occurs: i) when we have unidentifiable parameters. That is at locations, $\mathbf{x} = \mathbf{x}^*$, where $p(\mathbf{x}^*|\theta_1, \theta_2) = p(\mathbf{x}^*|\theta_1)$. In this case, at $\mathbf{x}^*$, $\partial_{\theta_1}\partial_{\theta_2} \log p(\mathbf{x}|\theta_1, \theta_2)|_{\mathbf{x}=\mathbf{x}^*} = 0$; ii) at change-points (changes in model parameter) locations in the data [253]. For instance, consider a force spectroscopy experiment monitoring the stepping motion of a molecular motor. After a single step, the signal appears to jump from a mean of μ to μ'. But, in practice, transitions may not be so discrete. In the extreme case, if hypothetically we collected data with an infinite time resolution, we may see the signal (i.e. a hypothetical noiseless time trace) pass

through a region of zero curvature as it continuously transitions from a region of negative to positive curvature. At this singular point, the AIC and BIC fail.

Despite vanishing curvatures at change-points, the AIC and BIC are commonly used in change-point analysis, as we have seen for our illustrative example. In practice, change-point methods ignore the point of zero curvature. That is, they treat the data as piecewise continuous as we had in our example.

Alternatively, to avoid vanishing leading order (quadratic) corrections, we may select a model by evaluating (often numerically) the full posterior rather than approximating it (as a BIC). Or, we can use a generalized frequentist information criterion (FIC) [224, 254] by starting with a biased estimator for T

$$T \sim \log p(\mathbf{x}|\hat{\boldsymbol{\theta}}(\mathbf{x})) \tag{2.74}$$

and quantifying its bias $E_x E_y \left[\log p(\mathbf{x}|\hat{\boldsymbol{\theta}}(\mathbf{y})) - \log p(\mathbf{x}|\hat{\boldsymbol{\theta}}(\mathbf{x}))\right]$.

Note on finite size effects

Both AIC and BIC are asymptotic statements valid in the large N limit. In change-point methods, we have found that they may be reliable for modest N, even below 50. In the case of the BIC, for instance – where Laplace's approximation is invoked – this is not surprising since the terms ignored are exponentially, not polynomially, subdominant in N. For smaller N, higher order corrections to the AIC – an example of which is called the AICc – depending only on K and N can be computed [212]. By contrast, higher order corrections to the BIC, to order N^0, explicitly capture features of the prior and these corrections have previously been applied to the detection of change-points from FRET data [48].

Select AIC and BIC applications

Model selection criteria, whether frequentist (AIC) or Bayesian (BIC), deal directly with likelihoods of observing the data. Treating the data directly using likelihoods avoids unnecessary data processing such as histogramming or data reduction into moments, cumulants, correlation functions or other heuristic point statistics that are otherwise common in the physics literature when dealing with macroscopic systems.

It is especially useful to treat the data directly in single molecule data analysis where the data, on the one hand, are plentiful (because of high acquisition rates) but where data, on the other hand, show too few discrete events to build a reliable histograms to fit a model to the data [45, 235].

The BIC has been widely used in biophysics [41, 52, 70, 255, 256] to determine: the number of intensity states in single molecule emission time traces [237, 257]; the number of steps in photobleaching data [238]; and the stepping dynamics of molecular machines [239]. The AIC, in turn, has been used to determine the number of diffusion coefficients sampled from labeled lipophilic dyes on a surface from single molecule trajectories [236]; and the number of ribosomal binding states of tRNA from single molecule FRET trajectories [233]. In addition, both the AIC and BIC have provided complementary insight on a number of problems including: photon arrival time kinetic parameters (such as rate of emission for different fluorophore states for an unknown number of fluorophore states and transition kinetics between them) [235]. And, when compared head-to-head in identifying trapping potentials for membrane proteins relevant to single particle tracking [234], the AIC and BIC perform differently across potentials. This is because, by contrast to simple Brownian motion, confining radial potentials, such as $V \propto r^4$, introduce nonlinear dynamics that are better ap-

proximated by complex models that are less heavily penalized by the AIC than by the BIC.

While the AIC and BIC have been useful, they are often used complementarily specifically because they are treated as model selection heuristics. As a result, their conceptual difference – and why their penalties differ – are rarely addressed [227] and it is therefore common, though ultimately incorrect, to treat the complexity penalty ($2K$ for the AIC or the $K \log N$ for the BIC) as an adjustable form.

We end this section on model selection with a note on additional methods used in single molecule analysis that have been directly inspired by the BIC [48, 52, 94, 200, 258]. For instance, Shuang *et al.* [48] have proposed an MDL heuristic – a method called STaSI (Step Transition and State Identification) – not only to find steps but subsequently identify states as well in time traces; see Fig. (2.10). The idea here is not only to penalize both the number of change-points detected but also to discretize the number of intensity levels (states) sampled in an smFRET trajectory. STaSI has subsequently been used to explore equilibrium transitions among N-methyl-D-aspartate receptor conformations by monitoring the distance across the glycine bound ligand binding domain cleft [241].

Maximum Evidence: Illustration on HMMs

While we have previously discussed HMMs, we have not addressed the important model selection challenge they pose [46, 47, 259]. That is, it appears paradoxical to assume that while we have no *a priori* knowledge of the HMM's model parameters, we have perfect knowledge of the underlying state-space. To address this challenge, it is possible to use a combination of maximum likelihood on the HMM and information criteria to select the number of states [41].

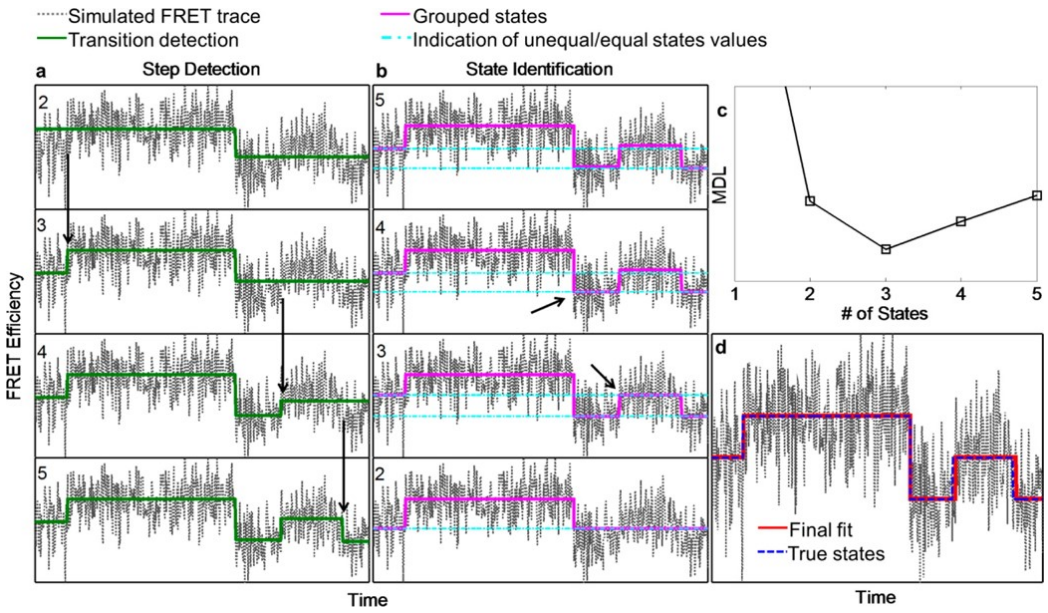

Fig. 2.10. **Identifying states can be accomplished while detecting steps.** STaSI is applied to synthetic smFRET data. STaSI works by first iteratively identifying change-points in the data (successive steps shown by arrows in panel (a). The mean of the data from change-point to change-point defines an intensity (FRET) state. An MDL heuristic is subsequently used to eliminate (or regroup) intensity levels (b). The MDL is plotted as a function of the number of states (c). The final analysis – with change-points and states identified – is shown in (d). For more details see Ref. [48].

Fig. (2.11) illustrates a different strategy to infer the total number of states (K) from a time trace. Fig. (2.11) compares the number of states inferred from a synthetic FRET time trace using maximum likelihood and maximum evidence [47]. The maximum likelihood, $p(\mathbf{y}|\boldsymbol{\theta}^*, K)$ – where $\boldsymbol{\theta}^*$ designates the parameter estimates that maximize the likelihood – increases monotonically with the number of states. By contrast, the maximum evidence – $p(\mathbf{y}|K)$ defined as the likelihood marginalized over all unknown parameter values – peaks at the theoretically expected number of states.

While the concept of maximum evidence is not limited to HMMs, here we use maximum evidence to illustrate model selection on HMMs.

Like the BIC, maximum evidence penalizes complexity by summing over all unknown parameters not all of which fit the data very well. So, to construct the evidence for our HMM example, we consider the joint likelihood for a sequence of observations, $\mathbf{y}$, and states populated at each time interval, $\mathbf{s}$,

$$p(\mathbf{y}, \mathbf{s}|\boldsymbol{\theta}, K) = \prod_{i=2}^{N} [p(y_i|\mathbf{s}_i, \boldsymbol{\theta}, K) p(\mathbf{s}_i|\mathbf{s}_{i-1}, K)] p(y_1|\mathbf{s}_1, \boldsymbol{\theta}, K) p(\mathbf{s}_1|K) \tag{2.75}$$

where $\boldsymbol{\theta}$ denotes a vector of parameters: the K-dimensional initial probability vector of states, $\boldsymbol{\pi} \equiv p(\mathbf{s}_1|K)$; the K-dimensional observation parameters such as means, $\boldsymbol{\mu}$, and standard deviations, $\boldsymbol{\sigma}$, for each state assuming Gaussian $p(y_i|\mathbf{s}_i, \boldsymbol{\theta}, K)$; and the $K \times K$ matrix, $\mathbf{A}$, of transition matrix elements $a_{ij} = p(\mathbf{s}_j|\mathbf{s}_i, K)$. Contrary to Eq. (2.9), we have made all parameter dependencies of the probabilities contained in Eq. (2.75) explicit.

The evidence then follows from Eq. (2.75)

$$p(\mathbf{y}|K) = \sum_{\mathbf{s}} \int d\boldsymbol{\theta} p(\mathbf{y}, \mathbf{s}|\boldsymbol{\theta}, K) p(\boldsymbol{\theta}|K) \tag{2.76}$$

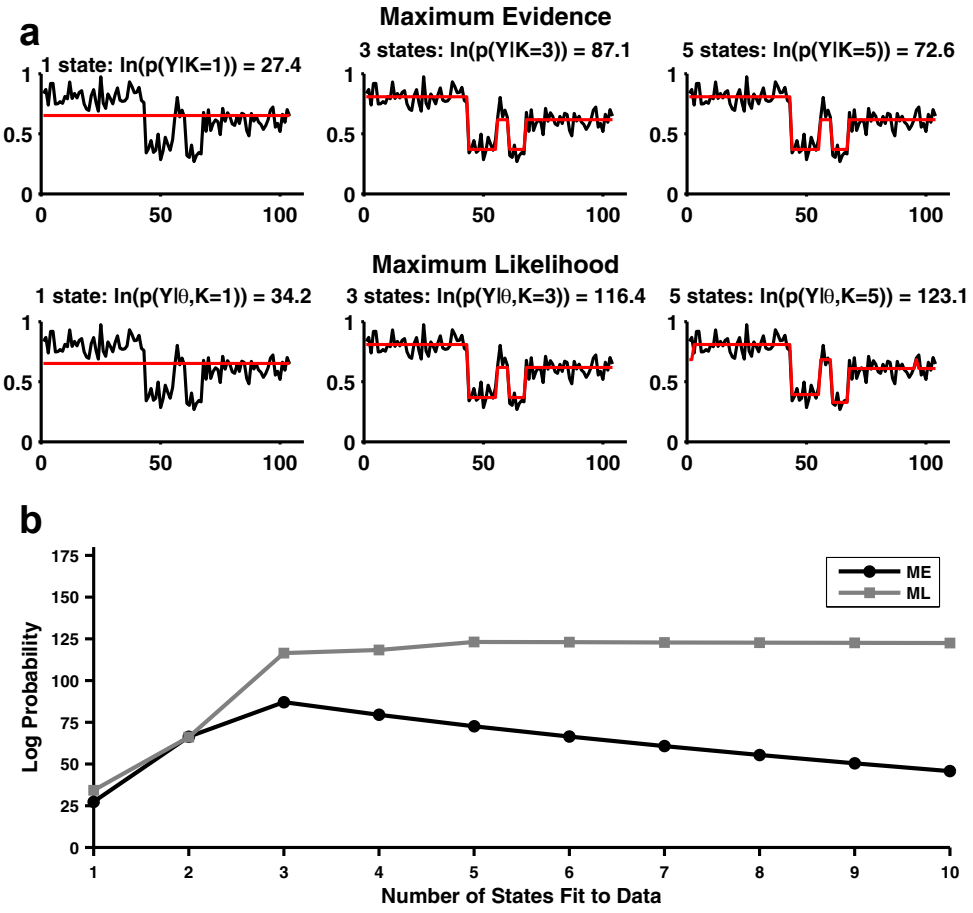

Fig. 2.11. **Maximum evidence can be used in model selection. a)** For this synthetic time trace, maximum likelihood (ML) will overfit the data. This is clear from **b)** where it is shown that the log likelihood or probability of the model – evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ – increases monotonically as we increase the number of states, K. By contrast, maximum evidence (ME) – obtained by marginalizing the likelihood over $\boldsymbol{\theta}$ – identifies the theoretically expected number of states, $K = 3$. Sample time traces are shown in (a) and the log probability is plotted in (b). See details in text and Ref. [47].

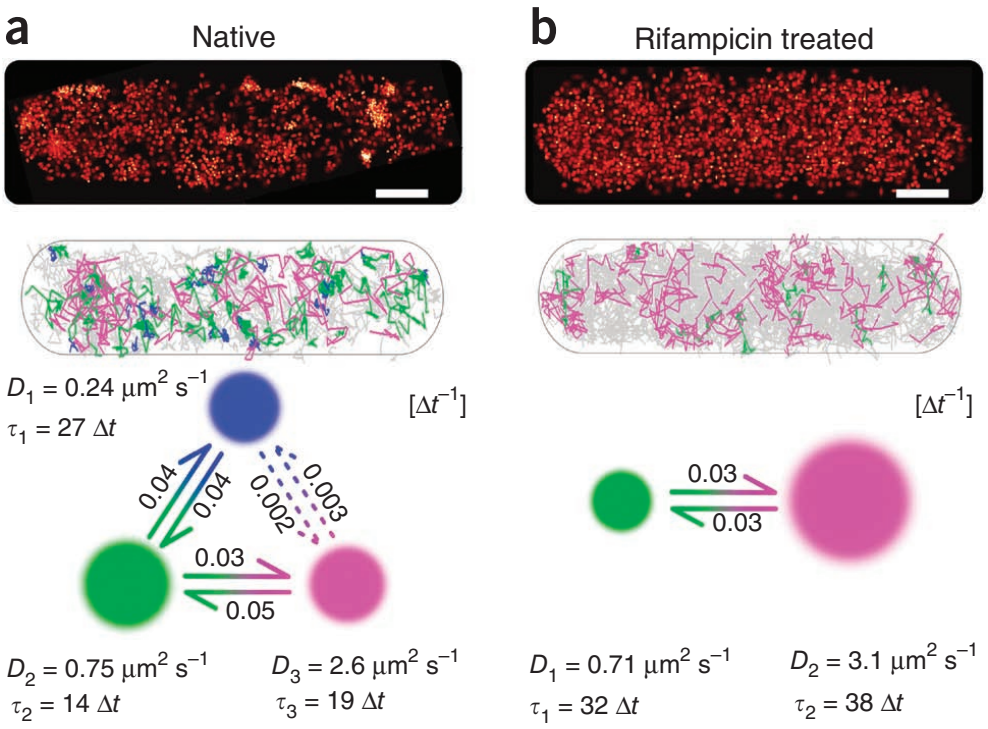

Fig. 2.12. **The number of diffusive states detected using maximum evidence can establish changes in interactions of Hfq upon treatment of** *E. coli* **cells with rifampicin. a)** vbSPT analysis of the RNA helper protein Hfq tracking data. Three distinct diffusive states are detected and sample trajectories are shown color-coded according to which state they belong. The kinetic scheme shows the diffusion coefficient in each state as well as transition rates between diffusion coefficients. **b)** When treated with a transcription inhibitor (rif), vbSPT finds that the slowest diffusive state vanishes suggesting that the slowest diffusive state of Hfq was related to an interaction of Hfq with RNA. $\Delta t = 300$Hz throughout the figure. The scale bar indicates 0.5 μm^2/s. See details in Ref. [46].

where the prior is $p(\boldsymbol{\theta}|K) = p(\boldsymbol{\pi}|K)p(\mathbf{A}|K)p(\boldsymbol{\mu}, \boldsymbol{\sigma}|K)$ and an example of how this prior is selected is given in Ref. [47].

A numerical variational Bayesian (vb) procedure called vbFRET [47] was implemented to evaluate the maximum evidence with sample results shown in Fig. (2.11). The method has gained traction because it learns the number of states by comparing the probability of observation given different values of K, $p(\mathbf{y}|K)$ [208, 260–262]. The interested reader should refer to a general discussion of variational approximations in Ref. [98].

Using a method similar to vbFRET, maximum evidence applied to a HMM model was used to extract the number of diffusion coefficients – "diffusive states" – sampled from intracellularly diffusing proteins as well as transition rates describing the hopping kinetics between the diffusive states [46]. Using single particle tracking (SPT) data, the method was also implemented using a variational Bayesian procedure called vbSPT and applied to infer the diffusive dynamics of an RNA helper protein, Hfq, mediating the interaction between small regulatory RNAs and their mRNA targets [46]. Fig. (2.12) details the analysis of the Hfq tracking done on a control *E. coli* cell and one treated with a transcription inhibitor [rifampicin (rif)]. Fig. (2.12b) shows the disappearance of the slow diffusion component for treated cells suggesting that this sluggish component was associated with Hfq-RNA interactions in the untreated cell.

Finally, using a method called variational Bayes HMM for time-stamp FRET (VB-HMM-TS-FRET), the methods above can be generalized to treat time stamp photon arrival (as opposed to assuming binned data in intervals Δt) as well as time-dependent rates [261].

2.5 An Introduction to Bayesian Nonparametrics

We have already seen how flexible (nested) models – models that can be refined by the addition or removal of parameters – were critical to change-point analysis and enumeration of states in HMMs. Likewise, we have also seen how free-form probability distributions, $\{p_1, \cdots, p_K\}$ could be inferred from MaxEnt even if K, the number of parameters, largely exceeded the number of measured data points. Inferring a large number of parameters, i.e. a distribution on a fine grid, is useful even if many probabilities inferred from MaxEnt have small numerical values. For example, these probabilities may predict the relative weight of sampling a biologically relevant albeit unusual protein fold – as compared to its native conformation – based on free energy estimates alone even if such conformations are highly unlikely.

These previous treatments went beyond parametric modeling where models have a given mathematical structure with a fixed number of parameters, such as Gaussians with means and variances.

While the maximum evidence methods we presented earlier provided model probabilities (marginal likelihoods $p(\mathbf{y}|K)$) for a fixed number of states or parameters (K), in this section we investigate the possibility of averaging over all acceptable K to find posteriors $p(\boldsymbol{\theta}|\mathbf{y})$.

These posteriors – $p(\boldsymbol{\theta}|\mathbf{y})$ obtained by averaging over all possible starting models – are the purview of Bayesian nonparametrics, a reasonably new (1973) approach to statistical modeling [263]. Bayesian nonparametrics are poised to play an important role in the analysis of single molecule data since so few model features – such as the number of states of a single molecule in any time trace – are known *a priori*.

Contrary to their name, nonparametric models are not parameter-free [263]. Rather, they have an *a priori* infinite number of parameters that are subsequently winnowed down – or, more precisely as we will see, selectively sampled – by the available

data [263–266]. This large initial model-space attempts to capture all reasonable starting hypotheses [267] and avoids potentially computationally costly model selection and model averaging [268]. In other words, they let the model complexity adapt to the information provided by the raw data and can efficiently and rigorously promote sparse models through the priors considered.

2.5.1 The Dirichlet process

An important object in Bayesian nonparametrics is the prior process and the most widely used process is the Dirichlet process (DP) prior [263]. Much, though not all of Bayesian nonparametrics, relies on generalizations of the DP and its representations; see first graph of Ref. [269]. These representations include the infinite limit of a Gibbs sampling for finite mixture models, the Chinese restaurant process and the stick-breaking construction [266,270]. We will later discuss the stick-breaking construction.

Samples from a DP are distributions much like samples from the exponential of the entropy that we saw earlier are distributions as well. Density estimation [271] and clustering [272] are natural applications of the DP.

To introduce the DP, we start with a parametric example and first consider a probability of outcomes indexed k, $\boldsymbol{\pi} = \{\pi_1, \pi_2, \cdots, \pi_k\}$ – with $\Sigma_k \pi_k = 1$ and $\pi_k \geq 0$ for all k – distributed according to a Dirichlet distribution, $\boldsymbol{\pi} \sim Dirichlet(\alpha_1, \cdots, \alpha_K)$. That is, with a distribution over $\boldsymbol{\pi}$ given by

$$p(\boldsymbol{\pi}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1}. \tag{2.77}$$

The Dirichlet distribution is conjugate to the multinomial distribution. Thus, a sequence of observations, $\mathbf{z} = \{z_1, z_2, \cdots z_N\}$, is distributed according to a multinomial having K unique bins with populations $\mathbf{n} = \{n_1, n_2, \cdots, n_K\}$

$$p(\mathbf{z}|\boldsymbol{\pi}) = \frac{\Gamma(\sum_k n_k + 1)}{\prod_k \Gamma(n_k + 1)} \prod_{k=1}^{K} \pi_k^{n_k}. \tag{2.78}$$

The resulting posterior obtained from the prior, Eq. (2.77), and likelihood, Eq. (2.78), is

$$p(\boldsymbol{\pi}|\mathbf{z}, \boldsymbol{\alpha}) = \frac{p(\mathbf{z}|\boldsymbol{\pi})p(\boldsymbol{\pi}|\alpha)}{\int d\boldsymbol{\pi} p(\mathbf{n}|\boldsymbol{\pi})p(\boldsymbol{\pi}|\alpha)} = \frac{\Gamma(\sum_k n_k + \sum_k \alpha_k)}{\prod_k \Gamma(n_k + \alpha_k)} \prod_{k=1}^{K} \pi_k^{n_k+\alpha_k-1}. \tag{2.79}$$

Now, imagine a sequence of $N-1$ observations, $\{z_1, z_2, \cdots z_{N-1}\}$. Using the posterior above, we can calculate the probability of adding an observation to a pre-existing cluster, j, with probability π_j given the occupation $\{n_1, \cdots, n_{K-1}\}$ of all pre-existing clusters $K-1$ clusters.

For simplicity we assume all α_k identical and equal to α/K. Then

$$\begin{aligned} p(z_N = j|\{z_1, \cdots, z_{N-1}\}, \alpha) &= \int d\boldsymbol{\pi}\ p(z_N = j|\boldsymbol{\pi})p(\boldsymbol{\pi}|\{z_1, \cdots, z_{N-1}\}, \alpha) \\ &= \int d\boldsymbol{\pi}\ \pi_j p(\boldsymbol{\pi}|\{z_1, \cdots, z_{N-1}\}, \alpha). \end{aligned} \tag{2.80}$$

We can evaluate both: i) the probability that our observation populates an existing cluster with n_j members; or ii) that our observation populates a new cluster. In the non-parametric limit – where we allow an *a priori* infinite number of clusters ($K \to \infty$) – these probabilities are [269]

$$\frac{n_j}{\alpha + N - 1} \qquad \text{vs.} \qquad \frac{\alpha}{\alpha + N - 1} \tag{2.81}$$

respectively. Thus α – predictably called a concentration parameter – measures the preference for creating a new cluster. The DP therefore tends to populate clusters according to the number of current members.

The DP describes the infinite dimensional ($K \to \infty$) generalization of the Dirichlet distribution and describes the distribution over $\boldsymbol{\pi}$, or equivalently, the distribution over G defined as

$$G \equiv \boldsymbol{\pi} \cdot \mathbf{1} = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k} \tag{2.82}$$

where the Kronecker delta, δ_{θ_k}, denotes a mass point for parameter value θ_k [270,273].

The θ_k themselves are iid (identical independently distributed) samples from a base distribution H (e.g. a Gaussian). In other words, the base distribution parametrizes the density from which the θ_k are sampled, i.e.

$$\begin{aligned} G &\sim DP(\alpha, H) \\ \theta_k &\sim H. \end{aligned} \tag{2.83}$$

Thus as $\alpha \to \infty$, we have $G \to H$. The idea is to use H as the hypothetical parametric model we would have started from and use α to relax this assumption [266].

The stick-breaking construction, which we mentioned earlier, is a representation of the DP that can be implemented. If we follow [269]

$$\begin{aligned} v_k &\sim Beta(1, \alpha) \\ \pi_k &= v_k \prod_{j=1}^{k-1} (1 - v_j) \\ \theta_k &\sim H \\ G &= \Sigma_{k=1}^{\infty} \pi_k \delta_{\theta_k} \end{aligned} \tag{2.84}$$

then, without proof, we find that $G \sim DP(\alpha, H)$.

The analogy to stick-breaking follows from the steps given in Eq. (2.84). We begin with a stick of unit length and break the stick at location, v_1 , sampled from a Beta distribution $v_1 \sim Beta(1, \alpha)$. We assign $\pi_1 = v_1$. The remainder of the stick has length $(1 - v_1)$. The value of θ_1 that we then assign to π_1 is sampled from H. We then reiterate to determine π_2.

The π_k sampled according to the stick-breaking construction are now decreasing on average but not monotonically so. In practice, the procedure is terminated when the remaining stick is below a predesignated threshold. In the statistics literature, it is said that $\boldsymbol{\pi}$ is sampled according to $\pi \sim GEM(\alpha)$ where GEM stands for Griffiths-Engen-McClosky [274].

2.5.2 The Dirichlet process mixture model

We have used the DP, thus far, to generate discrete sample distributions, G. In order to treat continuous random variables, like y, we generalize our treatment and introduce the Dirichlet Process Mixture Model (DPMM) [275,276] where a continuous parametric distribution, F, is convolved with G, given by Eq. (2.82),

$$y \sim \int G(\theta)F(y|\theta)d\theta = \Sigma_{k=1}^{\infty} \int \pi_k \delta_{\theta_k} F(y|\theta)d\theta = \sum_k \pi_k F(y|\theta_k). \tag{2.85}$$

In other words [277]

$$\begin{aligned} y|\theta_k &\sim F(y|\theta_k) \\ \theta_k &\sim H \\ G &\sim DP(\alpha, H). \end{aligned}$$

The full posterior we want to determine is now

$$p(\theta|y) \propto \int d\pi p(y|\theta,\pi,F)p(\theta|H)p(\pi|\alpha). \tag{2.86}$$

where, to be explicit, $\pi \sim GEM(\alpha)$. In practice, in order to sample from this posterior, we must first sample the distribution G, then, given this G, we construct the mixture model involving F. Then, to sample y, we must determine from which mixture component, k, the random variable y was selected and subsequently sample y from the designated $F(y|\theta_k)$. We must then repeat over multiple G's.

Many specific Markov chain Monte Carlo (MCMC) methods such as the simple approach above – called Gibbs sampling – are discussed for the DPMM in Ref. [277]. A more general discussion on sampling from posteriors using MCMC can be found in Ref. [98].

Fig. (2.13) illustrates the ability of DPMMs to infer rates from a density generated by sampling 500 data points, y, from $y \sim \sum_i \pi_i e^{-y\theta_i}$ with four exponential components [154]. The DPMM then tries to determine how many components there were and correctly converges to four mixture components after fewer than 200 MCMC iterations with values for the rates closely matching those used to generate the synthetic data; see Fig. (2.13) for more details.

In comparison to MaxEnt deconvolution methods applied to exponential processes discussed earlier, MaxEnt would have generated a very large number of components (since MaxEnt generates smooth distributions) typically tightly centered at the correct values for the rates at low levels of noise in the data. The DPMM, on the other hand, converges to four discrete components with broader distributions over rates instead.

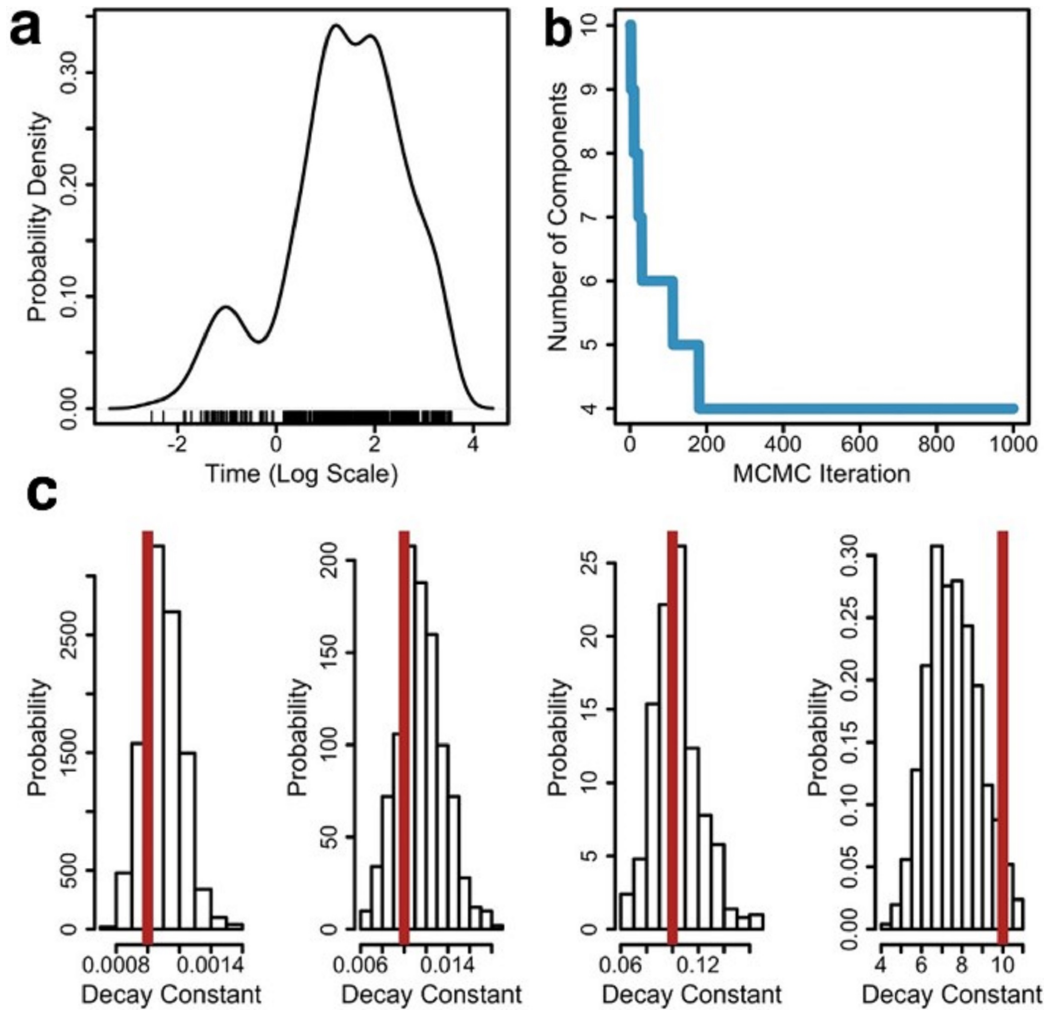

Fig. 2.13. **DPMMs can be used in deconvolution. a)** A density generated from $N = 500$ data points from the mixture of four exponential components. **b)** After fewer than 200 MCMC iterations, the DPMM has converged to four mixture components. **c)** The marginal distribution of the parameter for each mixture component is shown with the red line indicating the theoretical value used to generate the synthetic data $(0.001, 0.01, 0.1, 10)$. See Ref. [154] and main body for more details.

2.5.3 Dirichlet processes: An application to infinite hidden Markov model

As we mentioned earlier, one important challenge with HMMs is their reliance on a predefined number of states. To overcome this challenge, we may use the DP from

which to sample the HMM transition matrix, $p(s_t|s_{t-1})$ [270, 278, 279]. That is, the prior probability of starting from s_{t-1} and transitioning to any of an infinite number of states is sampled from a DP. This is the idea behind the infinite hidden Markov model (iHMM) which we now discuss.

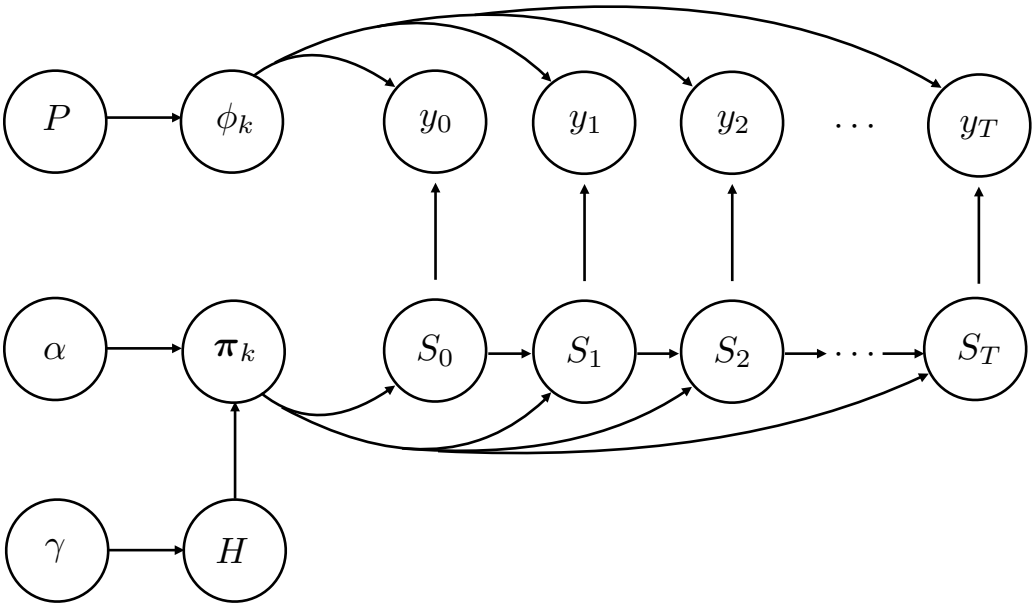

Fig. 2.14. **iHMM Graphical Model** [280].

The most naive formulation of the iHMM – one that samples from a DP the prior probability of the final state $\ell = s_t$ from a given initial state $m = s_{t-1}$, $p(\ell|m)$ – does not sufficiently couple states. In other words, the state ℓ is preferentially revisited under the DP process if transitions from m to ℓ have already occurred. But, because at every time step, m is a new state then, under the DP prior, the same states are never revisited.

To address this problem, the hierarchical DP (HDP) is used [270]. Briefly, under a HDP, we have

$$G \sim DP(\alpha, H)$$
$$H \sim GEM(\gamma)$$

where γ is a hyperparameter that plays the role of a concentration parameter on the prior of the base distribution of the DP. The HDP enforces that the probability of s_t starting from $m = s_{t-1}$ is sampled from a DP whose base has a common distribution amongst all transition probabilities. We summarize iHMM's as follows [259]

$$H \sim GEM(\gamma)$$
$$\boldsymbol{\pi}_k \sim DP(\alpha, H)$$
$$\phi_k \sim P$$
$$s_t|s_{t-1} \sim Multinomial(\boldsymbol{\pi}_{s_{t-1}})$$
$$y_t|s_t \sim F(y|\phi_{s_t})$$

where $\boldsymbol{\pi}_k$ are transitions out of state k, $F(\phi_{s_t})$ describes the probability of observing y_t under the condition we are in state s_t, P is a prior distribution over observation parameters. A graphical model illustrating the parameter inter-dependencies is shown in Fig. (2.14). While parameters can be inferred on an iHMM using Gibbs sampling [98], recent methods have been developed [85, 280], for example to increase the computational efficiency of implementing the iHMM by limiting the number of states sampled at each time point [280].

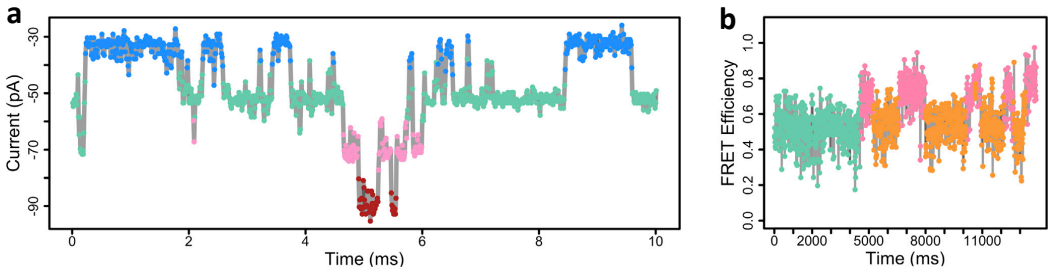

Fig. 2.15. **iHMM's can learn the number of states from a time series.** iHMMs not only parametrize transition probabilities as normal HMMs do. They also learn the number of states in the time series [154]. Here they have been used to find the number of states for **a)** ion (BK) channels in patch clamp experiments [with downward current deflections indicating channels opening]; **b)** conformational states of an agonist-binding domain of the NMDA receptor.

Most recently, the potential of iHMMs in biophysics has been illustrated by applying iHMMs for parameter determination and model selection in single and multichannel electrophysiology (Fig. (2.15a)), smFRET (Fig. (2.15b)) as well as single molecule photobleaching [154]. The colors in both time traces shown in Fig. (2.15) denote the different states that the system visits over the course of the time trace. While most transitions can be detected by eye in the first time trace, the second time trace demonstrates the potential of iHMMs to go beyond what is possible by visual inspection. There remain some clear challenges. For instance, it is conceivable that iHMM's willingness to introduce new states may over-interpret drift, say, in a time trace as the population of new states.

2.6 Information Theory: State Identification and Clustering

Single-molecule time-series measurements are often modeled using kinetic, or state-space, networks. As we have already seen, HMMs presuppose a kinetic model, and – once all parameters are determined – assign a probability of being in a partic-

ular state along a time series. Other methods such as DPMMs and iHMMs are more flexible and do not start with fixed kinetic models *a priori*. Information theory, by contrast, provides a alternative to non-parametric methods in identifying probabilities of states (thought of as data "clusters") populated along time series, that does not rely on prior processes.

Here we focus on information theoretic clustering [281–283] that follow directly from Shannon's rate-distortion theory (RDT) [75].

2.6.1 Rate-Distortion Theory: Brief outline

Shannon [75] conceived of rate-distortion theory (RDT) to quantify the amount of information that should be sent across noisy communication channels in order to convey a message within a set error margin. Although RDT was developed for both continuous as well as discrete transmissions, here – in the interest of single-molecule data analysis and for sake of brevity – we will restrict our discussion to discrete signals.

Shannon considered a transmission (a message) consisting of discrete signals from a source to a recipient. For example, if the message were intended to convey words in the English language, then each discrete signal would transmit a letter of the Latin alphabet. However, the letters comprising the words being transmitted cannot be sent directly; the communication channel requires that the transmitted information be encoded. That is, the set of letters being transmitted needs to be transformed into a set of codes that represent the letters being sent. For example, the letter A could be encoded by a set of binary symbols such as 0000. After transmission, the encoded signals are collected by the recipient and are then translated back into their representation in the original set of symbols. In this example, the average length of the binary sequences that encode the letters being sent corresponds roughly to the

"rate" (of information) and the potential for misinterpreting the encoded sequence by the recipient corresponds to the "distortion."

Put differently, the rate quantifies the amount of information about the intended message that is being transmitted across the channel. For example, consider encoding the letter A in two ways: a single binary character 0 and a binary sequence 00. Because the single binary character is of shorter length than the two-character sequence, it contains less information about the intended transmission (the letter A) than the longer sequence. Increasing the length of the encoded representation of the intended message thus increases the amount of information being transmitted.

The distortion, on the other hand, quantifies the potential for misinterpreting a transmission. For example, the encoded message may be transmitted as A = 00, but noise on the channel distorts the transmission, resulting in the signal being interpreted as 01, which may coincide with another letter, say B.

As a rudimentary example, consider a one-word message, "kangaroo", being transmitted from a source to a recipient as a set of binary symbol groups with each group representing a single letter of the alphabet. The intended message, "kangaroo", is first transformed at the source from the letters of the alphabet to a sequence of binary symbol groups, which are then transmitted to the recipient and decoded (translated) back into letters from the message.

Because the message will most likely be understood even if one or two letters is misinterpreted in the decoding process, some small level of distortion may be acceptable, e.g. "kongaroo" versus "kangaroo". On the other hand, if several letters of the message received are misinterpreted, then the correct translation of the intended message is unlikely. This latter situation is undesirable, and may be remedied by increasing the lengths of the binary sequences – i.e. by increasing the rate of informa-

tion – that encode the letters, thereby increasing the probability of correctly decoding the intended message and decreasing the level of distortion.

Then, given a level of acceptable distortion, such as one in eight letters ("konga-roo" vs. "kangaroo") we may then determine how short the binary symbol groups must be in order to accurately convey the intended message. Finding the optimal length of letters is the subject of RDT.

More formally, RDT poses the following question: what is the minimum rate of information required to convey the intended message at the desired level of distortion? RDT's main result is that a lower-bound on the rate of information is provided by the mutual information between the set of possible transmissions at the source and the set of observations at the recipient. In our example above, this quantity is the mutual information between the letters of the alphabet and the set of binary sequences that encode each letter. The minimum rate of information is then obtained by minimizing this mutual information given an acceptable level of distortion.

RDT and data clustering

We now discuss the relationship between RDT and data clustering. Data clustering is the grouping of elements of a data set into a subsets of elements, i.e. clusters, containing elements that have similar properties. This is often accomplished through the minimization of an average statistical distance between the elements assigned to particular clusters and their center by, for example, a method called k-means clustering [98]. Since there are typically fewer clusters than there are elements in the data set, clustering is a form of compression. In other words, data clustering seeks to compress the data with respect to some statistical distance.

In RDT, minimization of the rate of information is also a form of compression. In effect, by minimizing the rate we are minimizing the length of the encoded message to

be transmitted, thereby compressing the message. This compression is not performed directly, however, but with respect to a desired (or, as we will discuss, observed) level of distortion. For an infinite compression, there could be substantial distortion. Likewise, if every element belongs to its own cluster the distortion vanishes.

By contrast to "hard" partitioning algorithms – where probabilities of elements belonging to clusters are restricted to 0 or 1 – soft clustering algorithms allow elements to exist across all clusters with some probability of membership assigned to each cluster. These more general algorithms are particularly useful for cases where clusters may overlap. This is especially advantageous in the context of high-noise single-molecule measurement and allows estimation of errors associated with any parameter extracted from the experimental data. For instance, the smFRET example that we will explore later will have two important sources of error: empirical error (photon counting and fluctuations in the irradiance intensity) and sampling eror (the number of data points are finite).

Since, as we will see in the next section, RDT clustering directly returns conditional probabilities that each data point belongs to each cluster [283, 284], soft partitioning is directly built into the RDT framework.

RDT clustering: Formalism

RDT returns conditional probabilities, $p(C_k|s_i)$, of cluster k, selected from a set of n clusters $\mathbf{C} = \{C_1, \cdots, C_n\}$, given observation i selected from the set of N observations $\mathbf{s} = \{s_1, \cdots, s_N\}$.

As discussed above, the rate is the average amount of information needed to specify an observation s_i within the set of clusters $\mathbf{C}$, computed as the mutual information between the set of clusters $\mathbf{C}$ and observations $\mathbf{s}$ as follows

$$I(\mathbf{C}, \mathbf{s}) = \sum_{k=1}^{n} \sum_{i=1}^{N} p(C_k|s_i)p(s_i) \log \frac{p(C_k|s_i)}{p(C_k)}. \tag{2.87}$$

By minimizing the rate, we maximize the compression. However, if the minimization of the rate is not bounded, then we will over-compress and any information that clusters contain on the observations will vanish (i.e. $I(\mathbf{C}, \mathbf{s}) \to 0$). Our minimization of the rate thus needs to be informed, or constrained, by another quantity. This quantity is the mean distortion among the observations within the set of clusters $\mathbf{C}$, $\langle D(\mathbf{C}, \mathbf{s}) \rangle$, defined as the average of the pairwise distortions between all pairs of observations in $\mathbf{s}$ [281, 284]

$$\langle D(\mathbf{C}, \mathbf{s}) \rangle = \sum_{k=1}^{n} p(C_k) \sum_{i,j=1}^{N} p(s_i|C_k)p(s_j|C_k)d_{ij}. \tag{2.88}$$

The pairwise distortion between two observations, d_{ij}, is a measure of the dissimilarity between them, and its choice is problem-specific. For example, the dissimilarity between two probability (mass or density) functions can be measured as the area between their respective cumulative distribution functions, which corresponds to a metric known as the Kantorovich distance [285].

To obtain the minimum rate of information – constrained by the mean distortion – we define an objective function to be minimized

$$I(\mathbf{C}, \mathbf{s}) + \beta \langle D(\mathbf{C}, \mathbf{s}) \rangle \tag{2.89}$$

where β is a Lagrange multiplier that controls which term is favored in the minimization. A small value of β favors minimization of the rate over distortion and, thus, high compression. Conversely, large values of β will cause the minimization to favor the distortion, returning a less compressed clustering result in which the clusters contain a relatively large amount of information about the set of observations.

The formal solution to the minimization of Eq. (2.89) with respect to the conditional probabilities $p(C_k|s_i)$ is a Boltzmann-like distribution [283]

$$p(C_k|s_i) = \frac{p(C_k)}{Z(s_i,\beta)} \exp\left[-\beta \sum_{j=1}^{N} p(s_i|C_k) d_{ij}\right]. \tag{2.90}$$

The conditionals $p(s_i|C_k)$ of Eq. (2.90) are obtained, in turn, using Bayes' formula; $p(C_k)$ is the marginal probability of the cluster C_k

$$p(C_k) = \sum_{i=1}^{N} p(s_i) p(C_k|s_i) \tag{2.91}$$

and $Z(s_i,\beta)$ of Eq. (2.90) is the normalization [283]

$$Z(s_i,\beta) = \sum_{k=1}^{n} p(C_k) \exp\left[-\beta \sum_{j=1}^{N} p(s_i|C_k) d_{ij}\right]. \tag{2.92}$$

As can be seen from Eqns. (2.90)-(2.91), the conditional probabilities $p(C_k|s_i)$ and the marginal probabilities $p(C_k)$ must be self-consistent. This self-consistency is exploited to obtain numerical solutions to the variational problem through an iterative procedure (Blahut-Arimoto algorithm) [286, 287]. The procedure begins by randomly initializing each of the conditionals $p(C_k|s_i)$ followed by normalization over $\mathbf{C}$ and continuing with iterative calculations of the marginals and conditionals, via Eqns. (2.90)-(2.91), until the objective function, Eq. (2.89), has converged. In practice, this algo-

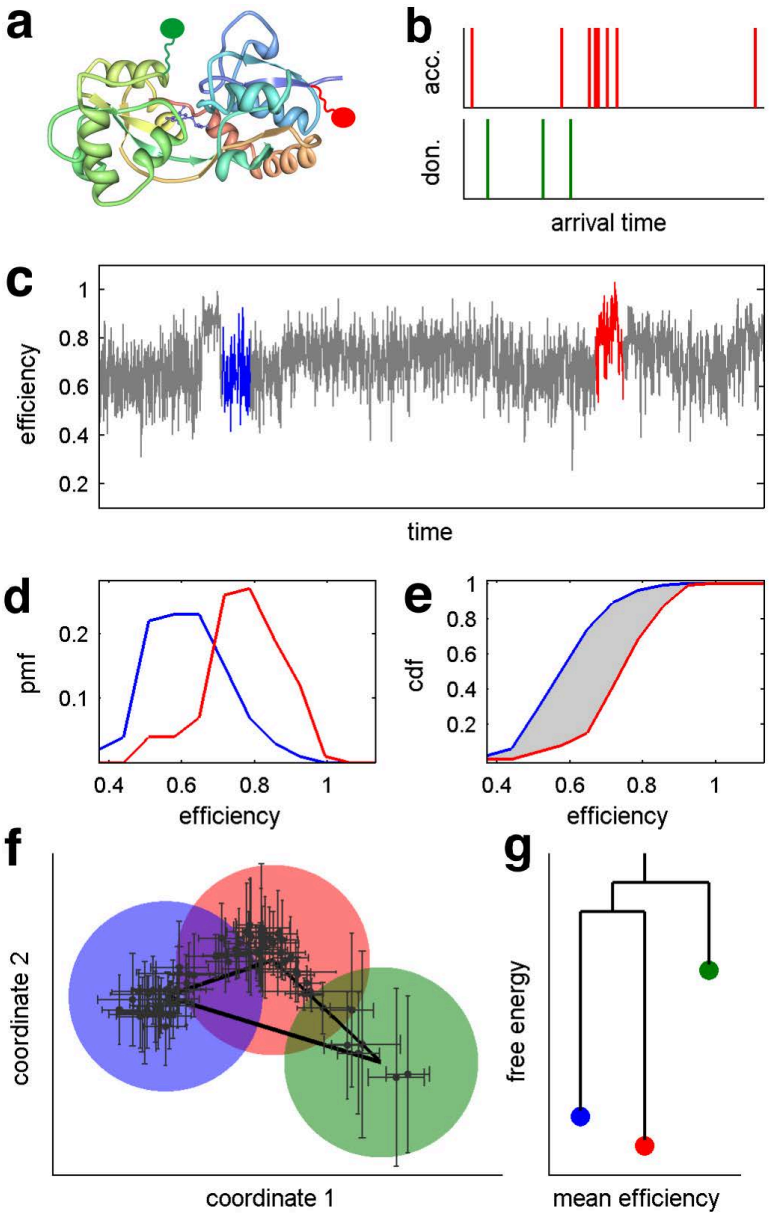

Fig. 2.16. **A soft clustering algorithm based on RDT is used to determine states from smFRET trajectories**. **a)** A crystal structure of the AMPA ABD. The green and red spheres represent the donor and acceptor fluorophores, respectively. **b)** Detection of photons emitted in an smFRET experiment. **c)** An experimental smFRET trajectory obtained by binning the data in (b). **d)** Probability mass functions (pmfs) of the blue and red segments highlighted in (c). **e)** Cumulative distribution function (cdfs) of the highlighted segments in (c). The shaded area represents the Kantorovich distance. **f)** Visual representation of clusters in (c) based on multidimensional scaling. **g)** Transition disconnectivity graph (TRDG) resulting from the trajectory in (c). See details in text and Ref. [284].

rithm may converge to a local, rather than the global, minimum requiring that it be re-initialized multiple times with random seeds.

RDT analysis on sample data

Both β and the number of clusters are inputs to RDT. We can use the distortion and rate of RDT as model selection tools to choose these input.

To do so, we may first obtain an appropriate estimate of the amount of distortion arising from errors in the data. The errors must be treated carefully on a case-by-case basis; this is detailed for the case of smFRET data in Ref. [284].

Once this estimate is obtained, it is used as a benchmark against which to compare the distortion, $\langle D(\mathbf{C}, \mathbf{s}) \rangle$, arising from other models (i.e. solutions obtained with different numbers of clusters and/or values of β). Any model having $\langle D(\mathbf{C}, \mathbf{s}) \rangle$ less than this benchmark is retained for model complexity comparison.

Model complexity is then assessed by comparing the values of the mutual information $I(\mathbf{C}, \mathbf{s})$ arising from each model. Specifically, the model satisfying the distortion criterion having the smallest value of $I(\mathbf{C}, \mathbf{s})$ is the least complex model retained for further analysis. Details are provided in Ref. [284].

Fig. (2.16) sketches key steps in using RDT to identify states (i.e. clusters) of an smFRET trajectory for a protein domain (AMPA ABD) that we will discuss shortly. Briefly, we segment the trajectory into "elements". Our goal is to cluster these elements shown as short stretches of data in Fig. (2.16c). Fig. (2.16d) shows the probability mass function (pmf) of these elements; while Fig. (2.16e) illustrates the Kantorovich distance that we use in our distortion. We subsequently use multidimensional scaling [288] to map clusters into two dimensions (Fig. (2.16f)). This gives us visual insight into cluster overlap as well as the breadth of the conditional distributions $p(C_k|s_i)$ detailed in Ref. [284].

Fig. (2.16g) captures another useful representation of the conformational space beyond clusters: the free energy landscape. A free energy landscape depicts the conformational motions of proteins (or their domains) such as the AMPA ABD, that we will discuss shortly, as diffusion on a multidimensional free energy surface, with conformational states represented as energy basins on the landscape and the transition times being characterized by the heights of the energy barriers among the set of basins [289, 290]. Because conformational motion in smFRET experiments is projected on a 1-dimensional coordinate, we approximate the free energy landscape with a transition disconnectivity graph (TRDG) [289, 290]. A simple, 3-state TRDG is shown in Fig. (2.16g). The nodes represent relative free energies of the conformational states while the horizontal lines represent free energies at the barriers for transitions among the conformational states. Briefly, the TRDG is constructed by first identifying the slowest transition (i.e. the highest energy barrier) between two disjoint sets in the network [284]. Each subsequently faster transition between disjoint sets is then identified resulting in the branching structure of the TRDG detailed in Refs. [284, 289, 290].

Application of RDT to single molecule time-series

We apply RDT clustering to extract kinetic models from smFRET time traces [284, 291, 292]. In the example we now discuss, smFRET monitors the conformational dynamics of binding domains of a single AMPA receptor, an agonist-mediated ion channel prevalent in the central nervous system. In the context of RDT, each smFRET trajectory is viewed as a noisy message received from the source, i.e. the underlying conformational network. RDT is used to decode the message sent by the source and to classify intervals along the time trace into underlying clusters (conformational states).

AMPA receptors are among the most abundant ion-channel proteins in the central nervous system [291]. They are comprised of extracellular N-termini and agonist binding domains (ABDs) [involved in ion channel activation [291]], transmembrane domains and intracellular C-terminal domains. They are agonist-mediated ion channels and interaction of the ABDs with neurotransmitters – such as glutamate – induces conformational motion in the protein which, in turn, triggers the activation of ion transmission through the cellular membrane.

X-ray crystal structures [293] show that ABD has two lobes that form a cleft containing the agonist-binding site [284]. X-ray studies also suggest that the degree of cleft closure controls the activation of the ion channel [293], with a closed cleft corresponding to an activated channel, but exceptions to this conjecture exist [294]. Molecular dynamics simulation [295,296] further suggest that ABD is capable of conformational motion even when bound to the full agonist glutamate, and that conformational fluctuations are increased in the absence of a bound agonist. What is more, smFRET studies of the apo and various agonist-bound forms of the AMPA ABD support this theoretical result [291] and, together, suggest that the activation mechanism is more complex. In order to gain deeper insight into this allosteric mechanism, we used RDT clustering along with time series segmentation and energy landscape theory to analyze the data [284].

Here we discuss the results of RDT clustering applied to smFRET trajectories of the AMPA ABD while bound to three different agonists: a full agonist, a partial agonist, and an antagonist [284]. The properties extracted from each of these systems, including population distributions and TRDGs, are shown in Fig. (2.17) [284]. Parameters estimated from the clustering results, including mean efficiencies, occupation probabilities, escape times and free energies of the basins are shown in Table (2.1) [284].

As shown by the transition networks, TRDGs, and state distributions in Figs. (2.17a)-(2.17c), the model selection process results in the assignment of 4, 5, and 6 states for the ABDs bound to the full agonist, partial agonist, and antagonist, respectively.

As shown in Fig. (2.17), and Table (2.1) [284], the most dominant state when the ABD is glutamate-bound has 74% occupation probability and a mean efficiency of 0.85, which corresponds to a relatively short interdye distance of $\sim$ 38 Å, in comparison to the apo form of the ABD [291]. Other states have smaller occupation probabilities ($<$ 10%) which, along with the relatively slow escape times from the states, suggest that, although conformational dynamics are observed, the glutamate-bound ABD possesses a relatively stable and closed ABD.

By contrast, the most populated state in the partial agonist-bound ABD system shown in Fig. (2.17b) has a smaller mean efficiency of 0.74 and a smaller occupation probability of 52%, indicating a longer interdye distance and a less stable conformation. Furthermore, the TRDG indicates a smaller transition barrier out of many states in the network which, along with the increase in the number of significantly populated states, suggests that ABD is more active when bound to the partial agonist.

Lastly, the results of the antagonist-bound ABD returned six states displaying a broader interdye range and a larger relative occupation at lower efficiencies. The most populous state (41%), however, has a high mean efficiency of 0.88, suggesting that the channel should be activated while occupying this closed cleft conformation as defined in Ref. [292]. Inspection of the TRDG shows relatively smaller ($\sim$ 1 kcal/mol) transition barriers and we found escape times for all states that are relatively faster (200-500 ms), suggesting a conformationally active ABD relative to the full and partial agonist-bound systems. It is this fast and frequent conformational motion that is the source of the ion channel's lack of activation.

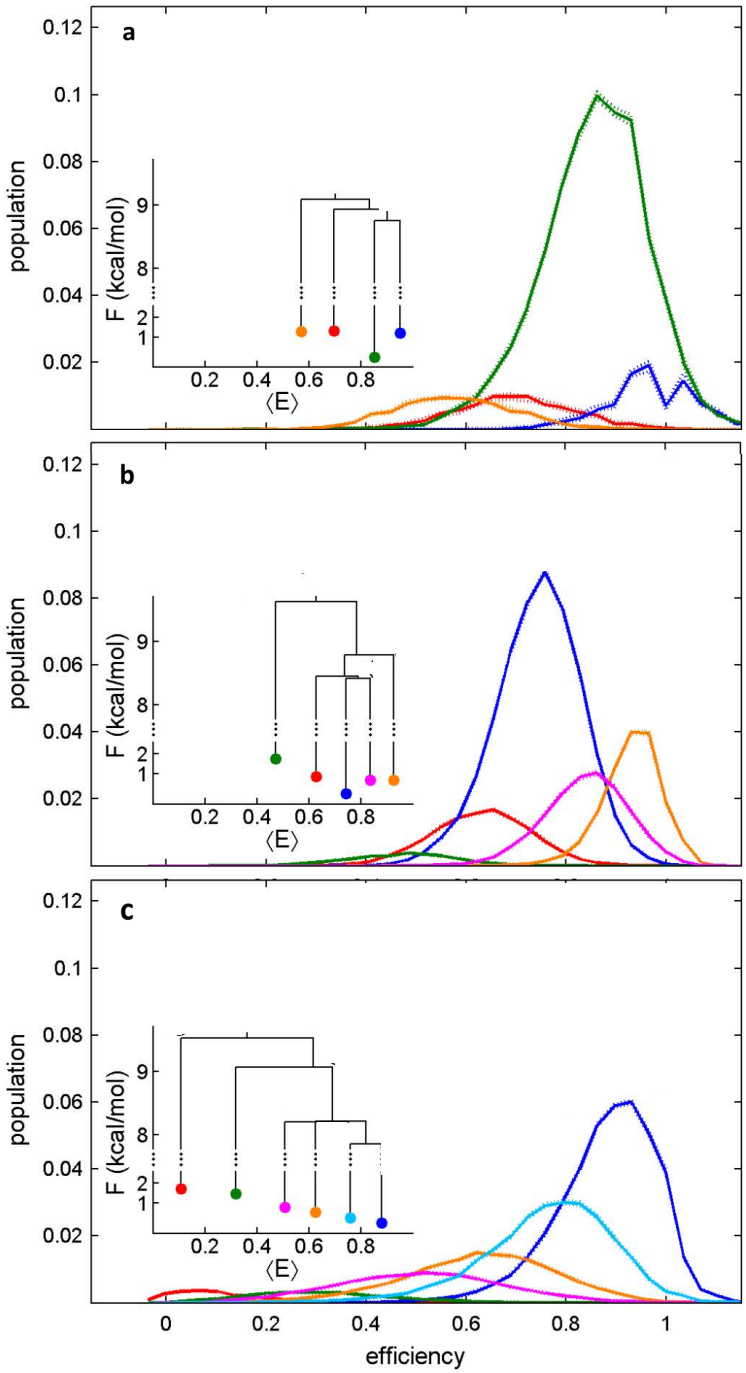

Fig. 2.17. **RDT clustering reveals differences in conformational dynamics for the AMPA ABD.** State distributions and TRDGs are given for the full agonist-bound ABD (a); the partial agonist-bound ABD (b); and the antagonist-bound ABD (c). $\langle E \rangle$ denotes the mean efficiency. See main body and Ref. [284] for details.

Table 2.1.

RDT clustering returns state properties for the AMPA ABDs. These properties include mean efficiencies ($\langle E \rangle$), occupation probabilities ($p(S_k)$), free energies (F_i), and escape times with 95% confidence intervals, for the full agonist (glutamate), the partial agonist (nitrowillardiine), and the antagonist (UBP282). See main body and ref [284] for details. Sapporo(Nick: Removed hypothesis test results from table.)

$\langle E \rangle$	$p(S_k)$ (%)	F_i (kcal/mol)	escape time (ms)
		Full Agonist	
0.97	10	1.21	308 (220,425)
0.85	74	0	674 (589,753)
0.75	8	1.31	185 (145,220)
0.64	8	1.29	310 (240,384)
		Partial Agonist	
0.93	16	0.68	807 (690,928)
0.84	17	0.68	288 (260,300)
0.74	52	0	664 (618,698)
0.66	12	0.88	290 (260,308)
0.47	3	1.75	502 (388,634)
		Antagonist	
0.88	41	0	490 (458,519)
0.76	27	0.26	223 (207,236)
0.62	16	0.56	207 (193,217)
0.51	11	0.78	220 (203,236)
0.32	3	1.47	250 (217,283)
0.11	2	1.71	512 (420,619)

In summary, RDT clustering applied to these three systems provides broader insight into the activation mechanism of the AMPA receptor [284]. The antagonist-bound ABD exhibits fast conformational fluctuations and a relatively unstable structure that explores a broad range of interdye distances while the most stable conformation of the partial agonist-bound ABD displays a relatively large interdye distance, indicating a weaker, and/or sterically distorted structure. By comparison, the full agonist-bound ABD displays a relatively stable and static structure with a small interdye distance, suggesting a strong and stable interaction of the full agonist with the ABD. It is the ability of the full agonist to hold the cleft of the ABD closed in a stable manner that causes the full activation of, i.e. the maximum ionic current through, the ion channel.

2.6.2 Variations of RDT: Information-based clustering

Similar in spirit to RDT is a general method known as information-based clustering that has been used on gene expression data [281]. Information-based clustering uses a similarity measure – rather than a distortion measure – to quantify how alike elements are [281].

However, unlike in RDT, the quantity that plays the role of RDT's distortion in information-based clustering – a quantity termed "multi-information"– is a multidimensional mutual information.

An advantage of information-based clustering is that multidimensional relationships among the data to be clustered are naturally incorporated into the clustering algorithm. The objective function to be maximized is

$$\langle S(\mathbf{C}, \mathbf{s}) \rangle - TI(\mathbf{C}, \mathbf{s}) \tag{2.93}$$

where $\mathbf{C}$ are again the set of clusters and $\mathbf{s}$ the set of observations and T is a "trade-off" parameter.

In Eq. (2.93), $\langle S(\mathbf{C}, \mathbf{s}) \rangle$ is the average multidimensional similarity among the set of observations within the set of clusters [281], $I(\mathbf{C}, \mathbf{s})$ is the mutual information between the set of clusters and the set of observations.

2.6.3 Variations of RDT: The information bottleneck method

There exists another variation of RDT proposed by Tishby *et al.* [282], termed the information bottleneck (IB) method, which focuses on how well the compressed description of the data, i.e. the clusters or states, can predict the outcome of another observation, say $\mathbf{u}$. It is similar to information-based clustering however its focus is on predicting the outcome of another variable and the "distortion" term is given by the mutual information between the clusters and $\mathbf{u}$.

In other words, the information contained in $\mathbf{s}$ is squeezed (compressed) through the "bottleneck" of clusters $\mathbf{C}$ which is then used to explain $\mathbf{u}$. The advantage with IB is that there is no need for a problem-specific distortion.

Just like in RDT, minimizing the mutual information between $\mathbf{s}$ and $\mathbf{C}$ generates broad overlapping clusters. However maximizing the mutual information between $\mathbf{u}$ and $\mathbf{C}$ tends to create sharper clusters.

Mathematically the objective function in IB to be maximized is

$$I(\mathbf{C}, \mathbf{s}) - \beta I(\mathbf{C}, \mathbf{u}) \tag{2.94}$$

where, as before, $I(\mathbf{C},\mathbf{s}) = \sum_{k,i} p(C_k, s_i)\log(p(C_k|s_i)/p(C_k))$ [likewise for $I(\mathbf{C},\mathbf{u})$] and β is a trade-off parameter. The maximization of Eq. (2.94) yields [282]

$$\begin{aligned} p(C_k|s_i) &= \frac{p(C_k)}{Z(s_i,\beta)}\exp\left[-\beta\sum_{j=1}^{M} p(u_j|s_i)\log\frac{p(u_j|s_i)}{p(u_j|C_k)}\right] \\ &\equiv \frac{p(C_k)}{Z(s_i,\beta)}\exp\left[-\beta D_{\mathrm{KL}}[p(u|s_i)\|p(u|C_k)]\right] \end{aligned} \tag{2.95}$$

where $D_{\mathrm{KL}}[p(u|s_i)\|p(u|C_k)]$ is the KL divergence, and

$$Z(s_i,\beta) = \sum_k p(C_k)\exp[-\beta D_{\mathrm{KL}}[p(u|s_i)\|p(u|C_k)]] \tag{2.96}$$

is the normalization. When comparing Eq. (2.95) with the formal solution of the conventional RDT, Eq. (2.90), we see that the KL divergence, $D_{\mathrm{KL}}[p(u|s_i)\|p(u|C_k)]$, serves as an effective distortion function in the IB framework. This means that by minimizing the distortion $D_{\mathrm{KL}}[p(u|s_i)\|p(u|C_k)]$, one obtains a compression of $\mathbf{s}$ (through $\mathbf{C}$) that preserves as much as possible the information provided by the relevant observable $\mathbf{u}$.

For illustrative purposes, we consider two special limiting values for the trade-off parameter: $\beta \to 0$ and $\beta \to \infty$. As $\beta \to 0$, we have $Z(s_i, \beta \to 0) = \sum_k p(C_k) = 1$. Eq. (2.95) then implies $p(C_k|s_i) = p(C_k)$. Using Bayes' rule, one then obtains $p(s_i|C_k) = p(s_i)$. This means that the probability to find s_i in a cluster is the same for all clusters, implying that all clusters overlap or, effectively, we have just one cluster.

The opposite extreme, $\beta \to \infty$, is the limit of hard-clustering where, to leading order, the normalization becomes $Z(s_i,\beta) \to p(C_{k*})\exp[-\beta D_{\mathrm{KL}}[p(u|s_i)\|p(u|C_{k*})]]$, where C_{k*} is the cluster with $p(u_j|C_{k*}) = p(u_j|s_i)$. Substituting this approximate normalization back into Eq. (2.95), we arrive at: $p(C_k|s_i) = 1$ if $p(u_j|s_i) = p(u_j|C_k)$,

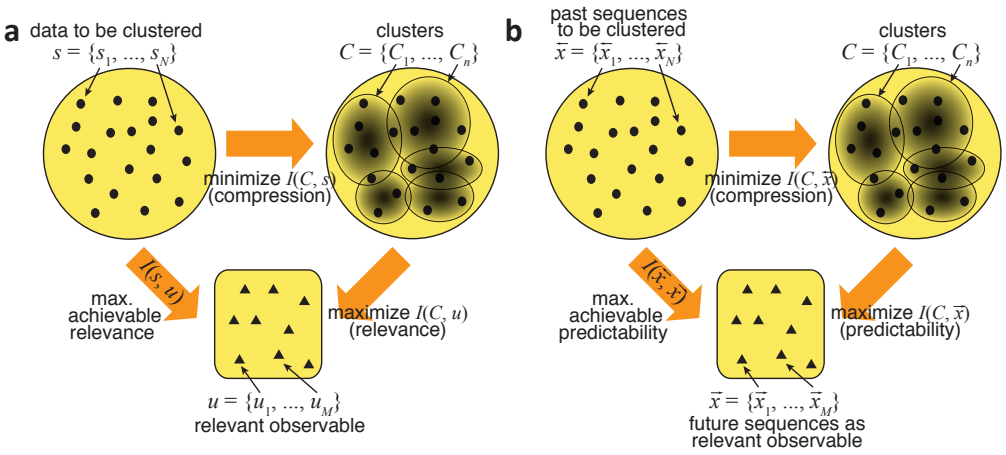

Fig. 2.18. **The IB method can be used to construct dynamical models. a)** The IB method starts from the data to be clustered $\mathbf{s}$ (top left), clustering then compresses the information contained in $\mathbf{s}$ by minimizing the rate $I(\mathbf{C}, \mathbf{s})$ (from top left to top right). Instead of introducing an *a priori* distortion measure, the IB compression maximizes $I(\mathbf{C}, \mathbf{u})$ quantifying how well another observable, $\mathbf{u}$, is predicted (from top right to bottom). The maximum achievable "relevance", predicting $\mathbf{u}$ from $\mathbf{s}$, is given by $I(\mathbf{s}, \mathbf{u})$. **b)** To construct a predictive dynamical model from time series data, we may define past sequences (top left) as the data to be clustered $\mathbf{s}$ and future sequences (bottom) as the relevant observables $\mathbf{u}$.

and $p(C_k|s_i) = 0$ otherwise. Since the conditional probability is either one or zero, the limit $\beta \to \infty$ coincides precisely with the hard clustering case.

Information bottleneck method: Application to dynamical state-space networks

The IB framework has been applied directly to time series data to construct dynamical state-space network models [62, 63]. Here we briefly describe important features of the IB-based state-space network construction detailed elsewhere [50, 62].

Intuitively, state-space networks derived from the IB construction, must preserve maximum information relevant to predicting the future outcome of the time series. More precisely, they must be minimally complex but, simultaneously, most predictive [50, 282, 283].

As an illustration, we consider a time series $\mathbf{x}$ sampled at discrete times. We have both past, $\overleftarrow{\mathbf{x}} = \{\overleftarrow{\mathbf{x_1}}, \cdots, \overleftarrow{\mathbf{x_N}}\}$, and future, $\overrightarrow{\mathbf{x}} = \{\overrightarrow{\mathbf{x_1}}, \cdots, \overrightarrow{\mathbf{x_M}}\}$, sequences of different total length. We have bolded the elements of $\overleftarrow{\mathbf{x}}$ and $\overrightarrow{\mathbf{x}}$ because each element can itself be a vector of some length, say L and L' respectively.

To construct a minimal state-space network with maximal predictability from the IB framework, we identify the data to be clustered $\mathbf{s}$ as the past sequences $\overleftarrow{\mathbf{x}}$ and the relevant observable $\mathbf{u}$ as the future sequences $\overrightarrow{\mathbf{x}}$ (see Fig. (2.18a)-(2.18b)). The clusters $\mathbf{C}$ obtained in Fig. (2.18b) represent constructed network states. In principle, the length of the past and future sequences L and L' should be chosen to be long enough as compared to all dynamical correlations in the time series. However, this may cause some practical sampling problems if L and L' are too large. These problems are addressed using multiscale wavelet based CM described in detail in Refs. [50, 62, 63].

Multiscale state-space networks developed from IB methods may capture dynamical correlations of conformational fluctuations covering a wide range of timescales (from millisecond to second) [62,63]. Moreover, the topographical features of the networks constructed, including the number connections among the states and the heterogeneities in the transition probabilities among the states, depend on the timescale of observation, namely, the longer the timescale, the simpler and more random the underlying state-space network becomes. These insights provides us with a network topography perspective to understand dynamical transitions from anomalous to normal diffusion [62, 63].

We end this brief section by mentioning that clustering past sequences to form state-space networks in which the relevant variable are future sequences was proposed separately by Crutchfield *et al.* [297,298], and termed computational mechanics (CM). The states and resulting state-space network are called causal states and epsilon machine, respectively. We refer the interested readers to an excellent review of CM [299].

2.7 Final Thoughts on Data Analysis and Information Theory

We have previously seen how information theory can be used in deconvolution, model selection and clustering. In this purely theoretical section, we discuss efforts to use information theory in experimental design and end with some considerations on the broader applicability of information theory.

2.7.1 Information theory in experimental design

Just as information theory can be used in model selection after an experiment has been performed, it may also be used to suggest an experimental design, labeled ξ.

The goal, in this so far theoretical endeavor, is to find a design that optimizes information gain [300–302]. For instance, a choice of design may involve tuning data collection times, bin sizes, choice of variables under observation and sample sizes [303,304]. Fig. (2.19) illustrates – for a concrete example we will discuss shortly – how the number of observations can be treated as a design variable and how information gained grows as we tune this variable (repeat trials).

To quantify the information gained, we first consider the expected utility, $U(\xi)$ – depending on the experimental design ξ – defined as the mutual information between the data, $\mathbf{y}$, and model, $\boldsymbol{\theta}$ [300]

$$U(\xi) \equiv I(\mathbf{y}, \boldsymbol{\theta}|\xi) = \int d\boldsymbol{\theta} d\mathbf{y} p(\mathbf{y}, \boldsymbol{\theta}|\xi) \log \left(\frac{p(\mathbf{y}, \boldsymbol{\theta}|\xi)}{p(\mathbf{y}|\xi) p(\boldsymbol{\theta}|\xi)} \right) \tag{2.97}$$

which we must now maximize with respect to our choice of experiment, ξ.

More concretely, the choice of experiments dictates the mathematical form for $p(\mathbf{y}|\xi)$ and $p(\mathbf{y}, \boldsymbol{\theta}|\xi)$. Thus maximizing with respect to ξ may imply comparing different mathematical forms dictated by the experimental design for $p(\mathbf{y}|\xi)$ and $p(\mathbf{y}, \boldsymbol{\theta}|\xi)$.

Our utility function, $U(\xi)$, is simply the difference in Shannon information before and after the data $\mathbf{y}$ was used to inform the model

$$I(\mathbf{y}, \boldsymbol{\theta}|\xi) = I(\boldsymbol{\theta}|\mathbf{y}, \xi) - I(\boldsymbol{\theta}|\xi) \tag{2.98}$$

where

$$\begin{aligned} I(\boldsymbol{\theta}|\mathbf{y}, \xi) &\equiv \int d\boldsymbol{\theta} d\mathbf{y} p(\mathbf{y}, \boldsymbol{\theta}|\xi) \log p(\boldsymbol{\theta}|\mathbf{y}, \xi) \\ I(\boldsymbol{\theta}|\xi) &\equiv \int d\boldsymbol{\theta} d\mathbf{y} p(\mathbf{y}, \boldsymbol{\theta}|\xi) \log p(\boldsymbol{\theta}|\xi). \end{aligned} \tag{2.99}$$

As an illustration of this formalism, we can now quantify whether future experiments – repeated trials yielding data $\mathbf{y}'$ – appreciably change the expected information gain. We begin by iterating Eq. (2.98) and write

$$I(\mathbf{y}', \boldsymbol{\theta}|\mathbf{y}, \xi) = I(\boldsymbol{\theta}|\mathbf{y}', \mathbf{y}, \xi) - I(\boldsymbol{\theta}|\mathbf{y}, \xi). \tag{2.100}$$

As a concrete example, suppose we monitor photon arrival times in continuous time. We write the likelihood, i.e. the probability of observing a sequence of photon arrivals with arrival times $\mathbf{t} = \{t_1, t_2, \cdots, t_N\}$,

$$p(\mathbf{y} = \mathbf{t}|\boldsymbol{\theta} = r, \xi) = re^{-rt_1} \times re^{-rt_2} \times \cdots \times re^{-rt_N} \tag{2.101}$$

where r is the rate of arrival and ξ is implicitly specified by our choice of experiment (and thus by the form of our likelihood). For sake of concreteness, we take a simple exponential prior distribution over r, $p(r|\xi) = \phi e^{-r\phi}$ where ϕ is a hyperparameter. Now, our joint distribution, $p(\mathbf{t}, r|\xi) = p(\mathbf{t}|r, \xi)p(r|\xi)$, as well as our marginal distribution over data, $p(\mathbf{t}|\xi) = \int drp(\mathbf{t}, r|\xi)$, are fully specified. We tune the design here by selecting N.

The expected information gained can now be explicitly calculated from Eq. (2.100). The integrals are over all allowed r and arrival times. All, but the last time, t_N, are considered as $\mathbf{y}$. The last time, t_N, is $\mathbf{y}'$.

The expected information gained, $I(\mathbf{y}', \boldsymbol{\theta}|\mathbf{y}, \xi)$, increases monotonically but sub-linearly with the number of trials and, for our specific example, independently of ϕ. See Fig. (2.19). The monotonicity is expected because we have averaged over all possible outcomes (i.e. photon arrival times). The sub-linearity however quantifies that future experiments result in diminishing returns [305]. Put differently, insofar

as the expected information gained allows to build a predictive model, most of the information gathered has little predictive value.

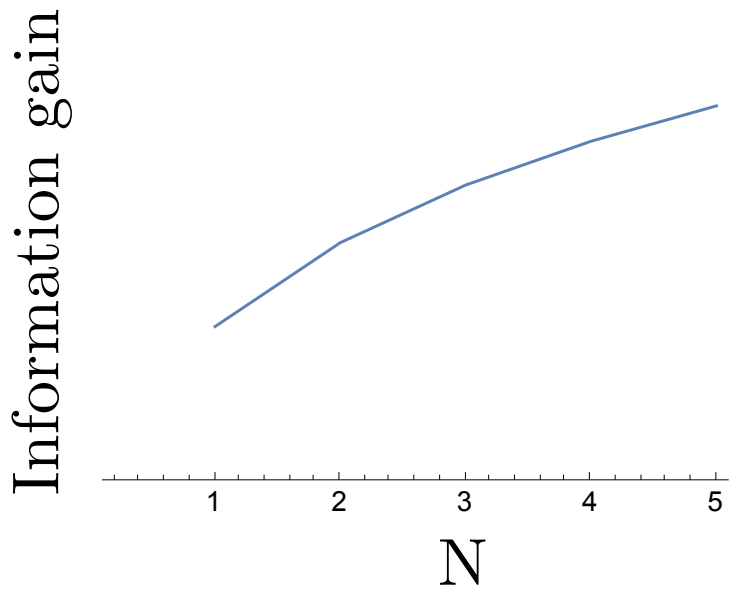

Fig. 2.19. **Diminishing returns: most data collected from additional experiments does not result in information gain.** The expected information gained, Eq. (2.100), grows sub-linearly with the number of photon arrival measurements.

2.7.2 Predictive information and model complexity

The result from Fig. (2.19) in the previous section suggested that most repeated experiments yield little information gain.

Here we want to quantify this concept by calculating the predictive value of previously collected data on future data. We consider the mutual information linking past, i.e. previously collected, data, $\mathbf{y}_p$, and future, $\mathbf{y}_f$, data [300, 305]

$$I_{pred}(\mathbf{y}_f, \mathbf{y}_p|\xi) \equiv \int d\mathbf{y}_f d\mathbf{y}_p p(\mathbf{y}_f, \mathbf{y}_p|\xi) \log\left(\frac{p(\mathbf{y}_f, \mathbf{y}_p|\xi)}{p(\mathbf{y}_f|\xi)p(\mathbf{y}_p|\xi)}\right). \tag{2.102}$$

We wish to simplify Eq. (2.102) and derive $I_{pred}(\mathbf{y}_f, \mathbf{y}_p|\xi)$'s explicit dependence on the total number of data points collected in the past, N_p. We call N_f the total number of data points collected in the future.

To simplify Eq. (2.102), we re-write Eq. (2.102) as follows

$$\begin{aligned} I_{pred}(\mathbf{y}_f, \mathbf{y}_p|\xi) &= I_{joint} - I_{future} - I_{past} \\ &= \int d\mathbf{y}_f d\mathbf{y}_p p(\mathbf{y}_f, \mathbf{y}_p|\xi) \log p(\mathbf{y}_f, \mathbf{y}_p|\xi) \\ &- \int d\mathbf{y}_f d\mathbf{y}_p p(\mathbf{y}_f, \mathbf{y}_p|\xi) \log p(\mathbf{y}_f|\xi) - \int d\mathbf{y}_f d\mathbf{y}_p p(\mathbf{y}_f, \mathbf{y}_p|\xi) \log p(\mathbf{y}_p|\xi) \end{aligned} \tag{2.103}$$

and express $p(\mathbf{y}_f, \mathbf{y}_p|\xi)$ as

$$p(\mathbf{y}_f, \mathbf{y}_p|\xi) = \int d^{K}\grave{} \; p(\mathbf{y}_f, \mathbf{y}_p, \boldsymbol{\theta}|\xi) \tag{2.104}$$

where K enumerates model parameters. We can similarly re-write both $p(\mathbf{y}_f|\xi)$ and $p(\mathbf{y}_p|\xi)$. Next, we simplify $p(\mathbf{y}_f, \mathbf{y}_p|\xi)$ by assuming that individual data points, both past and future, are sufficiently independent. As the number of total data point, $N_f + N_p$, grows we invoke Laplace's method – and using the notation for the rescaled logarithm of the likelihood f introduced earlier, Eq. (2.5) – to simplify $p(\mathbf{y}_f, \mathbf{y}_p|\xi)$. This yields

$$\begin{aligned} p(\mathbf{y}_f, \mathbf{y}_p|\xi) &= \int d^K\theta \; p(\mathbf{y}_f, \mathbf{y}_p, \boldsymbol{\theta}|\xi) \equiv \int d^K\theta \; e^{(N_f+N_p)\log f(\mathbf{y}_f, \mathbf{y}_p, \boldsymbol{\theta}|\xi)} \\ &\sim \frac{1}{\sqrt{det\left((N_f+N_p)[-(\log f(\mathbf{y}_f, \mathbf{y}_p, \boldsymbol{\theta}^*|\xi))\right)'']}} \times e^{(N_f+N_p)\log f(\mathbf{y}_f, \mathbf{y}_p, \boldsymbol{\theta}^*|\xi)} \\ &\propto \frac{1}{(N_f+N_p)^{K/2}} e^{(N_f+N_p)\log f(\mathbf{y}_f, \mathbf{y}_p, \boldsymbol{\theta}^*|\xi)} \end{aligned} \tag{2.105}$$

where $\boldsymbol{\theta}^*$ is the value of $\boldsymbol{\theta}$ maximizing the integrand. Similarly, $p(\mathbf{y}_f|\xi)$ and $p(\mathbf{y}_p|\xi)$ can be re-written as follows

$$p(\mathbf{y}_f|\xi) \propto \frac{1}{N_f^{K/2}} e^{N_f \log f(\mathbf{y}_f, \boldsymbol{\theta}_f^*|\xi)} \tag{2.106}$$

$$p(\mathbf{y}_p|\xi) \propto \frac{1}{N_p^{K/2}} e^{N_p \log f(\mathbf{y}_p, \boldsymbol{\theta}_p^*|\xi)} \tag{2.107}$$

where $\boldsymbol{\theta}_f^*$ and $\boldsymbol{\theta}_p^*$ are those parameter values that maximize their respective integrands. But, for sufficiently large enough N_f and N_p, $\log f(\mathbf{y}_f, \boldsymbol{\theta}_f^*|\xi)$ and $\log f(\mathbf{y}_p, \boldsymbol{\theta}_p^*|\xi)$ are both well-approximated by $\log f(\mathbf{y}_f, \mathbf{y}_p, \boldsymbol{\theta}^*|\xi)$. This is true to the extent that $\mathbf{y}_f$ and $\mathbf{y}_p$ are typical. Thus [305]

$$p(\mathbf{y}_f|\xi) \propto \frac{1}{N_f^{K/2}} e^{N_f \log f(\mathbf{y}_f, \mathbf{y}_p, \boldsymbol{\theta}^*|\xi)} \tag{2.108}$$

$$p(\mathbf{y}_p|\xi) \propto \frac{1}{N_p^{K/2}} e^{N_f \log f(\mathbf{y}_f, \mathbf{y}_p, \boldsymbol{\theta}^*|\xi)}. \tag{2.109}$$

Inserting these simplified probabilities, Eq. (2.105), (2.108) and (2.109), into Eq. (2.103) yields an extensive part that scales with the number of data points (first two lines) and, to next order, a portion scaling with the logarithm of the number of data points (third line) [305]

$$\begin{aligned} I_{pred}(\mathbf{y}_f, \mathbf{y}_p|\xi) \sim\ & (N_f + N_p) \int d\mathbf{y}_f d\mathbf{y}_p p(\mathbf{y}_f, \mathbf{y}_p|\xi) \log f(\mathbf{y}_f, \mathbf{y}_p, \boldsymbol{\theta}^*|\xi) \\ & - N_p \int d\mathbf{y}_f d\mathbf{y}_p p(\mathbf{y}_f, \mathbf{y}_p|\xi) \log f(\mathbf{y}_f, \mathbf{y}_p, \boldsymbol{\theta}^*|\xi) \\ & - N_f \int d\mathbf{y}_f d\mathbf{y}_p p(\mathbf{y}_f, \mathbf{y}_p|\xi) \log f(\mathbf{y}_f, \mathbf{y}_p, \boldsymbol{\theta}^*|\xi) \\ & - \frac{K}{2} \log(N_f + N_p) + \frac{K}{2} \log N_f + \frac{K}{2} \log N_p + \mathcal{O}(N_f^0) + \mathcal{O}(N_p^0). \end{aligned} \tag{2.110}$$

The extensive portion of I_{pred}, Eq.(2.110), cancels to leading order. This directly implies that the vast majority of data collected provides no predictive information. If we then ask what the past data collected tells us about the entirety of future observations ($N_f \to \infty$), upon simplifying Eq. (2.110), we find

$$I_{pred} = \frac{K}{2} \log N_p. \tag{2.111}$$

In other words, the predictive information grows logarithmically with the data collected and linearly with K.

Asymptotically, the predictive information is directly related to features of the model (in this case, the number of parameters) drawn from the data. In addition, Eq. (2.111) provides an interpretation to the penalty term of the BIC [73] as twice the predictive information.

2.7.3 The Shore & Johnson axioms

While we've discussed the Shannon entropy in the context of its information theoretic interpretation, the SJ axioms provide a complementary way to understand the central role of information theory in model inference.

The key mathematical steps in deriving the Shannon entropy from the SJ axioms concretely highlight what assumptions are implicit when using $H = -\sum_i p_i \log p_i$ that go beyond the illustration of the kangaroo example with eye-color and handedness. Conversely, they clarify which assumptions must be violated in rejecting $H = -\sum_i p_i \log p_i$ in inferring models for $\{p_i\}$ [151, 306].

Briefly, SJ wanted to devise a prescription to infer a probability distribution, $\{p_i\}$. Thus they constructed an objective function which, when maximized, would guarantee that inferences drawn from their model – the probability distribution, $\{p_i^*\}$, which

maximizes their objective function – would satisfy basic self-consistency conditions that we now define.

SJ suggested that the maximum of their objective function must be: 1) unique; 2) coordinate transformation invariant; 3) subset-independent (i.e. if data are provided on subsets of a system independently, then the relative probabilities on two subsets of outcomes within a system should be independent of other subsets); 4) system-independent (i.e. if data are provided for systems independently, the joint probability for two independent systems should be the product of their marginal probabilities).

The starting point is a function H (to be determined by the axioms) constrained by data (using Lagrange multipliers). SJ considered general equality and inequality constraints for the data.

For concreteness, here we consider a single constraint on an average $\bar{a}$ of a quantity a. Then SJ's starting point is the following objective function

$$H(\{p_i, q_i\}) - \lambda \left(\sum_i a_i p_i - \bar{a} \right). \tag{2.112}$$

To find the specific form for H, we first invoke SJ's axiom on subset independence. Subset independence states that unless the data are coupled, then the maximum of Eq. (2.112) with respect to each p_i can only depend on this index, namely i. This is only guaranteed if H is a sum over i's, i.e. outcomes. That is,

$$H = \sum_i f(p_i, q_i). \tag{2.113}$$

To further specify the function f, we must apply SJ's second axiom of coordinate invariance. To do this, we use a continuum representation for the probabilities and write Eq. (2.113) as

$$H = \int \mathcal{D}[x] f(p(x), q(x)) \tag{2.114}$$

where $\mathcal{D}[x]$ denotes the integration measure. Our goal is to show that the maximum of $H - \lambda\left(\int \mathcal{D}[x]p(x)a(x) - \bar{a}\right)$ with respect to $p(x)$ – where we have used an average constraint only as a matter of simplicity– is equivalent to the maximum of the coordinate transformed $H' - \lambda'\left(\int \mathcal{D}'[y]p'(y)a'(y) - \bar{a}\right)$ with respect to $p'(y)$ where the primes denote a coordinate transform from $x \to y$. That is

$$\frac{\delta}{\delta p(x)}\left(H - \lambda\left(\int \mathcal{D}[x]p(x)a(x) - \bar{a}\right)\right) = \frac{\delta}{\delta p'(y)}\left(H' - \lambda'\left(\int \mathcal{D}'[y]p'(y)a'(y) - \bar{a}\right)\right) \tag{2.115}$$

where the δ denotes a functional derivative. To simplify Eq. (2.115), we note that, under coordinate transformation

$$\mathcal{D}'[y] = \mathcal{D}[x]J \tag{2.116}$$

where J is the corresponding Jacobian. It then follows that one acceptable relationship between the transformed and untransformed probabilities and observables is [141]: $p' = J^{-1}p$, $q' = J^{-1}q$ (from normalization of the coordinate transformed distributions); and $a' = a$ (from the conservation of $\int \mathcal{D}[x]p(x)a(x)$ under coordinate transformation). Selecting these relations, we find $H' = \int \mathcal{D}[x]Jf(J^{-1}p(x), J^{-1}q(x))$.

Thus Eq. (2.115) simplifies to

$$-\lambda a(x) + g(p(x), q(x)) = -\lambda' a(x) + g(J^{-1}p(x), J^{-1}q(x)) \tag{2.117}$$

where $g(p, q) = \delta f(p, q)/\delta p$. Since $a(x)$ and the Jacobian, J, are arbitrary functions of x, then Eq. (2.117) can only be true if $g(p, q) = g(p/q)$ and $\lambda = \lambda'$. By integrating g, it then follows that $f(p, q) = ph(p/q)$ up to an arbitrary constant in q where h is some function of p/q. The fourth axiom on system independence ultimately fixes the functional form for h.

To see this, we consider independent constraints on two systems described by coordinates x_1 and x_2 as follows

$$\int \mathcal{D}[\mathbf{x}] a_k(x_k) p(x_1, x_2) = \bar{a}_k \quad (k = 1, 2). \tag{2.118}$$

We then define $H = \int \mathcal{D}[\mathbf{x}] p(\mathbf{x}) h(r)$ where $r(\mathbf{x}) \equiv p(\mathbf{x})/q(\mathbf{x})$ and $\mathbf{x} \equiv \{x_1, x_2\}$. Variation of H with respect to p under the constraints given in Eq. (2.118) yields

$$\begin{aligned}
&\frac{\delta}{\delta p(\mathbf{x})} \left(H - \lambda_1 \int \mathcal{D}[\mathbf{x}] p(x_1, x_2) a_1(x_1) - \lambda_2 \int \mathcal{D}[\mathbf{x}] p(x_1, x_2) a_2(x_2) \right) \\
&= h(r(\mathbf{x})) + r(\mathbf{x}) h'(r(\mathbf{x})) - \lambda_1 a_1(x_1) - \lambda_2 a_2(x_2) \\
&= h(r_1(x_1) r_2(x_2)) + r_1(x_1) r_2(x_2) h'(r_1(x_1) r_2(x_2)) - \lambda_1 a_1(x_1) - \lambda_2 a_2(x_2) = 0
\end{aligned} \tag{2.119}$$

where, from system independence, we've set $r(\mathbf{x})$ to $r_1(x_1) r_2(x_2)$ and $h' = \delta h / \delta r$. To obtain a simple differential equation in terms of $h(r)$, we take derivatives of the last line of Eq. (2.119) with respect to both x_1 and x_2 which yields

$$r_1'(x_1) r_2'(x_2) (r_1^2 r_2^2 h'''(r_1 r_2) + 4 r_1 r_2 h''(r_1 r_2) + 2 h'(r_1 r_2)) = 0 \tag{2.120}$$

which further simplifies to

$$r^2 h'''(r) + 4 r h''(r) + 2 h'(r) = 0 \tag{2.121}$$

from which we find $h(r) = -K \log(r) + B + C/r$ with constant K, B and C. From $H = \int \mathcal{D}[x] p(x) h(r)$, we find that H assumes the following form

$$H = -K \int \mathcal{D}[\mathbf{x}] p(\mathbf{x}) \log(p(\mathbf{x})/q(\mathbf{x})) \tag{2.122}$$

up to a positive multiplicative factor K, and additive constants (independent of our model parameters $p(\mathbf{x})$) that we are free to set to zero. In discrete form, this becomes [141]

$$H = -K \sum_i p_i \log(p_i/q_i). \tag{2.123}$$

Thus any function with a maximum identical to that of H can be used in making self-consistent inferences – that is, inferences that satisfy the SJ axioms – about probability distributions. In the absence of any constraint, maximizing H with respect to each p_i returns the corresponding hyperparmeter q_i up to a normalization constant. The hyperparameters are therefore understood to be a probability distribution.

Finally, a note on axiom 1 is in order: while our derivation was for a special type of constraint, our arguments above hold and, in particular, maximizing the constrained H returns a unique set of $\{p_i\}$ if the constraints do not change the overall convexity of the objective function.

The mathematics above help clarify the following important points about the principle of MaxEnt and, more broadly, the philosophy of data analysis.

1) While historically MaxEnt has been closely associated to thermodynamics and statistical mechanics in physics, nothing in H's derivation limits the applicability of MaxEnt to equilibrium phenomena. In fact, MaxEnt is a general inference scheme valid for any probability distributions whether they be probability distributions over trajectories or equilibrium states [49]. In fact, MaxEnt's application to dynamical system is reviewed in Ref. [49]. To reiterate, MaxEnt is no more tied to equilibrium than are Bayes' theorem or even the concept of probability itself.

2) While the H derived from SJ' axioms is additive, in that $H(\{p_{ij} = u_i v_j\}) = H(\{u_i\}) + H(\{v_j\})$, H can be used to infer probability distributions for either additive or non-additive systems [151]. The SJ axioms only enforce that if no couplings are imposed by the data, then no couplings should be imposed by hand. Put differently,

the function H that SJ have derived is not only valid for independent systems. It only says that if the data do not couple two outcomes i and j, then the probabilities inferred must also be independent and satisfy the normal addition and multiplication rule of normal probabilities. In this way, spurious correlations unwarranted by the data are not introduced into the model inferred. Conversely, if data couples two outcomes, then this H constrained by coupled data generates coupled outcomes.

This fundamentally explains why it is incorrect to use other unconventional entropies in model inference which specifically enforce couplings by hand [151]. What is more, parametrizing *ad hoc* couplings (the q-parameter, say, in the Tsallis entropy [307]) in a prior (equivalently the entropy) from data is tautological since the data are used to then inform the prior and, simultaneously, the likelihood [306].

3) As a corollary to 2, H is not a thermodynamic entropy [49]. The thermodynamic entropy, S, is a number not a function. S is H evaluated at its maximum, $\{p_i^*\}$, under equilibrium (thermodynamic) constraints. While H is additive, the thermodynamic entropy S, derived from H, may not be. The non-additivity of S originates from the non-additivity of the constraints.

4) Historically, constraints used on H to infer probabilistic models were limited to means and variances [116, 117] and, in order to infer more complex models, exotic *ad hoc* constraints were developed leading to problems detailed in Refs. [308–311]. But, as we have explained earlier, H is the logarithm of a prior and constraints on H are the logarithm of a likelihood. Selecting constraints should be no more arbitrary than selecting a likelihood. Classical thermodynamics, as it arises from MaxEnt, is therefore atypical. It is a very special (extreme) example where data (such as average energy or, equivalently, temperature) is provided with vanishingly small error bar and the likelihoods are delta-functions.

2.7.4 Repercussions to rejecting MaxEnt

Since the time when Jaynes provided a justification for the exponential distribution in statistical mechanics from Shannon's information theory [116, 117], other entropies have been invoked to justify more complex models in the physical and social sciences such as power laws [307, 312–320]. The most widely used of these entropies is the Tsallis entropy [307].

It has been argued that the Tsallis entropy generalizes statistical mechanics because it is not additive [317, 319, 321]. That is, $H(\{p_{ij} = u_i v_j\}) = H(\{u_i\}) + H(\{v_j\}) + \epsilon H(\{u_i\})H(\{v_j\})$, where ϵ measures the deviation from additivity (though the choice of ϵ has been criticized because it is selected in an *ad hoc* manner by fitting data [306, 322, 323]).

Non-additive entropies do not follow from the SJ axioms. In fact, the Tsallis entropy explicitly violates the fourth axiom [151]. Thus, we can ascertain that the resulting H no longer generates self-consistent inferences about probability distributions.

To see this, we start from the discrete analog of Eq. (2.114) dictated by the third axiom

$$H(\mathbf{p}) = \sum_k f(p_k) \tag{2.124}$$

and, for simplicity only, we assume a uniform prior q_j which we exclude from the calculation. Now we consider bringing together two systems, indexed i and j, with probability p_{ij}.

The fourth axiom – system independence – says that bringing together two systems having marginal probabilities $\mathbf{u} = \{u_i\}$ and $\mathbf{v} = \{v_j\}$ gives new probabilities p_{ij} that are factorizable as $p_{ij} = u_i v_j$ unless the data couples the systems.

That is, under independent constraints, on $\{u_i\}$ and $\{v_j\}$ we have

$$H(\mathbf{p}) - \lambda_a \left(\sum_{i,j} p_{ij} a_i - \bar{a} \right) - \lambda_b \left(\sum_{i,j} p_{ij} b_j - \bar{b} \right). \tag{2.125}$$

Taking a derivative with respect to $p_{ij} = u_i v_j$ then yields

$$f'(p_{ij}) - \lambda_a a_i - \lambda_b b_j = 0. \tag{2.126}$$

Subsequently taking two more derivatives of Eq. (2.126) (with respect to u_i and v_j) yields

$$f''(p_{ij}) + p_{ij} f'''(p_{ij}) = 0. \tag{2.127}$$

Defining $f''(p_\alpha) \equiv g(p_\alpha)$, where $\alpha \equiv (i,j)$, yields from Eq. (2.127) $g(p_\alpha) = -1/p_\alpha$ from which we obtain $f(p_\alpha) = -p_\alpha \log p_\alpha + p_\alpha$ and, ultimately, $H = -\sum_\alpha p_\alpha \log p_\alpha + C$, where C is a constant.

By contrast, the Tsallis entropy is defined as

$$H \equiv \frac{K}{1-q} \left(\sum_k p_k^q - 1 \right). \tag{2.128}$$

This entropy satisfies SJ's third axiom (by virtue of still being a sum over outcomes) but not the fourth axiom. That is, even if data do not couple systems indexed i and j, it is no longer true that $p_{ij} = u_i v_j$. Rather, the Tsallis entropy assumes a coupling $p_{ij} = p(u_i, v_j)$ even if no such coupling has yet been imposed by the data. What is more, the constant q is fitted to the data from which it (problematically) follows that both prior and likelihood are informed by the data [306].

To find precisely what the form of this coupling is, we repeat steps that lead us from Eqns. (2.125)-(2.127) except treating p_{ij} as a general function of $p(u_i, v_j)$ and using the $f(p_{ij})$ dictated by Eq. (2.128). This yields

$$(2-q)^{-1} p_{ij} \frac{\partial^2 p_{ij}}{\partial u_i \partial v_j} = \frac{\partial p_{ij}}{\partial u_i} \frac{\partial p_{ij}}{\partial v_j}. \tag{2.129}$$

As a sanity check, in the limit that q approaches 1 (i.e. when the Tsallis entropy approaches the usual Shannon information) we immediately recover $p_{ij} = u_i v_j$.

The solution to Eq. (2.129) is [151]

$$p_{ij} = \left(u_i^{q-1} + v_j^{q-1} - 1\right)^{1/(q-1)}. \tag{2.130}$$

This exercise can be repeated for other entropies (such as the Burg entropy [324]) and the explicit form for the correlations such entropies impose can be calculated [151].

Eq. (2.130) captures the profound consequence of invoking the Tsallis entropy, or other entropies not consistent with the SJ axioms, in probabilistic model inference. By virtue of violating SJ's fourth axiom the Tsallis entropy imposes coupling between events where none are yet warranted by the data. By contrast, couplings can be introduced in models from the normal Shannon entropy by either systematically selecting a prior distribution, $\{q_k\}$, with couplings or letting the data impose those couplings.

2.8 Concluding Remarks and the Danger of Over-interpretation

Throughout this review we have discussed multiple modeling strategies. We began by investigating how model parameters may be inferred from data. We then explored more sophisticated formalisms that have allowed us to infer not only model parameters, but models themselves starting from broader model classes. We discussed how

information theory is useful across data analysis and how it connects to Bayesian methods.

While the formalisms we have presented are powerful and perform well on test (synthetic) data sets used to benchmark the data, it is difficult to determine how well they perform on real data.

For instance, it may be difficult to quantify if Bayesian-inspired parameter averaging – which is critical in model selection – ultimately rejects models on the basis of parameter values that may be unphysical (such as infinite standard deviations) and should not have been considered in the first place.

What is more, we often do not know the exact likelihood in any analysis either. Thus, the more we ask of a model inference scheme, the more sensitive we become to over-interpretation because noise properties may not be well captured by our approximate likelihood. We gave as a concrete example that inference methods that do not start with a fixed number of states in the analysis of time series data, may interpret drift as the occupation of new states over the course of the time trace.

Likewise, inferences made under one choice of noise model are biased by this choice. So, in practice, non-parametric mixture models are still limited by the parametric choice for their distribution over observations which then determine the states that will be populated.

Despite these apparent shortcomings, the analysis methods presented here outperform methods from the recent past and broaden our thinking. Furthermore, the mechanistic insights provided from statistical modeling may ultimately help inspire new theoretical frameworks to describe biological phenomena.

The mathematical frameworks we've described here are helpful and worth investigating in their own right. They suggest what model features should be extractable from data and, in this sense, may even help inspire new types of experiments.

3. PITCHING SINGLE-FOCUS CONFOCAL DATA ANALYSIS ONE PHOTON AT A TIME WITH BAYESIAN NONPARAMETRICS

Tavakoli, Meysam, Sina Jazani, Ioannis Sgouralis, Omer M. Shafraz, Sanjeevi Sivasankar, Bryan Donaphon, Marcia Levitus, and Steve Pressé. "Pitching Single-Focus Confocal Data Analysis One Photon at a Time with Bayesian Nonparametrics." Physical Review X 10, no. 1 (2020): 011021.
DOI:https://doi.org/10.1103/PhysRevX.10.011021
Contribution: MT analyzed data and developed analysis software; MT, SJ, IS developed computational tools; OS, SS, BD, ML contributed experimental data; MT, SJ, IS, SP conceived research; SP oversaw all aspects of the projects.

3.1 Abstract

Fluorescence time traces are used to report on dynamical properties of molecules. The basic unit of information in these traces is the arrival time of individual photons, which carry instantaneous information from the molecule, from which they are emitted, to the detector on timescales as fast as microseconds. Thus, it is theoretically possible to monitor molecular dynamics at such timescales from traces containing only a sufficient number of photon arrivals. In practice, however, traces are stochastic and in order to deduce dynamical information through traditional means–such as fluorescence correlation spectroscopy (FCS) and related techniques–they are collected and temporally autocorrelated over several minutes. So far, it has been impossible

to analyze dynamical properties of molecules on timescales approaching data acquisition without collecting long traces under the strong assumption of stationarity of the process under observation or assumptions required for the analytic derivation of a correlation function. To avoid these assumptions, we would otherwise need to estimate the instantaneous number of molecules emitting photons and their positions within the confocal volume. As the number of molecules in a typical experiment is unknown, this problem demands that we abandon the conventional analysis paradigm. Here, we exploit Bayesian nonparametrics that allow us to obtain, in a principled fashion, estimates of the same quantities as FCS but from the direct analysis of traces of photon arrivals that are significantly smaller in size, or total duration, than those required by FCS.

3.2 Introduction

Methods to capture static molecular structures, such as super-resolution microscopy [325–327], provide only snapshots of life in time. Yet life is dynamical and obtaining a picture of life in action–one that captures diffraction-limited biomolecules as they move, assemble into and disassemble from larger bimolecular complexes–remains an important challenge [328]. In fact, the creative insights directly leading to fluorescence correlation spectroscopy (FCS) [329, 330]–and related methods such as FCS-FRET [331, 332] and FCCS [333]–have shown that deciphering dynamical information from molecules, often biomolecules, does not demand spatial resolution or spatial localization. Rather, the key is to inhomogeneously illuminate a sample over a small volume.

As fluorescently-labeled molecules diffuse across this inhomogeneously illuminated volume, they emit photons (*i.e.*, they fluoresce) in a way that is proportional to the illumination at their respective locations [334]. Single photon detectors, often photo-

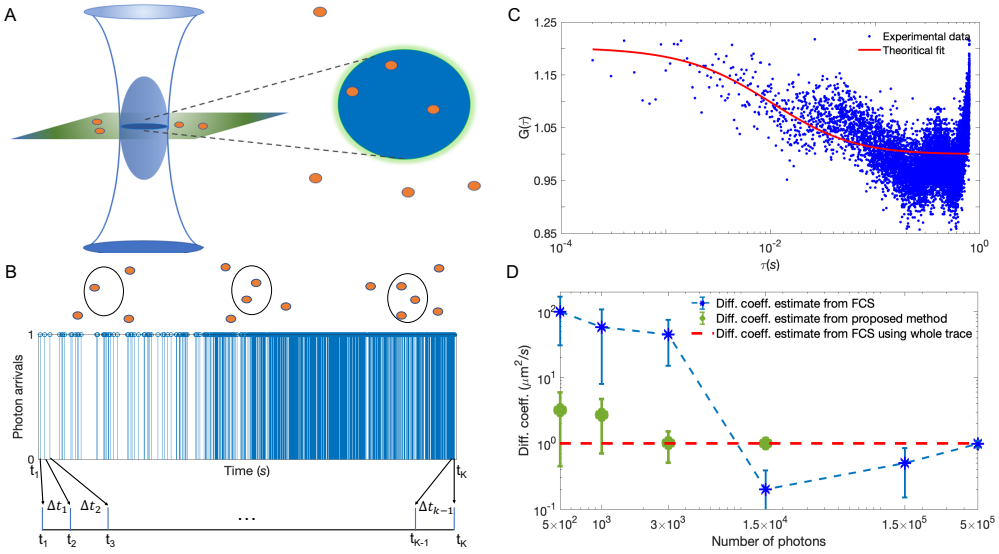

Fig. 3.1. **Photon arrival times can characterize dynamical properties of molecules on fast, photon-detection, timescales**. (A) Schematic of an illuminated confocal volume (blue) with fluorescent molecules emitting photons based on their location within that volume. (B) Synthetic trace containing $\approx$ 1500 photon arrivals produced by 4 molecules diffusing at 1 $\mu m^2/s$ for a total time of 30 ms under background and molecule photon emission rates of 10^3 photons/s and 4×10^4 photons/s, respectively. (C) Autocorrelation curve, $G(\tau)$, of the trace in (B), binned at 100 μs. On account of the limited data available in the trace, any reasonable fit is impossible. Normally, in FCS analysis, much longer traces are used to generate smoother $G(\tau)$ that are fitted to determine a diffusion coefficient. In Fig. 3.14 of the Appendix, we show that the quality of the fit does not improve considerably by fitting to a semi-logarithmic curve. (D) Comparison between diffusion coefficient estimates using our proposed method (detailed later) and FCS as a function of the number of photon arrivals in the analyzed trace. Since by 1.5×10^4 photon arrivals our method has converged, we avoid analyzing larger traces.

multiplier tubes or avalanche photodiodes, are then used to record these photons. In principle, with the appropriate electronics, photons can be recorded within μs-ms. This suggests that information on the molecules' motion could be drawn from the data on fast timescales that approach data acquisition, *i.e.* no more than a few μs-ms.

The fundamental quantities measured in a confocal optical setup are individual photon arrival times, from which photon inter-arrival times, *i.e.*, the intervals between adjacent photon arrivals, can be readily obtained [335]. When imaging molecules fixed in space and under homogeneous (uniform) illumination, these inter-arrival times–excluding other experimental and label artifacts such as detector noise, background photons, and label photo-physical kinetics–are independent and identically distributed and so uncorrelated with each other. However, inter-arrival times measured in conventional confocal experiments encode the number of molecules in the vicinity of the confocal volume, their diffusion dynamics, their position with respect to the confocal center in addition to an array of experiment specific artifacts such as detector characteristics and label photo-kinetics. Consequently, inter-arrival times are correlated with each other and, in principle, these correlations can be exploited to characterize the dynamics of the underlying molecular system.

Thus far, correlations in the inter-arrival times are exploited by collecting photons over long periods [336] and temporally autocorrelating the resulting fluorescence intensity measurements [329, 330, 337, 338]. For sufficiently long intensity traces, the stochasticity in the number of labeled molecules contributing photons, as well as their positions in the illuminated volume and their instantaneous photon emission rates, are averaged out. As such, the mathematical expression for the fluorescence intensity time-autocorrelation function takes a simple form that–under strong assumptions on

the illuminated volume's geometry and the molecules' photon emission rate–can be summarized in analytic formulas that are fitted on the acquired measurements.

However, despite the elegance and simplicity of the mathematics involved in the derivation of the time-autocorrelation function [329,330,337,338], a critical limitation of autocorrelative methods, including all those within the FCS framework, remains the stark timescale separation between data collection (*e.g.*, typical time between successive photon arrivals) and the timescale required to deduce a meaningful dynamical interpretation (*e.g.*, typical duration between first and last photon arrivals used); see Fig. (3.1). A method that takes direct advantage of single photon arrivals, without using intensity traces (*i.e.,* downsampled photon arrivals), has the potential to reveal dynamical information on timescales several orders of magnitude faster than traditional FCS analysis. As a result, rapid or non-equilibrium processes and, as such, abrupt changes in molecular chemistry, could be studied. Furthermore, provided such a method can utilize substantially shorter traces, the total duration of experiments can be shrunk and the phototoxic damage induced on biological samples can be reduced substancially [327, 339–341]. This is especially relevant for *in vivo* FCS applications [342–345].

Previously proposed methods to analyze single photon measurements [66,208,235, 346–351] make assumptions that render them inappropriate for imaging molecules moving through inhomogeneously illuminated volumes [349]. For example, for the analysis of single molecule fluorescence resonance energy transfer (FRET), existing methods assume that the photon inter-arrival times reflect only biomolecular conformational transitions [346,349,352] but not diffusive motion of the entire biomolecule [346,353–355], and so are appropriate only for experiments on immobilized molecules. Along the same lines, existing methods combine FRET with FCS [356] to quantify ns dynamics; however, they do not directly exploit single photon measurements. Rather,

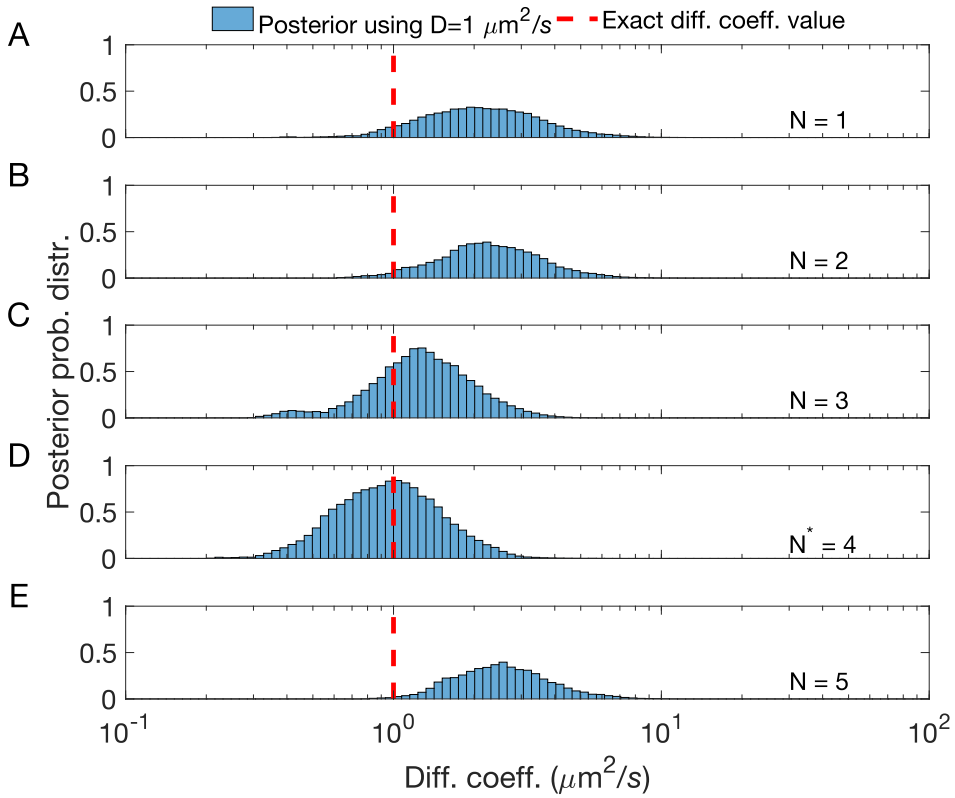

Fig. 3.2. **Estimates of diffusion coefficients from photon arrival traces strongly depend on the number of molecules assumed to be contributing to the trace.** The trace analyzed contained $\approx$ 1800 photon arrivals produced by 4 molecules diffusing at 1 $\mu m^2/s$ for a total time of 30 ms under background and molecule photon emission rates of 10^3 photons/s and 4×10^4 photons/s, respectively. To estimate D *parametrically,* we assumed a fixed number of molecules, $N = 1$ (A); $N = 2$ (B); $N = 3$ (C); $N = 4$ (D); and $N = 5$ (E). The correct estimate in (D)–and the mismatch in all others–underscores why it is critical to estimate the number of molecules contributing to the trace to deduce quantities such as diffusion coefficients from single photon arrivals.

they operate on downsampled measurements, achieved through binning, similar to traditional FCS, and therefore inherit the same limitations and drawbacks.

To be able to use single photon arrival times to estimate the diffusion coefficient of labeled molecules in a confocal experiment, as in most biological applications, we must be able to determine the particular number of molecules responsible for the observed photon arrival time trace. Otherwise, naively, many molecules with low diffusion coefficients emitting photons at the periphery of the illuminated confocal volume could be mistaken for fewer molecules with higher diffusion coefficients in the center region which is most illuminated. As we illustrate in Fig. (3.2), misidentifying the number of molecules, or incorrectly assessing their positions, may give rise to incorrect diffusion coefficient estimates.

More concretely, to obtain quantitative estimates of the diffusion coefficient, we need to formulate a likelihood [357–359]. In turn, to formulate a likelihood for photon arrival data demands that we know the number of molecules contributing photons as well as their locations across time. As the number of molecules instantaneously located within the confocal volume is unknown, all reasonable possibilities need to be considered and rank-ordered using expensive pre- or post-processing model selection heuristics [79,328]. This has not been achieved yet, in part, because of the prohibitive computational cost it entails. Analyzing single photon arrivals from a confocal setup to derive dynamical information therefore demands fundamentally new tools.

The conceptually novel framework that we propose in this study can winnow down infinite possibilities (*i.e.*, infinite populations of molecules potentially contributing photons) to a finite, computationally manageable, number in a mathematically exact manner. Such a framework avoids compromising temporal resolution, as it requires no intensity trace to be formed (*i.e.,* no downsampling), and allows us to directly deduce dynamical quantities, such as diffusion coefficients, efficiently from raw single

photon arrivals. The underlying theory, Bayesian nonparametrics (BNPs) [263], is a powerful set of tools still under active development and largely unknown to the Physical Sciences [2–4, 79, 328, 360–364].

Mathematical devices within BNPs, such as the beta-Bernoulli process [365–367], allow us to place priors not only on parameters themselves, as traditional parametric Bayesian methods, but also on distributions over an infinite number of candidate models to which parameters are associated [368]. Concretely, for the case of our single photon time traces, BNPs and in particular beta-Bernoulli processes can be used to assign posterior probabilities over an array of quantities including all possible number of molecules responsible for producing the data and their associated locations at each photon arrival time. With these devices, as we describe herewith, we turn the otherwise difficult problem of model-selection–that is, determining how many molecules contribute photons–into a parameter estimation problem that remains computationally tractable [365–367].

3.3 Materials and Methods

Here, we describe the mathematical formulation of our BNPs method for the analysis of confocal single photon data. We begin with the overall input which consists of photon inter-arrival times, $\boldsymbol{\Delta t} = (\Delta t_1, \Delta t_2, \ldots, \Delta t_{K-1})$ where Δt_k represents the time interval between adjacent observations of photons, which occur at times t_k with $k = 1, \ldots, K$. We also use as input the illuminated confocal volume's shape and background photon emission rate which we can determine separately through calibration [369].

To derive estimates for the diffusion coefficient from $\boldsymbol{\Delta t}$, we need to determine intermediate quantities which include: i) photon emission rates of molecular labels; and, most importantly, ii) the unknown number of molecules contributing photons

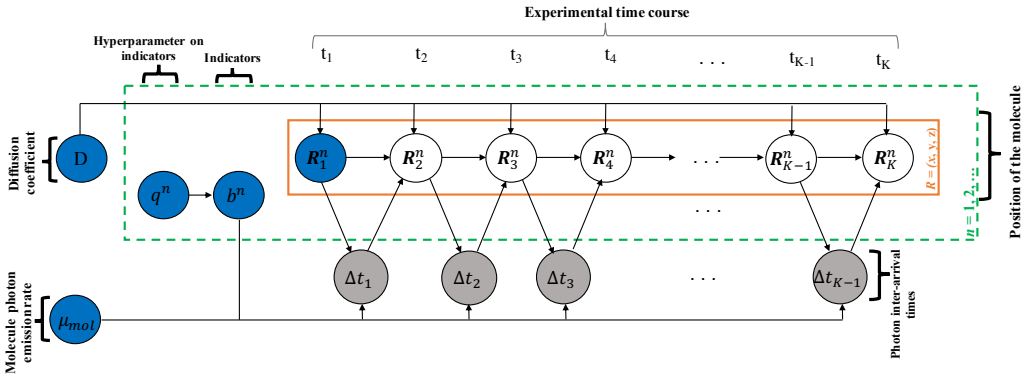

Fig. 3.3. **BNP formulation used for the analysis of photon arrival traces**. Molecules, indexed $n = 1, 2, \ldots,$ evolve over the experimental time course which is indexed by $k = 1, 2, \ldots, K$. Here, $R_k^n = (x_k^n, y_k^n, z_k^n)$ indicates the location of molecule n at time t_k. During the experiment, only a single observation (inter-arrival time) Δt_k is recorded, thereby combining photon emissions from every molecule and the background. The diffusion coefficient D determines the evolution of the molecular positions which influence the photon emission rates and eventually the recorded Δt_k. The indicator variables b^n are introduced to infer the unknown molecule population size. In the graphical model, the measured data are highlighted by grey shaded circles and the model variables, which require priors, are designated by blue circles.

to the trace $\boldsymbol{\Delta t}$, as well as their location with respect to the center of the confocal volume.

A graphical summary of our formulation is shown in Fig. (3.3). Below, we explain briefly each step involved. More details, and an implementation of the whole method, are available in the Appendix. In addition, source code and a GUI version of our implementation are provided through the Supplementary Materials.

3.3.1 Model Formulation

We begin with the distribution according to which the k^{th} observation, Δt_k, is derived

$$\Delta t_k \sim \text{Exponential}\left(\mu_k\right). \tag{3.1}$$

Accordingly, Δt_k follows an exponential probability distribution [352, 370] with rate μ_k. In fact, the rate μ_k gathers the photon emission rates of all molecules which depend on their respective locations relative to the confocal center (see below) [369]. In addition to the molecule photon emissions rates, μ_k also includes background photons

$$\mu_k = \mu_{back} + \sum_n \mu_k^n, \tag{3.2}$$

where $\sum_n \mu_k^n$ is the sum over photon emission rates μ_k^n gathered from the individual molecules, that we index with $n = 1, 2, \ldots,$ and μ_{back} is the background photon emission rate. In our formulation, μ_k^n and μ_{back} are the emission rates of photons that reach our detectors which, due to optical and detector limitations, are typically lower than the rates of actual photon emissions [371, 372].

Next, we incorporate the dependency of the emission rate μ_k^n on location [373–375] with other effects such as camera pinhole shape and size, the laser intensity, laser wavelength, and quantum yield [369] into a characteristic point spread function (PSF). To be more precise, a PSF characterizes the optical response of an imaging system [335,376,377]. Although this term is mostly used for wide-field microscopes to describe the emission PSF, here, we follow the FCS literature, and use it to describe the confocal microscope, *i.e.,* both emission and detection PSFs. Consistent with FCS [329, 330, 337, 338], we assume a 3D Gaussian geometry [334]

$$\mu_k^n = \mu_{mol} \exp\left(-2\frac{(x_k^n)^2 + (y_k^n)^2}{\omega_{xy}^2} - 2\frac{(z_k^n)^2}{\omega_z^2}\right), \tag{3.3}$$

where (x_k^n, y_k^n, z_k^n) is the position of the n^{th} molecule at time t_k and the parameter μ_{mol} indicates the brightness of a single molecule. This is the rate of detected photon emissions achieved when the molecule is at the center of the confocal volume where illumination is highest.

Finally, for a molecule diffusing along one direction, the probability distribution $p(x,t)$ of its position x at time t satisfies the diffusion equation [378–380]

$$\frac{\partial p}{\partial t} = D\frac{\partial^2 p}{\partial x^2}. \tag{3.4}$$

To solve this equation, we assume that the molecule is located at x_{k-1} at time t_{k-1} and we obtain

$$p(x,t) = \frac{\exp\left(-\frac{(x-x_{k-1})^2}{4(t-t_{k-1})D}\right)}{\sqrt{4\pi(t-t_{k-1})D}}, \tag{3.5}$$

which is the probability density of a normal random variable with mean x_{k-1} and variance $2(t - t_{k-1})D$. Therefore, at time $t = t_k$, we write

$$x_k \sim \text{Normal}\left(x_{k-1}, 2D\Delta t_{k-1}\right), \tag{3.6}$$

where $\Delta t_{k-1} = t_k - t_{k-1}$ and D is the molecule's diffusion coefficient. Similarly, solving the diffusion equation for molecules following isotropic diffusion in free space along all three Cartesian directions, we obtain

$$x_k^n \sim \text{Normal}(x_{k-1}^n, 2D\Delta t_{k-1}) \tag{3.7}$$

$$y_k^n \sim \text{Normal}(y_{k-1}^n, 2D\Delta t_{k-1}) \tag{3.8}$$

$$z_k^n \sim \text{Normal}(z_{k-1}^n, 2D\Delta t_{k-1}). \tag{3.9}$$

3.3.2 Model Inference

All quantities which we need to infer–such as the diffusion coefficient, D, locations of molecules through time, (x_k^n, y_k^n, z_k^n) and the molecule photon emission rate μ_{mol}– are formulated as model variables. We estimate these variables within the Bayesian paradigm [79, 328, 358]. The model parameters such as D and μ_{mol} require priors. Additionally, we have to consider priors on the initial molecule locations, *i.e.*, at the time of the very first photon arrival, (x_1^n, y_1^n, z_1^n). Options for these priors are straightforward and, for computational convenience, we adopt the distributions described in the Appendix.

Meanwhile, before we proceed any further with our BNPs formulation, we need to revise eq. (3.3) as follows

$$\mu_k^n = b^n \mu_{mol} \exp\left(-2\frac{(x_k^n)^2 + (y_k^n)^2}{\omega_{xy}^2} - 2\frac{(z_k^n)^2}{\omega_z^2}\right). \tag{3.10}$$

The variables b^n, defined for each model molecule, take only values 1 or 0. Specifically, we have $b^n = 0$ when the n^{th} model molecule *does not* contribute photons to the measurements as in this case the molecule is *decoupled* from the overall photon emission rate μ_k. This indicator variable allows us to operate on an arbitrarily large population of model molecules; technically, an infinite population. The ability to recruit, from a potentially infinite pool of model molecules, the precise number that contributes to the measured trace $\boldsymbol{\Delta t}$ is the chief reason we abandon the parametric Bayesian paradigm and adopt BNPs. After introducing the indicators b^n, we can estimate the number of molecules that contribute photons, *i.e.*, those molecules where $b^n = 1$, simultaneously with the remaining of the parameters simply by having each b^n as a separate parameter and estimating its value.

To estimate b^n, we consider a Bernoulli prior with a beta hyper-prior

$$b^n|q^n \sim \text{Bernoulli}(q^n), \tag{3.11}$$

$$q^n \sim \text{Beta}(A_q, B_q), \tag{3.12}$$

where A_q and B_q are (hyper-hyper-)parameters specifically chosen to allow for $n \to \infty$. In this limit, eqs. (3.11) and (3.12) can be combined resulting in a beta-Bernoulli process [365–367]; see Appendix for more details.

With the specified priors, we can now form a joint posterior probability including all unknown variables which we seek to determine, $p(D, \mu_{mol}, (x_k^n, y_k^n, z_k^n)_k^n, (b^n, q^n)^n|\boldsymbol{\Delta} t)$. Nevertheless, the nonlinear dependence of the PSF on the molecules' positions $(x_k^n, y_k^n, z_k^n)_k^n$ and the nonparametric prior on the indicators $(b^n)^n$ exclude a closed form for our posterior. For this reason, we develop a Markov Chain Monte Carlo scheme [358,381,382] that exploits results from the theory of Computational Statistics and Non-linear filtering to generate pseudo-random samples from this posterior that we use in obtaining our estimates [358, 381]. A technical description of this scheme can be found in the

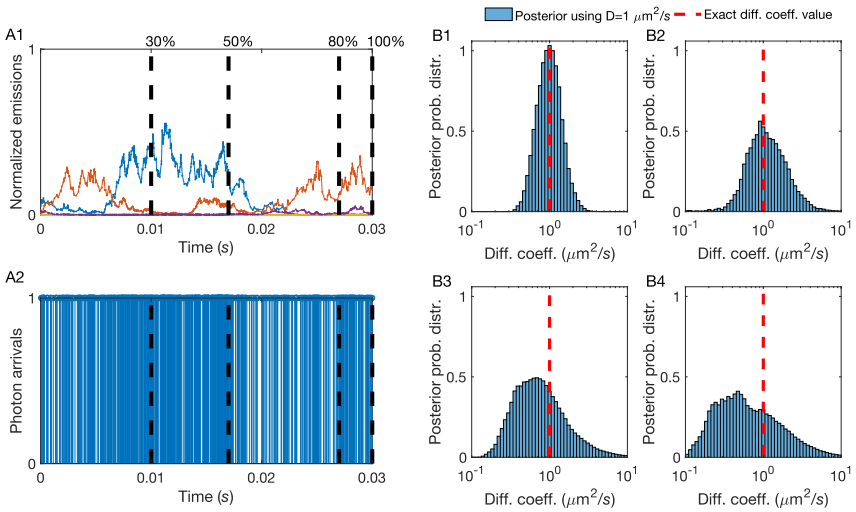

Fig. 3.4. **A higher number of total photon arrivals provide more photons per unit time and sharper diffusion coefficient estimates**. (A1) Instantaneous molecule photon emission rates μ_k^n, normalized by μ_{mol}. (A2) Photon arrival trace resulting from combining photon emissions from every molecule and the background. This synthetic trace contains ≈ 2000 photon arrivals produced by 4 molecules diffusing at 1 $\mu m^2/s$ for a total time of 30 ms under background and molecule photon emission rates of 10^3 photons/s and 4×10^4 photons/s, respectively. The dashed lines show the initial 30%, 50%, 80%, and 100% portions of the original trace containing ≈ 600, ≈ 1000, ≈ 1600, ≈ 2000 photon arrivals, respectively. (B1-B4) Posterior probability distributions drawn from traces with differing length (shown in (A2)). As expected, for the longer traces, the peak of the posterior matches with the exact value of D (dashed line). Gradually, as we decrease the total number of photon arrivals analyzed, the estimation becomes less reliable.

Appendix and a ready-to-use implementation is available through the Supplementary Materials.

3.3.3 Data Acquisition

Acquisition of Synthetic Data for Figs. (3.4)-(3.7)

We acquire the synthetic data shown in the Results section by computer simulations [383–387] that represent Brownian motion of point molecules moving through a typical illuminated confocal volume. We provide finer details and complete parameter choices in the Appendix.

Acquisition of Experiment data for Figs. (3.8)-(3.12)

For these experiments we used Cy3 fluoresent dyes. Solutions were made by suspending Cy3 dye in glycerol/buffer (pH 7.5, 10 mM Tris-HCl, 100 mM NaCl and 10 mM KCl, 2.5 mM $CaCl_2$) at various v/v, to a final concentration of either 100 pM or 1 nM. The solution was placed in a glass- bottomed fluid-cell, assembled on a custom designed confocal microscope [388] and a 532 nm laser beam was focused to a diffraction-limited spot on the glass coverslip of the fluid-cell using a 60x, 1.42 N.A., oil-immersion objective (Olympus). In our setup, the laser beam is focused at the glass-water/glycerol interface and the beam is refocused by visual inspection at the beginning of every measurement. Emitted fluorescence was collected from the same objective and focused onto a Single Photon Avalanche Diode (SPAD, Micro Photon Devices) with a maximum count rate of 11.8 Mc/s. A bandpass filter placed in front of the detector blocked all back-scattered excitation light and relayed only fluorescence from Cy3. Individual photon arrivals on the detector triggered TTL pulses and were both timestamped and registered at 80 MHz. This was achieved using a field programmable gate array (FPGA, NI Instruments) and custom LabVIEW software [389].

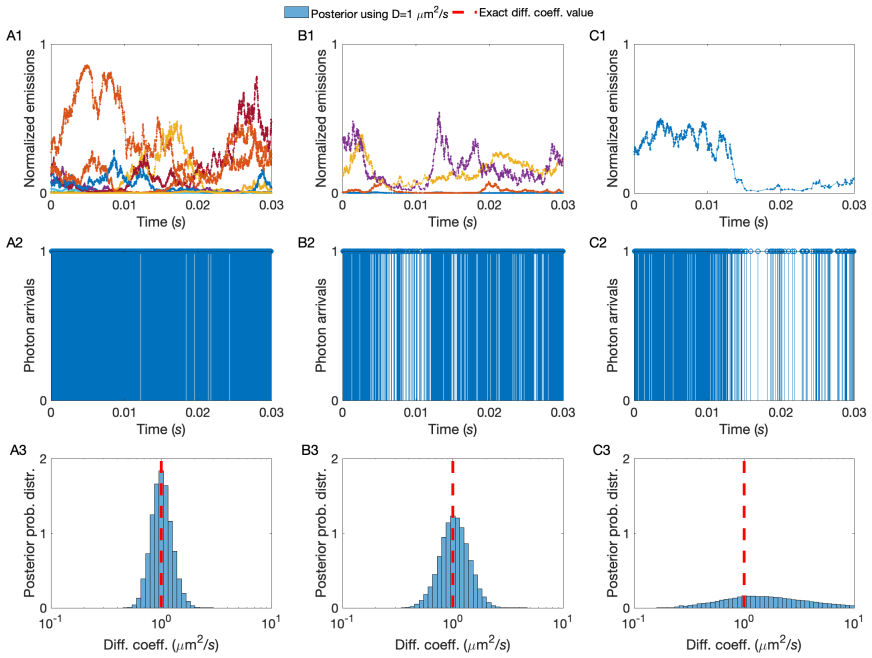

Fig. 3.5. **A higher molecular concentration provides more photons per unit time and sharper diffusion coefficient estimates**. (A1, B1, C1) Instantaneous molecule photon emission rates μ_k^n, normalized by μ_{mol}. (A2, B2, C2) Photon arrival traces resulting from combining photon emissions from every molecule and the background. These are produced by 10 molecules containing $\approx$ 3000 photon arrivals (A2), 4 molecules containing $\approx$ 2000 photon arrivals (B2), and 1 molecules containing $\approx$ 1000 photon arrivals (C2), diffusing at 1 $\mu m^2/s$ for a total time of 30 ms under background and molecule photon emission rates of 10^3 photons/s and 4×10^4 photons/s, respectively. (A3, B3, C3) Posterior probability distributions drawn from traces with differing number of molecules (shown in (A2, B2, C2)). As expected, for the traces with higher number of molecules, the peak of the posterior matches with the exact value of D (dashed line). Gradually, as we decrease the total number of molecules the estimation becomes less reliable.

Acquisition of Experimental Data for Fig. (3.13)

For these experiments we used 5-TAMRA fluorescent dyes. The excitation source was a supercontinuum fiber laser Fianium WhiteLase SC480 (NKT Photonics, Birkerod, Denmark) operating at a repetition rate of 40 MHz. The excitation wavelength (550 nm) was selected by an acousto-optic tunable filter (AOTF), and the exiting beam was collimated and expanded by approximately a factor of three to slightly overfill the back aperture of the objective lens. The light was reflected into the objective lens (Zeiss EC Plan-Neofluar 100x oil, 1.3 NA pol M27, Thornwood, NY, USA) by a dichroic mirror (Chroma 89016bs). The same objective was used to collect the fluorescence from the sample, and passed through a band pass filter (Chroma ET575/50m) before being focused into a position motorized pinhole wheel set at 25 μm. The output of the pinhole was focused on a multimode hybrid fiber optic patch cable (M18L01, Thorlabs, NJ, USA) which was coupled to a single-photon avalanche diode (SPCM AQRH-14, Excelitas Technologies, Quebec, Canada). The detected photons were recorded by a TimeHarp 200 time-correlated single photon counting board (PicoQuant, Berlin, Germany) operating in T3 mode. The sample ($\approx$50 μL) was contained in a perfusion chamber gasket (CoverWell) adhered on a glass coverslip. The sample was 20 nM 5-Carboxytetramethylrhodamine (5-TAMRA, purchased from Sigma-Aldrich, USA) dissolved in doubly distilled water at room temperature.

3.4 Results

Our goal is to characterize quantities that describe molecular dynamics, especially dynamics encountered in biological samples, such as diffusion coefficients, at the data-acquisition timescales of conventional single-focus confocal setups. Our input consists

of: i) the measured photon inter-arrival times $\boldsymbol{\Delta t} = (\Delta t_1, \Delta t_2, ..., \Delta t_{K-1})$; ii) the background photon emission rate; and iii) the geometry of the illuminated volume specified through a characteristic PSF.

As we explain in the Methods section, in order to estimate the molecules' diffusion coefficient, D, we also estimate intermediate quantities (namely, molecule photon emission rates, molecule positions over time and the molecule numbers in the first place). These intermediate quantities demand that we use BNPs to determine quantities that *a priori* may be arbitrarily large such as the number of molecules contributing photons to our datasets $\boldsymbol{\Delta t}$.

Within the Bayesian paradigm [328, 359], our estimates take the form of posterior probability distributions over the unknown quantities. These distributions combine parameter values, probabilistic relations among different parameters, as well as the associated uncertainties. According to the common statistical interpretation [358, 359], the sharper the posterior, the more conclusive (and certain) the estimate. To quantify the uncertainty, we compute a posterior variance and use the square root of this variance to construct error-bars (*i.e.*, credible intervals) [358, 359]. In Table 3.2 in the Appendix, we summarize the mean values and error bars of our analyses.

Below, we validate first our method on synthetic data where the ground truth is available. For these, we use a confocal volume of typical size $\omega_{xy} = 0.3\ \mu m$ and $\omega_z = 1.5\ \mu m$ [335]. We then test our method on experimental data collected in two labs utilizing different FCS setups. For the latter cases, we demonstrate the advantages of our method by comparing our results to the results obtained from autocorrelative methods used in FCS analysis.

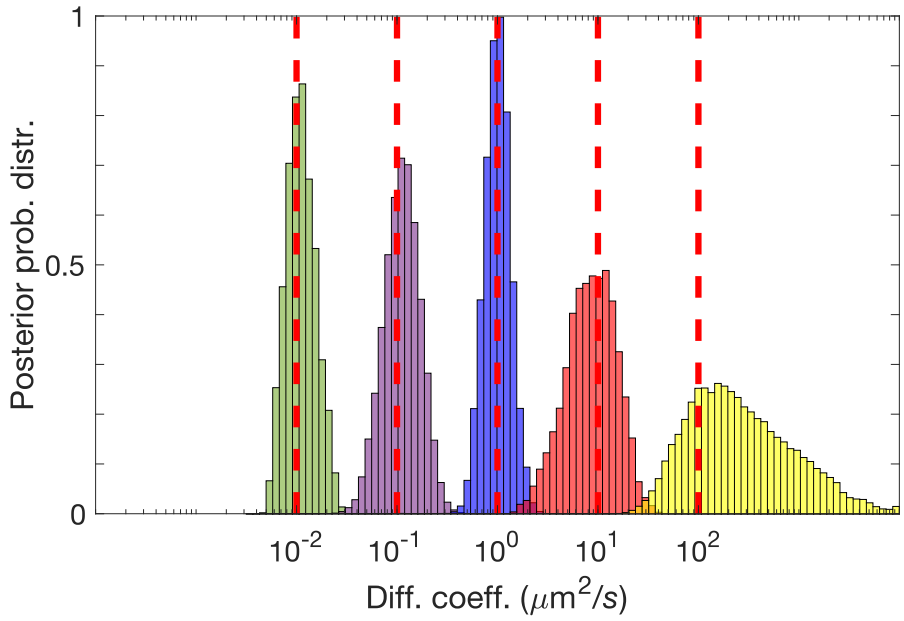

Fig. 3.6. **A lower diffusion coefficient provides more photons per unit time and sharper diffusion coefficient estimates**. Posterior probability distributions drawn from traces containing $\approx$ 2000 photon arrivals produced by 4 molecules diffusing at $D = 0.01, 0.1, 1, 10\ \mu m^2/s$ for a total time of 30 ms under background and molecule photon emission rates of 10^3 photons/s and 4×10^4 photons/s, respectively. For molecules diffusing at $D = 100\ \mu m^2/s$, under similar conditions, we used a trace containing $\approx$ 3000 photons for a total time of 50 ms, since we needed a longer trace to gather sufficient information for drawing a posterior.

3.4.1 Method Validation using Simulated Data

To demonstrate the robustness of our approach, we simulate raw single photon arrival traces under a broad range of: i) total photon arrivals, Fig. (3.4); ii) concentrations of labeled molecules, Fig. (3.5); iii) diffusion coefficients, Fig. (3.6); and iv) molecule photon emission rates, Fig. (3.7). The parameters not varied are held fixed at the following baseline values: diffusion coefficient of 1 $\mu m^2/s$ which is typical of slower *in vivo* conditions [345, 390–392], molecule photon emission rates of 4×10^4 photons/s [347, 393], and 4 as the number of labeled molecules contributing photons. We chose 4, a small number of molecules (as opposed to a larger number of molecules), because this scenario presents the greatest analysis challenge as very few photons, and thus little data, are gathered to aid the analysis.

As illustrated in Fig. (3.1), a critical and recurring point throughout this section is that the traces we analyze are shorter than those that could be meaningfully analyzed using FCS. While we focus on the diffusion coefficient estimation here, we note that our framework supports more detailed parameter estimation which we provide in the Appendix.

Total Photon Arrivals

We evaluate the robustness of our method with respect to the length of the trace (*i.e.*, the total number of photon arrivals recorded) at a fixed number of molecules, diffusion coefficient, and molecule photon emission rates. The first important finding is that, for the values of parameters selected, we need 2 orders of magnitude less data than FCS; see Fig. (3.1D). For instance, to obtain an estimate of the diffusion coefficient within 10% of the ground truth value, we require $\approx 10^3$ photons (directly emitted from the labeled molecule), while FCS requires $\approx 10^5$ photons. Under our

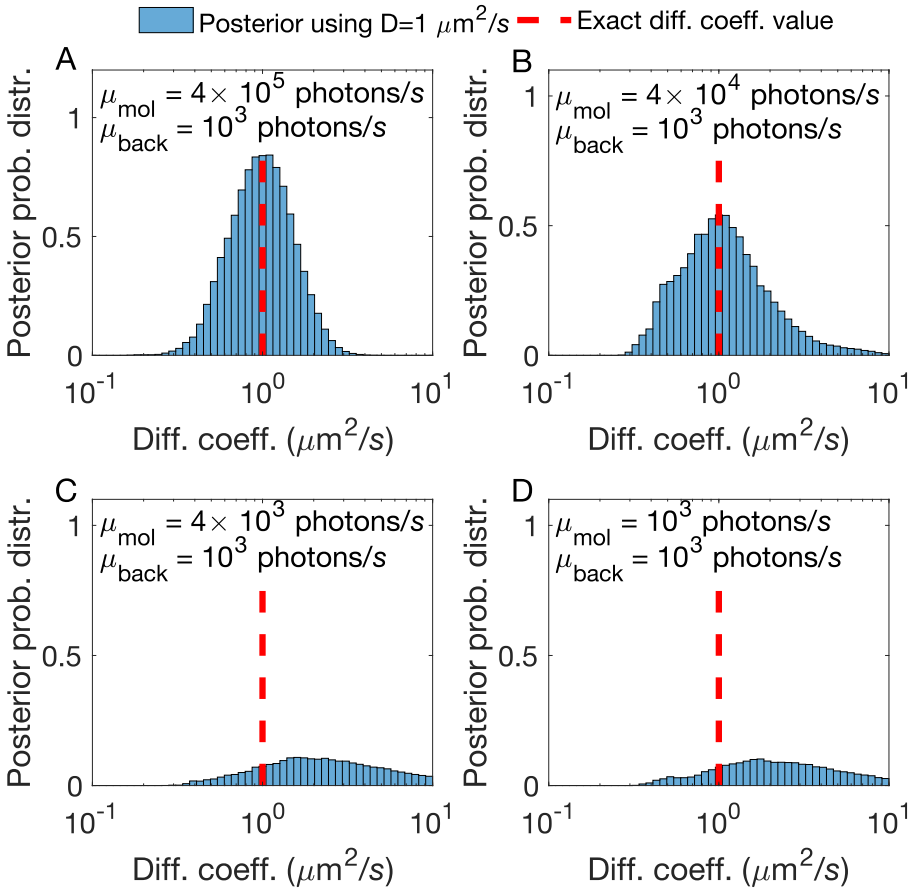

Fig. 3.7. **A higher molecule photon emission rate provides more photons per unit time and sharper diffusion coefficient estimates**. (A, B, C, D) Posterior probability distributions drawn from traces produced by 4 molecules diffusing at 1 $\mu m^2/s$ for a total time of 30 ms under background photon emission rate of 10^3 photons/s and molecule photon emission rates 4×10^5, 4×10^4, 4×10^3, 10^3 photons/s, respectively. As expected, under higher molecule photon emission rates, the peak of the posterior matches sharply with the exact value of D (dashed line). Gradually, as we decrease the molecule photon emission rate, the estimation becomes less reliable.

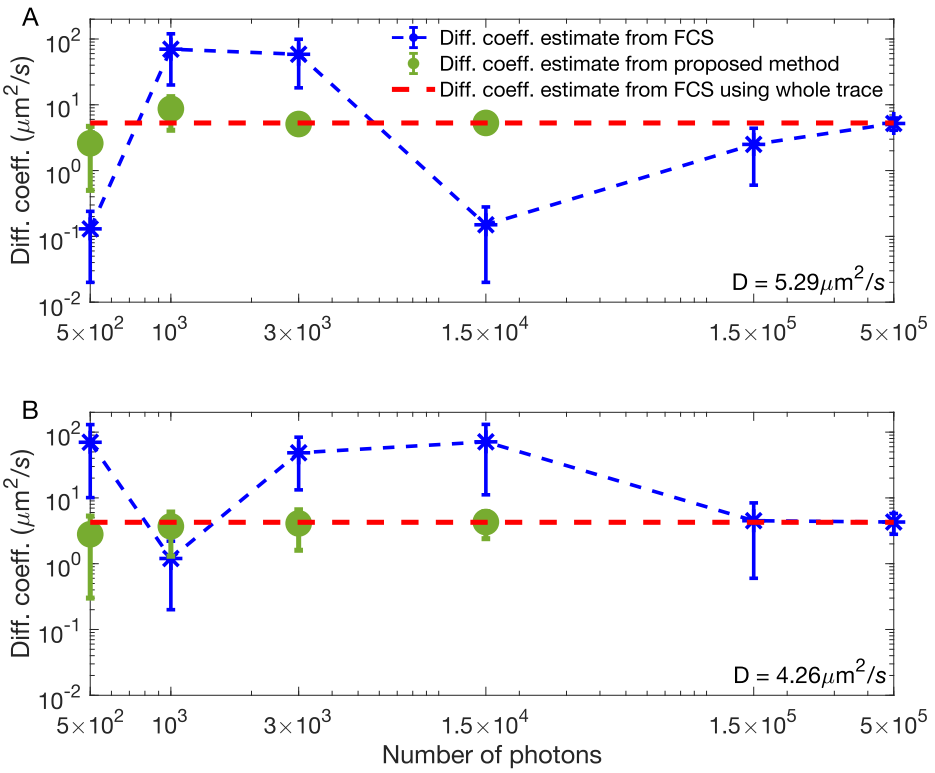

Fig. 3.8. **Higher molecular concentrations in experimental traces provide more photons per unit time resulting in sharper diffusion coefficient estimates**. Estimates shown are drawn from experimental traces with a low (100 pM) (A) and high (1 nM) (B) concentration of Cy3 dye molecules and 75% glycerol at a fixed laser power of 100 μW. Similarly to Fig. (3.1), we compare our method's diffusion coefficient estimate (circle green dots) to FCS (blue asterisk) as a function of the number of photons used in the analysis. Since by 1.5×10^4 photon arrivals our method has converged, we avoid analyzing larger traces. The red dash line is the FCS estimate produced from the entire 5 min trace containing $\approx 3\times10^6$ photon arrivals.

simulated scenario, these correspond to traces of total duration $30 - 50\ ms$ and $50\ s$, respectively. To determine our error, we chose the mean value of the diffusion coefficient's marginal posterior, $p(D|\mathbf{\Delta}t)$, and measure the percentage difference of this mean value to the ground truth.

In general, the precise photon numbers demanded by our method and traditional FCS depend on a broad range of experimental parameter settings. This is the reason, we explore different settings in subsequent subsections as well as the Appendix.

An important overarching concept is the concept of a photon arrival as a unit of information. The more photon arrivals we have in the analyzed trace, the sharper our diffusion coefficient estimates become. This is valid, as we see in Fig. (3.1D) and Fig. (3.4), for increasing total photon arrivals. Similarly, as we see in subsequent subsections, we also collect more photons as we increase the concentration of labeled molecules (and thus the number of molecules contributing photons to the trace), increase the molecule photon emission rates of molecular labels, or decrease diffusion coefficients of molecules. In the latter case, a slower diffusion coefficients provides more time for each molecule to traverse the illuminated region, in turn, resulting in more photon arrivals.

Molecule Concentration

To test the robustness of our method under different concentrations of labeled molecules at fixed diffusion coefficient, and molecule photon emission rates, we simulate molecules diffusing at 1 μm^2/s for a total time 30 ms with: i) average concentrations of 10 molecules/μm^3, Fig. (3.5A1, A2); ii) 4 molecules/μm^3, Fig. (3.5B1, B2); and iii) 1 molecule/μm^3, Fig. (3.5C1, C2). The molecule and background photon emission rates are taken to be 4×10^4 photons/s and 10^3 photons/s respectively, which are typical of confocal imaging [347].

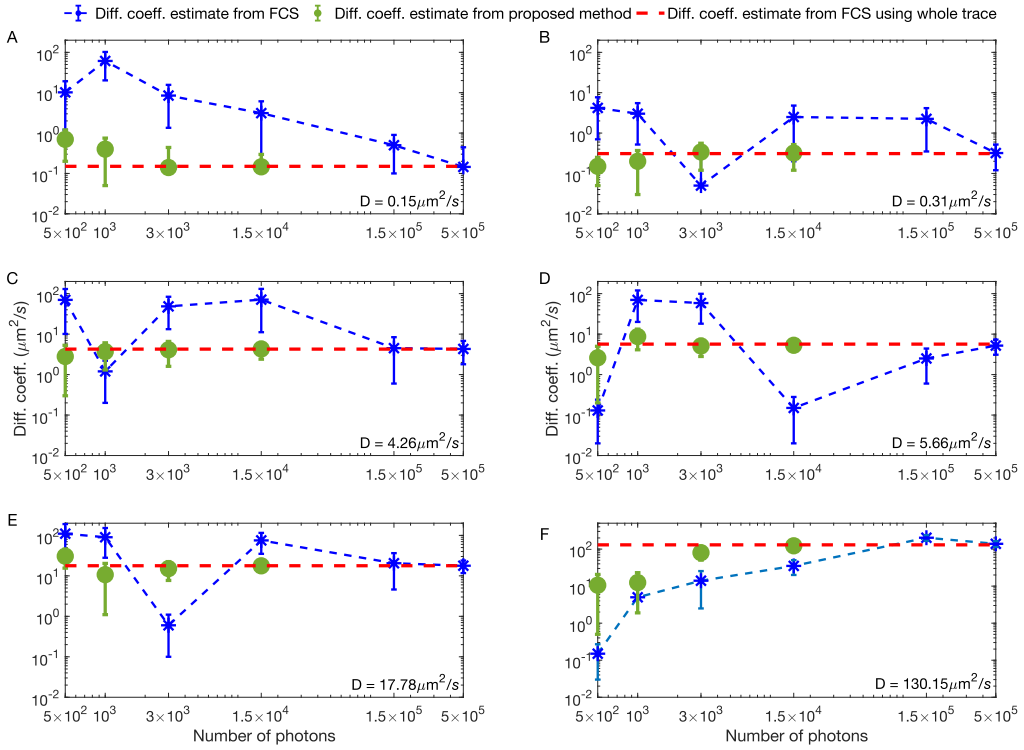

Fig. 3.9. **Lower diffusion coefficients in experimental traces provide more photons per unit time and sharper diffusion coefficient estimates**. Estimates shown are drawn from experimental traces with 99% glycerol (A), 94% glycerol (B), 75% glycerol (C), 67% glycerol (D), 50% glycerol (E), and 0% glycerol (F) with fixed concentration 1 nM of Cy3 dye molecules and laser power of 100 μW. Similarly to Fig. (3.1), we compare our method's diffusion coefficient estimate (circle green dots) to FCS (blue asterisk) as a function of the number of photons used in the analysis. Since by 1.5×10^4 photon arrivals our method has converged, we avoid analyzing larger traces. The red dash line is the FCS estimate produced from the entire 5 min trace containing $\approx 3 \times 10^6$ photon arrivals.

Figure (3.5) summarizes our results and suggests that posteriors over diffusion coefficients are broader–and thus the accuracy with which we can pinpoint the diffusion coefficient drops–when the concentration of labeled molecules is lower. Intuitively, we expect this result as fewer molecules within the confocal volume provide fewer photons arrivals.

Diffusion Coefficients

We repeat the simulations of the previous subsection to demonstrate, using synthetic data, the robustness of our method with respect to the diffusion coefficient magnitude at fixed number of molecules, and molecule photon emission rates; see Fig. (3.6). Intuitively, and again on the basis of the fact that photon arrivals carry information, we expect that faster moving molecules give rise to broader posterior distributions as these emit fewer photons, and thus provide less information, while they traverse the confocal volume.

Molecule Photon Emission Rates

Figure (3.7) illustrates the robustness of our method with respect to the molecule photon emission rates (*i.e.*, set by the laser power used in the experimental setting and the choice of fluorescent label) by fixing the number of molecules, diffusion coefficient ($1\ \mu m^2/s$), and background emission (10^3 photons/s). To accomplish this, we simulate increasingly dimmer molecules until the molecule signature is effectively lost in the background. As expected, dimmer molecules lead to broader posterior estimates over diffusion coefficients as these traces are associated with higher uncertainty.

3.4.2 Estimation of Physical Parameters from Experimental Data

To evaluate our BNPs method on real data, we used experimental single photon traces collected under a broad range of conditions. That is, we used measurements from two different experimental setups and different fluorescent dyes, that are commonly used in labeling biological samples, as well as diffusing labeled proteins. Additional differences between the setups include different numerical apertures (NA), laser powers, and overall detection instrumentation as detailed in the Methods section.

Figures (3.8)-(3.11) were collected using the Cy3 dye and these results were used to benchmark the robustness of our method on dye concentration, diffusion coefficients, and laser power. Moreover, to evaluate the proposed approach beyond free dyes, in Fig. (3.12), we used labeled proteins, namely freely diffusing streptavidin labeled with Cy3. For Fig. (3.13), photon arrivals were collected using 5-TAMRA dye in order to test the robustness of our method on a different fluorophore.

Benchmarking on Experimental Data using Cy3

We begin by verifying our method on mixtures of water and glycerol. While we only use short segments in our analysis, the collected traces are long enough ($\approx$5 min each) to be meaningfully analyzed by traditional autocorrelative analysis used in FCS for sake of comparison. The result of the analysis of the full trace by FCS yields a diffusion coefficient that we treat as an effective ground truth. We then ask how long of a trace our method requires, as compared to FCS, in order for our diffusion coefficient estimate to converge to this ground truth.

Our strategy addresses the following complication: we anticipate that the PSF may be distorted from the idealized shape assumed especially with increasing amounts of glycerol [394]. However, the same (possibly incorrect) PSF is used in both FCS and

our method in order to compare both methods head-to-head. Thus, concretely, we are asking: how many photon arrivals do we need to converge to the *same* result as FCS (irrespective of whether the FCS result is affected by PSF distortion artifacts)?

Our single photon traces are obtained under a range of conditions, namely different: i) dye concentrations, Fig. (3.8); ii) diffusion coefficients, Fig. (3.9); and iii) laser powers, Fig. (3.10). As before, longer traces, higher concentrations, lower diffusion coefficients, and higher laser powers result, on average, in sharper estimates with the results still converging with at least 2 orders of magnitude fewer photon arrivals than FCS for equal accuracy in Figs. (3.8), (3.9), and Fig. (3.10), respectively. We mention "on average" as individual traces are stochastic. Thus, some traces under higher concentrations of fluorescent molecules may happen to have fewer molecules contribute photons to the traces than experiments with lower concentrations.

Figures (3.8) recapitulates our expectations derived from the synthetic data shown earlier (Fig. (3.5)), where dye concentrations are low yielding a wider posterior for our diffusion coefficient and correspondingly sharper posteriors for the higher concentration. Here, similarly to Fig. (3.1), we compare our method's diffusion coefficient estimate to FCS as a function of the number of photon arrivals used in the analysis, Fig. (3.8A) and Fig. (3.8B), both in good agreement with FCS estimates, produced by the entire traces which is $\approx 10^3$ times longer.

Similar to the analysis of synthetic data, by comparing different diffusion coefficients, the slower a diffusing molecule is, the more time it spends within the confocal volume, the more photons are collected providing us with a sharper posterior estimate of its diffusion coefficient (see Fig. (3.9)).

Similarly to the synthetic data shown earlier (Fig. (3.7)), Fig. (3.10) illustrates the robustness of our method to lower laser power which, as expected, yields a wider posterior for our diffusion coefficient and correspondingly sharper posteriors for higher

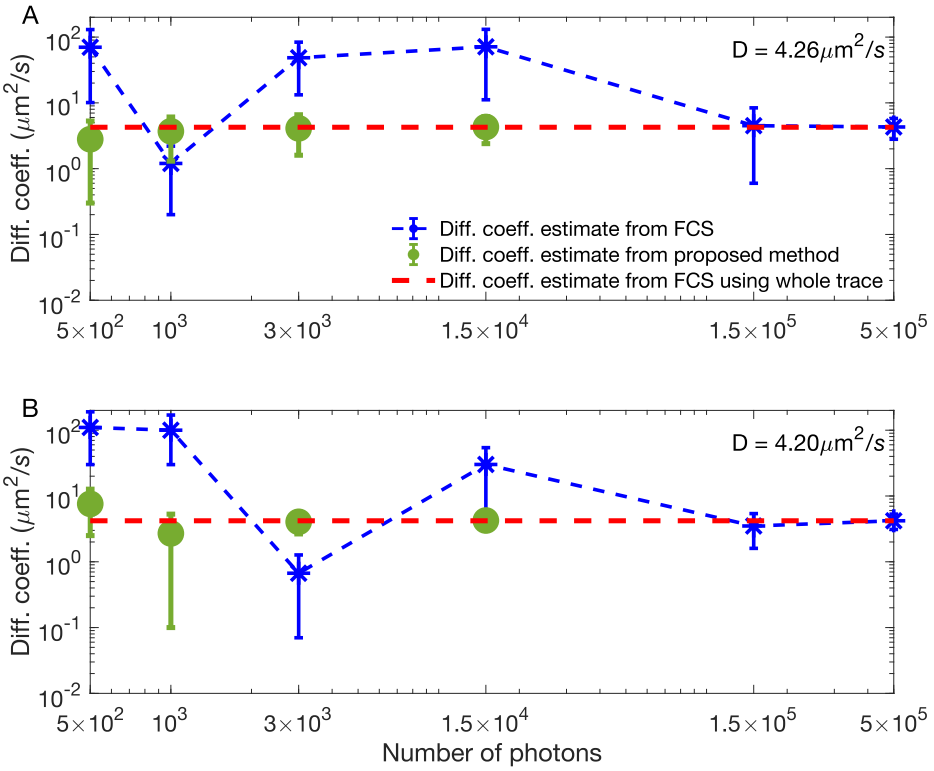

Fig. 3.10. **Higher laser powers in experimental traces provide more photons per unit time and sharper diffusion coefficient estimates**. Estimates shown are drawn from experimental traces with high (100 μW) (A) and low (25 μW) (B) laser power with fixed concentration 1 nM of Cy3 dye molecules and 75% glycerol. Similarly to Fig. (3.1), we compare our method's diffusion coefficient estimate (circle green dots) to FCS (blue asterisk) as a function of the number of photons used in the analysis. Since by 1.5×10^4 photon arrivals our method has converged, we avoid analyzing larger traces. The red dash line is the FCS estimate produced from the entire 5 min trace containing $\approx 3\times10^6$ photon arrivals.

laser power. Here, we compare our method's diffusion coefficient estimate to FCS as a function of the number of photon arrivals used in the analysis, Fig. (3.10A) and Fig. (3.10B), both in good agreement with FCS estimates, produced from the entire trace.

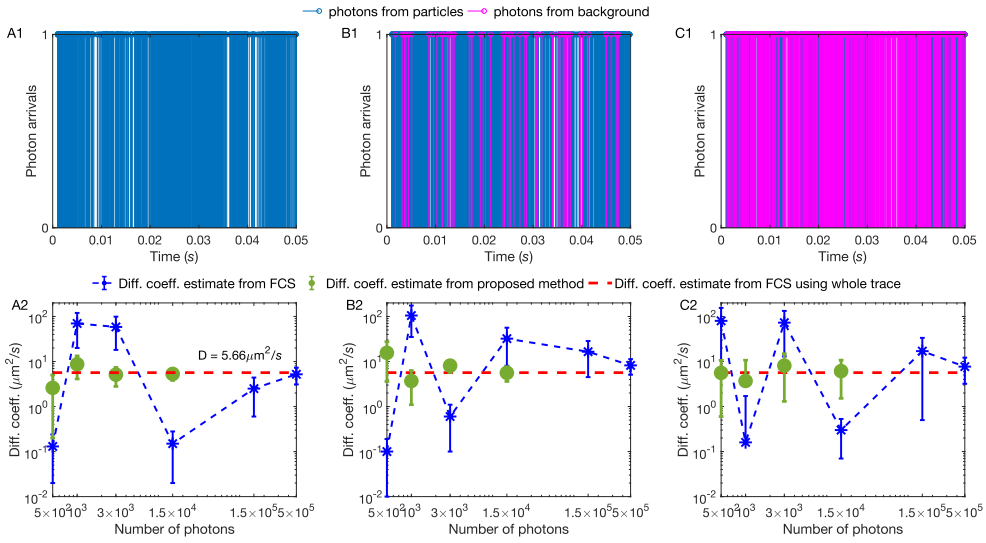

Fig. 3.11. **Background photon emission rates are artificially added to experimental traces yielding challenging imaging conditions and broader diffusion coefficient estimates**. Experimental traces with fixed concentration 1 nM of Cy3 dye molecules and 67% glycerol and fixed laser power 100 μW. The same total number of photons analyzed under differing (artificially increased) background photon emission rates (0 (A1), 500 (B1), 1000 (C1) photons/s). (A2, B2, C2) Similarly to Fig. (3.1), we compare our method's diffusion coefficient estimate (green dots) to FCS (blue asterisk) as a function of the number of photons used in the analysis. Since by 1.5×10^4 photon arrivals our method has converged, we avoid analyzing larger traces. The red dash line is the FCS estimate obtained from the entire, 5 min, trace containing $\approx 3 \times 10^6$ photon arrivals.

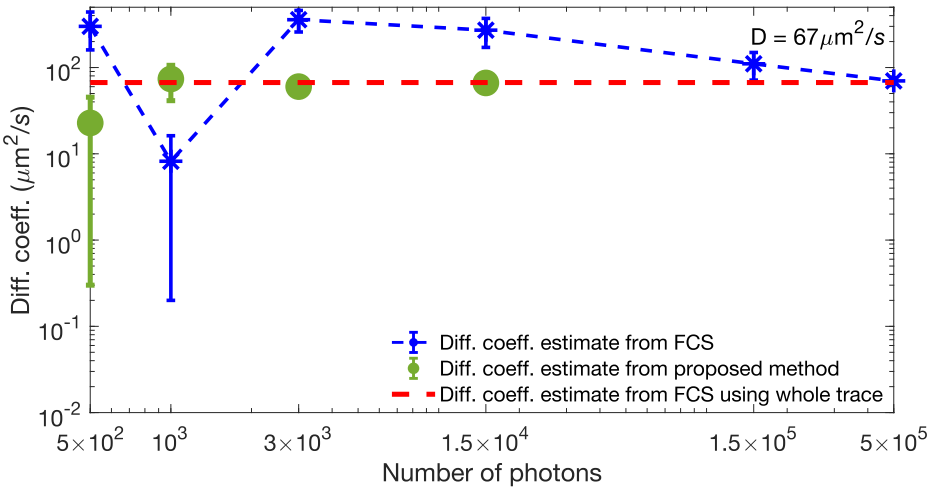

Fig. 3.12. **Diffusion coefficient estimates of labeled protein.** Estimates shown are drawn from experimental traces with fixed concentration 1 nM of Cy3-labeled streptavidin molecules and laser power 100 μW. Similarly to Fig. (3.1), we compare our method's diffusion coefficient estimate (green dots) to FCS (blue asterisk) as a function of the number of photons used in the analysis. Since by 1.5×10^4 photon arrivals our method has converged, we avoid analyzing larger traces. The red dash line is the FCS estimate obtained from the entire, 5 min, trace containing $\approx 3 \times 10^6$ photon arrivals.

As further controls, Fig. (3.11) demonstrates a set of analysis where the background photon emission rate is artificially added to real data. In these cases, we test the limits of our method on more challenging imaging conditons. Furthermore, we repeat our analysis on single photon traces produced by a labeled biomolecule. Specifically, in Fig. (3.12), we use streptavidin proteins labeled with Cy3.

Benchmarking on Experimental Data using 5-TAMRA

Finally, we switch to a different dye, different setup and acquisition electronics as detailed in the Methods section. Our sample contained 20 nM of 5-TAMRA dissolved

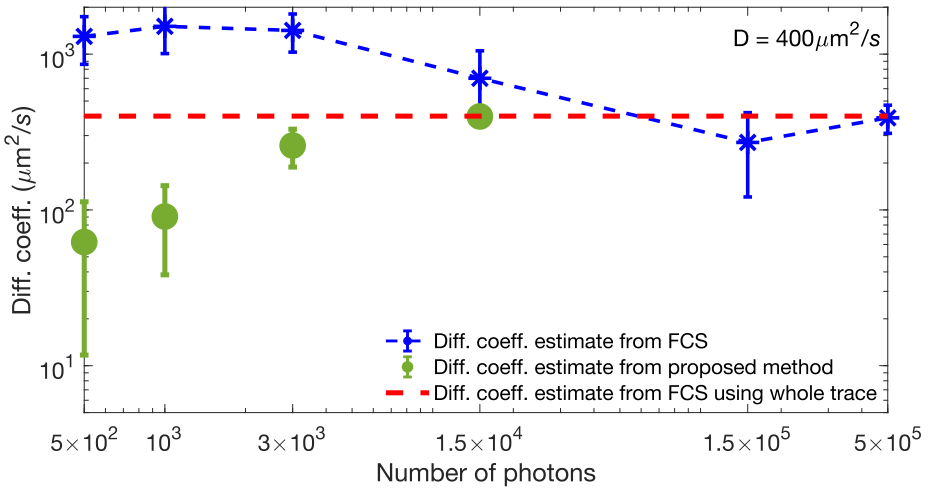

Fig. 3.13. **Diffusion coefficient estimates of 5-TAMRA dye**. Estimates shown are drawn from experimental traces with fixed concentration 20 nM of 5-TAMRA dye molecules. Similarly to Fig. (3.1), we compare our method's diffusion coefficient estimate (green dots) to FCS (blue asterisk) as a function of the number of photons used in the analysis. Since by 1.5×10^4 photon arrivals our method has converged, we avoid analyzing larger traces. The red dash line is the FCS estimate obtained from the entire, 10 min, trace containing $\approx 6\times10^6$ photon arrivals.

in water. As previously, we successfully benchmark our estimates of the diffusion coefficient versus the value obtained from FCS on much longer (≈10 min) traces, see Fig. (3.13).

3.5 Discussion

A single photon arriving at a detector mounted to a confocal microsocope encodes information that reports on the fastest timescale achievable for spectroscopic and imaging applications [335,395]. Directly exploiting this information can help uncover the dynamics of physical or biological systems at fast timescales with accuracy

superior to that obtained from *derived* quantities such as down-sampled intensity traces.

Our method takes a Bayesian nonparametrics (BNPs) approach to tackling single photon arrival data to characterize dynamical quantities from as few as hundreds to thousands of datapoints from confocal imaging. This is by contrast to conventional autocorrelative methods used in FCS [329–332] that require dramatically more data, *i.e.*, datasets several orders of magnitude larger in either total duration or total number of photon arrivals, to characterize dynamical quantities with similar accuracy.

There have been partial solutions to the challenge of interpreting single molecule data at the single photon level often outside FCS applications. Indeed, existing methods make assumptions that render them inapplicable to diffusion through inhomogenesouly illuminated volumes. For example, they assume uniform illumination [347, 349], apply downsampling or binning and thereby reduce temporal resolution to exploit existing mathematical frameworks such as the hidden Markov model [66, 208, 346, 396, 397], or focus on immobile molecules [346, 353–355]. More recently, fluorescence-based nanosecond FCS approaches, in which the data are still correlated under the assumption that the time trace reports on processes at equilibrium, have been used to obtain information on rapid fluctuations in proteins [356]. As such, correlative methods largely continue to dominate confocal data analysis almost half a century beyond their inception [329, 330, 334].

To take full advantage of single photon data, new Mathematics are required. These must treat the inherent non-stationarity between photon arrivals arising due to molecular diffusion in an inhomogeneously illuminated volume and the stochastic number of molecules contributing photons. In particular, analyzing data derived from mobile molecules within an illuminated confocal region breaks down the perennial parametric Bayesian paradigm that has been the workhorse of data analysis [79, 208,

328,345,396,398,399]. We argue here that BNPs–which provide principled extensions of the Bayesian methodology [263, 400]–show promise in Physics [1, 79, 328, 360–362, 401] and give us a working solution to fundamental parametric challenges.

Our new tools open up the possibility to explore at the single photon level non-equilibrium processes resolved on fast timescales [402,403], reaching ms or even below, that have been the focus of recent attention [404]. Moreover, and of immediate relevance for biophysical applications, if a single molecule photobleaches after emitting just a few hundred photons, then our novel method can still provide a diffusion coefficient estimate. Additionally, by analyzing single photon data pointwise, as we do in this study, we obtain a better handle on error bars than analyzing post-processed, such as correlated, data where the error bars can become difficult to compute or interpret [405, 406]. As such, a sharp diffusion coefficient posterior may not only suggest a good estimate of the diffusion coefficient but also suggest that the underlying model, such as normal diffusion, is appropriate and *vice versa* a broad posterior may suggest a poor estimate or an inappropriate motion model.

Furthermore, armed with a transformative framework, founded upon rigorous Statistics, it is now possible to extend the proof-of-principle study to treat effects that lie beyond the current scope of this work. In particular, we can extend our framework to treat multiple color imaging [407], triplet effect and complex molecule photophysics [408] (such as molecular blinking [363, 409] and photobleaching [410, 411]), more complex molecule motion models [77, 412] other than free diffusion [364], distorted or abberated PSF models [413], or even incorporate chemical reactions among the molecules [414, 415]. As our BNP framework explicitly represents the instantaneous position of each involved molecule throughout the experiment's time course, these are extensions that require modest modifications.

Appendix

Additional Analysis Results

In Fig. (3.15) we illustrate the weakness of FCS analysis when applied on limited datasets such as those in the scope of our method. Additionally, using synthetic data, in Fig. (3.16) we estimate diffusion coefficients faster than those in the Results section and in Fig. (3.17) we estimate photon emission rates. Finally, using experimental data of Cy3 and 5-TAMRA dyes obtained as described in the Methods section, in Figs. (3.18) and (3.19), respectively, we benchmark the same estimates on real data.

Detailed Methods Description

Description of Fluorescence Correlation Spectroscopy (FCS)

In FCS the primary quantity of interest is the spontaneously fluctuating fluorescence intensity [416,417]. Correlations in fluorescence intensities are used to determine physical parameters such as diffusion coefficients. The normalized time autocorrelation function of the fluorescence intensity is defined as

$$G(\tau) = \frac{\langle \delta I(t) \delta I(t+\tau) \rangle}{\langle \delta I(t) \rangle^2} = \frac{\langle I(t) I(t+\tau) \rangle}{\langle I(t) \rangle^2} - 1,$$

where $I(t)$ is the fluorescence intensity, $\delta I(t)$ is intensity fluctuations at time t, and τ is the lag time. The intensity fluctuations of the fluorescence intensity are defined as the deviations from the average of the intensity, $\delta I(t) = I(t) - \langle I(t) \rangle$. For freely

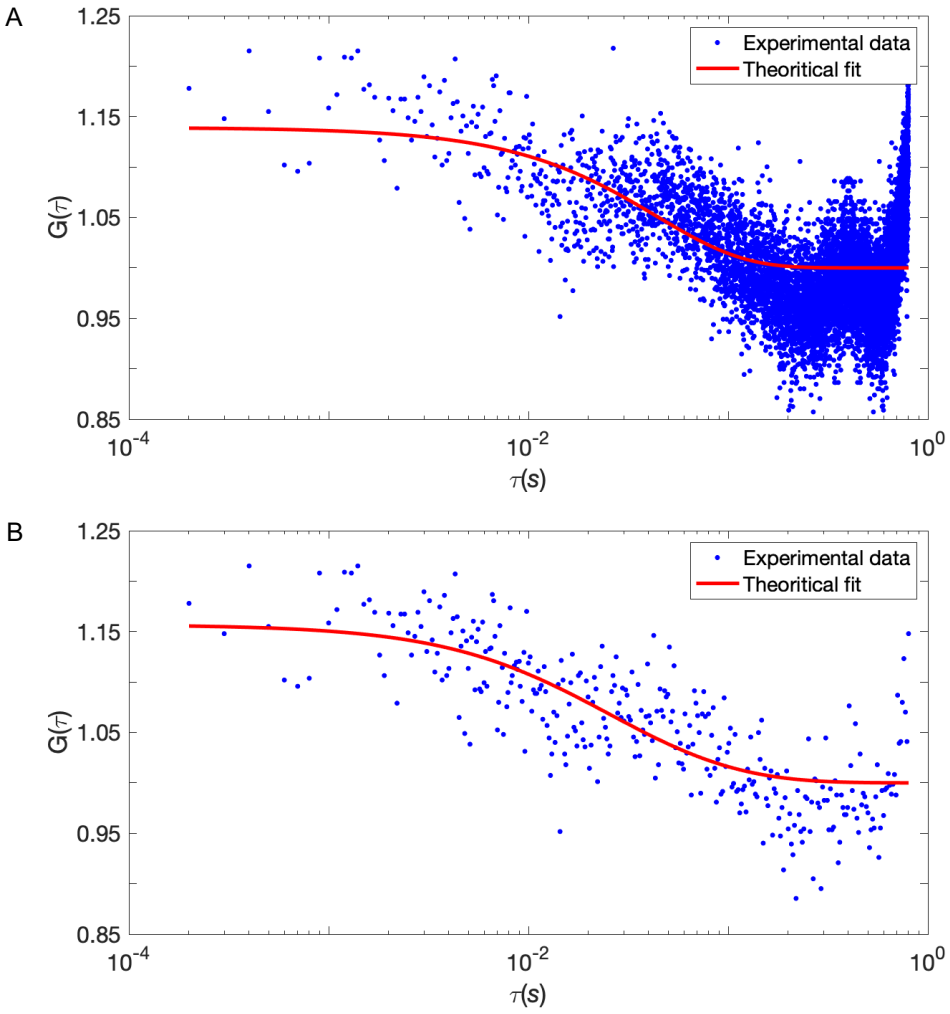

Fig. 3.14. **FCS curves resulting from exceedingly short traces (same synthetic data as Fig. 3.1) with linear (A) and semi-logarithmic (B) binning.** Due to the limited data, the quality of the fitted autocorrelation curve, $G(\tau)$, does not improve considerably for (B) as compared to (A).

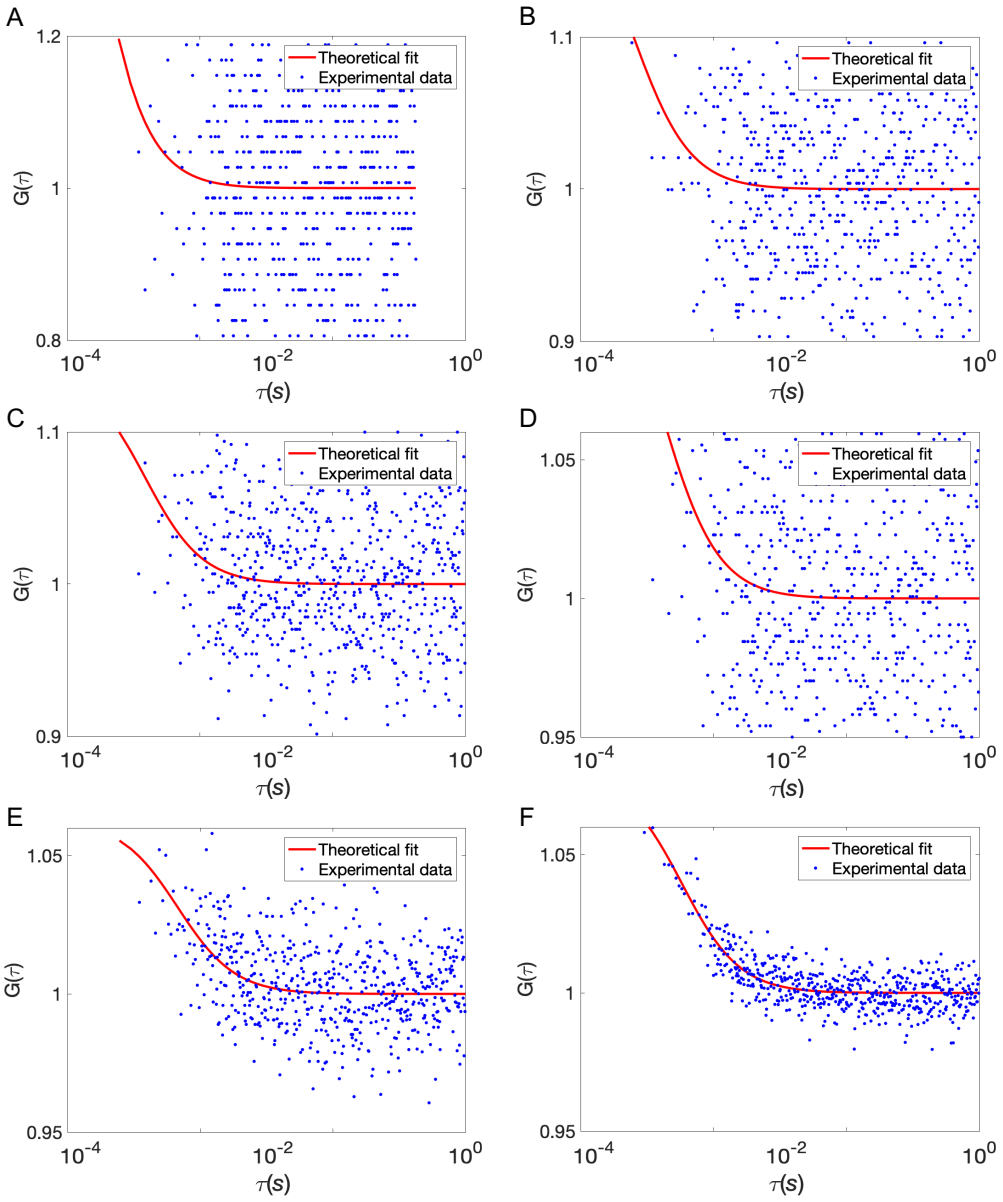

Fig. 3.15. **FCS curves resulting from exceedingly short traces.** Shown are autocorrelation curves, $G(\tau)$, of 5-TAMRA experimental traces, binned at 10 μs, for 100 ms and $\approx$ 500 photon arrivals (A); 200 ms and $\approx$ 1000 photon arrivals (B); 300 ms and $\approx$ 3000 photon arrivals (C); 2 s and $\approx$ 15000 photon arrivals (D); 30 s and $\approx 15 \times 10^5$ photon arrivals (E); 100 s and $\approx 15 \times 10^6$ photon arrivals (F). Even a visual inspection illustrates how poorly FCS applies on traces as sort as those analyzed by our BNP method.

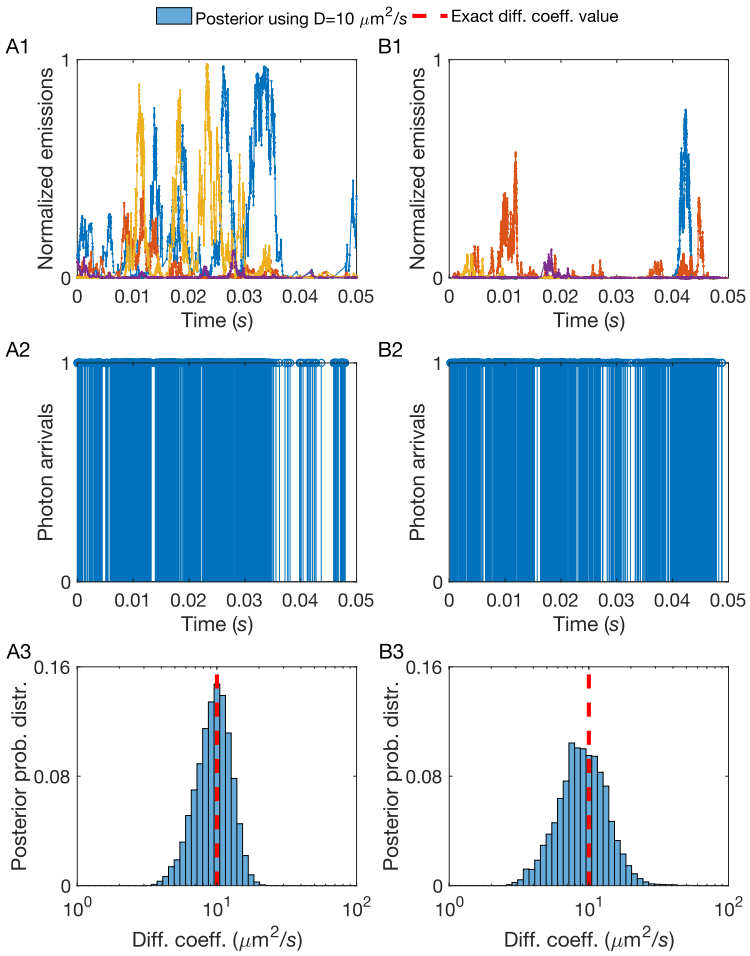

Fig. 3.16. **A larger molecule photon emission rate provides more photons per unit time and sharper diffusion coefficient estimates.** (A1, B1) Instantaneous molecule photon emission rates μ_k^n, normalized by μ_{mol}. (A2, B2) Photon arrival trace resulting from combining photon emissions from every molecule and the background. These traces are produced by 10 molecules diffusing at 10 $\mu m^2/s$ for a total time of 50 ms under background photon emission rate of 10^3 photons/s and molecule photon emission rate 4×10^5 photons/s containing $\approx$ 3000 photon arrivals (A2), and molecule photon emission rate 4×10^4 photons/s containing $\approx$ 2000 photon arrivals (B2). (A3, B3) Posterior probability distributions drawn from traces with differing molecule photon emission rates (shown in (A2, B2)). As expected, for the traces with higher molecule photon emission rate, the peak of the posterior sharply matches with the exact value of D (dashed line). Gradually, as we decrease the molecule photon emission rate, the estimation becomes less reliable.

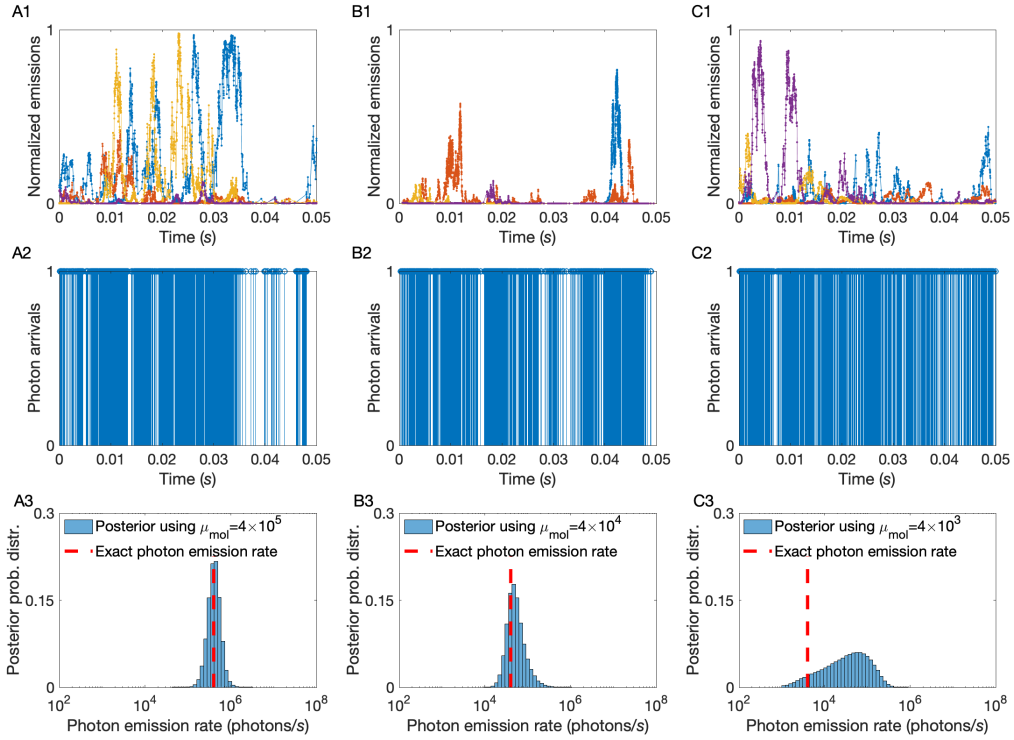

Fig. 3.17. **A higher molecule photon emission rate provides more photons per unit time and sharper emission rate estimates**. (A1, B1, C1) Instantaneous molecule photon emission rates μ_k^n, normalized by μ_{mol}. (A2, B2, C2) Photon arrival traces resulting from combining photon emissions from every molecule and the background. These traces produced by 10 molecules diffusing at 10 $\mu m^2/s$ for a total time of 50 ms under background photon emission rate of 10^3 photons/s and molecule photon emission rate 4×10^5 photons/s containing ≈ 3000 photon arrivals (A2), molecule photon emission rate 4×10^4 photons/s containing ≈ 2000 photon arrivals (B2), and molecule photon emission rate 4×10^3 photons/s containing ≈ 1000 photon arrivals (C2). (A3, B3, C3) Posterior probability distributions drawn from traces with differing molecule photon emission rates (shown in (A2, B2, C2)). As expected, for the traces with higher molecule photon emission rate, the peak of the posterior sharply matches with the exact value of μ_{mol} (dashed line). Gradually, as we decrease the molecule photon emission rate, the estimation becomes less reliable.

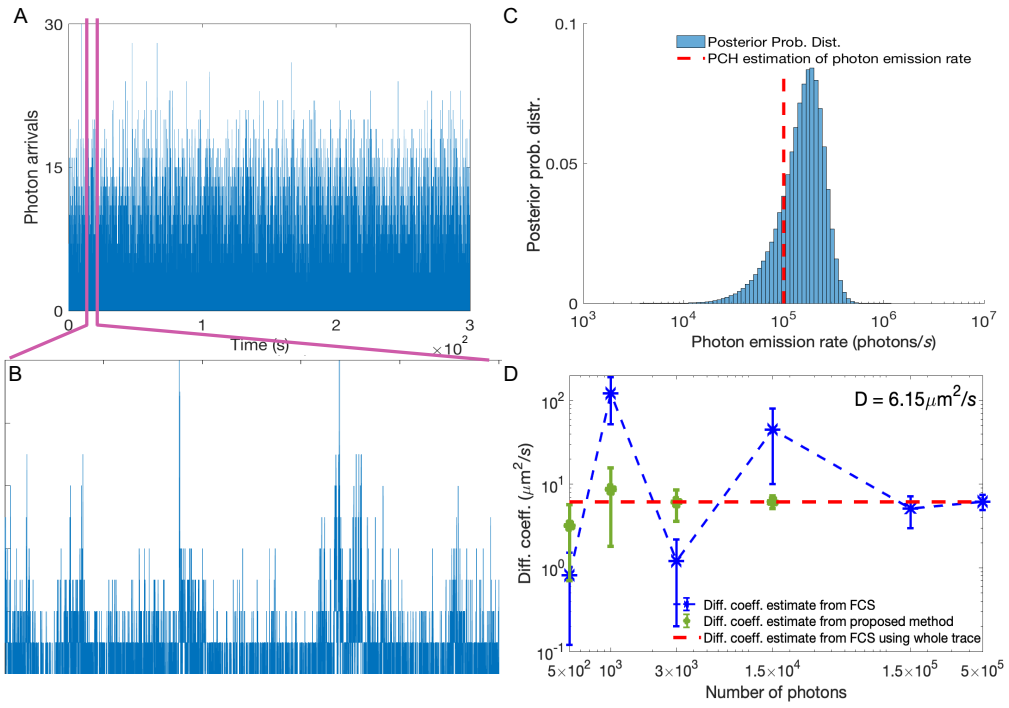

Fig. 3.18. **Estimation of the diffusion coefficient and molecule photon emission rate for Cy3 dyes**. (A) Experimental intensity trace (binned at 100 μs) with concentration 1 nM of Cy3 dye molecules and 61% glycerol. A background photon emission rate of 600 photons/s is known from calibration. (B) Analyzed portion of the trace containing $\approx$ 3000 photon arrivals. (C) Posterior probability distributions and the value (red dash line) of molecule photon emission rate determined by the photon counting histogram (PCH) method on the entire trace [374]. (D) Similarly to Fig. (3.1), we compare our method's diffusion coefficient estimate (green dots) to FCS (blue asterisk) as a function of the number of photons used in the analysis. Since by 1.5×10^4 photon arrivals our method has converged, we avoid analyzing larger traces. The red dash line is the FCS estimate obtained from the entire, 5 min, trace containing $\approx 3 \times 10^6$ photon arrivals.

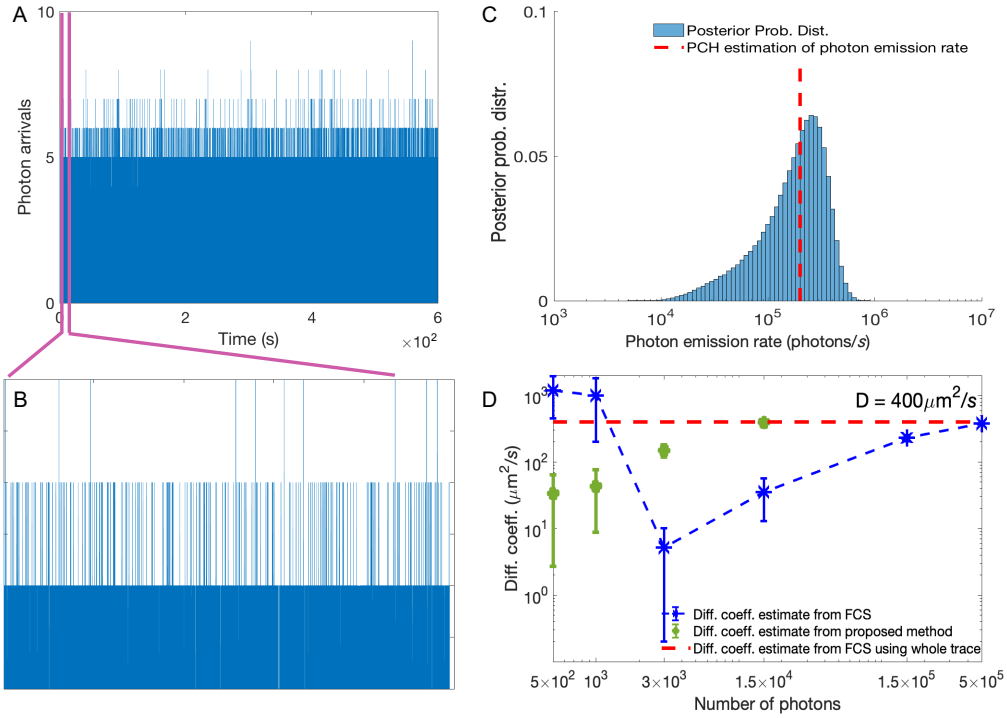

Fig. 3.19. **Estimation of the diffusion coefficient and molecule photon emission rate for 5-TAMRA dyes**. (A) Experimental intensity trace (binned at 10 μs) with concentration 20 nM of 5-TAMRA dye molecules. A background photon emission rate of 300 photons/s is known from calibration. (B) Analyzed portion of the trace containing $\approx$ 8000 photon arrivals. (C) Posterior probability distributions and the value (red dash line) of molecule photon emission rate determined by the photon counting histogram (PCH) method on the entire trace [374]. (D) Similarly to Fig. (3.1), we compare our method's diffusion coefficient estimate (green dots) to FCS (blue asterisk) as a function of the number of photons used in the analysis. Since by 1.5×10^4 photon arrivals our method has converged, we avoid analyzing larger traces. The red dash line is the FCS estimate obtained from the entire, 10 min, trace containing $\approx 6\times10^6$ photon arrivals.

diffusing molecules in a 3D Gaussian confocal volume, the autocorrelation, which we use in this study, is

$$G(\tau) = \frac{1}{\langle N \rangle} \frac{1}{1 + \frac{4D\tau}{\omega_{xy}^2}} \frac{1}{\sqrt{1 + \frac{4D\tau}{\omega_z^2}}},$$

where $\langle N \rangle$ is the average number of molecules in the confocal volume, D is the diffusion coefficient, ω_{xy} and ω_z are the confocal volume axes along the xy and z directions. Further details on correlative analysis are contained in the cited literature [329, 330, 337, 338, 416, 417].

Explanation of Data Simulation

To generate synthetic traces we simulate molecules moving through a three dimensional illuminated volume. The number of moving molecules, N, is predefined in each simulation. We apply periodic boundaries to our volume of L_{xy} and L_z parallel to the focal plane and optical axis, respectively, to keep a relatively stable concentration of molecules near the confocal volume.

We denote the locations of the molecules as x_k^n, y_k^n and z_k^n, where k labels time levels and $n = 1, 2, \ldots, N$ labels molecules. The total trace duration $T_{total} = t_K - t_0$, is predefined. The time intervals between successive recorded photons $\Delta t_{k-1} = t_k - t_{k-1}$, are generated through pseudo-random computer simulations and recorded for subsequent analysis.

The locations of the molecules x_0^n, y_0^n, z_0^n at the first evaluation time t_0 are randomly sampled from the uniform distribution with borders identical to the boundaries $\pm L_{xy}$ and $\pm L_z$ of the prescribed simulation region. Locations x_k^n, y_k^n, z_k^n, for $k = 1, \ldots, K$, at times t_k are generated according to the diffusion model explained above under a predefined diffusion coefficient D.

We obtain photon inter-arrival times, $\boldsymbol{\Delta t} = (\Delta t_1, \Delta t_2, \ldots, \Delta t_{K-1})$, by simulating exponential random variables of rate μ_k. For independent background and molecule photon emission rates, the corresponding exponential emission mean rates μ_k depend on a Gaussian PSF as eqs. (3.2)–(3.3). Both background, μ_{back}, and the molecule photon emission rate, μ_{mol}, are predefined.

Definition of Molecule Photon Emission Rate

In this study the emission rate of detected photons for a single fluorophore at position x, y, z is used. This is formulated as the product $\mu(x, y, z) = \mu_0 \varphi_d \varphi_{de} \varphi_f \sigma \times \mathrm{EXC}(x, y, z)\mathrm{CEF}(x, y, z)$. Here, μ_0 and φ_d are the maximum excitation intensity and the efficiency of the photon collection at the center of the confocal volume, respectively, φ_{de} is the efficiency of the detector, φ_f is the quantum efficiency of the fluorophore, σ is the fluorophore absorption cross-section, $\mathrm{EXC}(x, y, z)$ is the excitation profile and $\mathrm{CEF}(x, y, z)$ is the detection profile [418]. By revising the definition of $\mu(x, y, z)$, we obtain $\mu(x, y, z) = \mu_{mol}\mathrm{PSF}(x, y, z)$ where $\mu_{mol} = \mu_0 \varphi_d \varphi_{de} \varphi_f \sigma$ and $\mathrm{PSF}(x, y, z) = \mathrm{EXC}(x, y, z)\mathrm{CEF}(x, y, z)$.

To relate our *single molecule* photon emission rate μ_{mol} to the average photon count rate typically determined in *bulk* experiments, we compute a spatial average

$$
\begin{aligned}
\langle \mu(x, y, z) \rangle &= \mu_{mol} \langle \mathrm{PSF}(x, y, z) \rangle = \mu_{mol} \left\langle \exp\left(-2\frac{x^2 + y^2}{\omega_{xy}^2} - 2\frac{z^2}{\omega_z^2} \right) \right\rangle \\
&= \mu_{mol} \frac{\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \exp\left(-2\frac{x^2}{\omega_{xy}^2} - 2\frac{y^2}{\omega_{xy}^2} - 2\frac{z^2}{\omega_z^2} \right) dx dy dz}{V} \\
&= \mu_{mol} \sqrt{\frac{\pi}{2}\omega_{xy}^2} \sqrt{\frac{\pi}{2}\omega_{xy}^2} \sqrt{\frac{\pi}{2}\omega_z^2} \frac{1}{V}
\end{aligned}
$$

where V denotes our PSF's effective volume [335, 336] which is equal to $V = \pi^{\frac{3}{2}}\omega_{xy}^2\omega_z$. As a result, our molecule photon emission rate μ_{mol} is related to $\langle\mu(x,y,z)\rangle$ according to $\mu_{mol} = \sqrt{8}\ \langle\mu(x,y,z)\rangle$.

Description of Wilson-Hiferty Approximation

To perform the necessary computations of the next section, we use a Wilson-Hiferty transform [419] to approximate the probability density of exponential random variables. We use this approximation to sample the locations of the molecules within our overall Gibbs sampler (see next).

To apply the Wilson-Hiferty approximation, first we transform our observation random variable Δt_k to a new random variable ρ_k, where $\rho_k = 2\mu_k\Delta t_k$. A change of variables, indicates that $\Delta t_k|\mu_k \sim \text{Exponential}(\mu_k)$ implies $\rho_k|\mu_k \sim \chi^2(2)$, where $\chi^2(2)$ denotes the chi-square probability distribution with 2 degrees of freedom. By applying another transformation, where $\xi_k = \sqrt[3]{\rho_k/2}$, according to [419], ξ_k follows an *approximately* normal probability distribution $\xi_k \sim \text{Normal}\left(\frac{8}{9}, \frac{1}{9}\right)$. So, by $\xi_k = \sqrt[3]{\rho_k/2}$ and $\rho_k = 2\mu_k\Delta t_k$, we conclude $\sqrt[3]{\Delta t_k} = \xi_k/\sqrt[3]{\mu_k}$. Therefore, since $\sqrt[3]{\Delta t_k} = \xi_k/\sqrt[3]{\mu_k}$, we establish the approximation

$$\begin{aligned} p\left(\sqrt[3]{\Delta t_k}|\mu_k\right) &= p\left(\xi_k = \sqrt[3]{\Delta t_k}\sqrt[3]{\mu_k}\right)\sqrt[3]{\mu_k} \\ &\approx \sqrt[3]{\mu_k}\ \text{Normal}\left(\sqrt[3]{\Delta t_k}\sqrt[3]{\mu_k}; \frac{8}{9}, \frac{1}{9}\right) \\ &= \text{Normal}\left(\sqrt[3]{\Delta t_k}; \frac{8}{9\sqrt[3]{\mu_k}}, \left(\frac{1}{3\sqrt[3]{\mu_k}}\right)^2\right). \end{aligned}$$

Detailed Description of the Inference Framework

Prior Probability Distributions

Within the Bayesian paradigm, all unknown model parameters need priors. These parameters are: the diffusion coefficient D; the molecule photon emission rate μ_{mol}; the initial molecule locations x_1^n,y_1^n,z_1^n; as well as the indicator prior weights q^n.

Prior on the Diffusion Coefficient To make sure that D sampled in our formulation attains only positive values, we choose an Inverse-Gamma prior

$$D \sim \text{InvGamma}\left(\alpha_D, \beta_D\right). \tag{3.13}$$

This prior is conjugate to the motion model which simplifies the computations shown below.

Priors on Molecule Photon Emission Rate To guarantee that μ_{mol} sampled in our formulation also attains only positive values, we choose a Gamma prior

$$\mu_{mol} \sim \text{Gamma}\left(\alpha_{mol}, \beta_{mol}\right). \tag{3.14}$$

Priors on Initial Molecule Locations Because of the symmetries inherent to the confocal volume, *e.g.*, a molecule at a location (x, y, z) gives rise to the same photon emission rate as a molecule at location, $(-x, -y, -z)$, we use priors on the

initial locations that respect these symmetries. To simplify the computations, we use independent symmetric normal distributions

$$x_1^n \sim \text{SymNormal}\left(\mu_{xy}, \sigma_{xy}^2\right), \tag{3.15}$$

$$y_1^n \sim \text{SymNormal}\left(\mu_{xy}, \sigma_{xy}^2\right), \tag{3.16}$$

$$z_1^n \sim \text{SymNormal}\left(\mu_z, \sigma_z^2\right). \tag{3.17}$$

Priors and Hyperpriors for the Indicators To simplify the computations described in the next section, we use a finite, but large, model population of N molecules that contain contributing and noncontributing ones. These molecules are collectively indexed by $n = 1, 2, \ldots, N$. As described in the Methods section, inferring how many molecules are actually warranted by the data analyzed is the same as estimating how many of those N molecules are active, *i.e.*, $b^n = 1$, while the rest are inactive, *i.e.*, $b^n = 0$, and so have no influence and are applied just for computational reasons.

We use a Bernoulli prior of weight q^n to make sure that each indicator b^n takes only values 0 or 1. Moreover, on each weight q^n, we assign a beta hyperprior

$$b^n | q^n \sim \text{Bernoulli}\left(q^n\right), \tag{3.18}$$

$$q^n \sim \text{Beta}\left(A_q, B_q\right). \tag{3.19}$$

To make sure that the resulting formulation avoids overfitting, we make the specific selections $A_q = \alpha_q/N$ and $B_q = \beta_q(N-1)/N$. For these choices [365–367, 420], and in the limit that $N \to \infty$ (that is, when the assumed molecule population is large), this prior/hyperprior choice converges to a non-parametric beta-Bernoulli process. Therefore, for $N \gg 1$, the posterior is well defined and becomes *independent* of the selected value of N. In other words, provided N is large enough, its effect on the results is negligible; while its precise value has only computational implications.

Description of the Computational Implementation

Here, $p(D, \mu_{mol}, \{q^n, b^n, \overline{x}^n, \overline{y}^n, \overline{z}^n\}_n | \boldsymbol{\Delta t})$ is the joint probability distribution of our framework where molecular trajectories and measurements are gathered in

$$\begin{aligned}
\overline{x}^n &= (x_1^n, x_2^n, \ldots, x_K^n), \\
\overline{y}^n &= (y_1^n, y_2^n, \ldots, y_K^n), \\
\overline{z}^n &= (z_1^n, z_2^n, \ldots, z_K^n), \\
\boldsymbol{\Delta t} &= (\Delta t_1, \Delta t_2, \ldots, \Delta t_{K-1}).
\end{aligned}$$

Posterior samples are generated according to Gibbs sampling [79,358,359,381,421]. We achieve this by sampling a variable conditioned on all other variables and the given photon inter-arrival times $\boldsymbol{\Delta t}$. Conceptually, the steps in the generation of each posterior sample $(D, \mu_{mol}, \{q^n, b^n, \overline{x}^n, \overline{y}^n, \overline{z}^n\}_n)$ are:

1. For each n of the *active* molecules
   (a) Update trajectory $\overline{x}^n$ of active molecule n
   (b) Update trajectory $\overline{y}^n$ of active molecule n
   (c) Update trajectory $\overline{z}^n$ of active molecule n
2. Update jointly the trajectories $\overline{x}^n, \overline{y}^n, \overline{z}^n$ for all n of the *inactive* molecules
3. Update the diffusion coefficient D
4. Update jointly the prior weights q^n for all model molecules and simultaneously update jointly the indicators b^n for all model molecules
5. Update the molecule photon emission rate μ_{mol}

Sampling Active Molecules Locations To sample the location of an active molecule $(\overline{x}^n, \overline{y}^n, \overline{z}^n)$, we use forward filtering and backward sampling [364, 422–424].

In particular, we update each dimension sequentially from the following full conditional probability distributions $p(\overline{x}^n|D, \mu_{mol}, \{b^{n'}, \overline{y}^{n'}, \overline{z}^{n'}\}_{n'}, \{\overline{x}^{n'}\}_{n'\neq n}, \boldsymbol{\Delta t})$, $p(\overline{y}^n|D, \mu_{mol}, \{b^{n'}, \overline{x}^{n'}, \overline{z}^{n'}\}_{n'}, \{\overline{y}^{n'}\}_{n'\neq n}, \boldsymbol{\Delta t})$, and $p(\overline{z}^n|D, \mu_{mol}, \{b^{n'}, \overline{x}^{n'}, \overline{y}^{n'}\}_{n'}, \{\overline{z}^{n'}\}_{n'\neq n}, \boldsymbol{\Delta t})$. Below, we show in detail the calculation only for sampling $\overline{x}^n$, since for sampling $\overline{y}^n$ and $\overline{z}^n$ they are similar.

To sample the trajectory $\overline{x}^n$, we rely on the factorization

$$
\begin{aligned}
& p(\overline{x}^n|D, \mu_{mol}, \{b^{n'}, \overline{y}^{n'}, \overline{z}^{n'}\}_{n'}, \{\overline{x}^{n'}\}_{n'\neq n}, \boldsymbol{\Delta t}) \\
&= p(x_K^n|D, \mu_{mol}, \{b^{n'}, \overline{y}^{n'}, \overline{z}^{n'}\}_{n'}, \{\overline{x}^{n'}\}_{n'\neq n}, \boldsymbol{\Delta t}) \\
&\times p(x_{K-1}^n|x_K^n, D, \mu_{mol}, \{b^{n'}, \overline{y}^{n'}, \overline{z}^{n'}\}_{n'}, \{\overline{x}^{n'}\}_{n'\neq n}, \boldsymbol{\Delta t}) \\
&\times \cdots \\
&\times p(x_2^n|x_1^n, D, \mu_{mol}, \{b^{n'}, \overline{y}^{n'}, \overline{z}^{n'}\}_{n'}, \{\overline{x}^{n'}\}_{n'\neq n}, \boldsymbol{\Delta t}) \\
&\times p(x_1^n|D, \mu_{mol}, \{b^{n'}, \overline{y}^{n'}, \overline{z}^{n'}\}_{n'}, \{\overline{x}^{n'}\}_{n'\neq n}, \boldsymbol{\Delta t})
\end{aligned}
$$

and, according to this factorization, we sample individual locations x_k^n sequentially

$$
\begin{aligned}
x_K^n &\sim p(x_K^n|D, \mu_{mol}, \{b^{n'}, \overline{y}^{n'}, \overline{z}^{n'}\}_{n'}, \{\overline{x}^{n'}\}_{n'\neq n}, \boldsymbol{\Delta t}) \\
x_k^n &\sim p(x_k^n|x_{k+1}^n, D, \mu_{mol}, \{b^{n'}, \overline{y}^{n'}, \overline{z}^{n'}\}_{n'}, \{\overline{x}^{n'}\}_{n'\neq n}, \boldsymbol{\Delta t}),
\end{aligned}
$$

where, $k = 1, \ldots, K-1$. However, to be able to perform these steps, we first need to compute the involved probability distributions. We describe below a computationally efficient way to do so that proceeds in a forward filtering and a backward sampling step.

Before we start the sampling of the locations, we determine each one of the individual probability distributions that are needed. To do this in a computationally tractable manner [424,425], we compute filter distributions $p(x_k^n|D, \mu_{mol}, \{b^{n'}, \overline{y}^{n'}, \overline{z}^{n'}\}_{n'}$

$, \{\overline{x}^{n'}\}_{n' \neq n}, \{\Delta t_{k'}\}_{k'<k})$. In our case, both dynamic (eqs. (3.7)–(3.9))) and observation (eq. (3.1)) probability distributions provide equal probabilities for $+x_k^n$ and $-x_k^n$. Therefore, the filter distribution consists of two modes symmetrically placed across the origin [364]. Accordingly, we compute an *approximate* bimodal symmetric filter of the form

$$\begin{aligned} & p\left(x_k^n | D, \mu_{mol}, \{b^{n'}, \overline{y}^{n'}, \overline{z}^{n'}\}_{n'}, \{\overline{x}^{n'}\}_{n' \neq n}, \{\Delta t_{k'}\}_{k'<k}\right) \\ & \approx \text{SymNormal}\left(x_k^n; m_k^n, c_k^n\right) \end{aligned}$$

where $\text{SymNormal}\left(m_k^n, c_k^n\right)$ describes the symmetric normal distribution. The filter, that is the values of m_k^n and c_k^n, is updated iteratively according to

$$\begin{aligned} & p\left(x_k^n | D, \mu_{mol}, \{b^{n'}, \overline{y}^{n'}, \overline{z}^{n'}\}_{n'}, \{\overline{x}^{n'}\}_{n' \neq n}, \{\Delta t_{k'}\}_{k'<k}\right) \\ & \propto p\left(\Delta t_{k-1} | x_k^n, y_k^n, z_k^n, \mu_{mol}, \{b^{n'}, x^{n'}, y^{n'}, z^{n'}\}_n'\right) \\ \times & \int_{x_{k-1}^n} p\left(x_{k-1}^n | D, \mu_{mol}, \{b^{n'}, \overline{y}^{n'}, \overline{z}^{n'}\}_{n'}, \{\overline{x}^{n'}\}_{n' \neq n}, \{\Delta t_{k'}\}_{k'<k-1}\right) p\left(x_k^n | x_{k-1}^n, D\right) dx_{k-1}^n. \end{aligned} \tag{3.20}$$

To be able to carry out these computations efficiently, similar to [364], we work on an approximate model where the exponential emission equation, eq. (3.1), is replaced by a normal one using the Wilson-Hiferty approximation as we discussed earlier. Our *approximate* emission equation is

$$\begin{aligned} & T_{data}\left(\Delta t_k\right) | \{x_k^n, y_k^n, z_k^n, b^n\}_n, \mu_{mol} \\ & \sim \text{Normal}\left(T_{mean}(\mu_k), S(\mu_k)^2\right), \quad k = 1, \ldots, K-1. \end{aligned}$$

where μ_k is given in eq. (3.2); while $T_{data}\left(\Delta t_k\right)$, $T_{mean}(\mu_k)$ and $S^2(\mu_k)$ are given by the Wilson-Hiferty approximation [419]. As explained earlier, the approximation is given

by $T_{data}(\Delta t_k) = \Delta t_k^{1/3}$, $T_{mean}(\mu_k) = 8/(9\mu_k^{1/3})$, and $S^2(\mu_k) = 1/(9\mu_k^{2/3})$. Because of the specific choices of our problem (*i.e.,* diffusive molecules, symmetric normal filter at the proceeding time, and normal likelihood), eq. (3.20) reduces to

$$\begin{aligned} p\left(x_k^n | D, \mu_{mol}, \{b^{n'}, \overline{y}^{n'}, \overline{z}^{n'}\}_{n'}, \{\overline{x}^{n'}\}_{n' \neq n}, \{\Delta t_{k'}\}_{k'<k}\right) \\ = \text{Normal}\left(T_{data}(\Delta t_k); T_{mean}(\mu_k), S(\mu_k)^2\right) \\ \times \text{SymNormal}\left(x_k^n; m_{k-1}^n, c_{k-1}^n + 2D\Delta t_k\right). \end{aligned} \tag{3.21}$$

Finally, to obtain the values of m_k^n and c_k^n, we linearize the product in eq. (3.21) as described next. From eq. (3.21), we have

$$\begin{aligned} &\text{Normal}\left(T_{data}(\Delta t_k); T_{mean}(\mu_k), S(\mu_k)^2\right) \times \text{SymNormal}\left(x_k^n; m_{k-1}^n, c_{k-1}^n + 2D\Delta t_k\right) \\ &\propto \exp\left(\frac{\log \mu(x_k^n)}{3} - \frac{\left(\frac{8}{9} - \sqrt[3]{\mu(x_k^n)\Delta t_{k-1}}\right)^2}{\frac{2}{9}} - \frac{(x_k^n - m_{k-1}^n)^2}{2(c_{k-1}^n + 2D\Delta t_{k-1})}\right) \\ &+ \exp\left(\frac{\log \mu(x_k^n)}{3} - \frac{\left(\frac{8}{9} - \sqrt[3]{\mu(x_k^n)\Delta t_{k-1}}\right)^2}{\frac{2}{9}} - \frac{(x_k^n + m_{k-1}^n)^2}{2(c_{k-1}^n + 2D\Delta t_{k-1})}\right). \end{aligned} \tag{3.22}$$

The density in eq. (3.22) consists of two modes, one on the positive semi-axis of x and one on the negative semi-axis of x. Considering $f(x_k^n) = \sqrt[3]{\mu(x_k^n)\Delta t_{k-1}}$ and

$g(x_k^n) = \frac{1}{3}\log\mu(x_k^n)$ and linearizing them around the previous filter's mode, $+m_{k-1}^n$ or $-m_{k-1}^n$, the modes of eq. (3.22) are approximated by

$$\exp\left(\frac{\log\mu(x_k^n)}{3} - \frac{\left(\frac{8}{9} - \sqrt[3]{\mu(x_k^n)\Delta t_{k-1}}\right)^2}{\frac{2}{9}} - \frac{(x_k^n - m_{k-1}^n)^2}{2(c_{k-1}^n + 2D\Delta t_{k-1})}\right)$$
$$\approx \exp\left(v_1 x_k^n - \frac{(x_k^n - h_1)^2}{\frac{2\sigma^2}{f'(-m_{k-1}^n)^2}} - \frac{(x_k^n - m_{k-1}^n)^2}{2(c_{k-1}^n + 2D\Delta t_{k-1})}\right),$$
$$\exp\left(\frac{\log\mu(x_k^n)}{3} - \frac{\left(\frac{8}{9} - \sqrt[3]{\mu(x_k^n)\Delta t_{k-1}}\right)^2}{\frac{2}{9}} - \frac{(x_k^n + m_{k-1}^n)^2}{2(c_{k-1}^n + 2D\Delta t_{k-1})}\right)$$
$$\approx \exp\left(v_2 x_k^n - \frac{(x_k^n - h_2)^2}{\frac{2\sigma^2}{f'(-m_{k-1}^n)^2}} - \frac{(x_k^n + m_{k-1}^n)^2}{2(c_{k-1}^n + 2D\Delta t_{k-1})}\right).$$

Combining both approximations, the density of eq. (3.22), is approximated by

$$\exp\left(\frac{\log\mu(x_k^n)}{3} - \frac{\left(\frac{8}{9} - \sqrt[3]{\mu(x_k^n)\Delta t_{k-1}}\right)^2}{\frac{2}{9}} - \frac{(x_k^n - m_{k-1}^n)^2}{2(c_{k-1}^n + 2D\Delta t_{k-1})}\right)$$
$$+\exp\left(\frac{\log\mu(x_k^n)}{3} - \frac{\left(\frac{8}{9} - \sqrt[3]{\mu(x_k^n)\Delta t_{k-1}}\right)^2}{\frac{2}{9}} - \frac{(x_k^n + m_{k-1}^n)^2}{2(c_{k-1}^n + 2D\Delta t_{k-1})}\right)$$
$$\approx \exp\left(v_1 x_k^n - \frac{(x_k^n - h_1)^2}{\frac{2\sigma^2}{f'(+m_{k-1}^n)^2}} - \frac{(x_k^n - m_{k-1}^n)^2}{2(c_{n-1}^k + 2D\Delta t_{k-1})}\right)$$
$$+\exp\left(v_2 x_k^n - \frac{(x_k^n - h_2)^2}{\frac{2\sigma^2}{f'(-m_{k-1}^n)^2}} - \frac{(x_k^n + m_{k-1}^n)^2}{2(c_{k-1}^n + 2D\Delta t_{k-1})}\right) \tag{3.23}$$

where $h_1 = \frac{\frac{8}{9} - f(+m_{k-1}^n) + m_{k-1}^n f'(+m_{k-1}^n)}{f'(+m_{k-1}^n)}$, $v_1 = \frac{\mu'(+m_{k-1}^n)}{3\mu(+m_{k-1}^n)}$, and $h_2 = \frac{\frac{8}{9} - f(-m_{k-1}^n) - m_{k-1}^n f'(-m_{k-1}^n)}{f'(-m_{k-1}^n)}$, $v_2 = \frac{\mu'(-m_{k-1}^n)}{3\mu(-m_{k-1}^n)}$.

Equation (3.23) describes a symmetric normal distribution. Equating this distribution with our filter, *i.e.*, SymNormal (m_k^n, c_k^n), we obtain $c_k^n = \left(\frac{1}{c_{k-1}^n + 2D\Delta t_{k-1}} + \frac{9}{2} f'(m_{k-1}^n)^2\right)^{-1}$, and $m_k^n = \left(v_1 + 9h_1 f'(m_{k-1}^n)^2 + \frac{m_{k-1}^n}{c_{k-1}^n + 2D\Delta t_{k-1}}\right) c_k^n$. These apply for $k = 2, \ldots, K$ and are used to update the filter. To begin, we use eq. (3.15)–(3.17) and set $c_1^n = \sigma_{xy}^2$ and $m_1^n = \mu_{xy}$.

Having computed the filter distributions above, we are able to sample the individual locations by starting from x_K^n and moving backward towards x_1^n. In particular, by applying the Bayes' rule, each one of the individual distributions factorize as

$$p(x_k^n | x_{k+1}^n, D, \mu_{mol}, \{b^{n'}, \overline{y}^{n'}, \overline{z}^{n'}\}_{n'}, \{\overline{x}^{n'}\}_{n' \neq n}, \boldsymbol{\Delta t}) \propto \tag{3.24}$$
$$p\left(x_k^n | D, \mu_{mol}, \{b^{n'}, \overline{y}^{n'}, \overline{z}^{n'}\}_{n'}, \{\overline{x}^{n'}\}_{n' \neq n}, \{\Delta t_{k'}\}_{k' < k}\right) \times p\left(x_k^n | x_{k+1}^n, D\right).$$

The first term is given by the filter distribution which is replaced by our approximate SymNormal(m_k^n, c_k^n), and the second term is our motion model Normal $(x_k^n, 2D\Delta t_{k-1})$, all of which are known at this stage. Therefore, backward sampling starts at x_K^n and continues for x_{k-1}^n with

$$x_K^n \sim \frac{1}{2}\text{Normal}\left(+m_K^n, c_K^n\right) + \frac{1}{2}\text{Normal}\left(-m_K^n, c_K^n\right)$$
$$x_{k-1}^n \sim \frac{\exp\left[\frac{\left(x_k^n - m_k^n\right)^2}{c_k^n + 2D\Delta t_{k-1}}\right]}{2\sqrt{2\pi(c_k^n + 2D\Delta t_{k-1})}}\text{Normal}\left(\frac{x_k^n c_k^n + 2D\Delta t_{k-1} m_k^n}{c_k^n + 2D\Delta t_{k-1}}, \frac{2D\Delta t_{k-1} c_k^n}{c_k^n + 2D\Delta t_{k-1}}\right)$$
$$+ \frac{\exp\left[\frac{\left(x_k^n + m_k^n\right)^2}{c_k^n + 2D\Delta t_{k-1}}\right]}{2\sqrt{2\pi(c_k^n + 2D\Delta t_{k-1})}}\text{Normal}\left(\frac{x_k^n c_k^n - 2D\Delta t_{k-1} m_k^n}{c_k^n + 2D\Delta t_{k-1}}, \frac{2D\Delta t_{k-1} c_k^n}{c_k^n + 2D\Delta t_{k-1}}\right).$$

Sampling Inactive Molecule Trajectories To update the trajectories of the inactive molecules, we sample from the corresponding conditionals $p(\overline{x}^n, \overline{y}^n, \overline{z}^n | D, \mu_{mol}, \{q^n, b^n\}_n, \boldsymbol{\Delta t})$. Since the inactive molecules are not associated with the observations

in $\boldsymbol{\Delta t}$, these reduce to $p(\overline{x}^n, \overline{y}^n, \overline{z}^n | D, \{q^n, b^n\}_n)$ which we simulate as standard 3D Brownian motion [426].

Sampling the Diffusion Coefficient We sample the diffusion coefficient from the conditional probability distribution $p(D|\mu_{mol}, \{q^n, b^n, \overline{x}^n, \overline{y}^n, \overline{z}^n\}_n, \boldsymbol{\Delta t})$. Because of the specific dependencies of the variables in this formulation, *e.g.*, eq. (3.13) and eqs. (3.7)–(3.9), the conditional distribution simplifies to $p(D|\{\overline{x}^n, \overline{y}^n, \overline{z}^n\}_n, \boldsymbol{\Delta t})$. Using Bayes' rule, this distribution becomes $D \sim \text{InvGamma}(\alpha'_D, \beta'_D)$ where $\alpha'_D = \alpha_D + \frac{3N(K-1)}{2}$ and $\beta'_D = \beta_D + \sum_{k=2}^{K} \frac{\sum_{n=1}^{N}\left(\left(x_k^n - x_{k-1}^n\right)^2 + \left(y_k^n - y_{k-1}^n\right)^2 + \left(z_k^n - z_{k-1}^n\right)^2\right)}{4\Delta t_{k-1}}$.

Sampling Molecule Indicators For each molecule n we sample its indicator prior weight, q^n, from the corresponding conditional distribution $p(q^n|\mu_{mol}, D, \{b^n, \overline{x}^n, \overline{y}^n, \overline{z}^n\}_n, \boldsymbol{\Delta t})$, which simplifies to $p(q^n|b^n)$. For this we use eq. (3.19) and eq. (3.18). According to Bayes' rule, the latter distribution becomes $q^n \sim \text{Beta}(\alpha', \beta')$ where $\alpha' = \frac{\alpha}{N} + b^n$ and $\beta' = \beta\frac{N-1}{N} + 1 - b^n$. Subsequently, we update the indicators b^n by sampling from the corresponding conditional distribution $p(\{b^n\}_n | D, \mu_{mol}, \{q^n, \overline{x}^n, \overline{y}^n, \overline{z}^n\}_n, \boldsymbol{\Delta t})$ using a Methropolis-Hasting algorithm [427, 428]. For this, we use a proposal $b^n_{new} \sim \text{Bernoulli}(q^n)$. With this proposal, the acceptance ratio becomes

$$r = \prod_{k=1}^{K-1} \frac{\mu_{back} + \mu_{mol}\sum_{n=1}^{N} b^n_{new} \exp\left(-2\frac{(x_k^n)^2 + (y_k^n)^2}{\omega_{xy}^2} - 2\frac{(z_k^n)^2}{\omega_z^2}\right)}{\mu_{back} + \mu_{mol}\sum_{n=1}^{N} b^n_{old} \exp\left(-2\frac{(x_k^n)^2 + (y_k^n)^2}{\omega_{xy}^2} - 2\frac{(z_k^n)^2}{\omega_z^2}\right)} \times \frac{e^{-\left[\mu_{back} + \mu_{mol}\sum_{n=1}^{N} b^n_{new} \exp\left(-2\frac{(x_k^n)^2 + (y_k^n)^2}{\omega_{xy}^2} - 2\frac{(z_k^n)^2}{\omega_z^2}\right)\right]\Delta t_k}}{e^{-\left[\mu_{back} + \mu_{mol}\sum_{n=1}^{N} b^n_{old} \exp\left(-2\frac{(x_k^n)^2 + (y_k^n)^2}{\omega_{xy}^2} - 2\frac{(z_k^n)^2}{\omega_z^2}\right)\right]\Delta t_k}}.$$

Sampling the Molecule Photon Emission Rate In the last step, after updating the locations and indicators of the molecules, we sample the molecule photon emission rate from the corresponding conditional distribution $p(\mu_{mol}|D, \{q^n, b^n, \overline{x}^n, \overline{y}^n, \overline{z}^n\}_n, \boldsymbol{\Delta t})$. To sample this distribution, we also use a Metropolis-Hastings step. For this, we use

proposals of the form $\mu_{mol}^{new} \sim \text{Gamma}\left(\alpha, \mu_{mol}^{old}/\alpha\right)$ where μ_{mol}^{old} denotes the current sampled value. Using both eqs. (3.1) and (3.2), the acceptance ratio becomes

$$
\begin{aligned}
r = \exp & \left[\sum_{k=1}^{K-1} \log \left(\frac{\mu_{back} + \mu_{mol}^{new} \sum_{n=1}^{N} b^n \exp\left(-2\frac{(x_k^n)^2 + (y_k^n)^2}{\omega_{xy}^2} - 2\frac{(z_k^n)^2}{\omega_z^2} \right)}{\mu_{back} + \mu_{mol}^{old} \sum_{n=1}^{N} b^n \exp\left(-2\frac{(x_k^n)^2 + (y_k^n)^2}{\omega_{xy}^2} - 2\frac{(z_k^n)^2}{\omega_z^2} \right)} \right) \right] \\
& \times \exp \left[\sum_{k=1}^{K-1} \left[(\mu_{mol}^{old} - \mu_{mol}^{new}) \Delta t_k \sum_{n=1}^{N} b^n \exp\left(-2\frac{(x_k^n)^2 + (y_k^n)^2}{\omega_{xy}^2} - 2\frac{(z_k^n)^2}{\omega_z^2} \right) \right] \right] \\
& \times \exp \left[(\alpha_{mol} - 1) \log\left(\frac{\mu_{mol}^{new}}{\mu_{mol}^{old}} \right) + \frac{1}{\beta_{mol}} (\mu_{mol}^{old} - \mu_{mol}^{new}) + (2\alpha - 1) \log\left(\frac{\mu_{mol}^{old}}{\mu_{mol}^{new}} \right) \right] \\
& \times \exp \left[\alpha \left(\frac{\mu_{mol}^{new}}{\mu_{mol}^{old}} - \frac{\mu_{mol}^{old}}{\mu_{mol}^{new}} \right) \right].
\end{aligned}
$$

Table 3.1.

Here, we list point estimates of our analyses, which we obtain from the marginal posterior probability distributions $p(D|\boldsymbol{\Delta t})$ and $p(\mu_{mol}|\boldsymbol{\Delta t})$. Estimates are listed according to figure.

	D		μ_{mol}	
	mean	std	mean	std
	$\mu m^2/s$	$\mu m^2/s$	photons/s	photons/s
Fig. (3.2A)	4.54	4.49	-	-
Fig. (3.2B)	4.17	4.11	-	-
Fig. (3.2C)	1.14	1.12	-	-
Fig. (3.2D)	1.02	1.01	-	-
Fig. (3.2E)	4.75	4.64	-	-
Fig. (3.4B1)	1.03	0.25	-	-
Fig. (3.4B2)	0.95	0.63	-	-
Fig. (3.4B3)	0.75	0.68	-	-
Fig. (3.4B4)	0.45	0.77	-	-
Fig. (3.5A3)	1.01	0.27	-	-
Fig. (3.5B3)	1.09	0.51	-	-
Fig. (3.5C3)	1.65	1.59	-	-
Fig. (3.6)	1.05×10^{-2}	0.22×10^{-2}	-	-
	1.21×10^{-1}	0.34×10^{-1}	-	-
	1.06	0.19	-	-
	9.87	2.33	-	-
	117.62	35.13	-	-

Table 3.2.

Here, we continue above list point estimates of our analyses, which we obtain from the marginal posterior probability distributions $p(D|\boldsymbol{\Delta t})$ and $p(\mu_{mol}|\boldsymbol{\Delta t})$. Estimates are listed according to figure.

	D		μ_{mol}	
	mean	std	mean	std
	$\mu m^2/s$	$\mu m^2/s$	photons/s	photons/s
Fig. (3.6)	1.05×10^{-2}	0.22×10^{-2}	-	-
	1.21×10^{-1}	0.34×10^{-1}	-	-
	1.06	0.19	-	-
	9.87	2.33	-	-
	117.62	35.13	-	-
Fig. (3.7A)	0.99	0.34	-	-
Fig. (3.7B)	0.96	0.51	-	-
Fig. (3.7C)	2.92	2.68	-	-
Fig. (3.7D)	3.26	2.95	-	-
Fig. (3.16A3)	10.02	1.17	-	-
Fig. (3.16B3)	9.96	2.19	-	-
Fig. (3.17A3)	-	-	4.11×10^5	1.61×10^3
Fig. (3.17B3)	-	-	4.37×10^4	2.84×10^3
Fig. (3.17C3)	-	-	1.28×10^5	1.25×10^4

Table 3.3.

Summary of notation.

Description	Variable	Units
Diffusion coefficient	D	μm^2/s
α parameter of the diffusion coefficient prior	α_D	-
β parameter of the diffusion coefficient prior	β_D	μm^2/s
Photon inter-arrival time	Δt	s
Total trace duration	T_{total}	s
molecule photon emission rate (maximum)	μ_{mol}	photons/s
α parameter of the molecule photon emission rate's prior	α_{mol}	-
β parameter of the molecule photon emission rate's prior	β_{mol}	photons/s
Emission rate of molecule n at time t_k	μ_k^n	photons/s
Combined photon emission rate at time t_k	μ_k	photons/s
Background photon emission rate	μ_{back}	photons/s
Minor semi-axis of confocal PSF (focal plane)	ω_{xy}	μm
Major semi-axis of confocal PSF (optical axis)	ω_z	μm
Location of molecule n at time t_k in x-coordinate	x_k^n	μm
Location of molecule n at time t_k in y-coordinate	y_k^n	μm
Location of molecule n at time t_k in z-coordinate	z_k^n	μm
Recorded photon inter-arrival time between t_k and t_{k-1}	Δt_k	s
Indicator variable for molecule n	b^n	-
Prior weight for b_n	q^n	-
α parameter of prior weight q^n	α_q	-
β parameter of prior weight q^n	β_q	-

Table 3.4.

Summary of notation.

Description	Variable	Units
Upper bound for the number of model molecules	N	-
Mean value of initial molecule position's prior in the xy-plane	μ_{xy}	μm
Mean value of initial molecule position's prior on the z-axis	μ_z	μm
Variance of the initial molecule position's prior in the xy-plane	σ^2_{xy}	μm
Variance of the initial molecule position's prior on the z-axis	σ^2_z	μm
Periodic boundary in the xy-plane	L_{xy}	μm
Periodic boundary on the z-axis	L_z	μm

Table 3.5.

Probability distributions used and their densities. Here, the corresponding random variables are denoted by x. We use ";" to separate random variables from parameters. For example, Normal$(x;\mu,\sigma^2)$ means that x is the random variable (e.g. $\int_{-\infty}^{+\infty} dx\text{Normal}(x;\mu,\sigma^2) = 1$), and μ and σ^2 are parameters characterizing this density.

Distribution	Notation	Probability density function	Mean	Variance
Normal	Normal(μ,σ^2)	$\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2
Symmetric Normal	SymNormal(μ,σ^2)	$\frac{1}{2}\frac{e^{-\frac{(x+\mu)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} + \frac{1}{2}\frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}$	0	$\mu^2+\sigma^2$
Exponential	Exponential(μ)	$\mu e^{-\mu x}$	$\frac{1}{\mu}$	$\frac{1}{\mu^2}$
Chi-square	$\chi^2(\alpha,2)$	$\frac{1}{\Gamma(\frac{\alpha}{2})2^{\frac{\alpha}{2}}}x^{\frac{\alpha}{2}-1}e^{-\frac{x}{2}}$	α	2α
Gamma	Gamma(α,β)	$\frac{1}{\Gamma(\alpha)\beta^\alpha}x^{\alpha-1}e^{-\frac{x}{\beta}}$	$\alpha\beta$	$\alpha\beta^2$
Inverse-Gamma	InvGamma(α,β)	$\frac{\beta^\alpha}{\Gamma(\alpha)}x^{-\alpha-1}e^{-\frac{\beta}{x}}$	$\frac{\beta}{\alpha-1}$	$\frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$
Beta	Beta(α,β)	$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
Bernoulli	Bernoulli(q)	$(q-1)\delta_0(x) + q\delta_1(x)$	q	$q(1-q)$

Table 3.6.
Parameter values used in the generation of the synthetic traces. Choices are listed according to figures.

	L_{xy}	L_z	ω_{xy}	ω_z	N	D	μ_{mol}	μ_{back}	T_{total}
Units	μm	μm	μm	μm	-	$\mu m^2/s$	photons/s	photons/s	s
Fig. (3.2A)	1	2	0.3	1.5	4	1	4×10^4	10^3	0.03
Fig. (3.2B)	1	2	0.3	1.5	4	1	4×10^4	10^3	0.03
Fig. (3.2C)	1	2	0.3	1.5	4	1	4×10^4	10^3	0.03
Fig. (3.2D)	1	2	0.3	1.5	4	1	4×10^4	10^3	0.03
Fig. (3.2E)	1	2	0.3	1.5	4	1	4×10^4	10^3	0.03
Fig. (3.4A)	1	2	0.3	1.5	4	1	4×10^4	10^3	0.03
Fig. (3.5A)	1	2	0.3	1.5	10	1	4×10^4	10^3	0.03
Fig. (3.5C)	1	2	0.3	1.5	4	1	4×10^4	10^3	0.03
Fig. (3.5E)	1	2	0.3	1.5	1	1	4×10^4	10^3	0.03
Fig. (3.6)	1	2	0.3	1.5	4	10^{-2}	4×10^4	10^3	0.03
Fig. (3.6)	1	2	0.3	1.5	4	10^{-1}	4×10^4	10^3	0.03
Fig. (3.6)	1	2	0.3	1.5	4	1	4×10^4	10^3	0.03
Fig. (3.6)	1	2	0.3	1.5	4	10	4×10^4	10^3	0.03
Fig. (3.6)	1	2	0.3	1.5	4	100	4×10^4	10^3	0.03
Fig. (3.7A)	1	2	0.3	1.5	4	1	4×10^5	10^3	0.03
Fig. (3.7B)	1	2	0.3	1.5	4	1	4×10^4	10^3	0.03
Fig. (3.7C)	1	2	0.3	1.5	4	1	4×10^3	10^3	0.03

Table 3.7.
Here, we continue above parameter values used in the generation of the synthetic traces. Choices are listed according to figures.

	L_{xy}	L_z	ω_{xy}	ω_z	N	D	μ_{mol}	μ_{back}	T_{total}
Units	μm	μm	μm	μm	-	$\mu m^2/s$	photons/s	photons/s	s
Fig. (3.7D)	1	2	0.3	1.5	4	1	10^3	10^3	0.03
Fig. (3.16A)	1	2	0.3	1.5	10	10	4×10^5	10^3	0.05
Fig. (3.16C)	1	2	0.3	1.5	10	10	4×10^4	10^3	0.05
Fig. (3.17A)	1	2	0.3	1.5	10	10	4×10^5	10^3	0.05
Fig. (3.17C)	1	2	0.3	1.5	10	10	4×10^4	10^3	0.05
Fig. (3.17E)	1	2	0.3	1.5	10	10	4×10^3	10^3	0.05

Table 3.8.
Parameter values used in the analyses of the traces. Choices are listed according to figures.

	ω_{xy}	ω_z	N	α_D	β_D	α_{mol}	β_{mol}	α_q	β_q	μ_{xy}	μ_z	σ_{xy}^2	σ_z^2
Units	μm	μm	-	-	$\mu m^2/s$	-	phts/s	-	-	-	μm	μm	μm^2
Fig. (3.2A)	0.3	1.5	-	1	1	1	10^5	-	-	0.1	0.1	1	1
Fig. (3.2B)	0.3	1.5	-	1	1	1	10^5	-	-	0.1	0.1	1	1
Fig. (3.2C)	0.3	1.5	-	1	1	1	10^5	-	-	0.1	0.1	1	1
Fig. (3.2D)	0.3	1.5	-	1	1	1	10^5	-	-	0.1	0.1	1	1
Fig. (3.2E)	0.3	1.5	-	1	1	1	10^5	-	-	0.1	0.1	1	1
Fig. (3.4)	0.3	1.5	20	1	1	1	10^5	1	1	0.1	0.1	1	1
Fig. (3.5A3)	0.3	1.5	20	1	1	1	10^5	1	1	0.1	0.1	1	1
Fig. (3.5B3)	0.3	1.5	20	1	1	1	10^5	1	1	0.1	0.1	1	1
Fig. (3.5C3)	0.3	1.5	20	1	1	1	10^5	1	1	0.1	0.1	1	1
Fig. (3.6)	0.3	1.5	20	1	1	1	10^5	1	1	0.1	0.1	1	1
Fig. (3.6)	0.3	1.5	20	1	1	1	10^5	1	1	0.1	0.1	1	1
Fig. (3.6)	0.3	1.5	20	1	1	1	10^5	1	1	0.1	0.1	1	1
Fig. (3.6)	0.3	1.5	20	1	100	1	10^5	1	1	0.1	0.1	1	1
Fig. (3.7A)	0.3	1.5	20	1	1	1	10^5	1	1	0.1	0.1	1	1
Fig. (3.7B)	0.3	1.5	20	1	1	1	10^5	1	1	0.1	0.1	1	1
Fig. (3.7C)	0.3	1.5	20	1	1	1	10^5	1	1	0.1	0.1	1	1
Fig. (3.7D)	0.3	1.5	20	1	1	1	10^5	1	1	0.1	0.1	1	1
Fig. (3.8)	0.23	0.55	20	1	1	1	10^5	1	1	0.1	0.1	1	1
Fig. (3.9)	0.23	0.55	20	1	1	1	10^5	1	1	0.1	0.1	1	1

Table 3.9.
Here, we continue above parameter values used in the analyses of the traces. Choices are listed according to figures.

	ω_{xy}	ω_z	N	α_D	β_D	α_{mol}	β_{mol}	α_q	β_q	μ_{xy}	μ_z	σ^2_{xy}	σ^2_z
Units	μm	μm	-	-	$\mu m^2/s$	-	phts/s	-	-	-	μm	μm	μm^2
Fig. (3.10)	0.23	0.55	20	1	1	1	10^5	1	1	0.1	0.1	1	1
Fig. (3.11)	0.23	0.55	20	1	1	1	10^5	1	1	0.1	0.1	1	1
Fig. (3.12)	0.27	4.51	20	1	1	1	10^5	1	1	0.1	0.1	1	1
Fig. (3.13)	0.22	3.90	20	1	100	1	10^5	1	1	0.1	0.1	1	1
Fig. (3.16)	0.3	1.5	20	1	1	1	10^5	1	1	0.1	0.1	1	1
Fig. (3.17)	0.3	1.5	20	1	1	1	10^5	1	1	0.1	0.1	1	1
Fig. (3.18)	0.23	0.55	20	1	1	1	10^5	1	1	0.1	0.1	1	1
Fig. (3.19)	0.22	3.90	20	1	100	1	10^5	1	1	0.1	0.1	1	1

4. PHOTON-BY-PHOTON ANALYSIS OF TCSPC DATA WITH BAYESIAN NONPARAMETRICS

Meysam Tavakoli, Sina Jazani, Ioannis Sgouralis, Heo Wooseok, Kunihiko Ishii, Tahei Tahara, and Steve Pressé "Photon-by-photon Analysis of TCSPC Data with Bayesian Nonparametrics" Manuscript under review in Cell Reports Physical Science (2020).
Contribution: MT analyzed data and developed analysis software; MT, SJ, IS developed computational tools; HW, KI, TT contributed experimental data; MT, SJ, IS, SP conceived research; SP oversaw all aspects of the projects.

4.1 Abstract

Fluorescence Lifetime Imaging (FLIM) is an experimental imaging technique yielding excited state lifetimes of chemical species recorded over multiple pixels. Within one pixel, the determination of the number of species can be achieved either through fitting time correlated single photon counting (TCSPC) histograms or phasor analysis. Both methods yield lifetimes in a computationally efficient manner. However, they also have drawbacks that we address here. First, they do not yield the number of chemical species. Yet the number species is specifically encoded in the photon time of arrival. Next, even to determine lifetimes under the assumption of a known number of species, both methods rely on heavy data post-processing of the signal thereby requiring large amounts of data to retrieve lifetimes. As a result the sample is exposed to light orders of magnitude longer than required and temporal resolution is compromised. Here we propose a direct photo-by-photon analysis strategy to infer, simultaneously and self-consistently, the number of species and their associated lifetimes from as few as on the order of 3000 photons for two species. We do

so by leveraging new mathematical tools within the Bayesian nonparametric (BNP) paradigm that we have previously exploited in the analysis of single photon arrivals from single spot confocal. We benchmark our method on simulated as well as experimental data for one, two, three, and four species with both immobilized and freely diffusing molecule data sets.

4.2 Introduction

Fluorescence microscopy has provided us with the ability to monitor the dynamics of molecules by allowing for the selective detection of fluorophores or labeled molecules [429, 430]. A number of fluorescence approaches–such as confocal microscopy [431], two-photon microscopy [432] and super-resolution widefield applications [433]–use constant illumination to provide information on chemical kinetics [434–436], diffusional dynamics [401, 437, 438] or spatial locations of molecules [439, 440].

Other fluorescence methods use illumination that varies in time [414, 441–447] where the time of arrival of the photon now encodes critical information, say, on the excited state lifetime or the number of different chemical species. This is the basis of lifetime imaging [440]. Local variations in lifetimes across cells reveal information on the local pH [448, 449], oxygenation [448] and other metabolic traits [450, 451] of the cell.

There are different ways to achieve time-varying illumination [452–454]. The first is through pulsed illumination [455, 456]. Here the time of arrival of a photon can be analyzed directly [352, 457–459], under the assumption of a fixed and known number of molecular species, to determine the lifetime of each species. Methods of analysis include Bayesian approaches but always under the assumption of a known number of species [460–464]. The photon arrival times can also be histogrammed; an approach termed time-correlated single photon counting (TCSPC)

histogram method [430, 465–467]. These histograms are typically fitted using a multi-exponential fit [468, 469] to identify the lifetime of each species. Various metrics, depending on the experiment under investigation, are then minimized to improve fitting [462]. While such methods yield lifetimes in a computationally efficient manner, they have several drawbacks. Not least of which is the specification of the number of species that, while in principle encoded in the data, *cannot* be learned independently. Additionally, these methods are data *inefficient*. The latter drawback is especially problematic if temporal resolution is important, the sample is light sensitive, or multiple lifetimes are fairly similar requiring long photon arrival traces to discriminate one from the other.

A second way to illuminate a sample is by modulating the intensity at a fixed frequency [470–473]. As a result of the modulated excitation intensity, the emission is also modulated but otherwise phase shifted [452]. For this reason, phasor analysis [474] has been used to extract lifetimes from the modulated emission intensity. A variant of phasor analysis also holds for pulsed excitation [475–477]. The advantages and drawbacks here are similar to those of the methods we discussed in the previous paragraph. What is more, and perhaps more strikingly, is that the retrieval of lifetime information from phasor analysis requires independent knowledge of not only the number of species but also the lifetime of all but one unknown species whose lifetime is to be determined from a mixture of chemical species [477–479].

Fig. 4.1 captures just how sensitive the accuracy of TCSPC and phasor analysis are to the number of photons available to the analysis.

What we have available in the community for lifetime analysis are methods that can learn lifetimes at minimal computational cost. What we also know is that information on the number of species is encoded in the photon arrivals. At higher computational cost, we could learn these, and full distributions over species and their

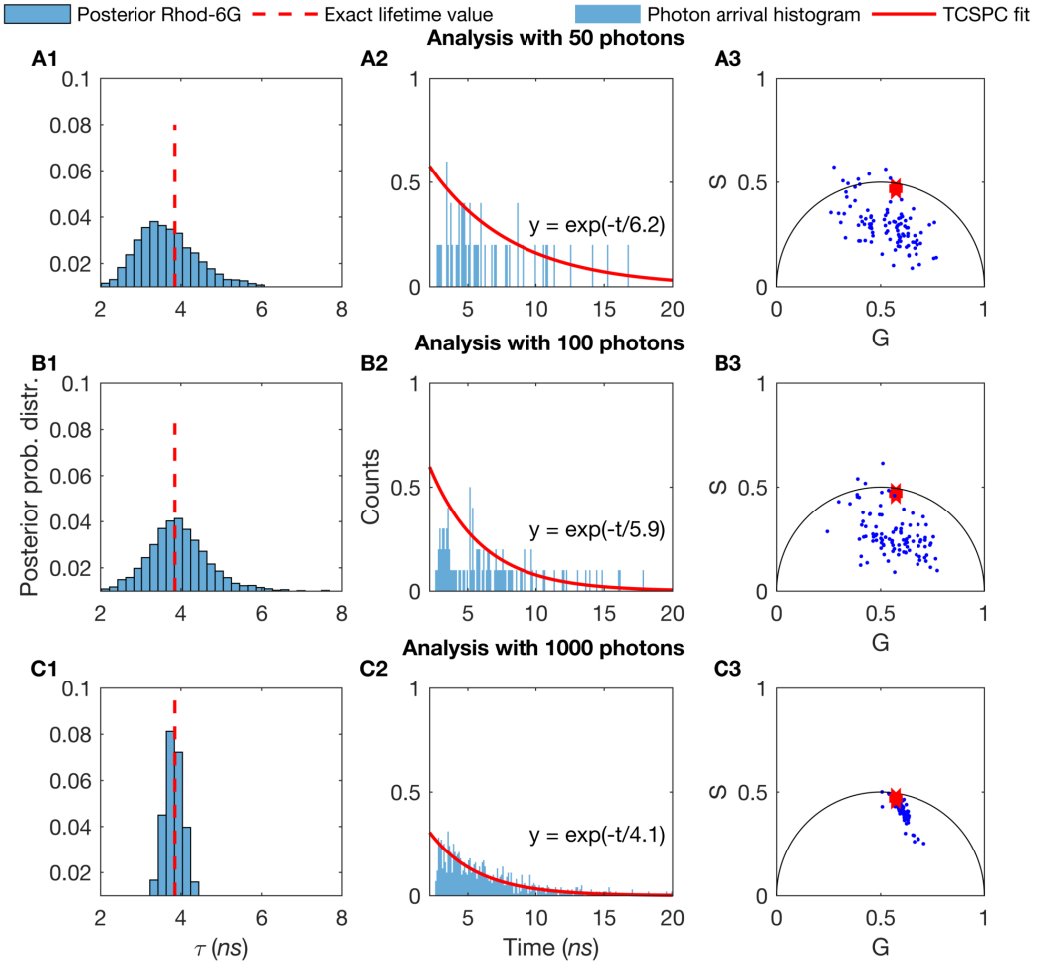

Fig. 4.1. **FLIM reveals excited state lifetimes provided a large number of photons are available.** Very preliminarily here we compare our method relying on Bayesian nonparametric (BNP) which we discuss in greater depth later to TCSPC and phasor analysis with limited data available for analysis. (A1-A3) Here we use just 50 photons from experimental time trace Rhod-6G to compare all three methods: (A1) BNPs, (A2) TCSPC, and (A3) phasor analysis. In (B1-B3) and (C1-C3) we repeat the analysis for 100 and then 1000 photons.

associated lifetimes as well, but it would require a different mathematical paradigm that goes beyond the parametric Bayesian paradigm.

We have previously exploited the Bayesian nonparametric (BNP) paradigm [263, 267] to analyze single photon arrival traces in order to learn diffusion coefficients from minimal photon numbers drawn from single spot confocal experiments [401, 480]. Traditionally, such photon arrivals were analyzed using tools from fluorescence correlation spectroscopy where very long traces were collected and auto-correlated in time. The direct photon-by-photon analysis demanded a different approach as the stochastic number of molecules contributing photons was unknown and an estimate of that number deeply impacted our diffusion coefficient estimate. It is for this reason that we invoked the nonparametric paradigm there.

Similarly the BNP paradigm is also required to infer the number of species and their associated lifetimes. This is because assuming an incorrect number of species leads to incorrect lifetime estimates for each species; see Fig. 4.2. BNP reshapes our interpretation of biophysical data as they fundamentally go beyond the parametric paradigm. In the "normal" (i.e., parametric) paradigm, we assume models and, given these models, write down likelihoods used in data analysis. Yet a growing number of biophysical applications, such as fluorescence correlation spectroscopy that we've published on [401, 480] and lifetime analysis which is the focus here, present a critical challenge where the model itself is unknown. In lifetime analysis, this model is the number of species. Just as we treat model parameters as random variables in the parametric Bayesian paradigm, we treat models themselves here as the random variables and try to learn full posterior distributions over the number of species.

Here we propose a protocol that exploits BNPs to learn species and their associated lifetimes with as few photons as possible. The advantages are four-fold: 1) we can learn the number of species; 2) by resolving lifetimes and species with fewer photons

we can minimize photo-damage; 3) we can monitor processes out-of-equilibrium where only few photons are available before chemical conversion into another species; 4) given long traces, we can exploit the additional data, if need be, to discriminate between species of similar lifetimes that could not otherwise be previously discerned.

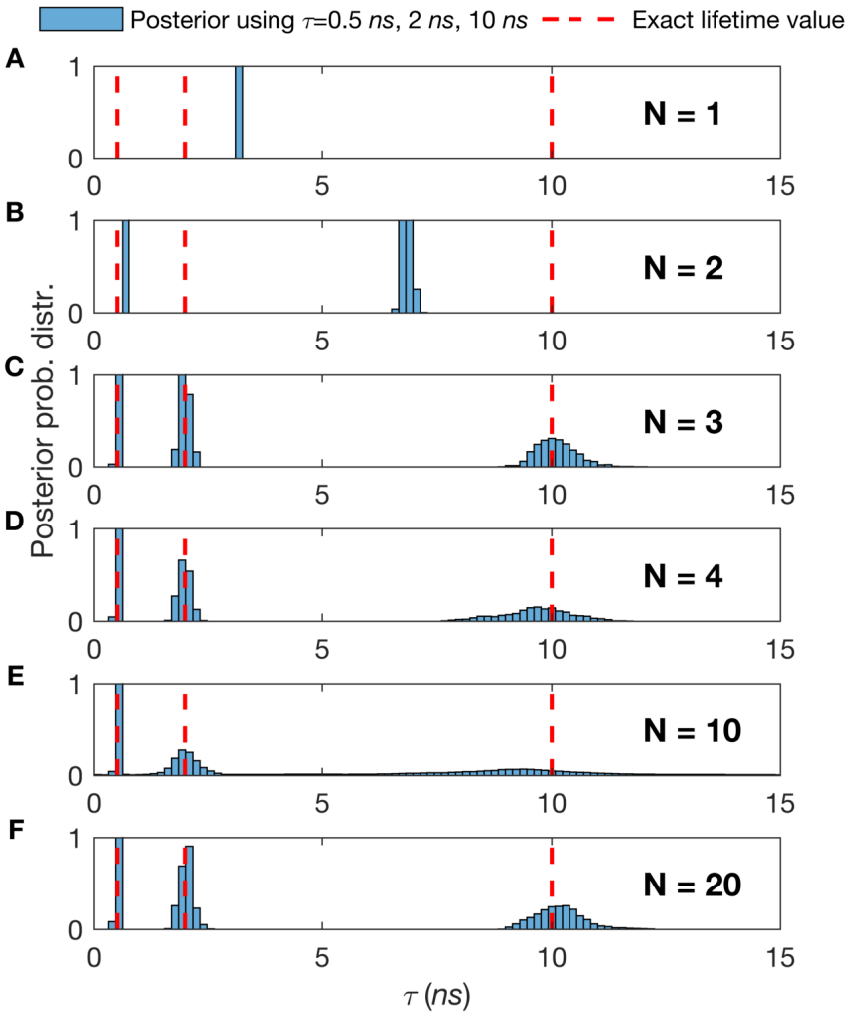

Fig. 4.2. **The number of species assumed in analysis directly impacts the lifetimes ascribed to those species. Thus, we need an independent method to estimate species numbers.** (A-F) We generate synthetic traces with three species with a total of 2×10^4 photon arrivals and lifetimes, τ, of 0.5 ns, 2 ns, and 10 ns. To estimate the τ within the normal (i.e., parametric) Bayesian paradigm, we start by assuming the following number of species, $N = 1$ (A), $N = 2$ (B), $N = 3$ (C), $N = 4$ (D), ..., $N = 10$ (E), ..., and $N = 20$ (F). The good fit provided by $N > 2$ and the mismatch in the peak of the posterior distribution over the lifetime and correct value of the lifetime (red dotted line) in all others underscores why it will be critical for us, or any method analyzing single photon data in the context of confocal microscope experiments, to correctly estimate the number of species contributing to the trace in order to deduce chemical parameters such as lifetime.

4.3 Methods

Here, we describe the mathematical formulation of our analysis method of TCSPC data. For clarity we focus on measurements obtained on a fluorescence setup that utilizes a train of identical excitation pulses. Following each pulse, one of more molecules located near the illuminated region may be excited from their ground state. As the excited molecules decay back to their ground state they may emit photons and we record the detection time. Below we describe how we analyze such recorded times.

We start from single photon detection times which consist of the raw output in a TCSPC experiment. Similarly, these are measured based on the time difference between excitation pulses, which are time stamped, and the detection time of the first photon arriving after each pulse [452, 455, 481]. Precisely, our raw input is $\boldsymbol{\Delta t} = (\Delta t_1, \Delta t_2, \ldots, \Delta t_K)$ where Δt_k is the time interval between the preceding pulse's time and the photon detection time of the k^{th} detection. In the literature, each Δt_k is often termed micro-time. Because, some pulses may not lead to a photon detection, in general the micro-times in $\boldsymbol{\Delta t}$ are fewer than the total number of pulses applied during an experiment.

4.3.1 Model description

We assume that, once excited, each molecule remains excited for a time period that is considerably lower (typically few nanoseconds) as compared to the time between two successive pulses (typically more than four times of the longest decay time in the sample [452]). This condition allows us to consider that any photon which is detected stems from an excitation caused by the very previous pulse and not from earlier pulses. Also, as excitation pulses in TCSPC experiments are weak [467, 482], and typically

one in ≈ 100 pulses results in a photon detection [452], we ignore, to a very good approximation, multiple photon arrivals. As the number of detected photons coming from the background is considerably lower than the number of detected photons coming from the excited molecules, typically one to ≈ 1000, we also ignore background photons. However, background photons can be dealt with straightforwardly as we show in the discussion, Sec. 4.5.

To analyze the recordings $\boldsymbol{\Delta t}$, we assume that the sample contains in total M different molecular species that are characterized by different lifetimes $\tau_1, \ldots, \tau_M$. Since molecules of each species may be excited by the pulses with different probabilities (because of different fraction of molecules contributing photons from different species), we consider a probability vector $\bar{\pi} = (\pi_1, \ldots, \pi_M)$ that gathers the probabilities of each species giving rise to a photon detection. Allowing s_k to be a tag attaining integer values $1, \ldots, M$, that indicates which species triggered the k^{th} detection, we may write

$$s_k|\bar{\pi} \sim \mathbf{Categorical}_{1:M}\left(\bar{\pi}\right). \tag{4.1}$$

With this convention, the lifetime of the molecule triggering the k^{th} detection is τ_{s_k}. Of course, the number of molecular species M and the precise values of the lifetimes $\tau_1, \ldots, \tau_M$ are unknown and our main task is to estimate them using the recordings in $\boldsymbol{\Delta t}$.

For clarity, we denote with $t_{pul,k}$ the application time of the pulse that triggers the k^{th} photon detection. More precisely, $t_{pul,k}$ is the time of the pulse's peak. Because, in general pulses last for some non-zero duration, and so they may excite the molecules at slightly different times, we denote with $t_{ext,k}$ the absorption time of the molecule triggering the k^{th} detection. Further, we denote with $t_{ems,k}$ the emission time of the photon triggering the k^{th} detection. Finally, due to the measuring electronics,

the detection time, which we denote with $t_{det,k}$, might be different from $t_{ems,k}$; see Fig. 4.3 for more details.

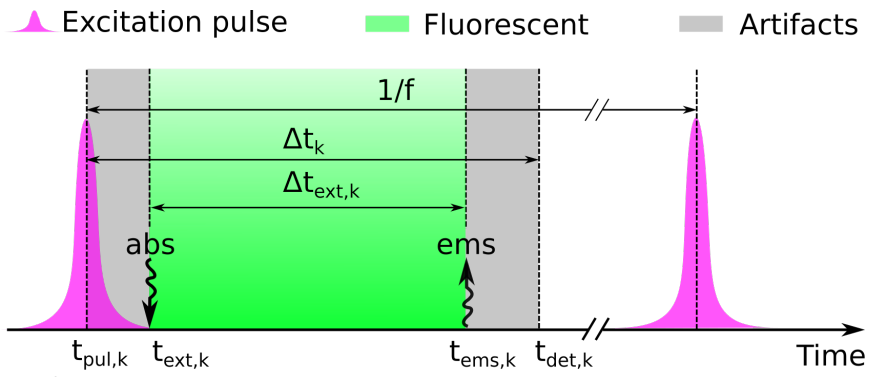

Fig. 4.3. **Cartoon of the factors that contribute to the recorded photon arrival times.** Here, $t_{pul,k}$ is the time of the pulse's peak. Since pulses last for some time, they may excite the molecules at slightly different times. As such, we denote with $t_{ext,k}$ the absorption time of the molecule triggering the k^{th} detection. Moreover, we denote with $t_{ems,k}$ the emission time of the photon triggering the k^{th} detection. At last, on account of electronics limitations, the detection time, which we denote with $t_{det,k}$, might be different from $t_{ems,k}$.

With this convention, our measured output consists of the time lags $\Delta t_k = t_{det,k} - t_{pul,k}$. These time lags include: (i) the time until absorption occurs, $t_{ext,k} - t_{pul,k}$; (ii) the time until fluorescence emission occurs, $t_{ems,k} - t_{ext,k}$; (iii) delays and errors introduced by the measuring electronic devices, $t_{det,k} - t_{ems,k}$. Below, we denote the middle period with $\Delta t_{ext,k} = t_{ems,k} - t_{ext,k}$; while, we denote with $\Delta t_{err,k} = (t_{ext,k} - t_{pul,k}) + (t_{det,k} - t_{ems,k})$ the sum of the others. From these two, $\Delta t_{ext,k}$ is the time the molecule spends in the excited state; while, $\Delta t_{err,k}$ gathers any artifacts caused by our setup either in the excitation or detection pathway. The advantages of considering these two periods separately, as we explain below, is that (i) these represent independent physical processes, and (ii) each one is theoretically and experimentally characterized well [452].

In particular, $\Delta t_{err,k}$ is characterized by the instrument response function (IRF) that, in each set-up, is readily obtained with calibration measurements. [452] In this study, we approximate the IRF as a Gaussian

$$\Delta t_{err,k} \sim \mathbf{Normal}(\tau_{\mathrm{IRF}}, \sigma^2_{\mathrm{IRF}}). \tag{4.2}$$

In this approximation, τ_{IRF} is the IRF's peak time and $\sigma_{\mathrm{IRF}} = \mathrm{FWHM}/2.355$ where FWHM is the IRF's full-width-at-half-maximum. In the supplementary information, we explain the IRF's calibration in detail.

Upon excitation, the time the molecule remains excited, $\Delta t_{ext,k}$, is memoryless [452], and so it follows the exponential distribution. Therefore,

$$\Delta t_{ext,k}|\lambda_{s_k} \sim \mathbf{Exponential}(\lambda_{s_k}) \tag{4.3}$$

where λ_{s_k} is the fluoresce rate of the molecule triggering the detection of $\Delta t_{ext,k}$. Of course, the fluorescence rate depends upon the lifetime by $\lambda_{s_k} = 1/\tau_{s_k}$.

Because $\Delta t_{ext,k}$ and $\Delta t_{err,k}$ are independent variables, the statistics of our measurements, which are given by $\Delta t_k = \Delta t_{ext,k} + \Delta t_{err,k}$, follow

$$\Delta t_k|\lambda_{s_k} \sim \mathbf{Normal}(\tau_{\mathrm{IRF}}, \sigma^2_{\mathrm{IRF}}) * \mathbf{Exponential}(\lambda_{s_k}) \tag{4.4}$$

where $*$ denotes a convolution [483], and specifically has the probability density

$$p\left(\Delta t_k|\lambda_{s_k}\right) = \frac{\lambda_{s_k}}{2} \exp\left[\frac{\lambda_{s_k}}{2}\left(2\left(\tau_{\mathrm{IRF}} - \Delta t_k\right) + \lambda_{s_k}\sigma^2_{\mathrm{IRF}}\right)\right] \operatorname{erfc}\left(\frac{\tau_{\mathrm{IRF}} - \Delta t_k + \lambda_{s_k}\sigma^2_{\mathrm{IRF}}}{\sigma_{\mathrm{IRF}}\sqrt{2}}\right) \tag{4.5}$$

where $\operatorname{erfc}(\cdot)$ denotes the complementary error function. In the supplementary information, we show analytically how Eq. (4.5) arises from Eqs. (4.2) and (4.3).

In the next section we describe how Eqs. (4.1) and (4.5) can be used in conjunction with BNP to obtain the estimates we are after.

4.3.2 Model inference

All quantities which we wish to infer, for example the species fluorescence rates $\lambda_1, \ldots, \lambda_M$ and excitation probabilities in $\bar{\pi}$, are represented by model variables in the preceding formulation. We infer values for these variables within the Bayesian paradigm [79, 328, 358]. Accordingly, on the fluorescence rates we place independent priors

$$\lambda_m \sim \textbf{Gamma}\left(\alpha_\lambda, \beta_\lambda\right), \qquad m = 1, \ldots, M \tag{4.6}$$

that ensure strictly positive values. As the total number of species contributing photon detections in an experiment is unknown, we consider a symmetric Dirichlet prior [79, 360] on $\bar{\pi}$ of the form

$$\bar{\pi} \sim \textbf{Dirichlet}_M\left(\frac{\alpha}{M}, \ldots, \frac{\alpha}{M}\right) \tag{4.7}$$

where α is a positive scalar hyper-parameter. A graphical summary of the whole formulation is shown on Fig. 4.4.

The distribution in Eq. (4.7) ensures that $\bar{\pi}$ are valid probability vectors. Further, Eq. (4.7) is specifically chosen to allow for a large, $M \to \infty$, number of species. This is particularly important because the total number of molecular species contributing to the detections in a FLIM experiment is typically unknown, and so choosing a finite M may lead to under-fitting. Specifically, at the limiting case $M \to \infty$, the prior on Eq. (4.7), combined with Eq. (4.1), results in a Dirichlet process [263, 271, 360, 484].

In other words, provided M is sufficiently large, the estimates obtained through our model are independent of the particular value chosen (i.e., overfitting cannot occur).

With the nonparametric model just presented, although the total number of model molecular species is infinite, the actual number of molecular species contributing photons to the measurements is finite. Specifically, the number of contributing species coincides with the number of different tags s_k associated with $\boldsymbol{\Delta t}$. In other words, instead of asking *how many species contribute to the measurements?*, with our model, we ask *how many of the represented species actually contribute at least one photon?* Further, instead of asking *what are the lifetimes of these species?* we ask *what are the lifetimes of the species contributing at least one photon?* Of course, as we estimate rates instead of lifetimes, we obtain the latter by $\tau_m = 1/\lambda_m$.

With these priors, we form $p\left(\bar{\pi}, s_1, \ldots, s_K, \lambda_1, \lambda_2, \ldots | \boldsymbol{\Delta t}\right)$ which is the joint posterior probability distribution that includes all unknown variables. To compute this posterior, we develop a Markov Chain Monte Carlo (MCMC) scheme [358, 381] that generates pseudo-random samples with the appropriate statistics. The scheme is described in the supplementary information and a working implementation is also provided.

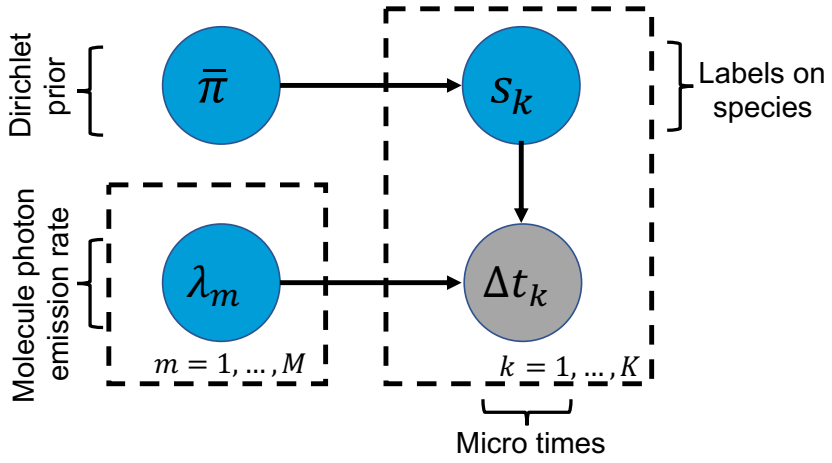

Fig. 4.4. **Graphical representation of the proposed model**. A simple graphical representation of the model, where Δt_k is the micro time k with $k = 1, \ldots, K$. The molecular emission rate of species m is shown by λ_m, $m = 1, \ldots, M$. The label s_k tells us which of the species is contributing the k^{th} photon. In the graphical model, the measured data are denoted by grey shaded circles and the model variables, which require priors, are designated by blue circles. Each one of the labels has a prior which is a Dirichlet probability $\overline{\pi}$.

4.3.3 Acquisition of Synthetic Data

The synthetic data presented in this study are obtained by standard pseudo-random computer simulations [383, 384, 386, 387, 485] that simulate a common fluorescence lifetime imaging modality with a conventional single-spot confocal setup. Further, in the simulations we consider confocal regions created with pulsed excitation. To generate data mimicking as closely as possible the measurements obtained in real experiments, we simulate freely diffusing molecules of different species characterized by different diffusion coefficients and lifetimes. Details and parameter choices are provide in the supplementary information, Tables 4.3 and 4.4.

4.3.4 Acquisition of Experiment Data

The synthetic data presented in this study are obtained as described below.

Sample preparation

Sample solutions of Rhodamine B (Rhod-B, Wako Pure Chemical Industries), Rhodamine 6G (Rhod-6G, Sigma-Aldrich), and tetramethylrhodamine-5-maleimide (TMR, Invitrogen), and Cy3 monofunctional NHS-ester (Cy3, GE Healthcare) were prepared with Milli-Q water at 1 μM concentration. Nonionic surfactant (0.01% Triton X-100) and 2 mM Trolox were added to prevent adsorption of dye molecules to the glass surface and reduce photophysical artifacts, respectively.

Experiments

Fluorescence lifetime measurements were carried out using a confocal fluorescence microscope with super continuum laser (Fianium SC-400-4, frequency of 40 MHz). The output of the laser was filtered by a bandpass filter (Chroma Technology D525/30 m), and focused onto the sample solution using a 60× objective lens (Nikon Plan Apo IR) with NA of 1.27. The excitation power was set to be 0.3 μW at the entrance port of the microscope. Fluorescence photons ware collected by the same objective lens and guided through a confocal pinhole as well as a bandpass filter (Chroma Technology D585/40 m), and then detected by a hybrid detector (Becker & Hickl HPM-100-40-C). For each photon signal detected, the routing information was appended by a router (Becker & Hickl HRT-82). The arrival time of the photon was measured by a TCSPC module (Becker & Hickl SPC-140) with the time-tagging mode [466]. The time resolution was evaluated by detecting the scattering of the incident laser light at a cover glass, and it was typically 180 ps at full width half maximum.

4.4 Results

Our goal is to characterize quantities that describe molecular chemistry at the data-acquisition timescales of FLIM with a focus on obtaining lifetime estimates. In order to estimate lifetimes, we also estimate intermediate quantities (namely molecule emission rates, and the fraction of interacting molecules) detailed in the method section.

Within the Bayesian nonparametric approach [328, 359, 360], our estimates take the form of posterior probability distributions over unknown quantities. These distributions combine parameter values, probabilistic relations among different parameters, as well as the associated uncertainties. To quantify this uncertainty, we calculate a posterior variance and use this variance to construct error-bars (i.e., credible intervals). According to the common statistical interpretation [358, 359], the sharper the posterior, the more conclusive (and certain) the estimate [480].

We first validate our approach on synthetic data where the ground truth is available. We then test our method on experimental data. For the latter case, we compare our analyses to the results obtained from both TCSPC and phasor plot methods used in FLIM.

4.4.1 Method Validation using Synthetic Data

To show the robustness of our method, we generate synthetic traces where molecules are immobilized under a broad range of: i) different number of photon arrivals, Fig. 4.5 with multiple species, Fig. 4.6; ii) different fraction of molecules contributing photons from different species, Fig. 4.7; and iii) resolution of lifetime (or the two closest lifetimes) as a number of photons obtained grows, Fig. 4.8. All parameters not explicitly varied are held constant across all figures. The parameters not varied are held fixed

at the following baseline values: lifetime between 1 ns and 10 ns which is the typical lifetime range of a fluorophore [452,486], two species which is most frequent in related studies [442, 451, 452], and fraction of molecules contributing photons from different species 50% : 50%.

Also, in the supplementary information, we worked cases with three and four different species (as opposed to a just one or even two species) as this scenario presents the greatest analysis challenge because very few photons, and thus little information, is gathered on each species. In a similar spirit, we also default to short traces that cannot meaningfully be analyzed using TCSPC and phasor approaches as illustrated in Fig. 4.1. Moreover, since the mathematics is identical, our proposed method applies also in the case when molecules are diffusing inside the confocal volume. We show in the supplementary information, Figs. 4.11 and 4.12, the results for freely diffusive molecules.

Number of photons

We benchmark the robustness of our approach with respect to the length of the trace (i.e., the total number of photon arrivals) at fixed number of species, lifetime, and molecule photon emission rate. The first important conclusion is that, for the values of parameters selected, we need at least one order of magnitude less data than both TCSPC and phasor analysis; see Fig. 4.1. For instance, to obtain an estimate of the lifetime within 10% of the correct result in the one species case, our method requires $\approx$ 100 photons (emitted from the species of interest), while both TCSPC and phasor require $\approx$ 1000 photons to determine the lifetime to within the same error bar. In the case of two species traditional approaches need at least 3×10^4 photons in comparison to $\approx$ 3000 for our proposed BNP approachs; see Figs. 4.5 and 4.6.

To determine how many photons were required for our method, we chose the mean value of the lifetime posterior, and measure the percentage difference of this mean to the ground truth known for these synthetic traces. For FLIM, in one species lifetime analysis case we require $\approx$ 1000 photons for same accuracy as our approach, and for the case of mixture of double species analysis this number arise to 10^4 or more photons [482, 487].

In general, the difference in the photon numbers demanded by our method and traditional analysis depend on a broad range of experimental parameter settings. This is the reason, we explore different settings–holding all other settings fixed–in subsequent subsections as well as the supplementary information.

Another important concept, illustrated in Figs. 4.1, 4.5, and 4.6 that will keep re-appearing in subsequent sections, is the concept of a photon as a *unit of information*. The more photons we have, the sharper our lifetime estimates. This is true, as we see in these figures, for increasing trace length. Similarly, as we will see in subsequent subsections, we also collect more photons as we increase the contribution of labeled molecules (and thus the number of molecules contributing photons to the trace).

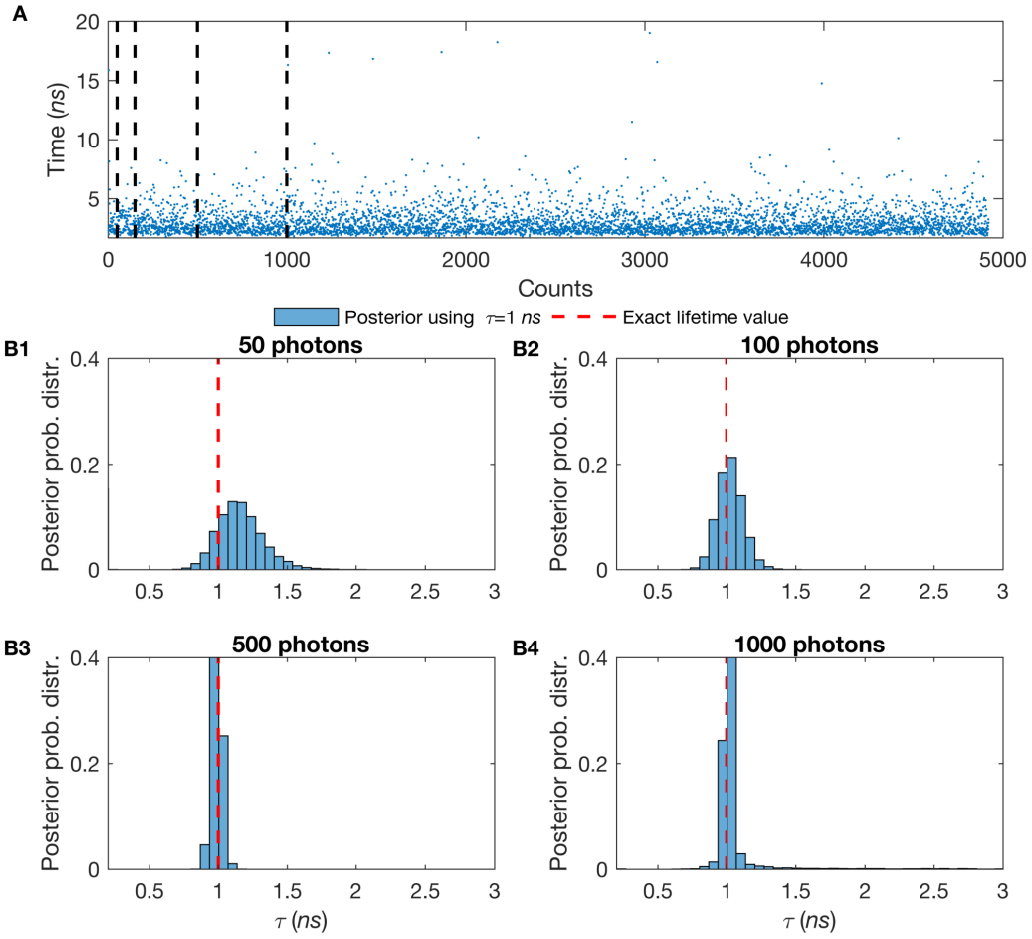

Fig. 4.5. **Effect of the number of detected photons on a single molecular lifetime estimation. The more photons per unit time and thus the sharper estimation of lifetime.** (A) Here, we work on single species lifetime while all molecules are immobilized. The synthetic trace generated by $\tau = 1\ ns$. The blue dot represents a single photon arrival time. The excitation pulses happen at frequency of 40 MHz and we consider then to have a Gaussian shape with standard deviation of 0.1 ns. We start with 50 photons (B1) and gradually increase the number of photons to 100 (B2), 500 (B3), and 1000 (B4) photons. The ground truth for the lifetime is known (as this is synthetic data) and it is shown by red dash line.

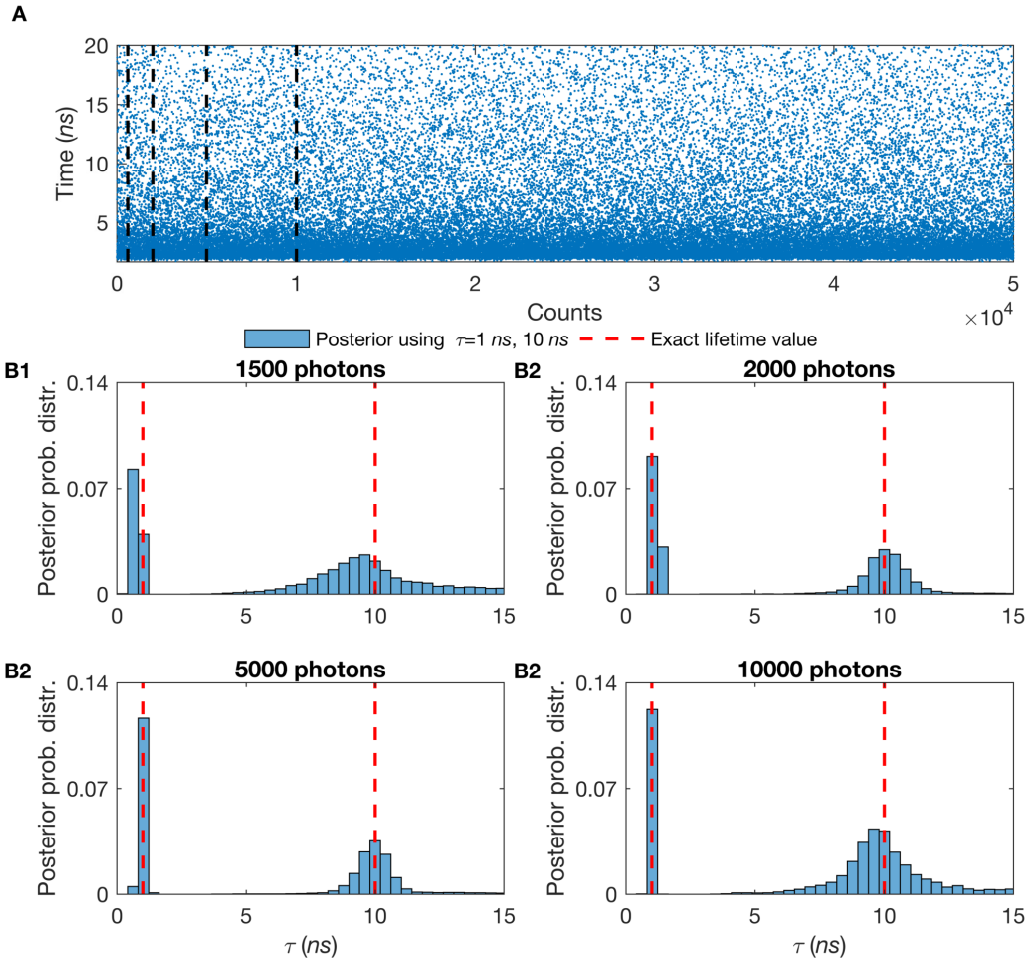

Fig. 4.6. **Effect of the number of detected photons on two molecular lifetimes estimation. The larger trace length has more photons per unit time and thus sharper estimation of lifetime for two species case.** (A) Here, we work on double species lifetimes while all molecules are immobilized. The synthetic trace generated by $\tau = 1\ ns$ and $\tau = 10\ ns$ with fraction of contributing molecules from different species of 50% for each of them (50% – 50%). The blue dot represents a single photon arrival time. We start with 1500 photons (B1) and gradually increase the number of photons to 2000 (B2), 5000 (B3), and 10000 (B4) photons. Here, all other features such as the frequency of acquisition and width of pulse are the same as in Fig. 4.5. Also, we follow the same red-dashed line convention. To see the results for more than two species see the supplementary information, Figs. 4.14 and 4.15.

Fraction of molecules contributing photons from different species

To test the robustness of our method when different species contribute an uneven number of photons, we simulate data with 70% of the population in species 1 and 30% in species 2 (Fig. 4.7A). We also considered fractions of contributing molecules from different species of 50% : 50% (Fig. 4.7 B), and 30% : 70% (Fig. 4.7 C). For all cases, the lifetimes were fixed at 1 ns and 10 ns for $\approx$ 3000 photon arrivals. Fig. 4.7 summarizes our results and suggests that posteriors over lifetimes are broader–and thus the accuracy with which we can pinpoint the lifetimes drops–when the contribution of labeled molecule is lower. Intuitively, we expect this result as fewer species within the confocal volume provide fewer photons and each photon carries with it information that helps refine our estimated lifetimes.

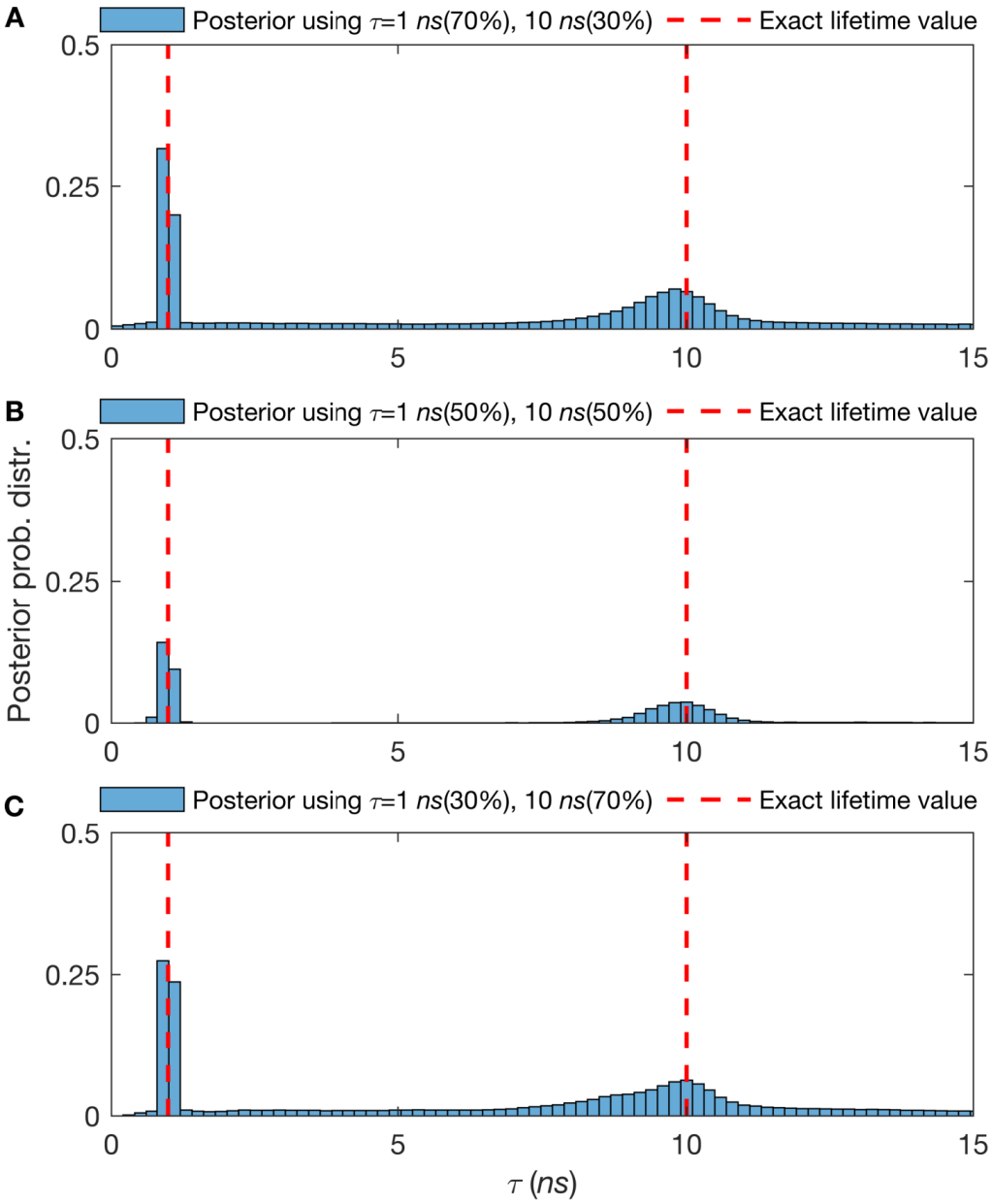

Fig. 4.7. **Effect of the relative fraction of contributing molecules from different species on molecular lifetime estimation. Higher molecular contributions provide more photons per unit time and thus sharper lifetimes estimates**. (A-C) The posterior probability distributions of traces with lifetimes of 1 ns and 10 ns, with 3000 total photons and fraction of contributing molecules from different species of $70\% - 30\%$, $50\% - 50\%$ and $30\% - 70\%$ respectively. Here, all other features such as the frequency of acquisition and width of pulse are the same as in Fig. 4.5. Also, we follow the same red-dashed line convention. For more details see supplementary information Fig. 4.16.

Lifetime resolution

We repeat the simulations with two species and ask about how many photons are required to resolve similar lifetimes. Here, we have shown the dependency of the time resolution to the number of collected photons in Fig. 4.8. As expected, the number of photons required to resolve increasingly similar lifetimes grows as the ratio of lifetimes approaches unity. However, this also suggests that if we were to resolve species of similar lifetimes, we could use the amount of data typically used in TCSPC or phasor analysis to resolve these while TCSPS or phasor analysis would still require an additional order of magnitude more data. As a note, they had to impose by hand how many species we have while in our method, number of species were learnt. Moreover, if we know number of species we require even less number of photons that we have claimed.

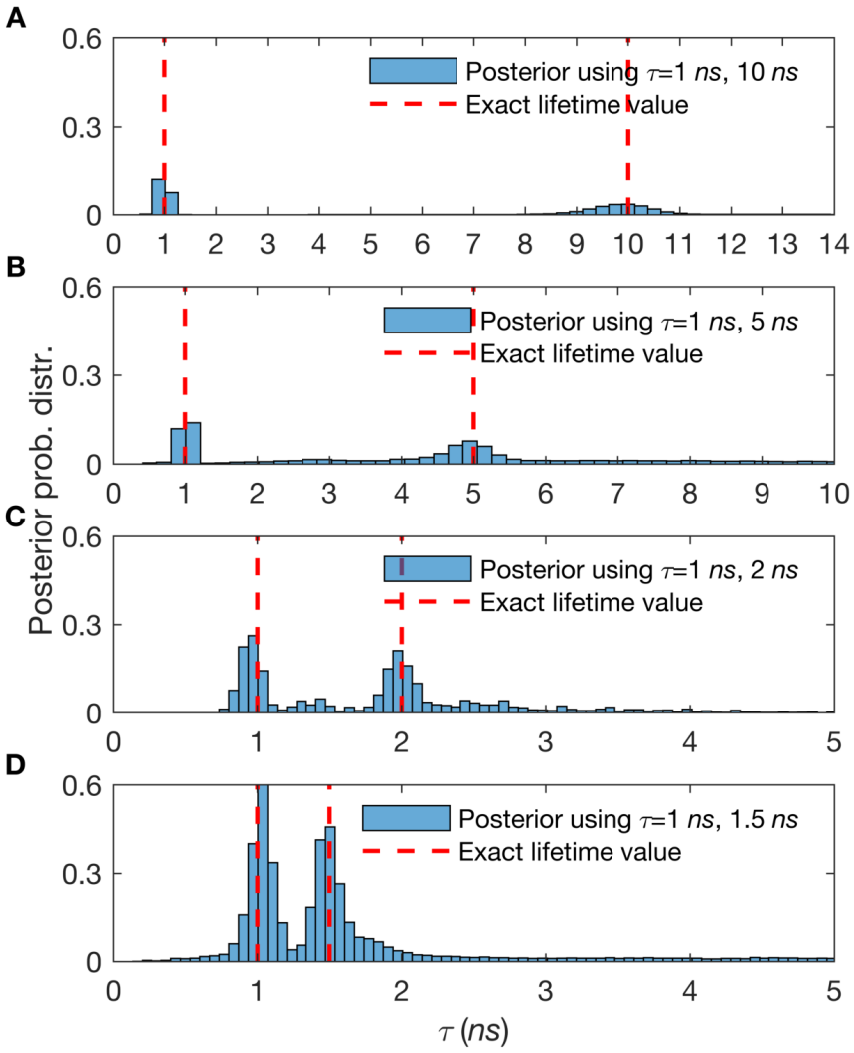

Fig. 4.8. **Lifetime resolution for double species lifetimes.** The synthetic traces are acquired for total of 3000 to 20000 photon arrivals and start with lifetimes of 1 ns and 10 ns ($\approx$ 3000 photons) and gradually make the lifetimes closer to each other. (B) 1 ns and 5 ns ($\approx$ 3000 photons), (C) 1 ns and 2 ns ($\approx$ 10000 photons), and (D) 1 ns and at last 1.5 ns ($\approx$ 20000 photons). The fraction of molecules contributing photons from different species in the total photon budget is equal ($50\% - 50\%$). Here, all other features such as the frequency of acquisition and width of pulse are the same as in Fig. 4.5. Also, we follow the same red-dashed line convention. Posterior probability distribution over the lifetimes estimated from the trace has been shown.

4.4.2 Estimation of physical parameters from experimental data

To evaluate our approach on real data, we used experimental data collected under a broad range of conditions. That is, we used measurements from different fluorophores, namely Cy3, TMR, Rhod-B, and Rhod-6G. Figs. 4.9 and 4.10 were collected using the Rhod-B and Rhod-6G dyes and these results were used to benchmark the robustness of our method on individual species as well as mixtures of species with different fraction of molecules contributing photons from different species. In the supplementary information, Fig. 4.17, we have shown more experimental result for the case of more than two species.

In Fig. 4.1, we verified our method on Rhod-6G with respect to the total number of photon arrivals. The first important conclusion is that we need at least one order of magnitude less data than both TCSPC and phasor analysis to obtain an estimate of the lifetime within 10% of the correct result. That is, we need $\approx$ 100 photons, while both TCSPC and phasor require $\approx$ 1000 photons. For two or more species the situation for both TCSPC and phasor grows more challenging for two reasons. First, the number of species cannot be independently determined and, even if known, at least $\approx$ 10000 photons are required to determine lifetimes to within 10% of the correct result. The percent to within the correct result is computed just as we had before in Sec. 4.4.1.

In general, the difference in the photon numbers demanded by our method and traditional analyses depend on a broad range of experimental parameter settings. This is the reason, we explore different settings–holding all other settings fixed–just as we did with synthetic data in subsequent subsections as well as the supplementary information.

Benchmarking on experimental data using a different number of photons for mixtures of Rhod-B and Rhod-6G

Similarly to the synthetic data analysis appearing in Fig. 4.6, we benchmark the robustness of our approach with respect to the length of the trace (i.e., the total number of photon arrivals) given fixed lifetimes and fraction of interacting molecules at 50% : 50%. Again the important conclusion is that, for the values of parameters selected, we need at least one order of magnitude less data than both TCSPC and phasor analysis; see Fig. 4.1 for the analysis of one species, Rhod-6G, and Fig. 4.9 for the analysis of two species. For instance, to obtain an estimate of the lifetime within 10% of the correct result for the case of two species, our method requires ≈ 3000 photons, while both TCSPC and phasor require $\approx 3 \times 10^4$ photons.

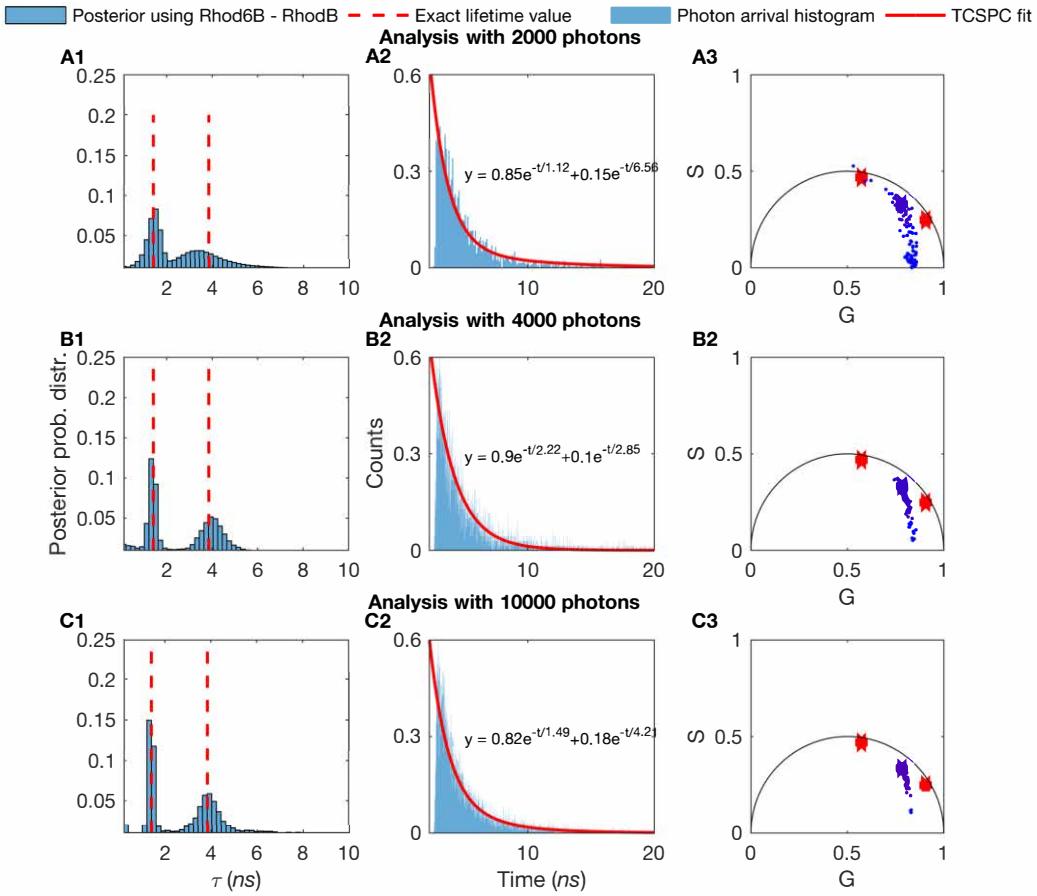

Fig. 4.9. **Comparison of number of photons needed to assess the lifetimes of mixtures of Rhod-B and Rhod-6G.** In (A1-A3) we use 2000 photons and compare all three methods: (A1) BNPs, (A2) TCSPC, and (A3) phasor analysis. In (B1-B3) and (C1-C3) we repeat the analysis for 4000 and then 10^4 photons.

Benchmarking on experimental data using different fractions of Rhod-B and Rhod-6G

We start by evaluating our method on mixtures of Rhod-B and Rhod-6G but present in different amounts. Similarly to Fig. 4.7 for the analysis of two species from synthetic data, we show estimates of the lifetimes for two species, Rhod-B and Rhod-6G, at 70% : 30% fraction (Fig. 4.10A), at 50% : 50% fraction (Fig. 4.10B),

and at 30% : 70% fraction (Fig. 4.10C). Fig. 4.10 summarizes our results and suggests that posteriors over lifetimes are broader–and thus the accuracy with which we can pinpoint the lifetimes drops–when the contribution from the dye concentration for that species is lower. Same as before, we need at least one order of magnitude less data than both TCSPC and phasor analysis; see Fig. 4.10. For instance, to obtain an estimate of the lifetime within 10% of the correct result, our method requires ≈ 3000 photons directly emitted from the dye, while both TCSPC and phasor analyses become more challenging and for two species case requires at least $\approx 3 \times 10^4$ photons (assuming the number of species is known). In the supplementary information we show additional results for the case of three and four species with different contribution where the number of photons required for analysis grows substantially in phasor and TCSPC analysis.

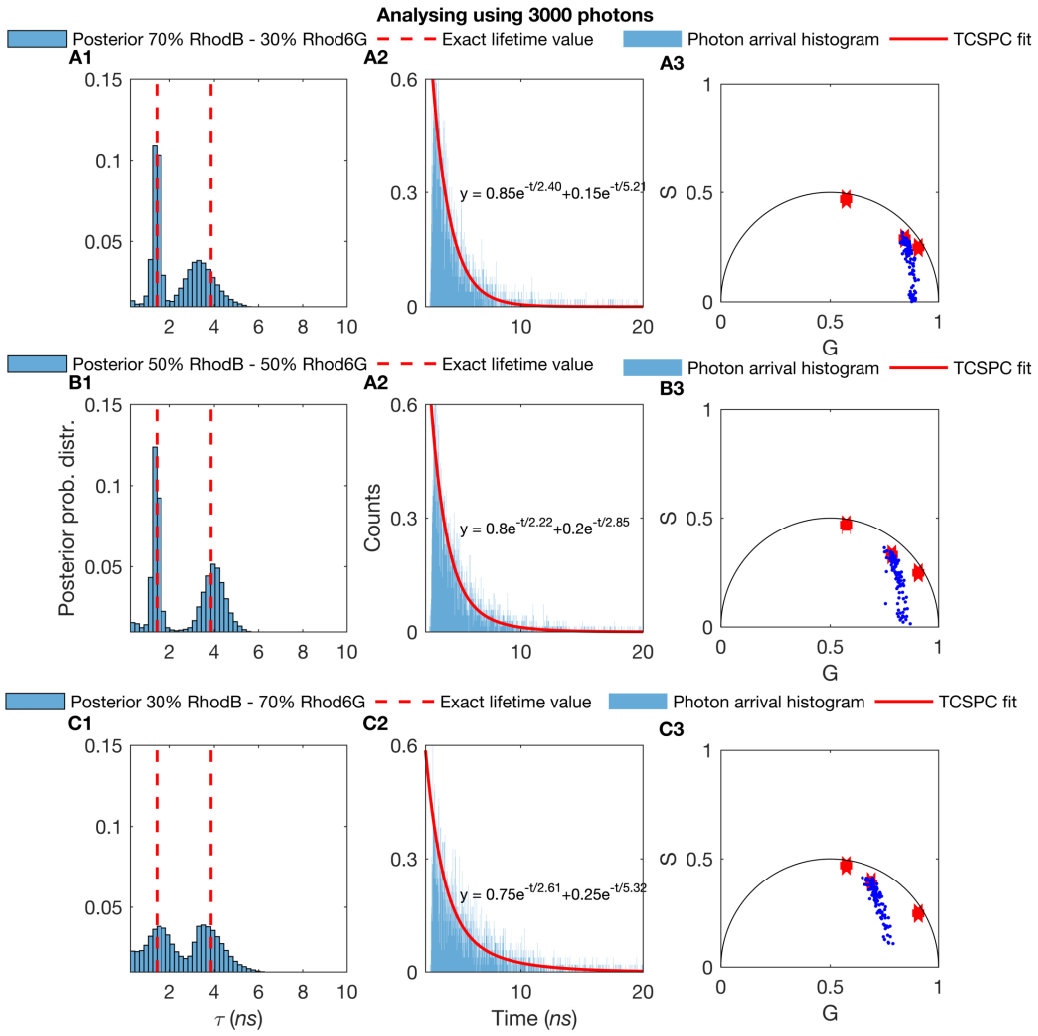

Fig. 4.10. **Effect of the fraction of molecules contributing photons from different species on molecular lifetime estimates. Higher molecular contributions provide more photons per unit time and thus sharper lifetime estimates**. (A1-A3) The experimental trace is selected using two species, Rhod-B and Rhod-6G, with a total of about 3000 photon arrivals with fraction of molecules contributing photons from different species $(70\% - 30\%)$. (A1) BNPs, (A2) TCSPC, and (A3) phasor estimations. In (B1-B3) and (C1-C3) we repeat the analysis for fraction of $(50\% - 50\%)$ and $(30\% - 70\%)$

4.5 Discussion

Across all spectroscopic and imaging applications, the photon is the basic unit of information [395, 480]. Decoding information directly from single photon arrivals, with as few photons as possible without binning or correlating or other pre-processing of the data, is the main focus of all data-centric analysis strategies. Yet decoding information directly from single photon arrivals presents fundamental model selection problems.

For example, in the case of FCS, if we are to learn diffusion coefficients directly from limited photon arrivals, we must know how to write down a likelihood or, put differently, we must know the number of molecules contributing photons that, in turn, dictate the form for the likelihood [480]. As we do not know how many molecules we have, and what the appropriate likelihood should be, we have a model selection problem. Similarly, for lifetime imaging, if we are to learn the lifetime of the chemical species contributing photons, we must also know the number of species in order to write down a conventional likelihood.

Traditional Bayesian methods do not have a direct solution to the model selection problem [79,328] as they also require us to be able to write down a likelihood. That is, they consider a fixed model (and a fixed likelihood) and treat the model's parameters as random variables of the posterior distribution. By contrast, BNP, which are a direct logical extension of parametric Bayesian methods, treat models alongside their parameters as random variables [263, 488–492].

This ability to treat models themselves as random variables is the key technical innovation that prompted the development of BNP in the first place. It makes it is possible to avoid the computationally infeasible task of first enumerating and second comparing all models for any associated parameter values to all other competing models and their associated parameter values.

The BNP approach to tackling lifetime image analysis that we propose here cannot replace phasor analysis [443, 451, 472, 474, 476, 477] or TCSPC [430, 445, 456, 467, 482] for simple one component systems on account of their computational efficiency. However, at an acceptable computational cost, BNP approaches provide an alternative. They give us the ability to: determine the number of species; use much less data to obtain lifetime estimates (and thus reduce photo-toxic damage to a light-sensitive sample); use longer photon arrival traces, if available, to tease out small differences in lifetimes between species as BNP-based methods are more data efficient; probe processes resolved on faster timescales (again, as we require fewer photons); exploit all information encoded in the photon arrivals (and thus not require separate control experiments, as needed in phasor approaches, for the measurement of the lifetime of one species to determine the lifetime of a second species when a mixture of two species, say, is present).

As for the computational cost, obtaining lifetimes (to within 10% of the ground truth lifetime for a one-species for the parameters we used in Figs. 4.5 and 4.1 requiring $\approx$100 photons) takes 5 minutes on a typical desktop (based on a system with 6G RAM, Core (TM) i7-2.67 GHz CPU). For a two-species mixture, Figs. 4.6 and 4.9 , under the same parameters and requiring 3000 photons, it was a modest increase to 15 minutes. The point, here, is that the analysis of single or multi-species data *can* be performed with an average desktop computer and it does not necessarily require high performance computing facilities.

The real strength of BNP becomes clear when we reach two, three, four or possibly even more species. Beyond being able to work with low photon counts, another key advantage of our method is its flexibility. The ability to use BNP, and treat models as random variables, in lifetime imaging is the real point here and, as such, our framework can be adapted to treat a range of experimental setups.

In particular, our framework can straightforwardly be adapted to treat: any IRF by modifying Eq. 4.2 as appropriate; and any background photon arrival statistics or detector dark counts by modifying Eq. 4.3 especially relevant to *in vivo* imaging. In the supplementary information, Fig. 4.13, we evaluated our method respect to different background levels to see how it behaves with different number background photons.More significant, and challenging, extensions of our work would be to consider lifetime changes over the timescale of data acquisition as may be expected in complex *in vivo* environments [493, 494].

We may not be able to provide answers for dealing with these types of complex questions yet. However, BNP gives us a genuinely different way to think about problems. They suggest productive paths forward to tentatively formulate inverse strategies to unravel processes of life that we already know to be encoded within *in vivo* photon arrival traces.

Supplementary Information

In this supplement, we present additional analyses and technical details expanding upon the material presented in the main text. These include: (i) additional analysis of synthetic and experimental traces that include the estimation of lifetimes and the fraction of interacting molecules; (ii) additional details on the theoretical approaches used; and (iii) a complete description of the inference framework developed that includes choices for prior probability distributions and a computational implementation.

Additional results

Analysis of additional synthetic data

In the main text we focused on the estimation of: lifetime, τ, with values less than 10 ns which are typical lifetime values in *in vivo* applications [493]. Here, we explore broader parameter ranges from freely diffusive molecules, Figs. 4.11 and 4.12 to the case when we have different background photons, Fig. 4.13, which we evaluated our method respect to different background levels to see how it behaves with different number background photons. Moreover, we evaluated our method in the cases with more than two species, Figs. 4.14 and 4.15, and estimate the fraction of molecules contributing photons from different species, Fig. 4.16, that we explain in the main text in Sec. 4.4.1.

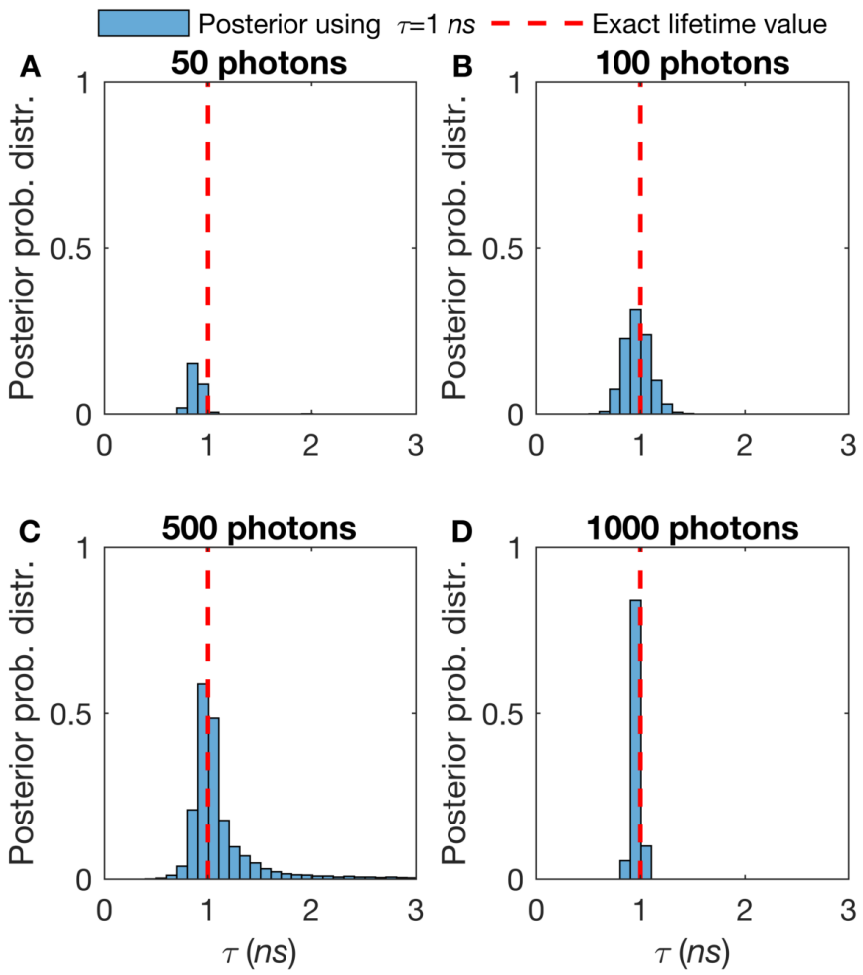

Fig. 4.11. **Effect of the number of detected photons on a single diffusive molecular lifetime estimation. The more photons per unit time and thus the sharper estimation of lifetime.** Here, we work on single species lifetime while all molecules are diffusing with diffusion coefficient, $D = 10\ \mu m^2/s$. The synthetic trace generated by $\tau = 1\ ns$. We start with 50 photons (A) and gradually increase the number of photons to 100 (B), 500 (C), and 1000 (D) photons. The excitation pulses occur at a frequency of 40 MHz and we assume that these pulses assume a Gaussian shape with standard deviation of 0.1 ns. The ground truth for the lifetimes are known (as this is a synthetic data) and they are shown by red dash lines.

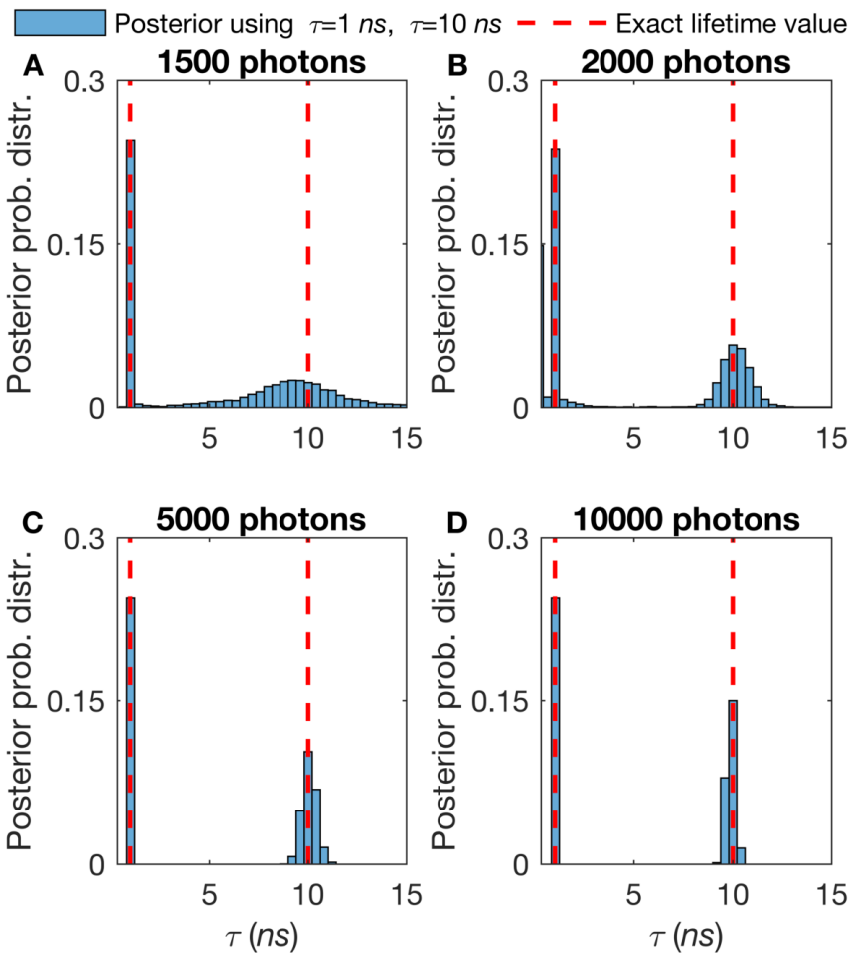

Fig. 4.12. **Effect of the number of detected photons on a double diffusive molecular lifetime estimation. The more photons per unit time and thus the sharper estimation of lifetime.** Here, we work on single species lifetime while all molecules are diffusing with diffusion coefficient, $D = 10\ \mu m^2/s$. The synthetic trace generated by $\tau = 1\ ns$ and $\tau = 10\ ns$. We start with 1500 photons (A) and gradually increase the number of photons to 2000 (B), 5000 (C), and 10000 (D) photons. Here, all other features such as the frequency of acquisition and width of pulse are the same as in Fig. 4.11. Also, we follow the same red-dashed line convention.

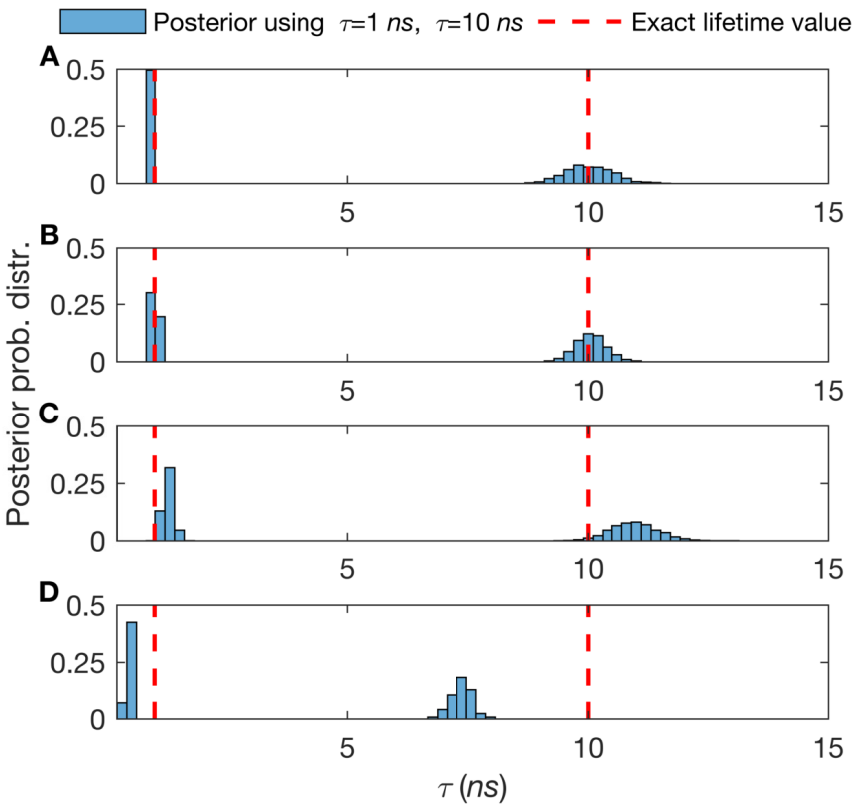

Fig. 4.13. **Effect of the number of background photons on a double diffusive molecular lifetimes estimation. The more background photons per unit time and thus the poorer estimation of lifetime.** Here, we work on double species lifetime while all molecules are diffusing with diffusion coefficient, $D = 10\ \mu m^2/s$. The synthetic trace generated by $\tau = 1\ ns$ and $\tau = 10\ ns$ with total 3000 photons. We start with 3 background photons (A) and gradually increase the number of background photons to 30 (B), 150 (C), and 300 (D) photons. Here, all other features such as the frequency of acquisition and width of pulse are the same as in Fig. 4.11. Also, we follow the same red-dashed line convention.

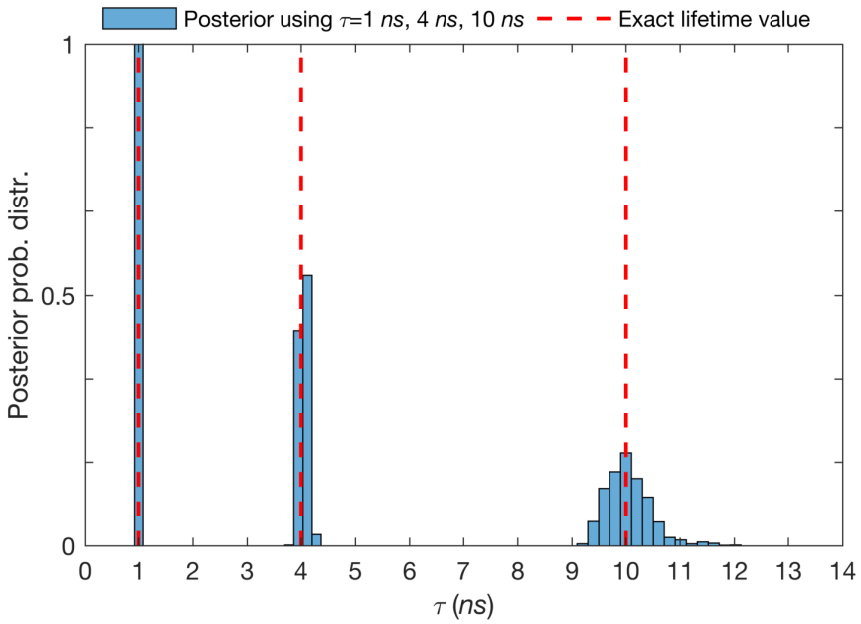

Fig. 4.14. **Lifetime estimation with three different species using synthetic data.** Here, we generate a synthetic trace with three species having lifetimes $\tau = 1\ ns$, $\tau = 4\ ns$ and $\tau = 10\ ns$ with equal fraction of molecules contributing photons from different species of 33% for each of them and total 2×10^5 photon arrivals. Here, all other features such as the frequency of acquisition and width of pulse are the same as in Fig. 4.11. Also, we follow the same red-dashed line convention.

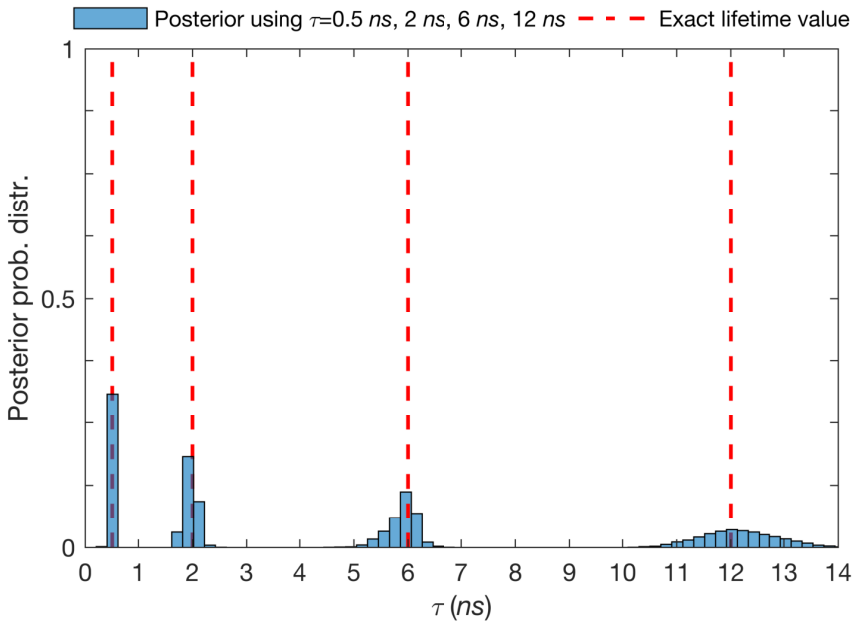

Fig. 4.15. **Lifetime estimation with four different species in synthetic data.** Here, we work with four species lifetimes while all molecules are immobilized. The synthetic trace generated by $\tau = 0.5\ ns$, $\tau = 2\ ns$, $\tau = 6\ ns$ and $\tau = 12\ ns$ with equal fraction of interacting molecules of 25% for each of them and total 3×10^5 photon arrivals. Here, all other features such as the frequency of acquisition and width of pulse are the same as in Fig. 4.11. Also, we follow the same red-dashed line convention.

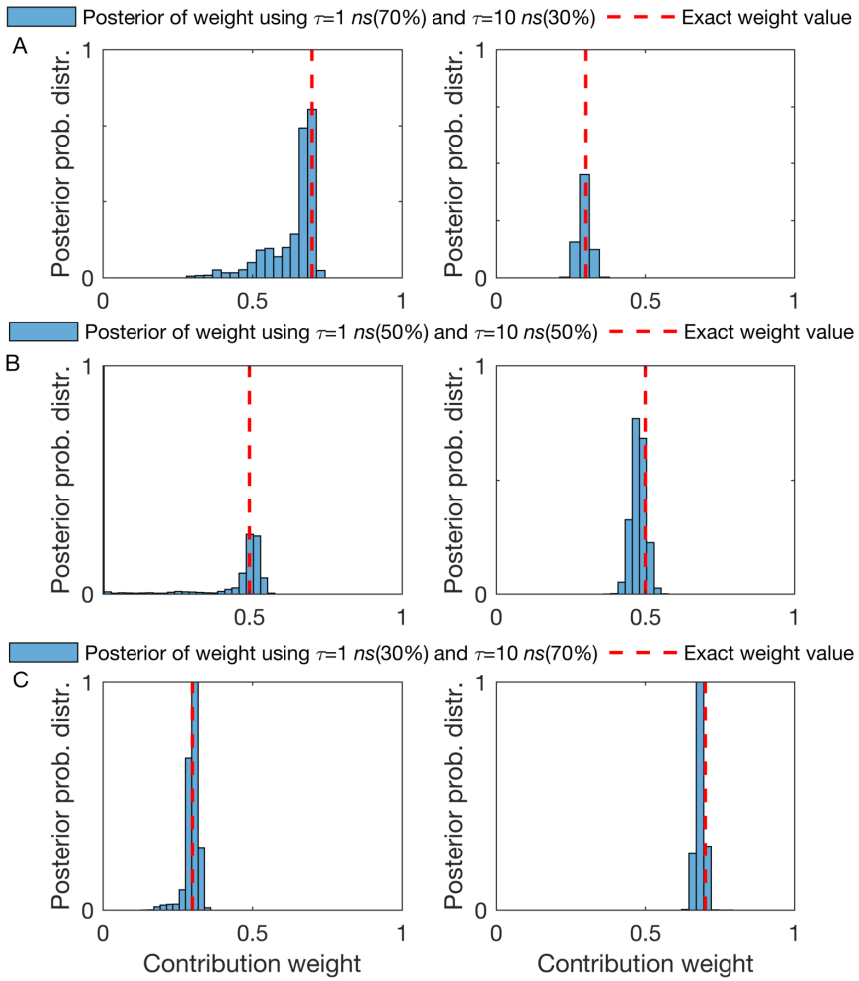

Fig. 4.16. **Estimation of the fraction of molecules contributing photons from different species**. (A-C) Using same synthetic traces as Fig. 4.7, the posterior probability distribution over the fraction of molecules contributing photons from different species (weight) with lifetimes of 1 ns and 10 ns, 3000 total number of detected photons and fraction of interacting molecules of $70\% - 30\%$, $50\% - 50\%$ and $30\% - 70\%$ respectively. Here, all other features such as the frequency of acquisition and width of pulse are the same as in Fig. 4.11. Also, we follow the same red-dashed line convention.

Analysis of additional experimental data

Here, we used real measurements, obtained as explained in the method section, from different fluorescent dyes, namely Cy3, TMR, Rhod-B, and Rhod-6G. In Fig. 4.17 we considered a mixture of all four species. In Fig. 4.18 we show that we can correctly identify the fraction of molecules contributing photons from different species.

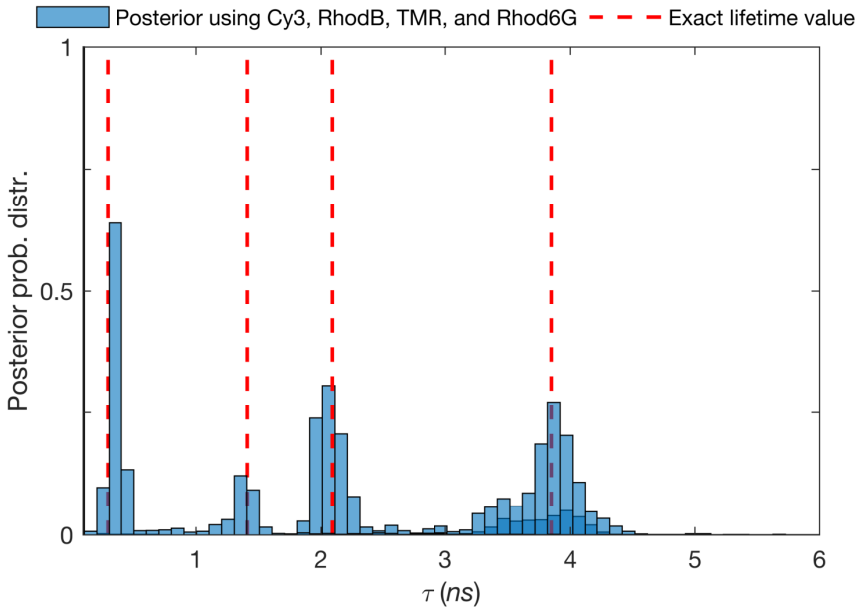

Fig. 4.17. **Lifetime estimation for the case of four different species from experimental data.** Here, we work on four species lifetimes while all molecules are immobilized. The experimental trace generated by four different dyes including Cy3, Rhod-B, TMR, and Rhod-6G with a total of $\approx 3 \times 10^5$ photon arrivals. The excitation pulses occur with a frequency of 40 MHz and we assume that these pulses assume a Gaussian shape with standard deviation of 0.1 ns. The ground truth estimates (as we do not have real ground truths for real data) for the lifetimes are determined using the whole trace which includes total 1.4×10^6 photon arrivals and they are shown by red dash lines.

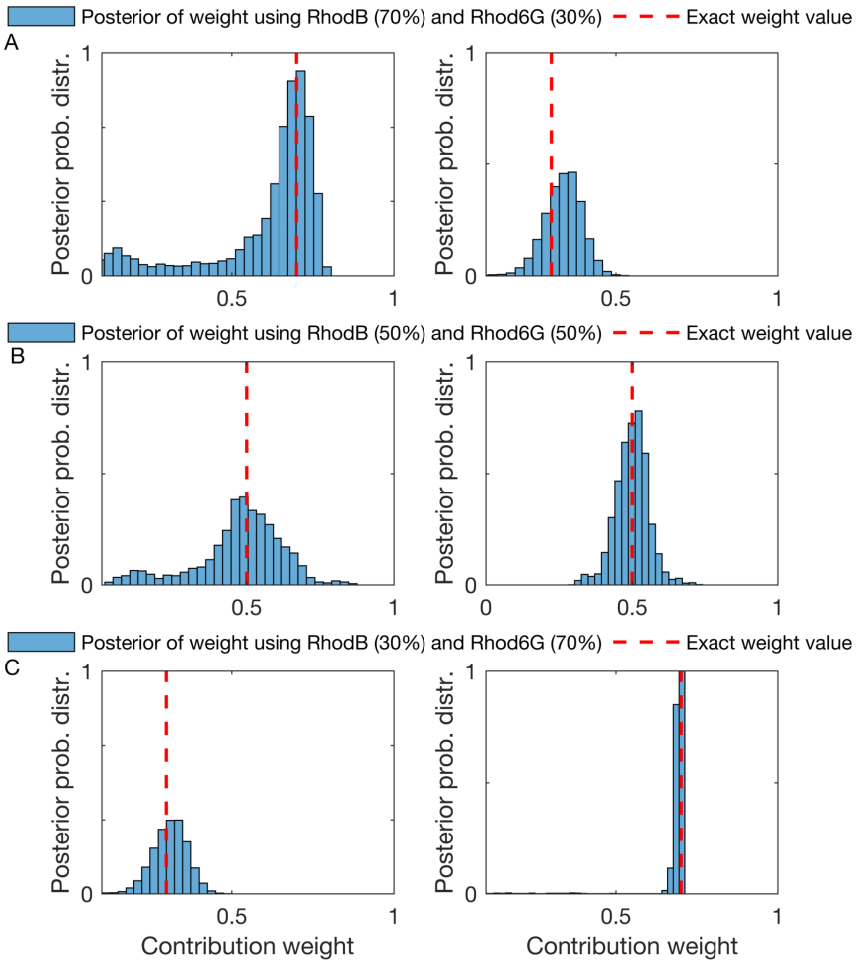

Fig. 4.18. **Estimation of the different fraction of molecules contributing photons from different species for the experimental trace**. (A-C) Using same traces as Fig. 4.10, the posterior probability distributions of fraction of interacting molecules (weight) for experimental dyes, RhodB and Rhod6G, with total ≈ 3000 total number of detected photons and fraction of interacting molecules of $70\% - 30\%$, $50\% - 50\%$ and $30\% - 70\%$ respectively. The excitation pulses happen at frequency of of 40 MHz and we consider then to have a Gaussian shape with standard deviation of 0.1 ns. The ground truth estimates (as we do not have real ground truths for real data) for the lifetimes are determined using the whole trace which includes total 1.4×10^6 photon arrivals and they are shown by red dash lines.

Brief description of FLIM analyses with TCSPC and phasor plots

Time domain

In typical time-domain lifetime imaging, a pulsed laser is used to excite the sample periodically, causing fluorescence emission for those pulses where a molecule is excited and decays back to the ground state radiatively. Experimentally, based on the data we presented in Fig. 4.19, this is typically 1 in 40 pulses [452].

The fluorescence decay of different species with distinct fluorescence lifetimes can be modeled by a mixture of exponential distributions though whew we need to be careful to convolve the fluorescence intensity with the measured IRF; see Eq. 4.16.

At present most of time-domain measurement analysis is performed using TCSPC [452, 467, 482].

Frequency Domain

Frequency-domain or phase-modulation experiments constitute an alternative way to measure excited state lifetimes. In this case, the sample is excited with an intensity-modulated light, typically a sine-wave modulation [452]. When a fluorescent sample is excited in this way, the emission intensity follows a shifted modulation (m) pattern with the phase shift (ϕ) and peak height that both encode information on the excited state lifetime [452].

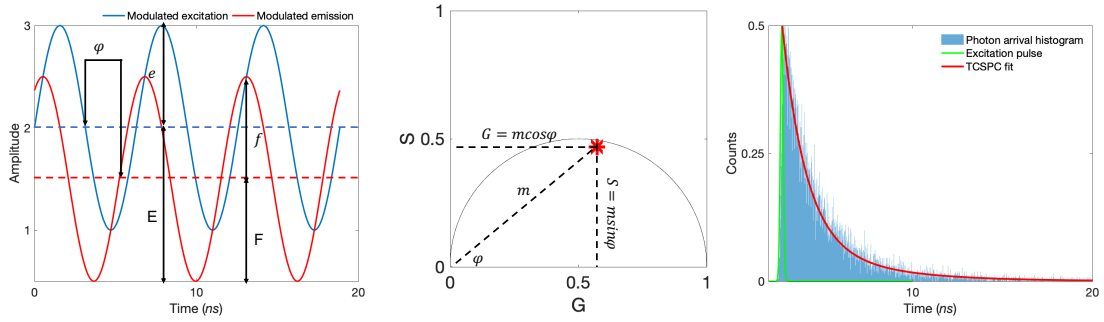

Fig. 4.19. **Comparison between time-domain and frequency-domain FLIM analysis.** Mapping frequency-domain (left) and TCSPC (right) data to the phasor plot (middle).

The modulation of the excitation is given by $\frac{e}{E}$, where e is the average intensity and E is the peak-to-peak height of the incident light (Fig. 4.19). The modulation of the emission is defined similarly, $\frac{f}{F}$, except using the intensities of the emission (Fig. 4.19). The shifted modulation between emission and excitation, $m = \frac{\frac{f}{F}}{\frac{e}{E}}$. The other experimental observable is the phase shift, (ϕ) which is the phase difference between excitation and emission. Both phase shift (ϕ) and the shifted modulation between emission and excitation (m) can be employed to calculate the lifetime using

$$\tan\phi = \omega\tau_\phi \tag{4.8}$$

$$m = \frac{1}{\sqrt{1+\omega^2\tau_m^2}} \tag{4.9}$$

These expressions can be also be used to calculate the phase (τ_ϕ) and shifted modulation (τ_m) lifetimes for the curves shown in Fig. 4.19. If the intensity decay is a single exponential, then Eqs. 4.8 and 4.9 yield the correct lifetime. In this case, both τ_ϕ and τ_m are equal. For more than one species these two are not same and we

have more calculation to extract the lifetimes. For more details regarding more than one species we refer interested readers to see Ref. [452]

Description of the pulsed excitation and microtimes simulation

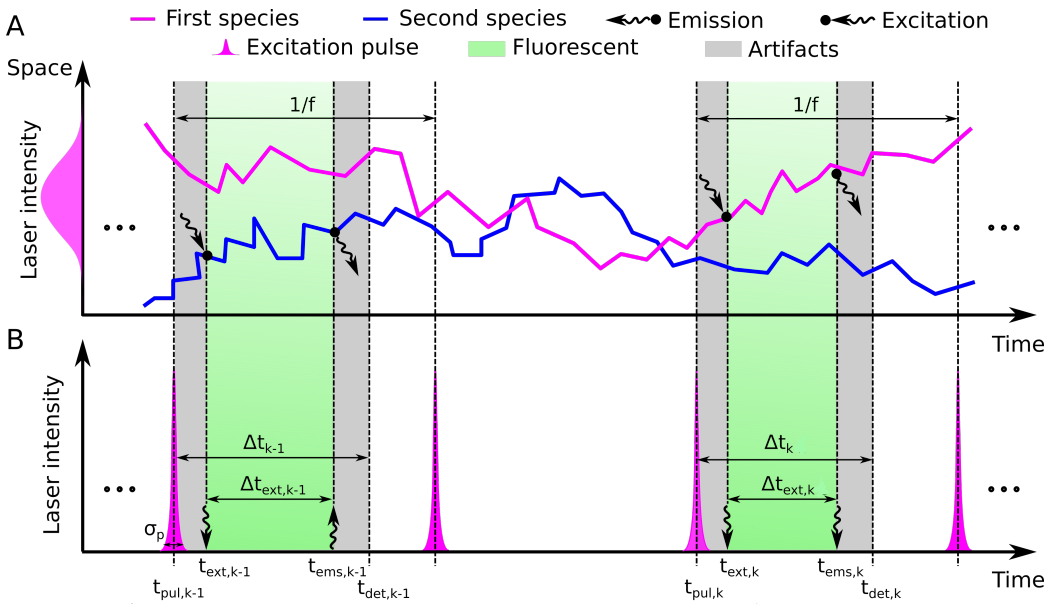

Fig. 4.20. **Pictorial representation of the experimental setup a sample with mixture of two species.** (A) The Brownian motion of two species in space versus time. Excitation and emission points are shown with different arrows. (B) The pulses and emission times will result in the micro-times as our observation which is the time between peak of pulse $t_{pul,k}$ that trigger the k^{th} photon detection and detection time $t_{det,k}$. The time between the excitation $t_{ext,k}$ and emission $t_{ems,k}$ of the molecule, $\Delta t_{ext,k}$ follows the molecular lifetime.

To simulate experimentally realistic microtimes, for mobile particles, we simulate diffusive molecules which freely traverse through an illuminated confocal volume. We define periodic boundaries [] ($\pm L_x$, $\pm L_y$, $\pm L_z$) which are much larger than the confocal radii to maintain a constant concentration of molecules. The confocal volume

itself is pulsed on and off and the probability of excitation of a molecule depends on its location within that volume during the pulse. Here we consider the confocal volume (the combined excitation and emission point spread function, PSF) to be a 3D Gaussian, with radii of $\omega_x = 0.3$ μm, $\omega_y = 0.3$ μm, $\omega_z = 3.5$ μm and centered at the point of origin. The precise formula for this PSF is

$$\mathbf{PSF}\left(x, y, z\right) = \exp\left(-2\left(\left(\frac{x}{\omega_x}\right)^2 + \left(\frac{y}{\omega_y}\right)^2 + \left(\frac{z}{\omega_z}\right)^2\right)\right). \tag{4.10}$$

So, the emission that received by molecule n of the m^{th} is species

$$\mu_{m,n} = \mu_{m,ext}\mathbf{PSF}\left(x, y, z\right) \tag{4.11}$$

where, $\mu_{m,ext}$ is the maximum excitation rate of the molecule n of species m which occurs when the molecule is at the center of the confocal volume [418].

Assuming that molecules do not move significantly over the duration of the pulse (of typical width 0.1 ns [495]), the probability of excitation of molecule n of species m is $q_{m,n} = \mu_{m,n}\delta t_p$ where, δt_p is the duration of the pulse. So, for any pulse excitation, we need to determine if the n^{th} molecule of species m is excited or not. We define the variable $b_{m,n}$ to be either 1 or 0 if the molecule emits or does not emit a photon and consider this variable to be Bernoulli distributed

$$b_{m,n} \sim \mathbf{Bernoulli}\left(q_{m,n}\right). \tag{4.12}$$

At the end, when a molecule is excited by each pulse $b_{m,n} = 1$, we need to consider the delays and errors introduced by the measuring electronic devices, $t_{det,k} - t_{ems,k}$. Since, we consider these errors follow a normal distribution, and the excitation time

is normal distributed as well, we denote both effects with $\Delta t_{err,k} = (t_{ext,k} - t_{pul,k}) + (t_{det,k} - t_{ems,k})$ and as the result, we sample it from a normal distribution

$$\Delta t_{err,k} \sim \mathbf{Normal}(\tau_{IRF}, \sigma^2_{IRF}) \tag{4.13}$$

where τ_{IRF} is the mean of IRF and σ_{IRF} is the standard deviation of the IRF (see Eq. 4.2 for comparison). In this simulation we considered $\sigma = \frac{\delta t_p}{2}$ as the width of the pulse.

After sampling the error time, we sample the emission time of each molecule from the exponential distribution with corresponding molecule emission rate belongs to species m

$$\Delta t_{ext,k} | \lambda_m \sim \mathbf{Exponential}\,(\lambda_m) \tag{4.14}$$

and as we have shown in the Fig. 4.20 the detection time of each molecule will be sum of these two times

$$\Delta t_k = \Delta t_{ext,k} + \Delta t_{err,k} \tag{4.15}$$

which is determined by the convolution of emission profile, Eq. 4.13, and excitation pulse, Eq. 4.14.

IRF approximation

To incorporated the effect of the IRF on the measured photon arrival times we approximated it with a Gaussian function [460] (See Fig. 4.21). Centrally symmetric pulses such as the Gaussian, are obtained from electronics as used in most modern instruments [462]. However, for non-symmetrical IRF it could be handled by proper modifications to Eq. 4.2 in the main text.

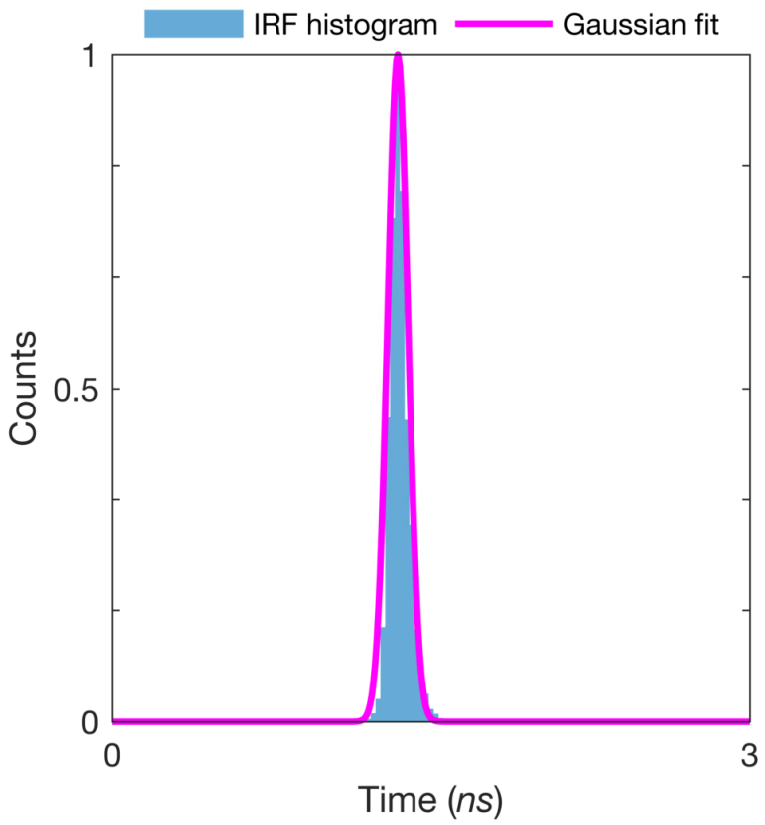

Fig. 4.21. **The actual IRF (blue color) fitted with with a Gaussian function (magenta color).** The fitted IRF is used for the analyses.

Derivation of model likelihood

As we mentioned in the main text, Sec. 4.3.1, measurements $\Delta t_k = \Delta t_{ext,k} + \Delta t_{err,k}$, follow

$$\Delta t_k | \lambda_{s_k} \sim \textbf{Normal}(\tau_{\text{IRF}}, \sigma^2_{\text{IRF}}) * \textbf{Exponential}(\lambda_{s_k}). \tag{4.16}$$

In this case we have

$$\begin{aligned}\Delta t_k|\lambda_{s_k} &\sim \int_{-\infty}^{\infty} \mathbf{Normal}(\tau_{\mathrm{IRF}}, \sigma^2_{\mathrm{IRF}})\mathbf{Exponential}(\lambda_{s_k})d\Delta t_{ext} \\ &= \frac{\lambda_{s_k}}{\sqrt{2\pi\sigma^2_{\mathrm{IRF}}}} \int_{-\infty}^{\infty} e^{-\frac{(\Delta t_k - \Delta t_{ext} - \tau_{\mathrm{IRF}})^2}{2\sigma^2_{\mathrm{IRF}}}} e^{\Delta t_{ext}\lambda_{s_k}} d\Delta t_{ext} \\ &= \frac{\lambda_{s_k}}{2} \exp\left[\frac{\lambda_{s_k}}{2}\left(2\left(\tau_{\mathrm{IRF}} - \Delta t_k\right) + \lambda_{s_k}\sigma^2_{\mathrm{IRF}}\right)\right] \mathrm{erfc}\left(\frac{\tau_{\mathrm{IRF}} - \Delta t_k + \lambda_{s_k}\sigma^2_{\mathrm{IRF}}}{\sigma_{\mathrm{IRF}}\sqrt{2}}\right)\end{aligned} \tag{4.17}$$

where $\mathrm{erfc}(\cdot)$ denotes the complementary error function.

Detailed description of the inference framework

Description of prior probability distributions

Within the Bayesian approach, all unknown model parameters need priors. The model parameters in our framework that require priors are: the molecular emission rates $\{\lambda_m\}_m$; labels on each species $\overline{s}$; and probability on the labels of species $\overline{\pi}$ (fraction of molecules contributing photons from different species). Our choices of priors are described below.

Molecular emission rate, $\{\lambda_m\}_m$ In FLIM we are faced with different species which have different lifetimes. In this study, for computational reasons, we consider emission rates instead of lifetimes, $\tau_m = \frac{1}{\lambda_m}$, where the τ_m is the molecular lifetime and λ_m is the molecular emission rate of species m.

To be able to learn emission rates, and to guarantee that their sampled values in our formulation also attain only positive values, we place a Gamma distribution prior over them as follows

$$\lambda_m \sim \mathbf{Gamma}\left(\alpha_\lambda, \beta_\lambda\right), \tag{4.18}$$

where, α_λ and β_λ are the prior parameters on the molecular emission rates.

Weights, $\overline{\pi}$ The weight on each species comes from the Dirichlet distribution

$$\overline{\pi} \sim \textbf{Dirichlet}_M \left(\frac{\alpha}{M}, \ldots, \frac{\alpha}{M}\right) \tag{4.19}$$

where α is the scalar parameter of the Dirichlet distribution. This prior is conjugate to the labeled species, s_k, which simplifies the computations shown below.

Labels on each species, s_k

Since we have many species, we define a label for each molecule which will tell us that molecule belongs to which species

$$s_k|\overline{\pi} \sim \textbf{Categorical}_{1:M} (\overline{\pi}) \tag{4.20}$$

where $\overline{\pi} = (\pi_1, \ldots, \pi_M)$ is the weight (which they actually are the fraction of molecules contributing photons from different species) on each species.

Summary of model equations

For concreteness, below we summarize all equations used in our framework, including a complete list of priors.

$$\lambda_m \sim \textbf{Gamma}\left(\alpha_\lambda, \beta_\lambda\right) \tag{4.21}$$

$$\overline{\pi} \sim \textbf{Dirichlet}_M\left(\frac{\alpha}{M}, \ldots, \frac{\alpha}{M}\right) \tag{4.22}$$

$$s_k|\overline{\pi} \sim \textbf{Categorical}_{1:M}\left(\overline{\pi}\right) \tag{4.23}$$

$$\Delta t_k|\lambda_m, s_k \sim \frac{\lambda_{s_k}}{2}\exp\left[\frac{\lambda_{s_k}}{2}\left(2\left(\tau_{\text{IRF}} - \Delta t_k\right) + \lambda_{s_k}\sigma_{\text{IRF}}^2\right)\right]\text{erfc}\left(\frac{\tau_{\text{IRF}} - \Delta t_k + \lambda_{s_k}\sigma_{\text{IRF}}^2}{\sigma_{\text{IRF}}\sqrt{2}}\right) \tag{4.24}$$

Inverse problem

Within the Bayesian paradigm, our goal is to sample from the following posterior probability distribution $\mathbb{P}\left(\{\lambda_m\}_m, \overline{s}, \overline{\pi}|\boldsymbol{\Delta t}\right)$. Since, it is not possible to directly compute this distribution, we will sample the random variables $\{\lambda_m\}_m$, $\overline{s}$, and $\overline{\pi}$ from their conditional distributions through a Gibbs sampling scheme [79,358,359,381,421]. Accordingly, posterior samples are generated by updating each one of the variables involved sequentially by sampling conditioned on all other variables and the measurements $\boldsymbol{\Delta t}$.

Conceptually, the steps involved in the generation of each posterior sample $(\{\lambda_m\}_m, \overline{s}, \overline{\pi})$ are:

Update the weights on each species $\overline{\pi}$

Update the labels on species $\overline{s}$

Update the molecular emission rates $\{\lambda_m\}_m$.

Sampling of the weights $\overline{\pi}$ To update the weights of the labels on the species $\overline{s}$, we sample them from the corresponding conditional probability $\mathbb{P}\left(\overline{\pi}|\{\lambda_m\}_m, \boldsymbol{\Delta t}, \overline{s}, \right)$, which simplifies to $\mathbb{P}\left(\overline{\pi}|\overline{s}\right)$.

$$
\begin{aligned}
\overline{\pi} &\sim \mathbb{P}\left(\overline{\pi}|\overline{s}\right) \propto \mathbb{P}\left(\overline{s}|\overline{\pi}\right)\mathbb{P}\left(\overline{\pi}\right) \\
&= \left[\prod_{k=1}^{K}\mathbb{P}\left(s_k|\overline{\pi}\right)\right]\mathbb{P}\left(\overline{\pi}\right) = \left[\prod_{k=1}^{K}\pi_{s_k}\right]\textbf{Dirichlet}_M\left(\frac{\alpha}{M},\ldots,\frac{\alpha}{M}\right) \\
&= \left[\prod_{k=1}^{K}\pi_{s_k}\right]\frac{\Gamma\left(\sum_{m=1}^{M}\frac{\alpha}{M}\right)}{\sum_{m=1}^{M}\Gamma\left(\frac{\alpha}{M}\right)}\prod_{m=1}^{M}\pi_m^{\frac{\alpha}{M}-1} \\
&= \textbf{Dirichlet}_M\left(\frac{\alpha}{M}+\sum_{k=1}^{K}\mathbb{I}(s_k=1),\ldots,\frac{\alpha}{M}+\sum_{k=1}^{K}\mathbb{I}(s_k=M)\right).
\end{aligned}
$$

Sampling of the labels $\overline{s}$ To sample the labels on species, we sample them from the conditional probability distribution $\mathbb{P}\left(s_k|\Delta t_k,\{\lambda_m\}_m,\overline{\pi}\right)$.

$$
\begin{aligned}
s_k &\sim \mathbb{P}\left(s_k|\Delta t_k,\{\lambda_m\}_m,\overline{\pi}\right) \propto \mathbb{P}\left(\Delta t_k|\{\lambda_m\}_m,s_k\right)\mathbb{P}\left(s_k|\overline{\pi}\right) \\
&= \textbf{Categorical}_{1:M}\Bigg(\pi_1\frac{\lambda_{s_k}}{2}\exp\left[\frac{\lambda_{s_k}}{2}\left(2\left(\tau_{\text{IRF}}-\Delta t_k\right)+\lambda_{s_k}\sigma_{\text{IRF}}^2\right)\right]\times \\
&\quad \text{erfc}\left(\frac{\tau_{\text{IRF}}-\Delta t_k+\lambda_{s_k}\sigma_{\text{IRF}}^2}{\sigma_{\text{IRF}}\sqrt{2}}\right), \\
&\quad \vdots \\
&\quad ,\pi_M\frac{\lambda_{s_k}}{2}\exp\left[\frac{\lambda_{s_k}}{2}\left(2\left(\tau_{\text{IRF}}-\Delta t_k\right)+\lambda_{s_k}\sigma_{\text{IRF}}^2\right)\right]\times \\
&\quad \text{erfc}\left(\frac{\tau_{\text{IRF}}-\Delta t_k+\lambda_{s_k}\sigma_{\text{IRF}}^2}{\sigma_{\text{IRF}}\sqrt{2}}\right)\Bigg), \quad k=1,\ldots,K
\end{aligned}
$$

Sampling the molecule emission rates $\{\lambda_m\}_m$ To sample λ_m, we sample from the corresponding conditional probability distribution $\mathbb{P}\left(\{\lambda_m\}_m|\boldsymbol{\Delta t}, \overline{s}\right)$.

$$\begin{aligned}\{\lambda_m\}_m \sim \mathbb{P}\left(\{\lambda_m\}_m|\boldsymbol{\Delta t}, \overline{s}\right) &\propto \mathbb{P}\left(\boldsymbol{\Delta t}|\{\lambda_m\}_m, \overline{s}\right)\left[\prod_{m=1}^{M}\mathbb{P}\left(\lambda_m\right)\right] \\ &= \left[\prod_{k=1}^{K}\frac{\lambda_{s_k}}{2}\exp\left[\frac{\lambda_{s_k}}{2}\left(2\left(\tau_{\text{IRF}}-\Delta t_k\right)+\lambda_{s_k}\sigma_{\text{IRF}}^2\right)\right]\text{erfc}\left(\frac{\tau_{\text{IRF}}-\Delta t_k+\lambda_{s_k}\sigma_{\text{IRF}}^2}{\sigma_{\text{IRF}}\sqrt{2}}\right)\right] \\ &\times\left[\prod_{m=1}^{M}\textbf{Gamma}\left(\lambda_m;\alpha_\lambda,\beta_\lambda\right)\right].\end{aligned} \tag{4.25}$$

Since, there is no close form to sample $\{\lambda_m\}_m$, we sample it using the Metropolis algorithm with the proposal

$$\lambda_m^{\text{prop}} \sim \textbf{Gamma}\left(\alpha_{\lambda_m}^{\text{prop}}, \frac{\lambda_m^{\text{old}}}{\alpha_{\lambda_m}^{\text{prop}}}\right), \qquad m=1,\ldots,M$$

where, the $\alpha_{\lambda_m}^{\text{prop}}$ is the parameter of the proposal distributions for the molecular emission rate. Then, the acceptance ratio is equal to

$$r_\lambda = \frac{\mathbb{P}\left(\{\lambda_m^{\text{prop}}\}_{\text{m}}|\boldsymbol{\Delta t}, \overline{s}\right)}{\mathbb{P}\left(\{\lambda_m^{\text{old}}\}_m|\boldsymbol{\Delta t}, \overline{s}\right)}\frac{\text{Proposal}\left(\{\lambda_m^{\text{old}}\}_m|\{\lambda_m^{\text{prop}}\}_m\right)}{\text{Proposal}\left(\{\lambda_m^{\text{prop}}\}_m|\{\lambda_m^{\text{old}}\}_m\right)}.$$

Also, to avoid numerical underflow, we work with the logarithm of the acceptance ratio

$$\begin{aligned}\log r_\lambda = &\left[\sum_{k=1}^{K}\log\left(\frac{\lambda_{s_{\text{k}}}^{\text{prop}}-\lambda_{s_{\text{k}}}^{\text{old}}}{2}\right)+\left(\Delta t_{\text{k}}-\tau_{\text{IRF}}\right)\left(\lambda_{s_k}^{\text{old}}-\lambda_{s_k}^{\text{prop}}\right)+\frac{\sigma^2}{2}\left(\lambda^{2\text{prop}}_{s_k}-\lambda^{2\text{old}}_{s_k}\right)\right] \\ &+\log\left(\frac{\text{erfc}\left(\frac{\tau_{\text{IRF}}-\Delta t_k+\lambda_{s_k}^{\text{prop}}\sigma_{\text{IRF}}^2}{\sigma_{\text{IRF}}\sqrt{2}}\right)}{\text{erfc}\left(\frac{\tau_{\text{IRF}}-\Delta t_k+\lambda_{s_k}^{\text{old}}\sigma_{\text{IRF}}^2}{\sigma_{\text{IRF}}\sqrt{2}}\right)}\right) \\ &+\left[\sum_{m=1}^{M}\left(2\alpha_{\lambda_m}^{\text{prop}}-\alpha_\lambda\right)\log\left(\frac{\lambda_m^{\text{old}}}{\lambda_m^{\text{prop}}}\right)+\left(\frac{\lambda_m^{\text{old}}-\lambda_m^{\text{prop}}}{\beta_\lambda}\right)+\alpha_{\lambda_m}^{\text{prop}}\left(\frac{\lambda_m^{\text{prop}}}{\lambda_m^{\text{old}}}-\frac{\lambda_m^{\text{old}}}{\lambda_m^{\text{prop}}}\right)\right].\end{aligned} \tag{4.26}$$

So, at the end we will accept or reject the proposal if

$$\log r_\lambda \geq 0 \Rightarrow \lambda_m^{\text{new}} = \lambda_{\text{m}}^{\text{prop}}, \qquad m = 1, \ldots, M$$

$$\log r_\lambda < 0 \Rightarrow \lambda_m^{\text{new}} = \lambda_m^{\text{old}}, \qquad m = 1, \ldots, M$$

Label switching correction of the molecular lifetimes Label switching is a well-known feature of BNP [496]. It arises when we are exploring complex posterior distributions by MCMC algorithms and the likelihood of the model is invariant to the relabelling of mixture components [497]. For example, here, due to exchangeability of the molecular lifetimes, at any iteration (i) of the Gibbs sampling scheme, the corresponding lifetime of the species m might switch with the molecule's lifetime of the species m'. These cases happen because the posterior probability of these events is equal, so, the sampler switches between the lifetimes. This label switching does not effect the joint posterior of all lifetimes. Since, at the end we need to report the posterior of each individual lifetime, we need to be sure that the sampler is not hopping from one mode to the other one.

To undo such label switching, at any iteration of the Gibbs sampling we compare the sampled lifetimes $\{\tau_m^{(i)}\}_m$ and their weights $\{\pi_m^{(i)}\}_m$ with a fixed set of lifetimes $\{\tau_m^*\}_m$ and weights $\{\pi_m^*\}_m$. Based on the distances of the lifetimes at iteration (i) from the fixed set of lifetimes, which we chose, we correct for label switching. The simple choice for this distance can be the distance between the lifetimes, but, since label switching happens in the sampled lifetimes, and subsequently the weights of each molecular lifetime, the particular distance we use incorporates the emission probability and the weights of each molecular lifetime

$$d_{m,m'} = |\pi_m \mathbf{Exp}\left(\tau_m\right) - \pi_m^* \mathbf{Exp}\left(\tau_m^*\right)| \tag{4.27}$$

and we solve the assignment problem is minimizing this distance over the species $\sum_{m=1}^{M} d_{m,m'}$. This problem and its computation can be done efficiently by applying the Hungarian algorithm [498–500].

Table 4.1.
Probability distributions used and their densities. Here, the corresponding random variables are denoted by x.

Distribution	Notation	Probability density function	Mean	Variance
Normal	Normal(μ, σ^2)	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2
Exponential	Exponential(μ)	$\mu e^{-\mu x}$	$\frac{1}{\mu}$	$\frac{1}{\mu^2}$
Gamma	Gamma(α, β)	$\frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-\frac{x}{\beta}}$	$\alpha\beta$	$\alpha\beta^2$

Table 4.2.

Here, we list point estimates of our analyses for synthetic data, which we obtain from the marginal posterior probability distributions $p(\tau|\boldsymbol{\Delta t})$. Estimates are listed according to figure.

	τ	
	mean	std
	ns	ns
Fig. 4.2C	0.51 , 2.19, 10.51	0.14 , 1.42 , 6.45
Fig. 4.2D	0.52 , 2.36 , 13. 01	0.26 , 1.65 , 12.59
Fig. 4.2E	0.52 , 2.51 , 9.74	0.31 , 2.33 , 15.74
Fig. 4.2F	0.51 , 2.10 , 11.06	0.32 , 0.65 , 6.71
Fig. 4.5B1	1.17	0.29
Fig. 4.5B2	1.03	0.23
Fig. 4.5B3	1.04	0.05
Fig. 4.5B4	1.01	0.03
Fig. 4.6B1	0.82 , 8.88	0.41 , 10.31
Fig. 4.6B2	1.10 , 10.37	0.33 , 6.31
Fig. 4.6B3	1.07 , 10.08	0.15 , 4.98
Fig. 4.6B4	1.01 , 10.1	0.05 , 5.23
Fig. 4.7A	0.95 , 9.21	0.21 , 8.91
Fig. 4.7B	1.10 , 10.13	0.35 , 7.11
Fig. 4.7C	1.07 , 10.08	0.15 , 10.18
Fig. 4.8A	1.05 , 10.12	0.14 , 3.84
Fig. 4.8B	1.10 , 5.11	0.25 , 3.11
Fig. 4.8C	0.87 , 2.18	0.98 , 2.06
Fig. 4.8D	1.13 , 1.48	0.26 , 0.68

Table 4.3.
Here, we continue above list point estimates of our analyses for synthetic data, which we obtain from the marginal posterior probability distributions $p(\tau|\boldsymbol{\Delta t})$. Estimates are listed according to figure.

	τ	
	mean	std
	ns	ns
Fig. 4.11A	0.85	0.31
Fig. 4.11B	1.03	0.39
Fig. 4.11C	0.99	0.48
Fig. 4.11D	1.01	0.11
Fig. 4.14	1.01 , 4.10 , 10.06	0.12 , 0.35 , 5.21
Fig. 4.15	0.51 , 1.97 , 6.16 , 12.25	0.14 , 0.55 , 3.41 , 7.43

Table 4.4.
Here, we list point estimates of our analyses for experimental data, which we obtain from the marginal posterior probability distributions $p(\tau|\boldsymbol{\Delta t})$. Estimates are listed according to figure.

	τ	
	mean	std
	ns	ns
Fig. 4.1A1	3.14	2.49
Fig. 4.1B1	3.84	1.84
Fig. 4.1C1	3.85	0.37
Fig. 4.9A1	1.44 , 3.39	1.14 , 1.52
Fig. 4.9B1	1.42 , 3.56	0.46 , 1.05
Fig. 4.9C1	1.41 , 3.81	0.30 , 1.10
Fig. 4.10A1	1.44 , 3.42	0.48 , 1.62
Fig. 4.10B1	1.42 , 3.91	0.39 , 1.24
Fig. 4.10C1	1.37 , 3.71	1.12 , 1.15
Fig. 4.17	0.21 , 1.37 , 2.06 , 3.89	0.25 , 0.72 , 1.41 , 2.44

5. SUMMARY

BNPs are proved to have a remarkable effect in the analysis of single molecule data since they give posterior probabilities over whole models consistent with the given data, not just model parameters of one preferred model. In this thesis, we employed BNPs and proposed a novel formulation to model and analyze fluorescence time traces. In fact, their unique characteristics make them ideal mathematical tools in modeling complex biomolecules as they suggest explicit physical interpretation and give full posterior probabilities over molecular models to be derived with minimum subjective choices.

In chapter 2, we introduced an overview on data analysis in single molecule biophysics. Here, we discussed statistical data-driven analysis approaches, and concentrated on parametric as well as more recent information theoretic and nonparametric statistical methods to biophysical data with a reliance on single-molecule applications. We talked about data analysis tools and model selection problem and mainly Bayesian approach. Moreover, we built a new theoretical framework to study BNPs. Here, we provided a description of the concepts and implementation of an important, and computational tool that extracts BNPs in the area of biophysics.

In chapter 3, we used our proposed BNPs to analyse fluorescence time traces to extract dynamical information (mainly diffusion coefficient) of molecules. Overall, the basis of every spectroscopic method is the detection of photons. Single photon arrivals encode complex dynamical information and approaches to analyze such arrivals have the capability to disclose dynamical processes on fast timescales. Here, we turned our attention to confocal methods where individual photons report on dynamics down

to the single molecule level. While photons reveal dynamics at ms timescales, or faster, data came from confocal methods are collected for many minutes to obtain stable fluorescence traces from which kinetic parameters are deduced. Here, instead we proposed a method to analyze single photon arrival traces using novel tools of BNPs. Using this method, we extract dynamical information efficiently with orders of magnitude less data than traditional correlative methods.

In chapter 4, we built a direct photo-by-photon analysis strategy to infer, simultaneously and self-consistently, the number of species and their associated lifetimes from as few as on the order of 3000 photons for two species. In general, Fluorescence Lifetime Imaging is an experimental imaging technique yielding excited state lifetimes of chemical species recorded over multiple pixels. Within one pixel, the determination of the number of species can be achieved either through fitting time correlated single photon counting histograms or phasor analysis. Both methods yield lifetimes in a computationally efficient manner. However, they also have drawbacks that we address here. First, they do not yield the number of chemical species. Yet the number species is specifically encoded in the photon time of arrival. Next, even to determine lifetimes under the assumption of a known number of species, both methods rely on heavy data post-processing of the signal thereby requiring large amounts of data to retrieve lifetimes. Here, we presented new mathematical tools within the BNPs paradigm we have previously exploited in the analysis of single photon arrivals from single spot confocal. We showed that the proposed approach is valid for both simulated as well as experimental data for one, two, three, and four species with both immobilized and freely diffusing molecule data sets.

REFERENCES

REFERENCES

[1] K. E. Hines, "A primer on Bayesian inference for biophysical systems," *Biophysical Journal*, vol. 108, no. 9, pp. 2103–2113, 2015.

[2] K. E. Hines, J. R. Bankston, and R. W. Aldrich, "Analyzing single-molecule time series via nonparametric Bayesian inference," *Biophysical Journal*, vol. 108, no. 3, pp. 540–556, 2015.

[3] K. Palla, D. A. Knowles, and Z. Ghahramani, "A reversible infinite HMM using normalised random measures," in *International Conference on Machine Learning*, 2014.

[4] C. P. Calderon and K. Bloom, "Inferring latent states and refining force estimates via hierarchical dirichlet process modeling in single particle tracking experiments," *PloS one*, vol. 10, no. 9, p. e0137633, 2015.

[5] T. Komatsuzaki (Ed.), M. Kawakami (Ed.), S. Takahashi (Ed.), H. Yang (Ed.), and R. Silbey (Ed.), *Advances in Chemical Physics. Vol. 146. Single Molecule Biophysics: Experiments and Theories*. John Wiley & Sons, 2011.

[6] C. Bouchiat, M. Wang, J. Allemand, T. Strick, S. Block, and V. Croquette, "Estimating the persistence length of a worm-like chain molecule from force-extension measurements," *Biophys. J.*, vol. **76**, no. 1, pp. 409–413, 1999.

[7] I. Chen, R. Roberts, and J. Szostak, "The emergence of competition between model protocells," *Science*, vol. **305**, no. 5689, pp. 1474–1476, 2004. [Online]. Available: 5689

[8] K. Lindorff-Larsen, S. Piana, R. Dror, and D. Shaw, "How fast-folding proteins fold," *Science*, vol. **334**, no. 6055, pp. 517–520, 2011. [Online]. Available: http://science.sciencemag.org/content/334/6055/517

[9] J. M. Beechem, "[2] global analysis of biochemical and biophysical data," in *Methods in enzymology*. Elsevier, 1992, vol. 210, pp. 37–54.

[10] K. Tsekouras, A. Siegel, R. Day, and S. Pressé, "Inferring diffusion dynamics from FCS in heterogeneous nuclear environments," *Biophys. J.*, vol. **109**, no. 1, pp. 7–17, 2015.

[11] S. Pressé, "A data-driven alternative to the fractional Fokker-Planck equation," *J. Stat. Mech: Th. and Expmt.*, vol. **2015**, no. 7, p. P07009, 2015.

[12] J. Gunawardena, "Models in biology: accurate descriptions of our pathetic thinking," *BMC Biol.*, vol. **12**, no. 1, pp. 29–40, 2014.

[13] I. Bronstein, Y. Israel, E. Kepten, S. Mai, Y. Shav-Tal, E. Barkai, and Y. Garini, "Transient anomalous diffusion of telomeres in the nucleus of mammalian cells," *Phys. Rev. Lett.*, vol. **103**, no. 1, p. 018102, 2009.

[14] S. Weber, A. Spakowitz, and J. Theriot, "Bacterial chromosomal loci move subdiffusively through a viscoelastic cytoplasm," *Phys. Rev. Lett.*, vol. **104**, no. 23, p. 238102, 2010.

[15] I. Golding and E. Cox, "Physical nature of bacterial cytoplasm," *Phys. Rev. Lett.*, vol. **96**, no. 9, p. 098102, 2006.

[16] G. Seisenberger, M. Ried, T. Endress, H. Buening, M. Hallek, and C. Braeuchle, "Real-time single-molecule imaging of the infection pathway of an adeno-associated virus," *Science*, vol. **294**, no. 5548, pp. 1929–1932, 2001.

[17] D. Banks and C. Fradin, "Anomalous diffusion of proteins due to molecular crowding," *Biophys. J.*, vol. **9**, no. 5, pp. 2960–2971, 2005.

[18] T. Feder, I. Brust-Mascher, J. Slattery, B. Baird, and W. Webb, "Constrained diffusion or immobile fraction on cell surfaces: a new interpretation," *Biophys. J.*, vol. **70**, no. 6, pp. 2767–2773, 1996.

[19] M. Konopka, I. Shkel, S. Cayley, M. Record, and J. Weisshaar, "Crowding and confinement effects on protein diffusion in vivo," *J. Bacteriol.*, vol. **188**, no. 17, pp. 6115–6123, 2006.

[20] S. McGuffee and A. Elcock, "Diffusion, crowding & protein stability in a dynamic molecular model of the bacterial cytoplasm," *PLoS Comput. Biol.*, vol. **6**, no. 3, p. e1000694, 2010.

[21] I. Tolić-Nørrelykke, E. Munteanu, G. Thon, L. Oddershede, and K. Berg-Sørensen, "Anomalous diffusion in living yeast cells," *Phys. Rev. Lett.*, vol. **93**, no. 7, p. 078102, 2004.

[22] M. Wachsmuth, W. Waldeck, and J. Langowski, "Anomalous diffusion of fluorescent probes inside living cell nuclei investigated by spatially-resolved fluorescence correlation spectroscopy," *J. Mol. Bio.*, vol. **298**, no. 4, pp. 677–689, 2000.

[23] M. Saxton, "A biological interpretation of transient anomalous subdiffusion. I. Qualitative model," *Biophys. J.*, vol. **92**, no. 4, pp. 1178–1191, 2007.

[24] A. Siegel, N. Hays, and R. Day, "Unraveling transcription factor interactions with heterochromatin protein using fluorescence lifetime imaging microscopy and fluorescence correlation spectroscopy," *J. Biomed. Opt.*, vol. **18**, no. 2, p. 25002, 2013.

[25] B. Parry, I. Surovtsev, M. Cabeen, C. O'Hern, E. Dufresne, and C. Jacobs-Wagner, "The bacterial cytoplasm has glass-like properties and is fluidized by metabolic activity," *Cell*, vol. **156**, no. 1, pp. 183–194, 2014.

[26] P. Schwille, J. Korlach, and W. Webb, "Fluorescence correlation spectroscopy with single-molecule sensitivity on cell and model membranes," *Cytometry*, vol. **36**, no. 3, pp. 176–182, 1999.

[27] A. Caspi, R. Granek, and M. Elbaum, "Enhanced diffusion in active intracellular transport," *Phys. Rev. Lett.*, vol. **85**, no. 26, pp. 5655–5658, 2000.

[28] L. Bruno, V. Levi, M. Brunstein, and M. Despósito, "Transition to superdiffusive behavior in intracellular actin-based transport mediated by molecular motors," *Phys. Rev. E*, vol. **80**, no. 1, p. 011912, 2009.

[29] P. Bressloff and J. Newby, "Stochastic models of intracellular transport," *Rev. Mod. Phys.*, vol. **85**, no. 1, pp. 135–196, 2013.

[30] B. Regner, D. Vucinić, C. Domnisoru, T. Bartol, M. Hetzer, D. Tartakovsky, and T. Sejnowski, "Anomalous diffusion of single particles in cytoplasm," *Biophys. J.*, vol. **104**, no. 8, pp. 1652–1660, 2013.

[31] J. Wu and K. Berland, "Propagators and time-dependent diffusion coefficients for anomalous diffusion," *Biophys. J.*, vol. **95**, no. 4, pp. 2049–2052, 2008.

[32] M. Tavakoli, S. Jazani, I. Sgouralis, and S. Presse, "Bayesian nonparametrics for fluorescence methods," *Biophysical Journal*, vol. 116, no. 3, p. 39a, 2019.

[33] M. Tavakoli, R. P. Shahri, H. Pourreza, A. Mehdizadeh, T. Banaee, and M. H. B. Toosi, "A complementary method for automated detection of microaneurysms in fluorescein angiography fundus images to assess diabetic retinopathy," *Pattern Recognition*, vol. 46, no. 10, pp. 2740–2753, 2013.

[34] M. Tavakoli, M. Nazar, and A. Mehdizadeh, "The efficacy of microaneurysms detection with and without vessel segmentation in color retinal images," in *Medical Imaging 2020: Computer-Aided Diagnosis*, vol. 11314. International Society for Optics and Photonics, 2020, p. 113143Y.

[35] M. Tavakoli, S. Jazani, I. Sgouralis, and S. Presse, "Single molecules dynamics learned from single photons-flim and fcs with bayesian nonparametrics," *Biophysical Journal*, vol. 118, no. 3, pp. 313a–314a, 2020.

[36] M. Tavakoli, S. Jazani, and M. Nazar, "Automated detection of microaneurysms in color fundus images using deep learning with different preprocessing approaches," in *Medical Imaging 2020: Imaging Informatics for Healthcare, Research, and Applications*, vol. 11318. International Society for Optics and Photonics, 2020, p. 113180E.

[37] M. Kaern, T. Elston, W. Blake, and J. Collins, "Stochasticity in gene expression: From theories to phenotypes," *Nat. Rev. Genet.*, vol. **6**, no. 6, pp. 451–464, 2005. [Online]. Available: http://www.nature.com/nrg/journal/v6/n6/abs/nrg1615.html

[38] W. Moerner and D. Fromm, "Methods of single-molecule fluorescence spectroscopy and microscopy," *Rev. Sci. Instrum.*, vol. **74**, no. 8, pp. 3597–3619, 2003.

[39] B. Schueler and W. Eaton, "Protein folding studied by single-molecule FRET," *Curr. Op. in Struct. Bio.*, vol. **18**, no. 1, pp. 16–26, 2008.

[40] Y. Taniguchi, P. Choi, G. Li, H. Chen, M. Babu, J. Hearn, A. Emili, and X. Xie, "Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells," *Science*, vol. **329**, no. 5991, pp. 533–538, 2010.

[41] S. McKinney, C. Joo, and T. Ha, "Analysis of single-molecule FRET trajectories using hidden Markov modeling," *Biophys. J.*, vol. **91**, no. 5, pp. 1941–1951, 2006. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0006349506719038

[42] B. Munsky, B. Trinh, and M. Khammash, "Listening to the noise: Random fluctuations reveal gene network parameters," *Mol. Syst. Biol.*, vol. **5**, no. 1, pp. 318–325, 2009. [Online]. Available: http://www.nature.com/msb/journal/v5/n1/full/msb200975.html

[43] R. Pourreza-Shahri, M. Tavakoli, and N. Kehtarnavaz, "Computationally efficient optic nerve head detection in retinal fundus images," *Biomedical Signal Processing and Control*, vol. 11, pp. 63–73, 2014.

[44] S. F. Gull and G. J. Daniell, "Image reconstruction from incomplete and noisy data," *Nature*, vol. **272**, pp. 686–690, 1978. [Online]. Available: http://www.nature.com/nature/journal/v272/n5655/abs/272686a0.html

[45] G. Rollins, J. Sin, C. Bustamante, and S. Pressé, "A stochastic approach to the molecular counting problem in super-resolution microscopy," *Proc. Natl. Acad. Sc.*, vol. **112**, no. 2, pp. E110–E118, 2015.

[46] F. Persson, M. Lindén, C. Unoson, and J. Elf, "Extracting intracellular diffusion states and transition rates from single-molecule tracking data," *Nat. Meth.*, vol. **10**, no. 3, pp. 265–269, 2013.

[47] J. Bronson, J. Fei, J. Hofman, R. Gonzalez Jr., and C. Wiggins, "Learning rates and states from biophysical time series: A Bayesian approach to model selection and single-molecule FRET data," *Biophys. J.*, vol. **97**, no. 12, pp. 3196–3205, 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0006349509015136

[48] B. Shuang, D. Cooper, J. Taylor, L. Kisley, J. Chen, W. Wang, C. Li, T. Komatsuzaki, and C. Landes, "Fast step transition and state identification (STaSI) for discrete single-molecule data analysis," *J. Phys. Chem. Lett.*, vol. **5**, no. 18, pp. 3157–3161, 2014.

[49] S. Pressé, K. Ghosh, J. Lee, and K. Dill, "Principles of maximum entropy and maximum caliber in statistical physics," *Rev. Mod. Phys.*, vol. **85**, no. 3, pp. 1115–1141, 2013.

[50] Y. Sako and M. Ueda, *Cell signaling reactions: Single-molecule kinetic analysis.* Springer, 2011.

[51] Y. Sako, *Theory and Evaluation of Single-Molecule Signals.* Springer, 2008.

[52] M. Greenfeld, D. Pavlichin, H. Mabuchi, and D. Herschlag, "Single molecule analysis research tool (SMART): An integrated approach for analyzing single molecule data," *PLoS ONE*, vol. **7**, no. 2, p. e30024, 2012.

[53] C. Bustamante, W. Cheng, and Y. Mejia, "Revisiting the central dogma one molecule at a time," *Cell*, vol. **144**, no. 4, pp. 480–497, 2011.

[54] G. Li and X. Xie, "Central dogma at the single-molecule level in living cells," *Nature*, vol. **475**, no. 7356, pp. 308–315, 2011.

[55] A. Baba and T. Komatsuzaki, "Construction of effective free energy landscape from single-molecule time series," *Proceedings of the National Academy of Sciences*, vol. 104, no. 49, pp. 19 297–19 302, 2007.

[56] A. Baba, "Extracting the underlying effective free energy landscape from single-molecule time series local equilibrium states and their network," *Phys. Chem. Chem. Phys.*, vol. **13**, no. 4, pp. 1395–1406, 2011.

[57] J. N. Taylor, C.-B. Li, D. R. Cooper, C. F. Landes, and T. Komatsuzaki, "Error-based extraction of states and energy landscapes from experimental single-molecule time-series," *Scientific reports*, vol. 5, p. 9174, 2015.

[58] M. Pirchi, G. Ziv, I. Riven, S. Sedghani Cohen, N. Zohar, Y. Barak, and G. Haran, "Single-molecule fluorescence spectroscopy maps the folding landscape of a large protein," *Nat. Comm.*, vol. **2**, p. 493, 2011.

[59] M. Woodside and S. Block, "Reconstructing folding energy landscapes by single-molecule force spectroscopy," *Ann. Rev. Biophys.*, vol. **43**, pp. 19–39, 2014.

[60] N. Harris, Y. Song, and C. Kiang, "Experimental free energy surface reconstruction from single-molecule force spectroscopy using Jarzynski's equality," *Phys. Rev. Lett.*, vol. **99**, no. 6, p. 068101, 2007.

[61] K. Kamagata, T. Kawaguchi, Y. Iwahashi, A. Baba, K. Fujimoto, T. Komatsuzaki, Y. Sambongi, Y. Goto, and S. Takahashi, "Long-term observation of fluorescence of free single molecules to explore protein-folding energy landscapes," *J. Am. Chem. Soc.*, vol. **134**, no. 28, pp. 11 525–11 532, 2012.

[62] C. Li, H. Yang, and T. Komatsuzaki, "Multiscale complex network of protein conformational fluctuations in single-molecule time series," *Proc. Natl. Acad. Sci.*, vol. **105**, no. 2, pp. 536–541, 2008.

[63] C. Li and T. Komatsuzaki, "New quantification of local transition heterogeneity of multiscale complex networks constructed from single-molecule time series," *J. Phys. Chem. B*, vol. **113**, no. 44, pp. 14 732–14 741, 2009.

[64] ——, "Aggregated Markov model using time series of single molecule dwell times with minimum excessive information," *Phys. Rev. Lett.*, vol. **111**, no. 5, p. 058301, 2013.

[65] T. Sultana, H. Takagi, M. Morimatsu, H. Teramoto, C. Li, Y. Sako, and T. Komatsuzaki, "Non-Markovian properties and multiscale hidden Markovian network buried in single molecule time series," *J. Chem. Phys.*, vol. **139**, no. 24, p. 245101, 2013.

[66] M. Andrec, R. Levy, and D. Talaga, "Direct determination of kinetic rates from single-molecule photon arrival trajectories using hidden Markov models," *J. Phys. Chem. A*, vol. **107**, no. 38, pp. 7454–7464, 2003.

[67] T. G. Terentyeva, H. Engelkamp, A. Rowan, T. Komatsuzaki, J. Hofkens, C. Li, and K. Blank, "Dynamic disorder in single-enzyme experiments: facts and artifacts," *ACS nano*, vol. **6**, no. 1, pp. 346–354, 2011.

[68] S. Pressé, J. Peterson, J. Lee, P. Elms, J. MacCallum, S. Marqusee, C. Bustamante, and K. Dill, "Single molecule conformational memory extraction: P5ab RNA hairpin," *J. Phys. Chem. B*, vol. **118**, no. 24, pp. 6597–6603, 2014.

[69] D. Glidden, "Robust inference for event probabilities with non-Markov event data," *Biometrics*, vol. **58**, no. 2, pp. 361–368, 2002. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1111/j.0006-341X.2002.00361.x/abstract

[70] B. Kalafut and K. Visscher, "An objective, model-independent method for detection of non-uniform steps in noisy signals," *Comp. Phys. Comm.*, vol. **179**, no. 10, pp. 716–723, 2008.

[71] J. He, S. Guo, and M. Bathe, "Bayesian approach to the analysis of fluorescence correlation spectroscopy data I: Theory," *Anal. Chem.*, vol. **84**, no. 9, pp. 3871–3879, 2012.

[72] P. Sengupta, K. Garai, J. Balaji, N. Periasamy, and S. Maiti, "Measuring size distribution in highly heterogeneous systems with fluorescence correlation spectroscopy," *Biophys. J.*, vol. **84**, no. 3, pp. 1977–1984, 2003.

[73] G. Schwartz, "Estimating the dimension of a model," *Ann. Stat.*, vol. **6**, no. 2, pp. 461–464, 1978.

[74] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Autom. Control.*, vol. **19**, no. 6, pp. 716–723, 1974.

[75] C. Shannon, "A mathematical theory of communication," *Bell Sys. Tech. J.*, vol. **27**, no. 1, p. 379, 1948.

[76] E. Jaynes, *Probability theory: The logic of science*. Cambridge University Press, 2003.

[77] M. Tavakoli and M. Neij, "Quantitative evaluation of the effect of attenuation correction in spect images with ct-derived attenuation," in *Medical Imaging 2019: Physics of Medical Imaging*, vol. 10948. International Society for Optics and Photonics, 2019, p. 109485U.

[78] M. Tavakoli, M. Naji, A. Abdollahi, and F. Kalantari, "Attenuation correction in spect images using attenuation map estimation with its emission data," in *Medical Imaging 2017: Physics of Medical Imaging*, vol. 10132. International Society for Optics and Photonics, 2017, p. 101324Z.

[79] M. Tavakoli, J. N. Taylor, C.-B. Li, T. Komatsuzaki, and S. Pressé, "Single molecule data analysis: an introduction," *arXiv preprint arXiv:1606.00403*, 2016.

[80] G. Casella and R. Berger, *Statistical inference*. Duxbury Press, 2002.

[81] H. Chung, K. McHale, J. Louis, and W. Eaton, "Single molecule fluorescence experiments determine protein folding transition path times," *Science*, vol. **335**, no. 6071, pp. 981–984, 2012.

[82] P. Elms, J. Chodera, C. Bustamante, and S. Marqusee, "The molten globule state is unusually deformable under mechanical force," *Proc. Natl. Acad. Sc.*, vol. **109**, no. 10, pp. 3796–3801, 2012.

[83] A. Berglund, "Statistics of camera-based single-particle tracking," *Phys. Rev. E*, vol. **82**, no. 1, p. 011917, 2010.

[84] L. Zhang, P. Mykland, and Y. Aït-Sahalia, *JASA*, vol. **100**, no. 472, pp. 1394–1411, 2010.

[85] C. Calderon and K. Bloom, "Inferring latent states and refining force estimates via hierarchical Dirichlet process modeling in single particle tracking experiments," *PloS ONE*, vol. **10**, no. 9, p. e0137633, 2015.

[86] J. Hamilton, *Time series analysis*. Economic Theory. II, Princeton University Press, 1995.

[87] J. Lee and S. Pressé, "A derivation of the master equation from path entropy maximization," *J. Chem. Phys.*, vol. **137**, no. 7, p. 074103, 2012.

[88] H. Ge, S. Pressé, K. Ghosh, and K. Dill, "Markov processes follow from the principle of maximum caliber," *J. Chem. Phys.*, vol. 136, no. 6, p. 064108, 2012. [Online]. Available: http://jcp.aip.org/resource/1/jcpsa6/v136/i6/p064108_s1?bypassSSO=1

[89] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. **77**, no. 2, pp. 257–286, 1989.

[90] A. McCallum, D. Freitag, and F. Pereira, "Maximum entropy Markov models for information extraction and segmentation." *ICML*, vol. **17**, pp. 591–598, 2000.

[91] Y. Liu, J. Park, K. Dahmen, Y. Chemla, and T. Ha, "A comparative study of multivariate and univariate hidden Markov modelings in time-binned single-molecule FRET data analysis," *J. Phys. Chem. B*, vol. **114**, no. 16, pp. 5386–5403, 2010.

[92] H. Chung and I. Gopich, "Fast single-molecule FRET spectroscopy: Theory and experiment," *Phys. Chem. Chem. Phys.*, vol. **21**, no. 35, pp. 18 644–18 657, 2014.

[93] B. Keller, A. Kobitski, A. Jaschke, G. Nienhaus, and F. Noé, "Complex RNA folding kinetics revealed by single-molecule FRET and hidden Markov models," *J. Am. Chem. Soc.*, vol. **136**, no. 12, pp. 4534–4543, 2014.

[94] M. Blanco and N. Walter, "Analysis of complex single-molecule FRET time trajectories," *Methods Enzymol.*, vol. **472**, pp. 153–178, 2010.

[95] T. Lee, "Extracting kinetics information from single-molecule fluorescence resonance energy transfer data using hidden Markov models," *J. Phys. Chem. B*, vol. **113**, no. 33, pp. 11 535–11 542, 2009.

[96] D. Kelly, M. Dillingham, A. Hudson, and K. Wiesner, "A new method for inferring hidden Markov models from noisy time sequences," *PloS ONE*, vol. **7**, no. 1, p. e29703, 2012.

[97] G. Forney, "The Viterbi algorithm," *Proc. IEEE*, vol. **61**, no. 3, pp. 268–278, 1973.

[98] C. Bishop, *Pattern recognition and machine learning*. Springer, 2006.

[99] F. Qin, A. Auerbach, and F. Sachs, "Maximum likelihood estimation of aggregated Markov processes," *Proc. R. Soc. B-Biol. Sci.*, vol. **264**, no. 1380, pp. 375–383, 1997.

[100] D. Colquhoun and A. Hawkes, "On the stochastic properties of bursts of single ion channel openings and of clusters of bursts," *Philos. Trans. R. Soc. Lond. Ser. B-Biol. Sci.*, vol. **300**, no. 1098, pp. 1–59, 1982.

[101] R. Horn and K. Lange, "Estimating kinetic constants from single channel data," *Biophys. J.*, vol. **43**, no. 2, pp. 207–223, 1983.

[102] F. Qin, A. Auerbach, and F. Sachs, "Estimating single-channel kinetic parameters from idealized patch-clamp data containing missed events," *Biophys. J.*, vol. **70**, no. 1, pp. 264–280, 1996.

[103] D. Fredkin and J. Rice, "On aggregated Markov processes," *J. App. Prob.*, vol. **23**, no. 1, pp. 208–214, 1986. [Online]. Available: http://www.jstor.org/discover/10.2307/3214130?uid=2&uid=4&sid=21102686795233

[104] P. Kienker, "Equivalence of aggregated Markov models of ion-channel gating," *Proc. R. Soc. B-Biol. Sci.*, vol. **236**, no. 1284, pp. 269–309, 1989.

[105] F. Ball and J. Rice, "Stochastic models for ion channels: Introduction and bibliography," *Math. Biosci.*, vol. **112**, no. 2, pp. 189–206, 1992.

[106] F. Qin, A. Auerbach, and F. Sachs, "A direct optimization approach to hidden Markov modeling for single channel kinetics," *Biophys. J.*, vol. **79**, no. 4, pp. 1915–1927, 2000.

[107] B. Roux and R. Sauvé, "A general solution to the time interval omission problem applied to single channel analysis," *Biophys. J.*, vol. **48**, no. 1, pp. 149–158, 1985.

[108] P. Lee, *Bayesian statistics: An introduction*. John Wiley & Sons, 2012.

[109] S. Kou, X. Xie, and J. Liu, "Bayesian analysis of single-molecule experimental data," *Appl. Statist.*, vol. **54**, no. 3, pp. 469–506, 2005.

[110] N. Monnier, Z. Barry, H. Park, K. Su, Z. Katz, B. English, A. Dey, K. Pan, I. Cheeseman, R. Singer, and M. Bathe, "Inferring transient particle transport dynamics in live cells," *Nat. Meth.*, vol. **12**, no. 9, pp. 838–840, 2015.

[111] N. Monnier, S.-M. Guo, M. Mori, J. He, P. Lénárt, and M. M. Bathe, "Bayesian approach to MSD-based analysis of particle motion in live cells," *Biophys. J.*, vol. **103**, no. 3, pp. 616–626, 2012.

[112] S. Tuerkcan, A. Alexandrou, and J. Masson, "A Bayesian inference scheme to extract diffusivity and potential fields from confined single-molecule trajectories," *Biophys. J.*, vol. **102**, no. 10, pp. 2288–2298, 2012.

[113] J. Witkoskie and J. Cao, "Single molecule kinetics. II. Numerical Bayesian approach," *J. Chem. Phys.*, vol. **121**, no. 13, pp. 6373–6379, 2004.

[114] M. Tavakoli, M. Nazar, and A. Mehdizadeh, "Effect of two different preprocessing steps in detection of optic nerve head in fundus images," in *Medical Imaging 2017: Computer-Aided Diagnosis*, vol. 10134. International Society for Optics and Photonics, 2017, p. 101343A.

[115] M. Tavakoli, M. Nazar, A. Golestaneh, and F. Kalantari, "Automated optic nerve head detection based on different retinal vasculature segmentation methods and mathematical morphology," in *2017 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC)*. IEEE, 2017, pp. 1–7.

[116] E. Jaynes, "Information theory and statistical mechanics," *Phys. Rev.*, vol. **106**, no. 4, pp. 620–630, 1957.

[117] T. Jaynes, "Information theory and statistical mechanics. II," *Phys. Rev.*, vol. **108**, no. 2, pp. 171–190, 1957.

[118] M. Tavakoli, P. Kelley, M. Nazar, and F. Kalantari, "Automated fovea detection based on unsupervised retinal vessel segmentation method," in *2017 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC)*. IEEE, 2017, pp. 1–7.

[119] A. Gelman, "Prior distribution," *Encyclopedia of environmetrics*, vol. **3**, pp. 1634–1637, 2002.

[120] H. Jeffreys, "An invariant form for the prior probability in estimation problems," *Proc. R. Soc. A*, vol. **186**, no. 1007, pp. 453–461, 1946.

[121] Jeffreys, *The theory of probability*. Oxford University Press, 1939.

[122] D. Eno, "Noninformative prior Bayesian analysis for statistical calibration problems," Ph.D. dissertation, Virginia Polytechnic Institute and State University, 1999.

[123] C. Fisher, O. Ullman, and C. Stultz, "Comparative studies of disordered proteins with similar sequences: application to aβ40 and aβ42," *Biophys. J.*, vol. **104**, no. 7, pp. 1546–1555, 2013.

[124] D. Ensign and V. Pande, "Bayesian detection of intensity changes in single molecule and molecular dynamics trajectories," *J. Phys. Chem. B*, vol. **114**, no. 1, pp. 280–292, 2009.

[125] D. Ensign, V. Pande, H. Andersen, and S. Boxer, *Bayesian statistics and single-molecule trajectories*. Stanford University Press, 2010.

[126] D. Ensign and V. Pande, "Bayesian single-exponential kinetics in single-molecule experiments and simulations," *J. Phys. Chem. B*, vol. **113**, no. 36, pp. 12 410–12 423, 2009.

[127] J. Skilling and R. Bryan, "Maximum entropy image reconstruction: General algorithm," *Month. Not. Roy. Astr. Soc.*, vol. **211**, no. 1, pp. 111–124, 1984.

[128] J. Skilling and S. Gull, "Bayesian maximum entropy image reconstruction," *Lecture Notes-Monograph Series*, vol. **20**, pp. 341–367, 1991.

[129] M. Tavakoli, F. Kalantari, and A. Golestaneh, "Comparing different preprocessing methods in automated segmentation of retinal vasculature," in *2017 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC)*. IEEE, 2017, pp. 1–8.

[130] Y. Chiang, P. Borbat, and J. Freed, "Maximum entropy: A complement to Tikhonov regularization for determination of pair distance distributions by pulsed ESR," *J. Magn. Res.*, vol. **177**, no. 2, pp. 184–196, 2005.

[131] A. J. W. G. Visser, S. Laptenok, N. Visser, A. van Hoek, D. Birch, J. Brochon, and J. Borst, "Time-resolved FRET fluorescence spectroscopy of visible fluorescent protein pairs," *Eur. Biophys. J.*, vol. **39**, no. 2, pp. 241–253, 2010.

[132] P. J. Steinbach, R. Ionescu, and C. Matthews, "Analysis of kinetics using a hybrid maximum-entropy/nonlinear-least-squares method: Application to protein folding," *Biophys. J.*, vol. **82**, no. 4, pp. 2244–2255, 2002.

[133] P. Steinbach, K. Chu, H. Frauenfelder, J. Johnson, D. Lamb, G. Nienhaus, T. Sauke, and R. Young, "Determination of rate distributions from kinetic experiments," *Biophys. J.*, vol. **61**, no. 1, pp. 235–245, 1992. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0006349592818301

[134] K. Hines, "A primer on Bayesian inference for biophysical systems," *Biophys. J.*, vol. **108**, no. 9, pp. 2103–2113, 2015.

[135] N. Durisic, L. Laparra-Cuervo, A. Sandoval-Álvarex, J. Borbely, and M. Lakadamyali, "Single-molecule evaluation of fluorescent protein photoactivation efficiency using an in vivo nanotemplate," *Nat. Meth.*, vol. **11**, no. 2, pp. 156–162, 2001.

[136] M. Tavakoli, A. Mehdizadeh, R. Pourreza, H. R. Pourreza, T. Banaee, and M. B. Toosi, "Radon transform technique for linear structures detection: application to vessel detection in fluorescein angiography fundus images," in *2011 IEEE Nuclear Science Symposium Conference Record*. IEEE, 2011, pp. 3051–3056.

[137] W. Hastings, "Monte Carlo sampling methods using Markov chains and their applications," *Biometrika*, vol. **57**, no. 1, pp. 97–109, 1970.

[138] A. F. M. Smith and G. O. Roberts, "Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods," *J. Roy. Stat. Soc. B*, vol. **55**, no. 1, pp. 3–23, 1993.

[139] M. Beckers, F. Drechsler, T. Eilert, J. Nagy, and J. Michaelis, "Quantitative structural information from single-molecule FRET," *Faraday Discuss.*, vol. **184**, pp. 117–129, 2015.

[140] M. Tavakoli, M. B. Toosi, R. Pourreza, T. Banaee, and H. R. Pourreza, "Automated optic nerve head detection in fluorescein angiography fundus images," in *2011 IEEE Nuclear Science Symposium Conference Record*. IEEE, 2011, pp. 3057–3060.

[141] J. Shore and R. Johnson, "Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy," *IEEE Trans. Inf. Theory*, vol. **26** , no. 1, pp. 26–37, 1980.

[142] S. Gull and J. Skilling, "Maximum entropy image reconstruction," *IEE Proc. F*, vol. **131**, no. 6, pp. 646–659, 1984.

[143] J. Skilling, "The axioms of maximum entropy," *Maximum-Entropy and Bayesian Methods in Science and Engineering*, vol. **1**, pp. 173–187, 1988.

[144] R. Bryan, "Maximum entropy analysis of oversampled data problems," *Eur. Biophys. J.*, vol. **18**, no. 3, pp. 165–174, 1990.

[145] A. Livesey and J. Brochon, "Analyzing the distribution of decay constants in pulse-fluorimetry using the maximum entropy method," *Biophys. J.*, vol. **52**, no. 5, pp. 693–706, 1987.

[146] M. Tavakoli, A. Mehdizadeh, R. Pourreza, T. Banaee, M. H. Bahreyni Toossi, and H. R. Pourreza, "Early detection of diabetic retinopathy in fluorescent angiography retinal images using image processing methods," *Iranian Journal of Medical Physics*, vol. 7, no. 4, pp. 7–14, 2010.

[147] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 1991.

[148] S. Kullback and R. Leibler, "On information and sufficiency," *Ann. Math. Stat.*, vol. **22**, no. 1, pp. 79–86, 1951.

[149] S. Amari and I. Amari, "Information geometry of the EM and em algorithms for neural networks," *Neural net.*, vol. **8**, no. 9, pp. 1379–1408, 1995.

[150] S. Amari, "Information geometry on hierarchy of probability distributions," *IEEE Trans. Inf. Theory*, vol. **47**, no. 5, pp. 1701–1711, 2001.

[151] S. Pressé, K. Ghosh, J. Lee, and K. Dill, "Nonadditive entropies yield probability distributions with biases not warranted by the data," *Phys. Rev. Lett.*, vol. **111**, no. 18, p. 180604, 2013.

[152] R. Feynman and R. Leighton, *The Feynman lectures on physics*. Addison-Wesley, 1964.

[153] D. Sivia, *Data analysis: A Bayesian tutorial*. Oxford University Press, 1996.

[154] K. Hines, J. Bankston, and R. Aldrich, "Analyzing single-molecule time series via nonparametric Bayesian inference," *Biophys. J.*, vol. **108**, no. 3, pp. 540–556, 2015.

[155] H. R. Pourreza, M. H. Bahreyni Toossi, A. Mehdizadeh, R. Pourreza, and M. Tavakoli, "Automatic detection of microaneurysms in color fundus images using a local radon transform method," *Iranian Journal of Medical Physics*, vol. 6, no. 1, pp. 13–20, 2009.

[156] G. Sun, S.-M. Guo, C. The, V. Korzh, M. Bathe, and T. Wohland, "Bayesian model selection applied to the analysis of fluorescence correlation spectroscopy data of fluorescent proteins in vitro and in vivo," *Anal. Chem.*, vol. **87**, no. 8, pp. 4326–4333, 2015.

[157] K. Bacia, S. Kim, and P. Schwille, "Fluorescence cross-correlation spectroscopy in living cells," *Nat. Meth.*, vol. **3**, no. 2, pp. 83–89, 2006.

[158] T. Torres and M. Levitus, "Measuring conformational dynamics: A new FCS-FRET approach," *J. Phys. Chem. B*, vol. **111**, no. 25, pp. 7392–7400, 2007. [Online]. Available: http://dx.doi.org/10.1021/jp070659s

[159] P. Kapusta, M. Wahl, A. Benda, M. Hof, and J. Enderlein, "Fluorescence lifetime correlation spectroscopy," *J. Fluoresc.*, vol. **17**, no. 1, pp. 43–48, 2007.

[160] A. Michelman-Ribeiro, D. Mazza, T. Rosales, T. Stasevich, H. Boukari, V. Rishi, C. Vinson, J. R. Knutson, and J. McNally, "Direct measurement of association and dissociation rates of DNA binding in live cells by fluorescence correlation spectroscopy," *Biophys. J.*, vol. **97**, no. 1, pp. 337–346, 2009.

[161] N. Kahya and P. Schwille, "Fluorescence correlation studies of lipid domains in model membranes," *Mol. Membr. Biol.*, vol. **23**, no. 1, pp. 29–39, 2006.

[162] S. Kim, K. Heinze, and P. Schwille, "Fluorescence correlation spectroscopy in living cells," *Nat. Meth.*, vol. **4**, no. 11, pp. 963–973, 2007.

[163] J. Szymanski and M. Weiss, "Elucidating the origin of anomalous diffusion in crowded fluids," *Phys. Rev. Lett.*, vol. **103**, no. 3, p. 038102, 2009.

[164] F. Hoefling and T. Franosch, "Anomalous transport in the crowded world of biological cells," *Rep. Prog. Phys.*, vol. **76**, no. 4, p. 046602, 2013.

[165] M. Weiss, H. Hashimoto, and T. Nilsson, "Anomalous protein diffusion in living cells as seen by fluorescence correlation spectroscopy," *Biophys. J.*, vol. **84**, no. 6, pp. 4043–4052, 2003.

[166] P. Schwille, U. Haupts, S. Maiti, and W. W. Webb, "Molecular dynamics in living cells observed by fluorescence correlation spectroscopy with one- and two-photon excitation," *Biophys. J.*, vol. **77**, no. 4, pp. 2251–2265, 1999.

[167] N. Malchus and M. Weiss, "Elucidating anomalous protein diffusion in living cells with fluorescence correlation spectroscopy – facts and pitfalls," *J. Fluoresc.*, vol. **20**, no. 1, pp. 19–26, 2010.

[168] O. Krichevsky and G. Bonnet, "Fluorescence correlation spectroscopy: The technique and its applications," *Rep. Prog. Phys.*, vol. **65**, no. 2, pp. 251–297, 2002.

[169] S. Guo, N. Bag, A. Mishra, T. Wohland, and M. Bathe, "Bayesian total internal reflection fluorescence correlation spectroscopy reveals hIAPP-induced plasma membrane domain organization in live cells," *Biophys. J.*, vol. **106**, no. 1, pp. 190–200, 2014.

[170] E. Elson and D. Magde, "Fluorescence correlation spectroscopy. I. Conceptual basis and theory," *Biopolymers*, vol. **13**, no. 1, pp. 1–27, 1974.

[171] J. Widengren, U. Mets, and R. Rigler, "Fluorescence correlation spectroscopy of triplet states in solution: A theoretical and experimental study," *J. Phys. Chem.*, vol. **99**, no. 36, pp. 13 368–13 379, 1995.

[172] K. Burnecki, E. Kepten, J. Janczura, I. Bronshtein, Y. Garini, and A. Weron, "Universal algorithm for identification of fractional Brownian motion. A case of telomere subdiffusion," *Biophys. J.*, vol. **103**, no. 9, pp. 1839–1847, 2012.

[173] A. Bancaud, S. Huet, N. Daigle, J. Mozziconacci, J. Beaudouin, and J. Ellenberg, "Molecular crowding affects diffusion and binding of nuclear proteins in heterochromatin and reveals the fractal organization of chromatin," *EMBO J.*, vol. **28**, no. 24, pp. 3785–3798, 2009.

[174] I. Wong, M. Gardel, D. Reichman, E. Weeks, M. Valentine, A. Bausch, and D. Weitz, "Anomalous diffusion probes microstructure dynamics of entangled F-actin networks," *Phys. Rev. Lett.*, vol. **92**, no. 17, p. 178101, 2004.

[175] J. Jeon, V. Tejedor, S. Burov, E. Barkai, C. Selhuber-Unkel, K. Berg-Sørensen, L. Oddershede, and R. Metzler, "In vivo anomalous diffusion and weak ergodicity breaking of lipid granules," *Phys. Rev. Lett.*, vol. **106**, no. 4, p. 048103, 2011.

[176] M. Saxton and K. Jacobson, "Single-particle tracking: Applications to membrane dynamics," *Ann. Rev. Biophys. Biomol. Struct.*, vol. **26**, no. 1, pp. 373–399, 1997.

[177] A. Ott, J. Bouchaud, D. Langevin, and W. Urbach, "Anomalous diffusion in "living polymers": A genuine Lévy flight?" *Phys. Rev. Lett.*, vol. **65**, no. 17, pp. 2201–2204, 1990.

[178] H. Jankevics, M. Prummer, P. Izewska, H. Pick, K. Leufgen, and H. Vogel, "Diffusion-time distribution analysis reveals characteristic ligand-dependent interaction patterns of nuclear receptors in living cells," *Biochemistry*, vol. **44**, no. 35, pp. 11 676–11 683, 2005.

[179] M. Woringer, X. Darzacq, and I. Izeddin, "Geometry of the nucleus: A perspective on gene expression regulation," *Curr. Opin. Chem. Biol.*, vol. **20**, pp. 112–119, 2014.

[180] M. Serag, M. Abadi, and S. Habuchi, "Single-molecule diffusion and conformational dynamics by spatial integration of temporal fluctuations," *Nat. Comm.*, vol. **5**, p. 5123, 2014.

[181] I. Izeddin, V. Récamier, L. Bosanac, I. Cissé, L. Boudarene, C. Dugast-Darzacq, F. Proux, O. Bénichou, R. Voituriez, O. Bensaude, M. Dahan, and X. Darzacq, "Single-molecule tracking in live cells reveals distinct target-search strategies of transcription factors in the nucleus," *Elife*, vol. **3**, p. e02230, 2014.

[182] J. Gorman, F. Wang, S. Redding, A. Plys, T. Fazio, S. Wind, E. Alani, and E. Greene, "Single-molecule imaging reveals target-search mechanisms during DNA mismatch repair," *Proc. Natl. Acad. Sci.*, vol. **109**, no. 45, pp. E3074–E3083, 2012.

[183] J. Gebhardt, D. Suter, R. Roy, Z. Zhao, A. Chapman, S. Basu, T. Maniatis, and X. Xie, "Single-molecule imaging of transcription factor binding to DNA in live mammalian cells," *Nat. Meth.*, vol. **10**, no. 5, pp. 421–426, 2013.

[184] G. Stormo and Y. Zhao, "Determining the specificity of protein-DNA interactions," *Nat. Rev. Genet.*, vol. **11**, no. 11, pp. 751–760, 2010.

[185] J. Elf, G.-W. Li, and X. Xie, "Probing transcription factor dynamics at the single-molecule level in a living cell," *Science*, vol. **316**, no. 5828, pp. 1191–1194, 2007.

[186] E. Marklund, A. Mahmutovic, O. Berg, P. Hammar, D. van der Spoel, D. Fange, and J. Elf, "Transcription-factor binding and sliding on DNA studied using micro- and macroscopic models," *Proc. Natl. Acad. Sci.*, vol. **110**, no. 49, pp. 19 796–19 801, 2013.

[187] A. Siegel, M. Baird, M. Davidson, and R. Day, "Strengths and weaknesses of recently engineered red fluorescent proteins evaluated in live cells using fluorescence correlation spectroscopy," *Int. J. Mo. Sci.*, vol. **14**, no. 10, pp. 20 340–20 358, 2013.

[188] M. Drobizhev, T. Hughes, Y. Stepanenko, P. Wnuk, K. O'Donnell, J. Scott, P. Callis, A. Mikhaylov, L. Dokken, and A. Rebane, "Primary role of the chromophore bond length alternation in reversible photoconversion of red fluorescence proteins," *Sci. Rep.*, vol. **2**, no. 688, pp. 1–6, 2012.

[189] B. Slaughter, J. Schwartz, and R. Li, "Mapping dynamic protein interactions in MAP kinase signaling using live-cell fluorescence fluctuation spectroscopy and imaging," *Proc. Natl. Acad. Sc.*, vol. **104**, no. 51, pp. 20 320–20 325, 2007.

[190] Z. Wang, J. Shah, Z. Chen, C.-H. Sun, and M. Berns, "Fluorescence correlation spectroscopy investigation of a GFP mutant-enhanced cyan fluorescent protein and its tubulin fusion in living cells with two-photon excitation," *J. Biomed. Opt.*, vol. **9**, no. 2, pp. 395–403, 2004.

[191] Z. Petrášek and P. Schwille, "Precise measurement of diffusion coefficients using scanning fluorescence correlation spectroscopy," *Biophys. J.*, vol. **94**, no. 8, pp. 1437–1448, 2008.

[192] Z. Wu, M. Zhao, L. Xia, Y. Yu, S. Shen, S. Han, H. Li, T. Wang, G. Chen, and L. Wang, "Pkcδ enhances C/EBPα degradation via inducing its phosphorylation and cytoplasmic translocation," *Biochem. Biophys. Res. Co.*, vol. **433**, no. 2, pp. 220–225, 2013.

[193] D. Ramji and P. Foka, "CCAAT/enhancer binding proteins: Structure, function and regulation," *Biochem. J.*, vol. **365**, no. 3, pp. 561–5751, 2002.

[194] P. Hemmerich, L. Schmiedeberg, and S. Diekmann, "Dynamic as well as stable protein interactions contribute to genome function and maintenance," *Chromosome Res.*, vol. **19**, no. 1, pp. 131–151, 2011.

[195] L. Schmiedeberg, K. Weisshart, S. Diekmann, G. Meyer zu Hoerste, and P. Hemmerich, "High- and low-mobility populations of HP1 in heterochromatin of mammalian cells," *Mol. Biol. Cell*, vol. **15**, no. 6, pp. 2819–2833, 2004.

[196] Y. Wang and R. Austin, "Single-molecule imaging of laci diffusing along nonspecific DNA." Williams (Ed.), M.C. and Maher (Ed.), J.L. Biophysics of DNA-protein interactions from single molecules to biological systems. Springer, 2010.

[197] J. de Rooi, C. Ruckebusch, and P. Eilers, "Sparse deconvolution in one and two dimensions: Applications in endocrinology and single-molecule fluorescence imaging," *Anal. Chem.*, vol. **86**, no. 13, pp. 6291–6298, 2014.

[198] Y. Brody, N. Neufeld, N. Bieberstein, S. Causse, E. Böhnlein, K. Neugebauer, X. Darzacq, and Y. Shav-Tal, "The in vivo kinetics of RNA polymerase II elongation during co-transcriptional splicing," *PLoS Biol*, vol. **9**, no. 1, p. e1000573, 2011.

[199] E. Anderson and A. Hoskins, "Single molecule approaches for studying spliceosome assembly and catalysis," *Spliceosomal Pre-mRNA Splicing: Methods and Protocols*, vol. **1126**, pp. 217–241, 2014.

[200] S. Arunajadai and W. Cheng, "Step detection in single-molecule real time trajectories embedded in correlated noise," *PLoS ONE*, vol. **8**, no. 3, p. e59279, 2013.

[201] T. Aggarwal, D. Materassi, R. Davison, T. Hays, and M. Salapaka, "Detection of steps in single molecule data," *Cell. Mol. Bioeng.*, vol. **5**, no. 1, pp. 14–31, 2012.

[202] G. Claeskens and N. Hjort, *Model selection and model averaging.* Cambridge University Press, 2008.

[203] H. Bozdogan, "Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions," *Psychometrika*, vol. **52**, no. 3, pp. 345–370, 1987.

[204] K. Yamaoka, T. Nakagawa, and T. Uno, "Application of Akaike's information criterion (AIC) in the evaluation of linear pharmacokinetic equations," *J. Pharmacokinet. Biopharm.*, vol. **6**, no. 2, pp. 165–175, 1978.

[205] D. Posada and T. Buckley, "Model selection and model averaging in phylogenetics: Advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests," *Syst. Biol.*, vol. **53**, no. 5, pp. 793–808, 2004.

[206] H. Bozdogan, "Akaike's information criterion and recent developments in information complexity," *J. Math. Psychol.*, vol. **44**, no. 1, pp. 62–91, 2000.

[207] M. Tavakoli, K. Tsekouras, K. W. Dunn, R. Day, S. Pressé, and M. T. Team, "Modeling and inference of hepatic transport kinetics," *APS*, vol. 2018, pp. L60–243, 2018.

[208] K. Okamoto and Y. Sako, "Variational bayes analysis of a photon-based hidden Markov model for single-molecule FRET trajectories," *Biophys. J.*, vol. **103**, no. 6, pp. 1315–1324, 2012.

[209] S. Kou, B. Cherayil, W. Min, B. P. English, and X. Xie, "Single-molecule Michaelis-Menten equations," *J. Phys. Chem. B*, vol. **109**, no. 41, pp. 19 068–19 081, 2005.

[210] P. Barber, S. Ameer-Beg, S. Pathmananthan, M. Rowley, and A. Coolen, "A Bayesian method for single molecule, fluorescence burst analysis," *Biomed. Opt. Express*, vol. **1**, no. 4, pp. 1148–1158, 2010.

[211] N. Zarrabi, S. Ernst, B. Verhalen, S. Wilkens, and M. Börsch, "Analyzing conformational dynamics of single P-glycoprotein transporters by Förster resonance energy transfer using hidden Markov models," *Methods*, vol. **66**, no. 2, pp. 168–179, 2014.

[212] K. Burnham and D. Anderson, *Model selection and multimodel inference: A practical information-theoretic approach*. Springer, 2003.

[213] M. Hansen and B. Yu, "Model selection and the principle of minimum description length," *J. Am. Stat. Assoc.*, vol. **96**, no. 454, pp. 746–774, 2001.

[214] J. Rissanen, "Fisher information and stochastic complexity," *IEEE Trans. Inf. Theory*, vol. **42**, no. 1, pp. 40–47, 1996.

[215] V. Balasubramanian, "MDL, Bayesian inference, and the geometry of the space of probability distributions," *Grunwald (Ed.), P. and Myung (Ed.), I.J. and Pitt (Ed.), M. Advances in minimum description length: Theory and applications. MIT Press*, 2005.

[216] I. Myung, V. Balasubramanian, and M. Pitt, "Counting probability distributions: Differential geometry and model selection," *Proc. Natl. Acad. Sci.*, vol. **97**, no. 21, pp. 11 170–11 175, 2000.

[217] R. Shibata, "Selection of the order of an autoregressive model by Akaike's information criterion," *Biometrika*, vol. **63**, no. 1, pp. 117–126, 1976.

[218] R. Nishii, "Asymptotic properties of criteria for selection of variables in multiple regression," *Ann. Stat.*, vol. **12**, no. 2, pp. 758–765, 1984.

[219] S. Vrieze, "Model selection and psychological theory: A discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC)." *Psychol. Meth.*, vol. **17**, no. 2, pp. 228–243, 2012.

[220] J. Kuha, "AIC and BIC comparisons of assumptions and performance," *Socio. Meth. Res.*, vol. **33**, no. 2, pp. 188–229, 2004.

[221] S. Chen and P. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the Bayesian information criterion," *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, vol. **8**, pp. 127–132, 1998.

[222] J. Chen and Z. Chen, "Extended Bayesian information criteria for model selection with large model spaces," *Biometrika*, vol. **95**, no. 3, pp. 759–771, 2008.

[223] A. Atkinson, "Likelihood ratios, posterior odds and information criteria," *J. Econometrics*, vol. **16**, no. 1, pp. 15–20, 1981.

[224] C. LaMont and P. Wiggins, "The frequentist information criterion (FIC): The unification of information-based and frequentist inference," *arXiv preprint arXiv:1506.05855*, 2015.

[225] G. Chow, "A comparison of the information and posterior probability criteria for model selection," *J. Econometrics*, vol. **16**, no. 1, pp. 21–33, 1981.

[226] R. Shibata, *Statistical aspects of model selection*. Springer, 1989.

[227] P. Wiggins, "An information-based approach to change-point analysis with applications to biophysics and cell biology," *Biophys. J.*, vol. **109**, no. 2, pp. 346–354, 2015.

[228] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *Petrov (Ed.), B.N. and Csaki (Ed.), F. International symposium on information theory*. Springer, 1998.

[229] R. Hogg and A. Craig, *Introduction to mathematical statistics*. Macmillan, 1978.

[230] J. Chen and A. K. Gupta, "On change point detection and estimation," *Commun. Stat. Simulat. C.*, vol. **30**, no. 3, pp. 665–697, 2001.

[231] P. Krishnaiah and B. Miao, "19 review about estimation of change points," *Handbook of statistics*, vol. **7**, pp. 375–402, 1988.

[232] M. Basseville and I. Nikiforov, *Detection of abrupt changes: Theory and application*. Prentice Hall, 1993.

[233] J. Munro, A. Vaiana, K. Sanbonmatsu, and S. Blanchard, "A new view of protein synthesis: Mapping the free energy landscape of the ribosome using single-molecule FRET," *Biopolymers*, vol. **89**, no. 7, pp. 565–577, 2008.

[234] S. Türkcan and J. Masson, "Bayesian decision tree for the classification of the mode of motion in single-molecule trajectories," *PloS ONE*, vol. **8**, no. 12, p. e82799, 2013.

[235] M. Hajdziona and A. Molski, "Maximum likelihood-based analysis of single-molecule photon arrival trajectories," *J. Chem. Phys*, vol. **134**, no. 5, p. 054112, 2011.

[236] L. Elliott, M. Barhoum, J. Harris, and P. Bohn, "Trajectory analysis of single molecules exhibiting non-Brownian motion," *Phys. Chem. Chem. Phys.*, vol. **13**, no. 10, pp. 4326–4334, 2011.

[237] L. Watkins and H. Yang, "Detection of intensity change points in time-resolved single-molecule measurements," *J. Phys. Chem. B*, vol. **109**, no. 1, pp. 617–628, 2005.

[238] Y. Chen, N. Deffenbaugh, C. Anderson, and W. Hancock, "Molecular counting by photobleaching in protein complexes with many subunits: Best practices and application to the cellulose synthesis complex," *Mol. Bio. Cell*, vol. **25**, no. 22, pp. 3630–3642, 2014.

[239] M. Little, B. Steel, F. Bai, Y. Sowa, T. Bilyard, D. Mueller, R. Berry, and N. Jones, "Steps and bumps: Precision extraction of discrete states of molecular machines," *Biophys. J.*, vol. **101**, no. 2, pp. 477–485, 2011.

[240] J. Chen and Y. Wang, "A statistical change point model approach for the detection of DNA copy number variations in array CGH data," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. **6**, no. 4, pp. 529–541, 2009.

[241] D. Cooper, D. Dolino, H. Jaurich, B. Shuang, S. Ramaswamy, C. Nurik, J. Chen, V. Jayaraman, and C. Landes, "Conformational transitions in the Glycine-Bound GluN1 NMDA Receptor LBD via single-molecule FRET," *Biophys. J.*, vol. **109**, no. 1, pp. 66–75, 2015.

[242] H. Yang, "Change-point localization and wavelet spectral analysis of single-molecule time series," *Single-Molecule Biophysics: Experiment and Theory, Vol. 146. John Wiley & Sons*, 2011.

[243] J. Taylor, D. Makarov, and C. Landes, "Denoising single-molecule FRET trajectories with wavelets and Bayesian inference," *Biophys. J.*, vol. **98**, no. 1, pp. 164–173, 2010.

[244] Y. Wang, "Jump and sharp cusp detection by wavelets," *Biometrika*, vol. **82**, no. 2, pp. 385–397, 1995.

[245] M. Little and N. Jones, "Generalized methods and solvers for noise removal from piecewise constant signals. I. Background theory." *Proc. R. Soc. A-Math. Phys. Eng. Sci.*, vol. **467**, no. 2135, pp. 3088–3114, 2011.

[246] M. Little and S. Jones, "Generalized methods and solvers for noise removal from piecewise constant signals. II. New methods." *Proc. R. Soc. A-Math. Phys. Eng. Sci.*, vol. **467**, no. 2135, pp. 3115–3140, 2011.

[247] M. Little and N. Jones, "Signal processing for molecular and cellular biological physics: An emerging field," *Phil. Trans. R. Soc. A*, vol. **371**, no. 1984, p. 20110546, 2013.

[248] R. Killick, P. Fearnhead, and I. A. Eckley, "Optimal detection of changepoints with a linear computational cost," *JASA*, vol. **107**, no. 500, pp. 1590–1598, 2012. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/01621459.2012.737745

[249] K. Frick, A. Munk, and H. Sieling, "Multiscale change point inference," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. **76**, no. 3, pp. 495–580, 2014.

[250] S. Hu, "Akaike information criterion," *Cent. Res. Sci. Comput.*, 2007.

[251] J. Chen, A. Gupta, and J. Pan, "Information criterion and change point problem for regular models," *Sankhya: The Indian Journal of Statistics*, vol. **68**, no. 2, pp. 252–282, 2006.

[252] A. Gelman, J. Hwang, and A. Vehtari, "Understanding predictive information criteria for Bayesian models," *Stat. Comput.*, vol. **24**, no. 6, pp. 997–1016, 2014.

[253] C. LaMont and P. Wiggins, "The development of an information criterion for change-point analysis," *Neural Comput.*, vol. **28**, no. 3, pp. 594–612, 2016.

[254] C. LaMont, H. LaMont, and P. Wiggins, "Information-based inference for sloppy and singular models," *arXiv preprint arXiv:1506.05855*, 2015.

[255] S. Watanabe, "A widely applicable Bayesian information criterion," *J. Mach. Learn. Res.*, vol. **14**, no. 1, pp. 867–897, 2013.

[256] S. Balakrishnan, H. Kamisetty, J. Carbonell, S. Lee, and C. Langmead, "Learning generative models for protein fold families," *Proteins: Struct., Funct., Bioinf.*, vol. **79**, no. 4, pp. 1061–1078, 2011.

[257] R. Goldsmith and W. Moerner, "Watching conformational-and photodynamics of single fluorescent proteins in solution," *Nat. Chem.*, vol. **2**, no. 3, pp. 179–186, 2010.

[258] R. Krishnan, M. Blanco, M. Kahlscheuer, J. Abelson, C. Guthrie, and N. Walter, "Biased Brownian ratcheting leads to pre-mRNA remodeling and capture prior to first-step splicing," *Nat. Struct. Mol. Biol.*, vol. **20**, no. 12, pp. 1450–1457, 2013.

[259] M. Beal, Z. Ghahramani, and C. Rasmussen, "The infinite hidden Markov model," in *Advances in neural information processing systems*. MIT Press, 2001.

[260] S. Preus, S. L. Noer, L. L. Hildebrandt, D. Gudnason, and V. Birkedal, "iSMS: Single-molecule FRET microscopy software," *Nat. Meth.*, vol. **12**, no. 7, pp. 593–594, 2015.

[261] J. van de Meent, J. Bronson, C. Wiggins, and R. Gonzalez, "Empirical bayes methods enable advanced population-level analyses of single-molecule FRET experiments," *Biophys. J.*, vol. **106**, no. 6, pp. 1327–1337, 2014.

[262] S. Johnson, J. van de Meent, R. Phillips, C. Wiggins, and M. Lindén, "Multiple LacI-mediated loops revealed by Bayesian statistics and tethered particle motion," *Nucl. Acids Res.*, vol. **1**, p. gku563, 2014.

[263] T. Ferguson, "A Bayesian analysis of some nonparametric problems," *Ann. Stat.*, vol. **1**, no. 2, pp. 209–230, 1973.

[264] P. Orbanz and Y. W. Teh, "Bayesian nonparametric models," in *Encyclopedia of Machine Learning*. Springer, 2011.

[265] Z. Ghahramani, "Bayesian non-parametrics and the probabilistic approach to modelling," *Phil. Trans. R. Soc. A*, vol. **371**, no. 1984, p. 20110553, 2013.

[266] Y. W. Teh, "Dirichlet process," in *Encyclopedia of machine learning*. Springer, 2011.

[267] S. Gershman and D. Blei, "A tutorial on Bayesian nonparametric models," *J. Math. Psychol.*, vol. **56**, no. 1, pp. 1–12, 2012.

[268] P. Yau, R. Kohn, and S. Wood, "Bayesian variable selection and model averaging in high-dimensional multinomial nonparametric regression," *J. Comp. Graph. Stat.*, vol. **12**, no. 1, pp. 23–54, 2012.

[269] E. Phadia, *Prior processes and their applications*. Springer, 2013.

[270] Y. Teh, M. Jordan, M. Beal, and D. Blei, "Hierarchical Dirichlet processes," *JASA*, vol. **101**, no. 476, pp. 1566–1581, 2012.

[271] R. Neal, "Markov chain sampling methods for Dirichlet process mixture models," *J. Comp. Graph. Stat.*, vol. **9**, no. 2, pp. 249–265, 2000.

[272] S. Kim, M. Tadesse, and M. Vannucci, "Variable selection in clustering via Dirichlet process mixture models," *Biometrika*, vol. **93**, no. 4, pp. 877–893, 2006.

[273] J. Sethuraman, "A constructive definition of Dirichlet priors," *Statistica sinica*, vol. **4**, no. 2, pp. 639–650, 1994.

[274] J. Pitman, "Poisson–Dirichlet and GEM invariant distributions for split-and-merge transformations of an interval partition," *Comb. Probab. Comput.*, vol. **11**, no. 5, pp. 501–514, 2002.

[275] A. Lo, "On a class of Bayesian nonparametric estimates: I. Density estimates," *Ann. Stat.*, vol. **12**, no. 1, pp. 351–357, 1984.

[276] C. Antoniak, "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems," *Ann. Stat.*, vol. **2**, no. 6, pp. 1152–1174, 1974.

[277] R. Neal, "Markov chain sampling methods for Dirichlet process mixture models," *J. Comp. Graph. Stat.*, vol. **9**, no. 2, pp. 249–265, 2000.

[278] E. Fox, E. Sudderth, M. Jordan, and A. Willsky, "An HDP-HMM for systems with state persistence," in *Proc. ICML*. ACM, 2008.

[279] E. Fox, E. Sudderth, and A. Willsky, "Bayesian nonparametric inference of switching dynamic linear models," *IEEE Trans. Signal Process*, vol. **59**, no. 4, pp. 1569–1585, 2011.

[280] J. Van Gael, Y. Saatci, Y. Teh, and Z. Ghahramani, "Beam sampling for the infinite hidden Markov model," *Proc. Int. Conf. Machine Learning*, 2008.

[281] N. Slonim, G. S. Atwal, G. Tkačik, and W. Bialek, "Information-based clustering," *Proc. Natl. Acad. Sci.*, vol. **102**, no. 2, pp. 18 297–18 302, 2005.

[282] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," *Proc. 37th Annu. Allert. Conf. Commu. Contl. Comput.*, 1999.

[283] S. Still and W. Bialek, "How many clusters? An information-theoretic perspective," *Neural Comput.*, vol. **16**, no. 12, pp. 2483–2506, 2004.

[284] J. Taylor, C. Li, D. Cooper, C. Landes, and T. Komatsuzaki, "Error-based extraction of states and energy Landscapes from experimental single-molecule time-series," *Sci. Rep.*, vol. **5**, p. 9174, 2015.

[285] L. Kantorovitch, "On the translocation of masses," *Manag. Sci.*, vol. **5**, no. 1, pp. 1–4, 1958.

[286] R. E. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Trans. Inf. Theory*, vol. **18**, no. 4, pp. 460–473, 1972.

[287] S. Arimoto, "An algorithm for computing the capacity of arbitrary discrete memoryless channels," *IEEE Trans. Inf. Theory*, vol. **18**, no. 1, pp. 14–20, 1972.

[288] I. Borg and P. Groenen, *Modern multidimensional scaling: Theory and applications*. Springer, 2010. [Online]. Available: https://books.google.co.jp/books?id=zVcRkgAACAAJ

[289] S. V. Krivov and M. J. Karplus, "Free energy disconnectivity graphs: Application to peptide models," *J. Chem. Phys.*, vol. **117**, no. 23, pp. 10 894–10 903, 2002.

[290] S. V. Krivov, V. Krivov, and M. J. Karplus, "Hidden complexity of free energy surfaces for peptide (protein) folding," *Proc. Nat. Acad. Sci.*, vol. **101**, no. 41, pp. 14 766–14 770, 2004.

[291] C. F. Landes, A. Rambhadran, J. N. Taylor, F. Salatan, and V. Jayaraman, "Structural landscape of isolated agonist-binding domains from single AMPA receptors," *Nat. Chem. Biol.*, vol. **7**, no. 3, pp. 168–173, 2011.

[292] S. Ramaswamy, D. R. Cooper, N. K. Poddar, D. M. MacLean, A. Rambhadran, J. N. Taylor, H. Uhm, C. F. Landes, and V. Jayaraman, "Role of conformational dynamics in α-amino-3-hydroxy-5-methylisoxazole-4-propionic acid (AMPA) receptor partial agonism," *J. Biol. Chem.*, vol. **287**, no. 52, pp. 43 557–43 564, 2012.

[293] N. Armstrong and E. Gouaux, "Mechanisms for Activation and Antagonism of an AMPA-Sensitive Glutamate Receptor: Crystal Structures of the GluR2 Ligand Binding Core," *Neuron*, vol. **28**, no. 1, pp. 165–181, 2000.

[294] K. Poon, A. H. Ahmed, L. M. Nowak, and R. E. Oswald, "Mechanisms of modal activation of GluA3 receptors," *Mol. Pharmacol.*, vol. **80**, no. 1, pp. 49–59, 2011.

[295] A. Lau and B. Roux, "The free energy landscapes governing conformational changes in a glutamate receptor ligand-binding domain," *Structure*, vol. **15**, no. 10, pp. 1203–1214, 2007.

[296] A. Y. Lau and B. Roux, "The hidden energetics of ligand binding and activation in a glutamate receptor," *Nat. Struct. Mol. Biol.*, vol. **18**, no. 3, pp. 283–287, 2011.

[297] J. P. Crutchfield and K. Young, "Inferring statistical complexity," *Phys. Rev. Lett.*, vol. **63**, no. 2, pp. 105–108, 1989.

[298] J. P. Crutchfield, "Between order and chaos," *Nat. Phys.*, vol. **8**, no. 1, pp. 17–24, 2012.

[299] C. R. Shalizi and J. P. Crutchfield, "Computational mechanics: Pattern and prediction, structure and simplicity," *J. Stat. Phys.*, vol. **194**, no. 3, pp. 817–879, 2001.

[300] K. Chaloner and I. Verdinelli, "Bayesian experimental design: A review," *Stat. Sci.*, vol. **10**, no. 3, pp. 273–304, 1995.

[301] P. Sebastiani and H. Wynn, "Bayesian experimental design and Shannon information," *Proc. Sec. Bay. Stat. Sci.*, vol. **44**, pp. 176–181, 1997.

[302] P. Sebastiani, H. Wynn, and H. Wynn, "Maximum entropy sampling and optimal Bayesian experimental design," *J. Roy. Stat. Soc. B*, vol. **62**, no. 1, pp. 145–157, 2000.

[303] D. Talaga, "Information theoretical approach to single-molecule experimental design and interpretation," *J. Phys. Chem. A*, vol. **110**, no. 31, pp. 9743–9757, 2006.

[304] D. Talaga and D. Talaga, "Information-theoretical analysis of time-correlated single-photon counting measurements of single molecules," *J. Phys. Chem. A*, vol. **113**, no. 17, pp. 5251–5263, 2009.

[305] W. Bialek, I. Nemenman, and N. Tishby, "Predictability, complexity, and learning," *Neural comp.*, vol. **13**, no. 11, pp. 2409–2463, 2001.

[306] S. Pressé, "Nonadditive entropy maximization is inconsistent with Bayesian updating," *Phys. Rev. E*, vol. **90**, no. 5, p. 052149, 2014.

[307] C. Tsallis, "Possible generalization of Boltzmann-Gibbs statistics," *J. Stat. Phys.*, vol. **52**, no. 1–2, pp. 479–487, 1988.

[308] S. Abe, "Generalized molecular chaos hypothesis and the H theorem: Problem of constraints and amendment of nonextensive statistical mechanics," *Phys. Rev. E*, vol. **79**, no. 4, p. 041116, 2009.

[309] S. Abe and S. Abe, "Instability of q-averages in nonextensive statistical mechanics," *Euro. Phys. Lett.*, vol. **84**, no. 6, p. 60006, 2008.

[310] R. Hanel, S. Thurner, and C. Tsallis, "On the robustness of q-expectation values and Rényi entropy," *Euro. Phys. Lett.*, vol. **85**, no. 2, p. 20005, 2009.

[311] S. Abe, "Anomalous behavior of q-averages in nonextensive statistical mechanics," *J. Stat. Mech.: Theor. Expmt.*, vol. **2009**, no. 7, p. P07027, 2009.

[312] A. Rényi, "On measures of entropy and information," *Proc. Sym. Math. Stat. Prob.*, vol. **4**, pp. 547–561, 1961.

[313] C. Hanel and S. Thurner, "A comprehensive classification of complex statistical systems and an axiomatic derivation of their entropy and distribution functions," *Eur. Phys. Lett.*, vol. **93**, p. 20006, 2011.

[314] P. Jizba and T. Arimitsu, "Towards information theory for q-nonextensive statistics with q-deformed distributions," *Physica A: Stat. Mech. App.*, vol. **365**, no. 2, pp. 76–84, 2006.

[315] M. Hotta and I. Joichi, "Composability and generalized entropy," *Phys. Lett. A*, vol. **262**, no. 4, pp. 302–309, 1999.

[316] C. Tsallis, "Non-extensive thermostatistics: Brief review and comments," *Physica A*, vol. **221**, no. 1, pp. 277–290, 1995.

[317] E. Curado and C. Tsallis, "Generalized statistical mechanics: Connection with thermodynamics," *J. Phys. A: Math. Gen.*, vol. **52**, no. 2, pp. L69–L72, 1991.

[318] R. dos Santos, "Generalization of Shannon's theorem for Tsallis entropy," *J. Math. Phys.*, vol. **38**, no. 8, pp. 4104–4107, 1997.

[319] S. Abe, "Axioms and uniqueness theorem for Tsallis entropy," *Phys. Lett. A*, vol. **271**, no. 1, pp. 74–79, 2000.

[320] C. Tsallis, M. Gell-Mann., and Y. Sato, "Asymptotically scale-invariant occupancy of phase space makes the entropy Sq extensive," *Proc. Natl. Acad. Sci.*, vol. **102**, no. 43, pp. 15 377–15 382, 2005.

[321] S. Abe, "General pseudoadditivity of composable entropy prescribed by the existence of equilibrium," *Phys. Rev. E*, vol. **63**, no. 3, p. 061105, 2001.

[322] G. Wilk and Z. Wlodarczyk, "Interpretation of the nonextensivity parameter q in some applications of Tsallis statistics and Lévy distributions," *Phys. Rev. Lett.*, vol. **84**, no. 13, pp. 2770–2773, 2000.

[323] C. Beck, "Dynamical foundations of nonextensive statistical mechanics," *Phys. Rev. Lett.*, vol. **87**, no. 18, p. 180601, 2001.

[324] A. Gorban and I. Karlin, "Family of additive entropy functions out of thermodynamic limit," *Phys. Rev. E*, vol. **67**, no. 1, p. 016104, 2003.

[325] B. Huang, M. Bates, and X. Zhuang, "Super-resolution fluorescence microscopy," *Annual Review of Biochemistry*, vol. 78, pp. 993–1016, 2009.

[326] A. Gahlmann and W. Moerner, "Exploring bacterial cell biology with single-molecule tracking and super-resolution imaging," *Nature Reviews Microbiology*, vol. 12, no. 1, p. 9, 2014.

[327] Z. Liu, L. D. Lavis, and E. Betzig, "Imaging live-cell dynamics and structure at the single-molecule level," *Molecular cell*, vol. 58, no. 4, pp. 644–659, 2015.

[328] A. Lee, K. Tsekouras, C. Calderon, C. Bustamante, and S. Pressé, "Unraveling the thousand word picture: An introduction to super-resolution data analysis," *Chemical Reviews*, 2017.

[329] E. L. Elson and D. Magde, "Fluorescence correlation spectroscopy. I. conceptual basis and theory," *Biopolymers*, vol. 13, no. 1, pp. 1–27, 1974.

[330] D. Magde, E. L. Elson, and W. W. Webb, "Fluorescence correlation spectroscopy. II. an experimental realization," *Biopolymers*, vol. 13, no. 1, pp. 29–61, 1974.

[331] K. Remaut, B. Lucas, K. Braeckmans, N. Sanders, S. De Smedt, and J. Demeester, "FRET-FCS as a tool to evaluate the stability of oligonucleotide drugs after intracellular delivery," *Journal of Controlled Release*, vol. 103, no. 1, pp. 259–271, 2005.

[332] T. Torres and M. Levitus, "Measuring conformational dynamics: a new FCS-FRET approach," *The Journal of Physical Chemistry B*, vol. 111, no. 25, pp. 7392–7400, 2007.

[333] P. Schwille, F.-J. Meyer-Almes, and R. Rigler, "Dual-color fluorescence cross-correlation spectroscopy for multicomponent diffusional analysis in solution," *Biophysical Journal*, vol. 72, no. 4, pp. 1878–1886, 1997.

[334] O. Krichevsky and G. Bonnet, "Fluorescence correlation spectroscopy: the technique and its applications," *Reports on Progress in Physics*, vol. 65, no. 2, p. 251, 2002.

[335] J. R. Lakowicz, *Principles of fluorescence spectroscopy*. Springer, 2006.

[336] R. Rigler and E. S. Elson, *Fluorescence correlation spectroscopy: theory and applications*. Springer Science & Business Media, 2012, vol. 65.

[337] G. R. Bright, G. W. Fisher, J. Rogowska, and D. L. Taylor, "Fluorescence ratio imaging microscopy," *Methods in Cell Biology*, vol. 30, pp. 157–192, 1989.

[338] J. A. Fitzpatrick and B. F. Lillemeier, "Fluorescence correlation spectroscopy: linking molecular dynamics to biological function in vitro and in situ," *Current Opinion in Structural Biology*, vol. 21, no. 5, pp. 650–660, 2011.

[339] M. Purschke, N. Rubio, K. D. Held, and R. W. Redmond, "Phototoxicity of hoechst 33342 in time-lapse fluorescence microscopy," *Photochemical & Photobiological Sciences*, vol. 9, no. 12, pp. 1634–1639, 2010.

[340] V. Magidson and A. Khodjakov, "Circumventing photodamage in live-cell microscopy," in *Methods in cell biology*. Elsevier, 2013, vol. 114, pp. 545–560.

[341] J.-Y. Tinevez, J. Dragavon, L. Baba-Aissa, P. Roux, E. Perret, A. Canivet, V. Galy, and S. Shorte, "A quantitative method for measuring phototoxicity of a live cell imaging microscope," in *Methods in enzymology*. Elsevier, 2012, vol. 506, pp. 291–309.

[342] P. Schwille, U. Haupts, S. Maiti, and W. W. Webb, "Molecular dynamics in living cells observed by fluorescence correlation spectroscopy with one-and two-photon excitation," *Biophysical Journal*, vol. 77, no. 4, pp. 2251–2265, 1999.

[343] P. Dittrich, F. Malvezzi-Campeggi, M. Jahnz, and P. Schwille, "Accessing molecular dynamics in cells by fluorescence correlation spectroscopy," *Biological Chemistry*, vol. 382, no. 3, pp. 491–494, 2001.

[344] R. D. Phair and T. Misteli, "Kinetic modelling approaches to in vivo imaging," *Nature Reviews Molecular Cell Biology*, vol. 2, no. 12, p. 898, 2001.

[345] K. Tsekouras, A. P. Siegel, R. N. Day, and S. Pressé, "Inferring diffusion dynamics from FCS in heterogeneous nuclear environments," *Biophysical Journal*, vol. 109, no. 1, pp. 7–17, 2015.

[346] I. V. Gopich and A. Szabo, "Decoding the pattern of photon colors in single-molecule FRET," *The Journal of Physical Chemistry B*, vol. 113, no. 31, pp. 10 965–10 973, 2009.

[347] M. Pirchi, R. Tsukanov, R. Khamis, T. E. Tomov, Y. Berger, D. C. Khara, H. Volkov, G. Haran, and E. Nir, "Photon-by-photon hidden markov model analysis for microsecond single-molecule FRET kinetics," *The Journal of Physical Chemistry B*, vol. 120, no. 51, pp. 13 065–13 075, 2016.

[348] M. Waligórska and A. Molski, "Maximum likelihood-based analysis of photon arrival trajectories in single-molecule FRET," *Chemical Physics*, vol. 403, pp. 52–58, 2012.

[349] I. V. Gopich and A. Szabo, "Single-molecule FRET with diffusion and conformational dynamics," *The Journal of Physical Chemistry B*, vol. 111, no. 44, pp. 12 925–12 932, 2007.

[350] M. Antonik, S. Felekyan, A. Gaiduk, and C. A. Seidel, "Separating structural heterogeneities from stochastic variations in fluorescence resonance energy transfer distributions via photon distribution analysis," *The Journal of Physical Chemistry B*, vol. 110, no. 13, pp. 6970–6978, 2006.

[351] E. Nir, X. Michalet, K. M. Hamadani, T. A. Laurence, D. Neuhauser, Y. Kovchegov, and S. Weiss, "Shot-noise limited single-molecule fret histograms: comparison between theory and experiments," *The Journal of Physical Chemistry B*, vol. 110, no. 44, pp. 22 103–22 124, 2006.

[352] I. V. Gopich and A. Szabo, "Theory of the energy transfer efficiency and fluorescence lifetime distribution in single-molecule FRET," *Proceedings of the National Academy of Sciences*, vol. 109, no. 20, pp. 7747–7752, 2012.

[353] H. S. Chung and I. V. Gopich, "Fast single-molecule FRET spectroscopy: theory and experiment," *Physical Chemistry Chemical Physics*, vol. 16, no. 35, pp. 18 644–18 657, 2014.

[354] I. V. Gopich, "Accuracy of maximum likelihood estimates of a two-state model in single-molecule FRET," *The Journal of Chemical Physics*, vol. 142, no. 3, p. 034110, 2015.

[355] H. S. Chung, F. Meng, J.-Y. Kim, K. McHale, I. V. Gopich, and J. M. Louis, "Oligomerization of the tetramerization domain of p53 probed by two-and three-color single-molecule FRET," *Proceedings of the National Academy of Sciences*, p. 201700357, 2017.

[356] B. Schuler, "Perspective: Chain dynamics of unfolded and intrinsically disordered proteins from nanosecond fluorescence correlation spectroscopy combined with single-molecule FRET," *The Journal of Chemical Physics*, vol. 149, no. 1, p. 010901, 2018.

[357] A. J. Berglund, "Statistics of camera-based single-particle tracking," *Physical Review E*, vol. 82, no. 1, p. 011917, 2010.

[358] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian data analysis*. CRC press Boca Raton, FL, 2014, vol. 2.

[359] U. Von Toussaint, "Bayesian inference in physics," *Reviews of Modern Physics*, vol. 83, no. 3, p. 943, 2011.

[360] I. Sgouralis and S. Pressé, "An introduction to infinite HMMs for single-molecule data analysis," *Biophysical Journal*, vol. 112, no. 10, pp. 2021–2029, 2017.

[361] I. Sgourallis and S. Pressé, "ICON: an adaptation of infinite hmms for time traces with drift," *Biophysical Journal*, vol. 112, no. 10, pp. 2117–2126, 2017.

[362] I. Sgouralis, M. Whitmore, L. Lapidus, M. J. Comstock, and S. Pressé, "Single molecule force spectroscopy at high data acquisition: A bayesian nonparametric analysis," *The Journal of Chemical Physics*, vol. 148, no. 12, p. 123320, 2018.

[363] I. Sgouralis, S. Madaan, F. Djutanta, R. Kha, R. F. Hariadi, and S. Pressé, "A bayesian nonparametric approach to single molecule forster resonance energy transfer," *The Journal of Physical Chemistry B*, vol. 123, no. 3, pp. 675–688, 2018.

[364] S. Jazani, I. Sgouralis, and S. Pressé, "A method for single molecule tracking using a conventional single-focus confocal setup," *The Journal of Chemical Physics*, vol. 150, no. 11, p. 114108, 2019.

[365] J. Paisley and L. Carin, "Nonparametric factor analysis with beta process priors," in *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 2009, pp. 777–784.

[366] L. Al Labadi and M. Zarepour, "On approximations of the beta process in latent feature models: Point processes approach," *Sankhya A*, vol. 80, no. 1, pp. 59–79, 2018.

[367] T. Broderick, M. I. Jordan, J. Pitman *et al.*, "Beta processes, stick-breaking and power laws," *Bayesian Analysis*, vol. 7, no. 2, pp. 439–476, 2012.

[368] N. L. Hjort, C. Holmes, P. Müller, and S. G. Walker, *Bayesian nonparametrics*. Cambridge University Press, 2010, vol. 28.

[369] G. Brakenhoff, K. Visscher, and H. Van der Voort, "Size and shape of the confocal spot: control and relation to 3D imaging and image processing," in *Handbook of biological confocal microscopy*. Springer, 1990, pp. 87–91.

[370] R. Phillips, J. Theriot, J. Kondev, and H. Garcia, *Physical biology of the cell*. Garland Science, 2012.

[371] S. P. Meyn and R. L. Tweedie, "Stability of markovian processes II: Continuous-time processes and sampled chains," *Advances in Applied Probability*, vol. 25, no. 3, pp. 487–517, 1993.

[372] S. P. Meyn and L. Tweedie, "Stability of markovian processes III: Foster–lyapunov criteria for continuous-time processes," *Advances in Applied Probability*, vol. 25, no. 3, pp. 518–548, 1993.

[373] T. Wohland, R. Rigler, and H. Vogel, "The standard deviation in fluorescence correlation spectroscopy," *Biophysical Journal*, vol. 80, no. 6, pp. 2987–2999, 2001.

[374] Y. Chen, J. D. Müller, P. T. So, and E. Gratton, "The photon counting histogram in fluorescence fluctuation spectroscopy," *Biophysical Journal*, vol. 77, no. 1, pp. 553–567, 1999.

[375] D. Axelrod, D. Koppel, J. Schlessinger, E. Elson, and W. W. Webb, "Mobility measurement by analysis of fluorescence photobleaching recovery kinetics," *Biophysical Journal*, vol. 16, no. 9, pp. 1055–1069, 1976.

[376] B. Zhang, J. Zerubia, and J.-C. Olivo-Marin, "Gaussian approximations of fluorescence microscope point-spread function models," *Applied Optics*, vol. 46, no. 10, pp. 1819–1829, 2007.

[377] M. Born and E. Wolf, *Principles of optics: electromagnetic theory of propagation, interference and diffraction of light*. Elsevier, 2013.

[378] M. E. Johnson and G. Hummer, "Free-propagator reweighting integrator for single-particle dynamics in reaction-diffusion models of heterogeneous protein-protein interaction systems," *Physical Review X*, vol. 4, no. 3, p. 031037, 2014.

[379] A. C. Barato and U. Seifert, "Cost and precision of brownian clocks," *Physical Review X*, vol. 6, no. 4, p. 041053, 2016.

[380] A. V. Chechkin, F. Seno, R. Metzler, and I. M. Sokolov, "Brownian yet non-gaussian diffusion: from superstatistics to subordination of diffusing diffusivities," *Physical Review X*, vol. 7, no. 2, p. 021002, 2017.

[381] C. Robert and G. Casella, *Introducing Monte Carlo Methods with R*. Springer Science & Business Media, 2009.

[382] L. Lacasa, I. P. Mariño, J. Miguez, V. Nicosia, É. Roldán, A. Lisica, S. W. Grill, and J. Gómez-Gardeñes, "Multiplex decomposition of non-markovian dynamics and the hidden layer reconstruction problem," *Physical Review X*, vol. 8, no. 3, p. 031038, 2018.

[383] H. C. Berg, *Random walks in biology*. Princeton University Press, 1993.

[384] O. C. Ibe, *Elements of Random Walk and Diffusion Processes*. John Wiley & Sons, 2013.

[385] J. Haile, I. Johnston, A. J. Mallinckrodt, S. McKay *et al.*, "Molecular dynamics simulation: elementary methods," *Computers in Physics*, vol. 7, no. 6, pp. 625–625, 1993.

[386] D. J. Higham, "An algorithmic introduction to numerical simulation of stochastic differential equations," *SIAM Review*, vol. 43, no. 3, pp. 525–546, 2001.

[387] R. Erban and S. J. Chapman, "Stochastic modelling of reaction–diffusion processes: algorithms for bimolecular reactions," *Physical Biology*, vol. 6, no. 4, p. 046001, 2009.

[388] H. Li, C.-F. Yen, and S. Sivasankar, "Fluorescence axial localization with nanometer accuracy and precision," *Nano Letters*, vol. 12, no. 7, pp. 3731–3735, 2012.

[389] P. D. Schmidt, B. H. Reichert, J. G. Lajoie, and S. Sivasankar, "Method for high frequency tracking and sub-nm sample stabilization in single molecule fluorescence microscopy," *Scientific Reports*, vol. 8, no. 1, p. 13912, 2018.

[390] D. Scherfeld, N. Kahya, and P. Schwille, "Lipid dynamics and domain formation in model membranes composed of ternary mixtures of unsaturated and saturated phosphatidylcholines and cholesterol," *Biophysical Journal*, vol. 85, no. 6, pp. 3758–3768, 2003.

[391] A. Benda, M. Beneš, V. Marecek, A. Lhotskỳ, W. T. Hermens, and M. Hof, "How to determine diffusion coefficients in planar phospholipid systems by confocal fluorescence correlation spectroscopy," *Langmuir*, vol. 19, no. 10, pp. 4120–4126, 2003.

[392] R. Macháň and M. Hof, "Lipid diffusion in planar membranes investigated by fluorescence correlation spectroscopy," *Biochimica et Biophysica Acta (BBA)-Biomembranes*, vol. 1798, no. 7, pp. 1377–1391, 2010.

[393] E. Lerner, A. Ingargiola, and S. Weiss, "Characterizing highly dynamic conformational states: the transcription bubble in RNAP-promoter open complex as an example," *The Journal of Chemical Physics*, vol. 148, no. 12, p. 123315, 2018.

[394] S. F. Gibson and F. Lanni, "Experimental test of an analytical model of aberration in an oil-immersion objective lens used in three-dimensional light microscopy," *JOSA A*, vol. 9, no. 1, pp. 154–166, 1992.

[395] X. Michalet, O. Siegmund, J. Vallerga, P. Jelinsky, J. Millaud, and S. Weiss, "Detectors for single-molecule fluorescence imaging and spectroscopy," *Journal of Modern Optics*, vol. 54, no. 2-3, pp. 239–281, 2007.

[396] J. E. Bronson, J. Fei, J. M. Hofman, R. L. Gonzalez Jr, and C. H. Wiggins, "Learning rates and states from biophysical time series: a Bayesian approach to model selection and single-molecule FRET data," *Biophysical Journal*, vol. 97, no. 12, pp. 3196–3205, 2009.

[397] S. Uphoff, K. Gryte, G. Evans, and A. N. Kapanidis, "Improved temporal resolution and linked hidden markov modeling for switchable single-molecule FRET," *ChemPhysChem*, vol. 12, no. 3, pp. 571–579, 2011.

[398] H. Liu, B. Jiu, H. Liu, and Z. Bao, "Superresolution isar imaging based on sparse bayesian learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 8, pp. 5005–5013, 2014.

[399] J. He, S.-M. Guo, and M. Bathe, "Bayesian approach to the analysis of fluorescence correlation spectroscopy data i: theory," *Analytical Chemistry*, vol. 84, no. 9, pp. 3871–3879, 2012.

[400] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Sharing clusters among related groups: Hierarchical dirichlet processes," in *Advances in neural information processing systems*, 2005, pp. 1385–1392.

[401] S. Jazani, I. Sgouralis, O. M. Shafraz, M. Levitus, S. Sivasankar, and S. Pressé, "An alternative framework for fluorescence correlation spectroscopy," *Nature Communications*, vol. 10, 2019.

[402] J. Andrews, W. Conway, W.-K. Cho, A. Narayanan, J.-H. Spille, N. Jayanth, T. Inoue, S. Mullen, J. Thaler, and I. Cissé, "qSR: A quantitative superresolution analysis tool reveals the cell-cycle dependent organization of RNA polymerase i in live human cells," *Scientific Reports*, vol. 8, no. 1, p. 7424, 2018.

[403] A. Narayanan, A. B. Meriin, M. Y. Sherman, and I. I. Cisse, "A first order phase transition underlies the formation of sub-diffractive protein aggregates in mammalian cells," *bioRxiv*, p. 148395, 2017.

[404] W.-K. Cho, J.-H. Spille, M. Hecht, C. Lee, C. Li, V. Grube, and I. I. Cisse, "Mediator and rna polymerase II clusters associate in transcription-dependent condensates," *Science*, vol. 361, no. 6400, pp. 412–415, 2018.

[405] P. Kask, R. Günther, and P. Axhausen, "Statistical accuracy in fluorescence fluctuation experiments," *European Biophysics Journal*, vol. 25, no. 3, pp. 163–169, 1997.

[406] S. Saffarian and E. L. Elson, "Statistical analysis of fluorescence correlation spectroscopy: the standard deviation and bias," *Biophysical Journal*, vol. 84, no. 3, pp. 2030–2042, 2003.

[407] D. T. Chiu, N. L. Jeon, S. Huang, R. S. Kane, C. J. Wargo, I. S. Choi, D. E. Ingber, and G. M. Whitesides, "Patterned deposition of cells and proteins onto surfaces by using three-dimensional microfluidic systems," *Proceedings of the National Academy of Sciences*, vol. 97, no. 6, pp. 2408–2413, 2000.

[408] T. Ha, A. Y. Ting, J. Liang, W. B. Caldwell, A. A. Deniz, D. S. Chemla, P. G. Schultz, and S. Weiss, "Single-molecule fluorescence spectroscopy of enzyme conformational dynamics and cleavage mechanism," *Proceedings of the National Academy of Sciences*, vol. 96, no. 3, pp. 893–898, 1999.

[409] S. Wu, G. Han, D. J. Milliron, S. Aloni, V. Altoe, D. V. Talapin, B. E. Cohen, and P. J. Schuck, "Non-blinking and photostable upconverted luminescence from single lanthanide-doped nanocrystals," *Proceedings of the National Academy of Sciences*, vol. 106, no. 27, pp. 10 917–10 921, 2009.

[410] S.-H. Lee, J. Y. Shin, A. Lee, and C. Bustamante, "Counting single photoactivatable fluorescent molecules by photoactivated localization microscopy (PALM)," *Proceedings of the National Academy of Sciences*, vol. 109, no. 43, pp. 17 436–17 441, 2012.

[411] J. Stricker, P. Maddox, E. Salmon, and H. P. Erickson, "Rapid assembly dynamics of the escherichia coli FtsZ-ring demonstrated by fluorescence recovery after photobleaching," *Proceedings of the National Academy of Sciences*, vol. 99, no. 5, pp. 3171–3175, 2002.

[412] G. Wu, H. Ji, K. Hansen, T. Thundat, R. Datar, R. Cote, M. F. Hagan, A. K. Chakraborty, and A. Majumdar, "Origin of nanomechanical cantilever motion generated from biomolecular interactions," *Proceedings of the National Academy of Sciences*, vol. 98, no. 4, pp. 1560–1564, 2001.

[413] J. Enderlein, I. Gregor, D. Patra, and J. Fitter, "Art and artefacts of fluorescence correlation spectroscopy," *Current Pharmaceutical Biotechnology*, vol. 5, no. 2, pp. 155–161, 2004.

[414] S. Weiss, "Fluorescence spectroscopy of single biomolecules," *Science*, vol. 283, no. 5408, pp. 1676–1683, 1999.

[415] R. B. Best and G. Hummer, "Reaction coordinates and rates from transition paths," *Proceedings of the National Academy of Sciences*, vol. 102, no. 19, pp. 6732–6737, 2005.

[416] P. Schwille and E. Haustein, "Fluorescence correlation spectroscopy: an introduction to its concepts and applications," *Biophysics textbook online*, vol. 1, no. 3, 2001.

[417] E. L. Elson, "Fluorescence correlation spectroscopy: past, present, future," *Biophysical Journal*, vol. 101, no. 12, pp. 2855–2870, 2011.

[418] J. Enderlein and W. P. Ambrose, "Optical collection efficiency function in single-molecule detection experiments," *Applied Optics*, vol. 36, no. 22, pp. 5298–5302, 1997.

[419] E. B. Wilson and M. M. Hilferty, "The distribution of chi-square," *Proceedings of the National Academy of Sciences*, vol. 17, no. 12, pp. 684–688, 1931.

[420] J. Paisley and M. I. Jordan, "A constructive definition of the beta process," *arXiv preprint arXiv:1604.00685*, 2016.

[421] H. Liu and H. Motoda, *Computational methods of feature selection*. CRC Press, 2007.

[422] S. L. Scott, "Bayesian methods for hidden markov models: Recursive computing in the 21st century," *Journal of the American Statistical Association*, vol. 97, no. 457, pp. 337–351, 2002.

[423] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.

[424] O. Cappé, E. Moulines, and T. Rydén, "Inference in hidden markov models," in *Proceedings of EUSFLAT Conference*, 2009, pp. 14–16.

[425] M. Briers, A. Doucet, and S. Maskell, "Smoothing algorithms for state–space models," *Annals of the Institute of Statistical Mathematics*, vol. 62, no. 1, p. 61, 2010.

[426] L. C.-L. Lin and F. L. Brown, "Brownian dynamics in fourier space: membrane simulations over long length and time scales," *Physical Review Letters*, vol. 93, no. 25, p. 256001, 2004.

[427] B. Calderhead, "A general construction for parallelizing metropolis- hastings algorithms," *Proceedings of the National Academy of Sciences*, vol. 111, no. 49, pp. 17 408–17 413, 2014.

[428] S. Chib and E. Greenberg, "Understanding the metropolis-hastings algorithm," *The American Statistician*, vol. 49, no. 4, pp. 327–335, 1995.

[429] J. W. Lichtman and J.-A. Conchello, "Fluorescence microscopy," *Nature Methods*, vol. 2, no. 12, p. 910, 2005.

[430] T. Niehörster, A. Löschberger, I. Gregor, B. Krämer, H.-J. Rahn, M. Patting, F. Koberling, J. Enderlein, and M. Sauer, "Multi-target spectrally resolved fluorescence lifetime imaging microscopy," *Nature Methods*, vol. 13, no. 3, p. 257, 2016.

[431] J. Pawley, *Handbook of biological confocal microscopy*. Springer Science & Business Media, 2010.

[432] W. Denk, J. H. Strickler, and W. W. Webb, "Two-photon laser scanning fluorescence microscopy," *Science*, vol. 248, no. 4951, pp. 73–76, 1990.

[433] M. J. Rust, M. Bates, and X. Zhuang, "Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM)," *Nature Methods*, vol. 3, no. 10, p. 793, 2006.

[434] A. Ghosh, N. Karedla, J. C. Thiele, I. Gregor, and J. Enderlein, "Fluorescence lifetime correlation spectroscopy: Basics and applications," *Methods*, vol. 140, pp. 32–39, 2018.

[435] S. T. Hess, S. Huang, A. A. Heikal, and W. W. Webb, "Biological and chemical applications of fluorescence correlation spectroscopy: a review," *Biochemistry*, vol. 41, no. 3, pp. 697–705, 2002.

[436] G. I. Redford and R. M. Clegg, "Polar plot representation for frequency-domain analysis of fluorescence lifetimes," *Journal of Fluorescence*, vol. 15, no. 5, p. 805, 2005.

[437] S. T. Hess and W. W. Webb, "Focal volume optics and experimental artifacts in confocal fluorescence correlation spectroscopy," *Biophysical Journal*, vol. 83, no. 4, pp. 2300–2317, 2002.

[438] U. Haupts, S. Maiti, P. Schwille, and W. W. Webb, "Dynamics of fluorescence fluctuations in green fluorescent protein observed by fluorescence correlation spectroscopy," *Proceedings of the National Academy of Sciences*, vol. 95, no. 23, pp. 13 573–13 578, 1998.

[439] S. Jazani, I. Sgouralis, and S. Pressé, "A method for single molecule tracking using a conventional single-focus confocal setup," *The Journal of Chemical Physics*, vol. 150, no. 11, p. 114108, 2019.

[440] P. I. Bastiaens and A. Squire, "Fluorescence lifetime imaging microscopy: spatial resolution of biochemical processes in the cell," *Trends in Cell Biology*, vol. 9, no. 2, pp. 48–52, 1999.

[441] M. A. Digman and E. Gratton, "Lessons in fluctuation correlation spectroscopy," *Annual review of physical chemistry*, vol. 62, pp. 645–668, 2011.

[442] R. Ankri, A. Basu, A. Can Ulku, C. Bruschini, E. Charbon, S. Weiss, and X. Michalet, "Single photon, time-gated, phasor-based fluorescence lifetime imaging through highly scattering medium," *ACS Photonics*, 2019.

[443] A. Ulku, A. Ardelean, M. Antolovic, S. Weiss, E. Charbon, C. Bruschini, and X. Michalet, "Wide-field time-gated SPAD imager for phasor-based FLIM applications," *Methods and Applications in Fluorescence*, vol. 8, no. 2, p. 024002, 2020.

[444] M. A. Digman, M. Stakic, and E. Gratton, "Raster image correlation spectroscopy and number and brightness analysis," in *Methods in Enzymology*. Elsevier, 2013, vol. 518, pp. 121–144.

[445] R. Duncan, A. Bergmann, M. Cousin, D. K. Apps, and M. J. Shipston, "Multi-dimensional time-correlated single photon counting (TCSPC) fluorescence lifetime imaging microscopy (FLIM) to detect FRET in cells," *Journal of Microscopy*, vol. 215, no. 1, pp. 1–12, 2004.

[446] X. Michalet, S. Weiss, and M. Jäger, "Single-molecule fluorescence studies of protein folding and conformational dynamics," *Chemical Reviews*, vol. 106, no. 5, pp. 1785–1813, 2006.

[447] M. A. Digman, R. Dalal, A. F. Horwitz, and E. Gratton, "Mapping the number of molecules and brightness in the laser scanning microscope," *Biophysical Journal*, vol. 94, no. 6, pp. 2320–2332, 2008.

[448] M. OLeary, D. Boas, X. Li, B. Chance, and A. Yodh, "Fluorescence lifetime imaging in turbid media," *Optics Letters*, vol. 21, no. 2, pp. 158–160, 1996.

[449] A. Orte, J. M. Alvarez-Pez, and M. J. Ruedas-Rama, "Fluorescence lifetime imaging microscopy for the detection of intracellular ph with quantum dot nanosensors," *ACS Nano*, vol. 7, no. 7, pp. 6387–6395, 2013.

[450] R. Datta, A. Alfonso-García, R. Cinco, and E. Gratton, "Fluorescence lifetime imaging of endogenous biomarker of oxidative stress," *Scientific Reports*, vol. 5, p. 9848, 2015.

[451] N. Ma, N. R. de Mochel, P. D. Pham, T. Y. Yoo, K. W. Cho, and M. A. Digman, "Label-free assessment of pre-implantation embryo quality by the fluorescence lifetime imaging microscopy (FLIM)-phasor approach," *Scientific Reports*, vol. 9, no. 1, pp. 1–13, 2019.

[452] J. R. Lakowicz, *Principles of fluorescence spectroscopy*. Springer Science & Business Media, 2013.

[453] T. W. Gadella Jr, T. M. Jovin, and R. M. Clegg, "Fluorescence lifetime imaging microscopy (FLIM): spatial resolution of microstructures on the nanosecond time scale," *Biophysical Chemistry*, vol. 48, no. 2, pp. 221–239, 1993.

[454] E. Gratton, S. Breusegem, J. D. Sutin, Q. Ruan, and N. P. Barry, "Fluorescence lifetime imaging for the two-photon microscope: time-domain and frequency-domain methods," *Journal of Biomedical Optics*, vol. 8, no. 3, pp. 381–391, 2003.

[455] E. B. van Munster and T. W. Gadella, "Fluorescence lifetime imaging microscopy (FLIM)," in *Microscopy techniques*. Springer, 2005, pp. 143–175.

[456] D. Elson, J. Requejo-Isidro, I. Munro, F. Reavell, J. Siegel, K. Suhling, P. Tadrous, R. Benninger, P. Lanigan, J. McGinty *et al.*, "Time-domain fluorescence lifetime imaging applied to biological tissue," *Photochemical & Photobiological Sciences*, vol. 3, no. 8, pp. 795–801, 2004.

[457] H. S. Chung, J. M. Louis, and I. V. Gopich, "Analysis of fluorescence lifetime and energy transfer efficiency in single-molecule photon trajectories of fast-folding proteins," *The Journal of Physical Chemistry B*, vol. 120, no. 4, pp. 680–699, 2016.

[458] J. Yoo, J. M. Louis, I. V. Gopich, and H. S. Chung, "Three-color single-molecule fret and fluorescence lifetime analysis of fast protein folding," *The Journal of Physical Chemistry B*, vol. 122, no. 49, pp. 11 702–11 720, 2018.

[459] K. A. Merchant, R. B. Best, J. M. Louis, I. V. Gopich, and W. A. Eaton, "Characterizing the unfolded states of proteins using single-molecule fret spectroscopy and molecular simulations," *Proceedings of the National Academy of Sciences*, vol. 104, no. 5, pp. 1528–1533, 2007.

[460] M. I. Rowley, P. R. Barber, A. C. Coolen, and B. Vojnovic, "Bayesian analysis of fluorescence lifetime imaging data," in *Multiphoton Microscopy in the Biomedical Sciences XI*, vol. 7903. International Society for Optics and Photonics, 2011, p. 790325.

[461] B. Kaye, P. J. Foster, T. Y. Yoo, and D. J. Needleman, "Developing and testing a Bayesian analysis of fluorescence lifetime measurements," *PloS One*, vol. 12, no. 1, p. e0169337, 2017.

[462] M. I. Rowley, A. C. Coolen, B. Vojnovic, and P. R. Barber, "Robust Bayesian fluorescence lifetime estimation, decay model selection and instrument response determination for low-intensity FLIM imaging," *PloS One*, vol. 11, no. 6, p. e0158404, 2016.

[463] K. Santra, E. A. Smith, X. Song, and J. W. Petrich, "A Bayesian approach for extracting fluorescence lifetimes from sparse data sets and its significance for imaging experiments," *Photochemistry and Photobiology*, vol. 95, no. 3, pp. 773–779, 2019.

[464] S. Wang, J. V. Chacko, A. K. Sagar, K. W. Eliceiri, and M. Yuan, "Nonparametric empirical Bayesian framework for fluorescence-lifetime imaging microscopy," *Biomedical Optics Express*, vol. 10, no. 11, pp. 5497–5517, 2019.

[465] K. Ishii and T. Tahara, "Two-dimensional fluorescence lifetime correlation spectroscopy. 1. principle," *The Journal of Physical Chemistry B*, vol. 117, no. 39, pp. 11 414–11 422, 2013.

[466] T. Tahara and K. Ishii, "Two-dimensional fluorescence lifetime correlation spectroscopy. 2. application," *The Journal of Physical Chemistry B*, vol. 117, no. 39, pp. 11 423–11 432, 2013.

[467] W. Becker, *Advanced time-correlated single photon counting applications*. Springer, 2015, vol. 111.

[468] K. Ishii, T. Otosu, and T. Tahara, "Lifetime-weighted FCS and 2D FLCS: Advanced application of time-tagged TCSPC," in *Advanced Photon Counting*. Springer, 2014, pp. 111–128.

[469] I. Gregor and M. Patting, "Pattern-based linear unmixing for efficient and reliable analysis of multicomponent TCSPC data," in *Advanced Photon Counting*. Springer, 2014, pp. 241–263.

[470] A. Clayton, Q. Hanley, and P. Verveer, "Graphical representation and multicomponent analysis of single-frequency fluorescence lifetime imaging microscopy data," *Journal of Microscopy*, vol. 213, no. 1, pp. 1–5, 2004.

[471] M. Štefl, N. G. James, J. A. Ross, and D. M. Jameson, "Applications of phasors to in vitro time-resolved fluorescence measurements," *Analytical Biochemistry*, vol. 410, no. 1, pp. 62–69, 2011.

[472] C. Stringari, A. Cinquin, O. Cinquin, M. A. Digman, P. J. Donovan, and E. Gratton, "Phasor approach to fluorescence lifetime microscopy distinguishes different metabolic states of germ cells in a live tissue," *Proceedings of the National Academy of Sciences*, vol. 108, no. 33, pp. 13 582–13 587, 2011.

[473] W. Becker, "Fluorescence lifetime imaging–techniques and applications," *Journal of Microscopy*, vol. 247, no. 2, pp. 119–136, 2012.

[474] S. Ranjit, L. Malacrida, D. M. Jameson, and E. Gratton, "Fit-free analysis of fluorescence lifetime imaging data using the phasor approach," *Nature Protocols*, vol. 13, no. 9, pp. 1979–2004, 2018.

[475] M. A. Digman, V. R. Caiolfa, M. Zamai, and E. Gratton, "The phasor approach to fluorescence lifetime imaging analysis," *Biophysical Journal*, vol. 94, no. 2, pp. L14–L16, 2008.

[476] E. Hinde, M. A. Digman, C. Welch, K. M. Hahn, and E. Gratton, "Biosensor förster resonance energy transfer detection by the phasor approach to fluorescence lifetime imaging microscopy," *Microscopy Research and Technique*, vol. 75, no. 3, pp. 271–281, 2012.

[477] F. Fereidouni, A. Esposito, G. Blab, and H. Gerritsen, "A modified phasor approach for analyzing time-gated fluorescence lifetime images," *Journal of Microscopy*, vol. 244, no. 3, pp. 248–258, 2011.

[478] F. Fereidouni, A. N. Bader, and H. C. Gerritsen, "Spectral phasor analysis allows rapid and reliable unmixing of fluorescence microscopy spectral images," *Optics Express*, vol. 20, no. 12, pp. 12 729–12 741, 2012.

[479] Y. Sun, R. N. Day, and A. Periasamy, "Investigating protein-protein interactions in living cells using fluorescence lifetime imaging microscopy," *Nature Protocols*, vol. 6, no. 9, p. 1324, 2011.

[480] M. Tavakoli, S. Jazani, I. Sgouralis, O. M. Shafraz, S. Sivasankar, B. Donaphon, M. Levitus, and S. Pressé, "Pitching single-focus confocal data analysis one photon at a time with Bayesian nonparametrics," *Physical Review X*, vol. 10, no. 1, p. 011021, 2020.

[481] D. O'Connor, *Time-correlated single photon counting*. Academic Press, 2012.

[482] W. Becker, A. Bergmann, M. Hink, K. König, K. Benndorf, and C. Biskup, "Fluorescence lifetime imaging by time-correlated single-photon counting," *Microscopy Research and Technique*, vol. 63, no. 1, pp. 58–66, 2004.

[483] I. I. Hirschman and D. V. Widder, *The convolution transform*. Courier Corporation, 2012.

[484] A. E. Gelfand, A. Kottas, and S. N. MacEachern, "Bayesian nonparametric spatial modeling with dirichlet process mixing," *Journal of the American Statistical Association*, vol. 100, no. 471, pp. 1021–1035, 2005.

[485] J. M. Haile, *Molecular dynamics simulation: elementary methods*. Wiley New York, 1992, vol. 1.

[486] M. Y. Berezin and S. Achilefu, "Fluorescence lifetime measurements and biological imaging," *Chemical Reviews*, vol. 110, no. 5, pp. 2641–2684, 2010.

[487] M. Köllner and J. Wolfrum, "How many photons are necessary for fluorescence-lifetime measurements?" *Chemical Physics Letters*, vol. 200, no. 1-2, pp. 199–204, 1992.

[488] E. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky, "Nonparametric Bayesian learning of switching linear dynamical systems," in *Advances in neural information processing systems*, 2009, pp. 457–464.

[489] E. Fox and E. B. Sudderth, "Bayesian nonparametric inference of switching dynamic linear models," *IEEE Transactions on Signal Processing*, vol. 59, no. 4, pp. 1569–1585, 2011.

[490] P. Orbanz and Y. W. Teh, "Bayesian nonparametric models." *Encyclopedia of Machine Learning*, no. 1, 2010.

[491] Y. W. Teh and M. I. Jordan, "Hierarchical Bayesian nonparametric models with applications," *Bayesian Nonparametrics*, vol. 1, pp. 158–207, 2010.

[492] Y. W. Teh, "Dirichlet process," *Encyclopedia of Machine Learning*, pp. 280–287, 2010.

[493] Y. Sun, J. Phipps, D. S. Elson, H. Stoy, S. Tinling, J. Meier, B. Poirier, F. S. Chuang, D. G. Farwell, and L. Marcu, "Fluorescence lifetime imaging microscopy: in vivo application to diagnosis of oral carcinoma," *Optics Letters*, vol. 34, no. 13, pp. 2081–2083, 2009.

[494] M. C. Skala, K. M. Riching, D. K. Bird, A. Gendron-Fitzpatrick, J. Eickhoff, K. W. Eliceiri, P. J. Keely, and N. Ramanujam, "In vivo multiphoton fluorescence lifetime imaging of protein-bound and free nicotinamide adenine dinucleotide in normal and precancerous epithelia," *Journal of Biomedical Optics*, vol. 12, no. 2, p. 024014, 2007.

[495] T. Otosu, K. Ishii, H. Oikawa, M. Arai, S. Takahashi, and T. Tahara, "Highly heterogeneous nature of the native and unfolded states of the B domain of protein a revealed by two-dimensional fluorescence lifetime correlation spectroscopy," *The Journal of Physical Chemistry B*, vol. 121, no. 22, pp. 5463–5473, 2017.

[496] A. Jasra, C. C. Holmes, and D. A. Stephens, "Markov chain monte carlo methods and the label switching problem in Bayesian mixture modeling," *Statistical Science*, pp. 50–67, 2005.

[497] L. Egidi, R. Pappada, F. Pauli, and N. Torelli, "Relabelling in Bayesian mixture models by pivotal units," *Statistics and Computing*, vol. 28, no. 4, pp. 957–969, 2018.

[498] G. A. Mills-Tettey, A. Stentz, and M. B. Dias, "The dynamic hungarian algorithm for the assignment problem with changing costs," 2007.

[499] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[500] J. Munkres, "Algorithms for the assignment and transportation problems," *Journal of the Society for Industrial and Applied Mathematics*, vol. 5, no. 1, pp. 32–38, 1957.