# Single Photon smFRET. III. Application to Pulsed Illumination

Matthew Safar[1,2], Ayush Saurabh[1], Bidyut Sarkar[3,4], Mohamadreza Fazel[1], Kunihiko Ishii[3,4], Tahei Tahara[3,4], Ioannis Sgouralis[5], and Steve Pressé[1,6]

[1]Center for Biological Physics, Department of Physics, Arizona State University, Tempe, AZ, USA
[2]Department of Mathematics and Statistical Science, Arizona State University, Tempe, AZ, USA
[3]Molecular Spectroscopy Laboratory, RIKEN, 2-1 Hirosawa, Wako, Saitama 351-0198, Japan
[4]Ultrafast Spectroscopy Research Team, RIKEN Center for Advanced Photonics (RAP), 2-1 Hirosawa, Wako, Saitama 351-0198, Japan
[5]Department of Mathematics, University of Tennessee Knoxville, Knoxville, TN, USA
[6]School of Molecular Sciences, Arizona State University, Phoenix, AZ, USA

July 21, 2022

## Contents

**1 Terminology Convention** **2**

**2 Introduction** **2**

**3 Forward Model and Inverse Strategy** **3**
3.1 Inference Procedure: Parametric Sampler . . . . . 6
3.1.1 Inference Procedure: Nonparametrics Sampler . . . . . 7

**4 Results** **8**
4.1 Simulated Data Analysis . . . . . 8
4.2 Experimental Data Analysis: Holliday Junction . . . . . 11
4.3 Experimental Data Acquisition . . . . . 11

**5 Discussion** **14**

**6 Acknowledgments** **15**

**Bibliography** **16**

**Abstract**

Förster resonance energy transfer (FRET) using pulsed illumination has been pivotal in probing complex single molecule dynamics within subcellular environments. However, there are still major challenges in quantitative single photon, single molecule FRET (smFRET) data analysis under pulsed illumination including: 1) simultaneously deducing kinetics and number of system states; 2) providing uncertainties over estimates, particularly uncertainty over state numbers; 3) taking into account all experimental details such as crosstalk and instrument response function contributing to uncertainty; in addition to 4) background. Here, we implement the Bayesian nonparametric framework described in the first companion paper that addresses all the aforementioned issues in smFRET data analysis specialized for the case of pulsed illumination. Furthermore, we apply our method to both synthetic as well as experimental data acquired using Holliday junction.

# 1 Terminology Convention

To be consistent throughout our three part manuscript, we precisely define some terms as follows:

1. a molecular complex labeled with a FRET dye pair is always referred to as a *system*,
2. the configurations through which a system transitions are termed *system states*,
3. FRET dyes undergo quantum mechanical transitions between *photophysical states*,
4. a system-FRET combination is always referred to as a *composite*, and
5. a composite undergoes transitions among its *superstates*.

# 2 Introduction

Among the many fluorescence methods available [1–7], single molecule Förster resonance energy transfer (smFRET) has been useful in probing interactions and conformational variations within complex cellular environments at the single molecule scale [8–12].

In such smFRET experiments, the data collected typically involves a set of photon arrivals from donor and acceptor fluorophores excited using either pulsed or continuous illumination techniques [8]. Here, we focus on pulsed illumination where the sample is illuminated at regularly-spaced short pulses and the photon arrival times are recorded with respect to the previous pulse. The set of acquired arrival times contains information on fluorophore lifetimes, FRET rates associated with system states, and system transition rates. This

information is often decoded using FRET data analysis methods by: histogram methods [13–15]; bulk correlative methods [16–18]; and single photon methods [15, 19]. However, these methods are limited in learning the system kinetics rather than deducing the number of system states and its uncertainty by taking experimental details such as the IRF into account. In particular, the uncertainty over the number of states is ignored when using model selection methods methods such as the Bayesian information criterion (BIC) [20, 21].

In this paper, we adapt the general smFRET analysis framework presented in the first companion paper [22] for the case of pulsed illumination to learn full distributions over the system kinetics and photophysical rates, *i.e.*, donor and acceptor relaxation and FRET rates, while 1) inferring full distributions over the number of system states; and while 2) taking into account experimental factors such as IRF and crosstalk. As our main concern is deducing system state numbers using single photon arrivals while incorporating detector effects, we leverage the formalism of infinite hidden Markov models (iHMM) [23–28] within the Bayesian nonparametric (BNP) paradigm [23, 24, 29–36]. The iHMM framework assumes an *a priori* infinite number of system states with associated probabilities where the number of system states warranted by input data is enumerated by non-zero probabilities.

In what follows, we first briefly describe our adaptation of the mathematical framework presented in the first companion manuscript leveraging BNPs. Next, we demonstrate that our BNP analysis framework and its software implementation BNP-FRET [37] can robustly learn the system state numbers, associated system transitions and FRET rates while providing full distributions over all estimated quantities.

The synthetic and experimental smFRET data analyzed are acquired using a single confocal microscope with pulsed illumination. The excited donor then relaxes back to the ground state either radiatively by emitting a photon or non-radiatively via FRET leading, in turn, to an acceptor photon emission. As there are two detection channels, photons can be detected in the incorrect channel through crosstalk or not detected at all due to imperfect detector efficiency. Furthermore, recorded photon arrival times are corrupted by the IRF as well as from background photon sources [8, 38, 39]. Through our BNP paradigm, we rigorously propagate uncertainties by accounting for all such sources of errors.

To do so, we employ a broad range of synthetic data as well as empirical data acquired using Holliday junctions (HJ) with an array of different kinetic rates due to varying buffer concentration of $MgCl_2$ [40–43].

# 3 Forward Model and Inverse Strategy

In this section, we first briefly illustrate the adaptation of the general formalism described in our first companion paper [22] to the pulsed illumination case. Next, we present an specialized inference procedure for pulsed illumination. The details of the framework not provided herein can be found in the Supplementary Information.

We begin by considering a molecular complex labeled with a donor-acceptor FRET pair. As the molecular complex transitions through its $M_\sigma$ system states indexed by $\sigma_{1:M_\sigma}$, laser pulses separated by time $\tau$ may excite either the donor or acceptor to drive transitions among the photophysical states, $\psi_{1:M_\psi}$, as defined in the first companion manuscript. Such photophysical transitions lead to photon emissions that may be detected in the donor or

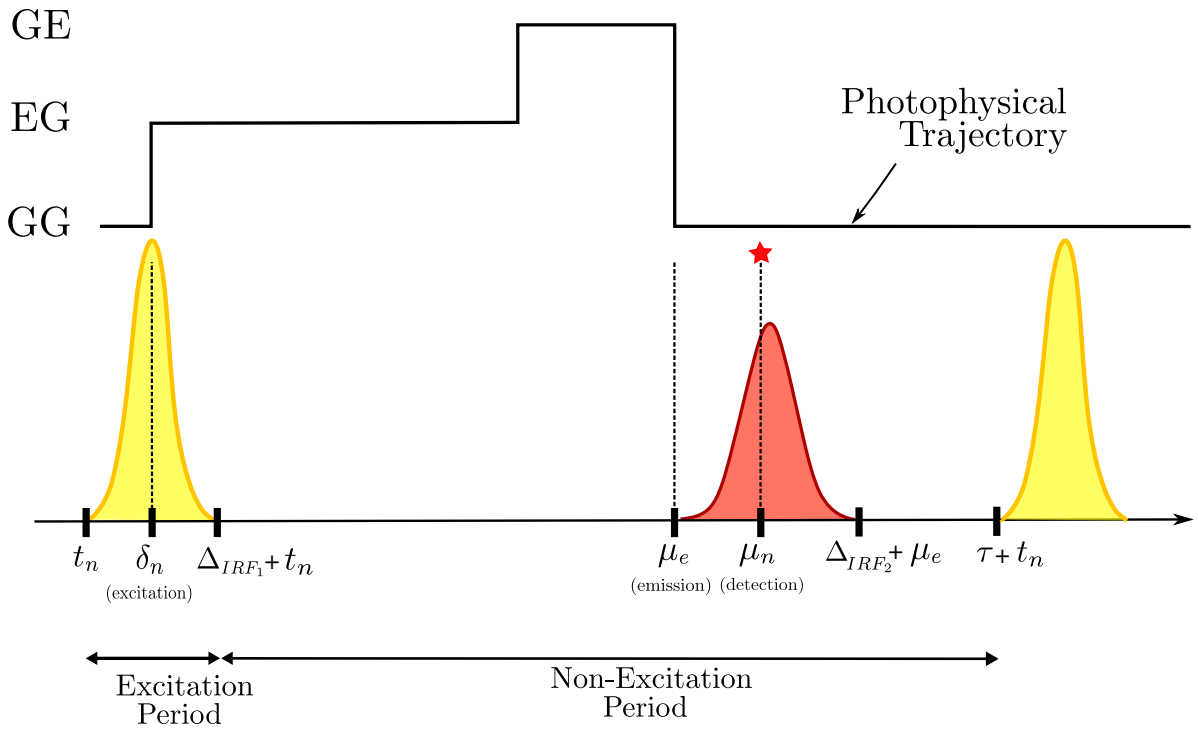


Figure 1: **Events over a pulsed illumination experiment pulse window**. Here, the beginning of the $n$-th interpulse window of size $\tau$ is marked by time $t_n$. The sample is then excited by a high intensity burst (shown by the yellow Gaussian) for a very short time $\Delta_{IRF_1}$. If excited, the fluorophore then emits a photon at $\mu_e$. However, detection, highlighted with a red star, occurs at time $\mu_n$ dictated by the IRF (shown by the red Gaussian). The acronyms GG, EG and GE denote the photophysical states of a FRET pair where $G$ and $E$, respectively, stand for ground and excited states. The first letter indicates the photophysical state of the donor.

acceptor channel. The arrival times of the detected photons are recorded as

$$w_{1:N} = \{w_1, w_2, \ldots, w_N\}. \tag{1}$$

Here, each individual measurement is a pair $w_n = (\mu_n^d, \mu_n^a)$, where $\mu_n^d$ and $\mu_n^a$ are the recorded arrival times (also known as microtimes) after the $n$-th pulse in both donor and acceptor channels, respectively. In cases where there is no photon detection, we denote the absent microtimes with $\mu_n^d = \varnothing$ and $\mu_n^a = \varnothing$ for donor and acceptor channels, respectively.

As is clear from Fig. 1, smFRET traces are inherently stochastic due to the nature of photon emission and noise introduced by detector electronics. To analyze such stochastic systems, we begin with the most generic likelihood derived in the first companion manuscript Eq. (20)

$$L \propto \boldsymbol{\rho}_{start}\ \mathbf{Q}_1\ \ldots\ \mathbf{Q}_n\ \ldots \mathbf{Q}_N\ \boldsymbol{\rho}_{norm}^T, \tag{2}$$

where $\boldsymbol{\rho}_{start}$ is the initial probability vector for the system-FRET composite to be in one of the $M$ $(= M_\psi \times M_\sigma)$ superstates, and $\boldsymbol{\rho}_{norm}$ is a vector that sums the elements of the propagated probability vector. Here, $\mathbf{Q}_n$ is the transition probability matrix between pulses $n$ and

$n+1$, characterizing system-FRET composite transitions among superstates. This transition probability matrix adopts different forms depending on whether a photon is detected or not during the associated period. In the case of photon detection, it is given by Eq. (22) in the first companion manuscript as

$$\mathbf{Q}_n = \mathbf{exp}\left(\int_0^{\Delta_{IRF1}} d\delta_n \mathbf{G}^{non}(\delta_n)\right) \int_0^{\Delta_{IRF2}} d\epsilon_n \, \mathbf{exp}\left(\left(\mu_n - \Delta_{IRF1} - \epsilon_n\right)\mathbf{G}^{dark}\right)\mathbf{G}^{rad} \times \mathbf{exp}\left(\left(\tau - \mu_n\right)\mathbf{G}^{dark}\right), \tag{3}$$

where $\mathbf{G}^{non}$ is the generator matrix in cases where only nonradiative superstate transitions are allowed, and is computed from the full generator $\mathbf{G}$ given in the first companion manuscript Sec. 2.7. Similarly, $\mathbf{G}^{dark}$ is the generator matrix when no excitation occurs and $\mathbf{G}^{rad}$ corresponds to when only radiative transitions take place. As shown in Fig. 1, $\Delta_{IRF1}$ and $\Delta_{IRF2}$ correspond to the width of the pulse and the IRF distribution, respectively, and the variables of integration, $\delta_n$ and $\epsilon_n$, correspond to the times of excitation and emission, respectively. The explicit form of $\mathbf{Q}_n$ for no photon detection is derived in the first companion manuscript Eq. (21).

So far, we have summarized the results already discussed in the first companion manuscript. However, now, we go beyond the first companion paper [22] by introducing a few realistic approximations and developing a specialized sampling scheme for these physically motivated approximations. These approximations include: 1) since the typical system kinetic timescales ($\approx$ 1 ms) are much longer than interpulse periods ($\approx$ 100 ns), we assume that the system state remains the same over an interpuls period; 2) the interpulse period ($\approx$ 100 ns) is longer than the donor and acceptor lifetimes ($\approx$ a few ns) so that they relax to the ground state before the next pulse.

The immediate implications of assumption (1) are that the system transitions may now to a good approximation only occur at the beginning of each pulse. Consequently, the evolution of the FRET pair between two consecutive pulses is now exclusively photophysical as the system state remains the same during interpulses. As such, the system now evolves in equally spaced discrete time steps of size $\tau$ where the system state trajectory can be written as

$$s_{1:N} = \{s_1, s_2, \ldots, s_n, \ldots, s_{N-1}, s_N\}.$$

where $s_n$ is the system state between pulses $n$ and $n+1$. The stochastic evolution of the system states in such discrete steps is determined by the transition probability matrix designated by $\mathbf{\Pi}$. For example, in the simplest case of a molecular complex with two system states $\sigma_{1:2}$, this matrix is computed as

$$\mathbf{\Pi} = \mathbf{exp}\left(\tau \begin{bmatrix} * & \lambda_{\sigma_1\to\sigma_2} \\ \lambda_{\sigma_2\to\sigma_2} & * \end{bmatrix}\right) = \begin{bmatrix} \pi_{\sigma_1\to\sigma_1} & \pi_{\sigma_1\to\sigma_2} \\ \pi_{\sigma_2\to\sigma_1} & \pi_{\sigma_2\to\sigma_2} \end{bmatrix}, \tag{4}$$

where the matrix in the exponential contains the transition rates among the system states and the $*$ represents the negative row sum.

Before proceeding to derive the transition probability matrix for a pulse, we repeat that, in general, the evolution of a system-FRET composite is described by the evolution of its system and photophysical states. This evolution is governed by the generator matrix $\mathbf{G}$

collecting both photophysical and system tramsition rates. However, for the pulse illumination case, the system state is fixed during interpulses by assumption (2) and is given by $s_n$ for the $n$th interpulse period. As such the evolution of the system-FRET composite during this interpulse window is completely described by the evolution of the photophysical state governed by the photophysical portion of the generator matrix. Therefore, the generic $\mathbf{Q}_n$ given by Eq. 3 reduces to $\mathbf{Q}_n^{\psi}$, denoting the transition probability matrix between only photophysical states, by restricting the generators to the photophysical portion associated with a fixed system state $s_n$.

We can now further simplify the problem by supposing that the fluorophores always start in the ground state at the beginning of every pulse by assumption (2). As a result, we can treat the pulses independently and write the likelihood as a product of individual pulse likelihoods:

$$L(w_{1:N}|\vartheta) = \prod_{n=1}^{N} L_n(w_n|\vartheta) = \prod_{n=1}^{N} \left(\boldsymbol{\rho}_{ground}\mathbf{Q}_n^{\psi}(s_n)\boldsymbol{\rho}_{norm}^{T}\right), \tag{5}$$

where $\boldsymbol{\rho}_{ground}$ denotes the probability vector when the FRET pair is in the ground state at the beginning of each pulse. The explicit form of the likelihood for individual pulses is derived in Supplementary Information Sec. S2. Here, $\vartheta$ is the set of parameters we wish to estimate including: the number of system states, $M_\sigma$, FRET rates, $\lambda_{FRET}^{1:M_\sigma}$, donor and acceptor relaxation rates, $\lambda_d$ and $\lambda_a$, donor excitation probability, $\pi_{ex}$, the system state trajectory, $s_{1:N}$, and the system transition probabilities $\pi_{\sigma_i\to\sigma_j}$. This form of the likelihood is advantageous in that it allows empty pulses to be computed together with themselves, greatly decreasing computational cost.

In the following, we first illustrate a parametric inference procedure assuming a given number of system states. We next generalize the developed procedure to a nonparametric case to deduce the number of system states along the rest of parameters.

## 3.1 Inference Procedure: Parametric Sampler

Now, with the likelihood at hand, we proceed to construct the object of prime interest in Bayesian inference, the posterior. Essentially, the posterior is the probability distribution obtained by updating a preliminary distribution over parameters, termed prior, as more data is incorporated through the likelihood function. More formally, updating is performed using Bayes' rule as

$$p(\vartheta|w_{1:N}) \propto L(w_{1:N}|\vartheta)p(\vartheta). \tag{6}$$

where $p(\vartheta)$ is the prior.

Here, the first most notable prior is the categorical prior on the system states, $s_n$,

$$s_n \sim \mathbf{Categorical}_{1:M}(\pi_{1:M}), \tag{7}$$

which is an extension of the Bernoulli distribution for more that two system states. The second important prior is the Dirichlet prior on the system transition probabilities assuming a given number of system states $M_\sigma$

$$\boldsymbol{\pi}_{1:M_\sigma} \sim \mathbf{Dirichlet}(\boldsymbol{\alpha}), \tag{8}$$

where $\boldsymbol{\alpha}$ is a vector of $M_\sigma$ elements called the concentration parameter. For the remaining parameters, we opt for priors that are either physically or computationally motivated provided in Supplementary Information Sec. S3.

After constructing the posterior, we can make inferences on the parameters by drawing samples from the posterior. However, as the resulting posterior has a non-analytical form, it cannot be directly sampled. Therefore, we develop a Markov chain Monte Carlo sampling (MCMC) procedure [36, 44–48] to draw samples from the posterior.

Our MCMC scheme follows a Gibbs sampling technique that sweeps through updates of the set of parameters in the following order: 1) donor and acceptor relaxation rates, $\lambda_d$ and $\lambda_a$, using the Metropolis Hasting (MH) procedure; 2) FRET rates, $\lambda_{FRET}^{1:M}$, for each system state using MH; 3) per-pulse donor excitation probability, $\pi_{ex}$ by directly sampling from the posterior; 4) transition probabilities between system states, $\pi_{1:M}$ by directly drawing samples from the posterior; 5) the system states trajectory, $s_{1:N}$, using forward backward sampling procedure [49]. In the end, the chains of samples drawn can be used for subsequent numerical analysis.

### 3.1.1 Inference Procedure: Nonparametrics Sampler

The smFRET data analysis method illustrated above assumes a given number of system states, $M_\sigma$. However, in many applications the number of system states is not specified *a priori*. Here, we describe a generalization of our parametric method to address this shortcoming and estimate the number of system states simultaneously along with the other unknown parameters.

We accomplish this by modifying our previously introduced parametric posterior as follows.

First, we suppose an infinite number of system states ($M_\sigma \to \infty$) for the likelihood introduced previously and learn the transition matrix $\mathbf{\Pi}$. The number of system states is then enumerated as those with nonzero transition probabilities.

To incorporate this infinite system state space into our inference strategy, we leverage the iHMM [23, 24, 26–28] from the BNP repertoire, placing a hierarchical Dirichlet process prior over the infinite set of system states instead of simply using a Dirichlet prior as described in the first companion manuscript. However, as dealing with an infinite number of random variables is not computationally feasible, we approximate this infinite value with a large number $M_\sigma^{max}$, reducing our hierarchical Dirichlet process prior to

$$\begin{aligned} \boldsymbol{\beta} &\sim \textbf{Dirichlet}\left(\frac{\gamma}{M_\sigma^{max}}, \ldots, \frac{\gamma}{M_\sigma^{max}}\right), \\ \boldsymbol{\pi}_m &\sim \textbf{Dirichlet}(\alpha\boldsymbol{\beta}), \qquad\qquad m = 1, ..., M_\sigma^{max}. \end{aligned}$$

Here $\boldsymbol{\beta}$ denotes the $M_\sigma^{max}$ long base probability vector serving itself as a prior on the probability transition matrix $\mathbf{\Pi}$ to reduce overfitting, and $\boldsymbol{\pi}_m$ is the $m$-th row of $\mathbf{\Pi}$. Moreover, $\gamma$ is the parameter of the Dirichlet process prior [26, 27].

Now, equipped with the nonparametric posterior, we proceed to simultaneously make inferences on the number of system states as well all remaining parameters. To do so, we employ the aforementioned Gibbs sampling scheme, except that we must now also sample

the base distribution $\boldsymbol{\beta}$. More details on the overall sampling scheme are found in the SI in section S4.

# 4 Results

The main objective of our method is to learn full distributions over the: 1) number of system states $M_\sigma$; 2) FRET rates, $\lambda_{1:M}^{FRET}$; 3)fluorophores' relaxation rates (inverse of lifetimes), $\lambda_a$ and $\lambda_d$; and 4) transition probabilities; $\pi_{\sigma_i \to \sigma_j}$. To sample over distributions over these parameters, our method requires input data comprised of photon arrival time traces from both donor and acceptor channels as well as a set of input parameters that can be precalibrated including: elements of the crosstalk matrix; background emission; detection efficiency; and IRF.

Here, we first show that our method samples posteriors over set of parameters employing realistic synthetic data generated using Gillespie's algorithm [50] to simulate system and photophysical state dynamics. The list of parameters used in data generation for all the figures is provided in Supplementary Information Table S2. Furthermore, the parameters used in the analysis of synthetic and experimental data are listed in the Supplementary Information Sec. S3.

We first show that our method works for the simplest case of slow transitions with two system states using synthetic data, see Fig. 2. Next, we proceed to tackle more challenging synthetic data with three system states and higher transition rates and temporal resolution. We show that our nonparametric algorithm correctly infers the number of system states and the corresponding transition rates; see Fig. 3.

After demonstrating the performance of our method using synthetic data, we use experimental data to investigate the dynamics of the HJ under different concentrations of magnesium chloride ($MgCl_2$) in the buffer; see Fig. 4. As expected, decreasing concentrations of $Mg^{2+}$ decrease screening between the negatively charged arms of the HJ resulting in longer low FRET dwells where arms are further apart [42, 51].

## 4.1 Simulated Data Analysis

To help validate BNPs on smFRET single photon data, we start with a simple case of a two state system and select kinetics similar to those of the experimental data sets, *cf.* the HJ in 10 mM $MgCl_2$, which has escape rates, *i.e.*, the rate of transitions pointing out of system states, at $40\,s^{-1}$ [52]. The generated system state trajectory and photon traces over a period of 500 ms from both channels are shown in Fig. 2 (a).

Fig. 2 (b) shows the posterior distribution over FRET efficiencies (FRET efficiency is defined as $\epsilon_{FRET} = \lambda_{FRET}/(\lambda_{FRET} + \lambda_d)$) and system state escape rates with two peaks corresponding to the two system states. If this were to be the HJ, the escape rates would coincide with escape from low FRET to high FRET states and vice versa.

Furthermore, the ground truth, designated by red dots, falls well within the high posterior region. The results for the remaining parameters, including donor and acceptor transition rates, FRET transition rates and system transition probabilities, are presented in Supplementary Information Section S7.

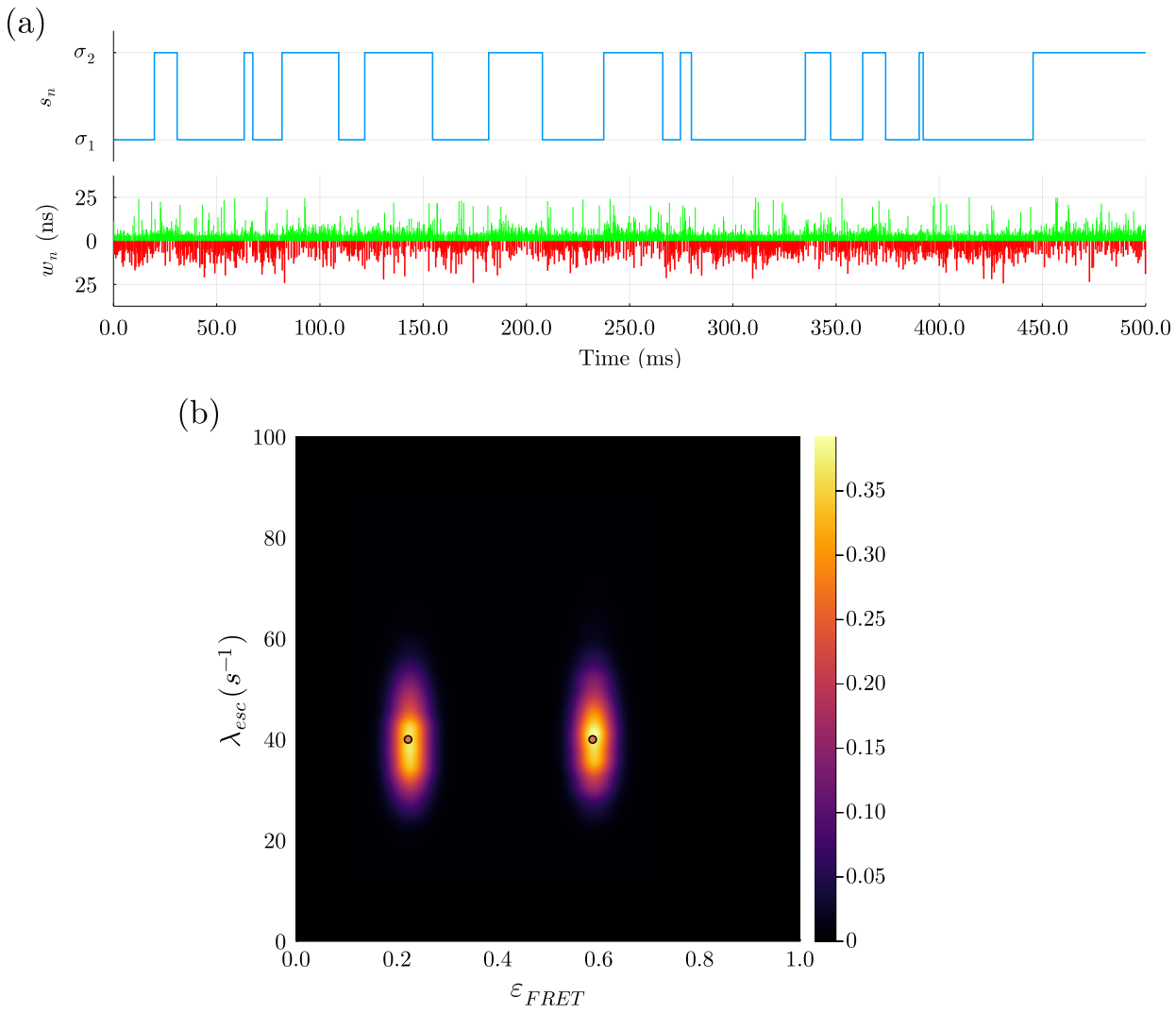


Figure 2: **Analysis on synthetic data for a system with two system states**. In panel (a), we show a section of synthetic data produced with the values in Table S2, which can be found in the SI. Furthermore, the system state trajectory is shown in blue. Below this, the arrival times of donor and acceptor photons $\mu_n^d$ and $\mu_n^a$ are shown in green and red, respectively. In panel (b), we show the bivariate posterior for the system transition rates $\lambda_{esc}$ and FRET efficiencies $\varepsilon_{FRET}$. The ground truth is shown with red dots. We see that we are able to clearly distinguish two system states and locate the ground truth values for the associated escape rates and FRET efficiencies, which are near the peaks of the learned posterior distribution. We have smoothed the distributions using Julia's Distribution.jl kernel density estimation (KDE).

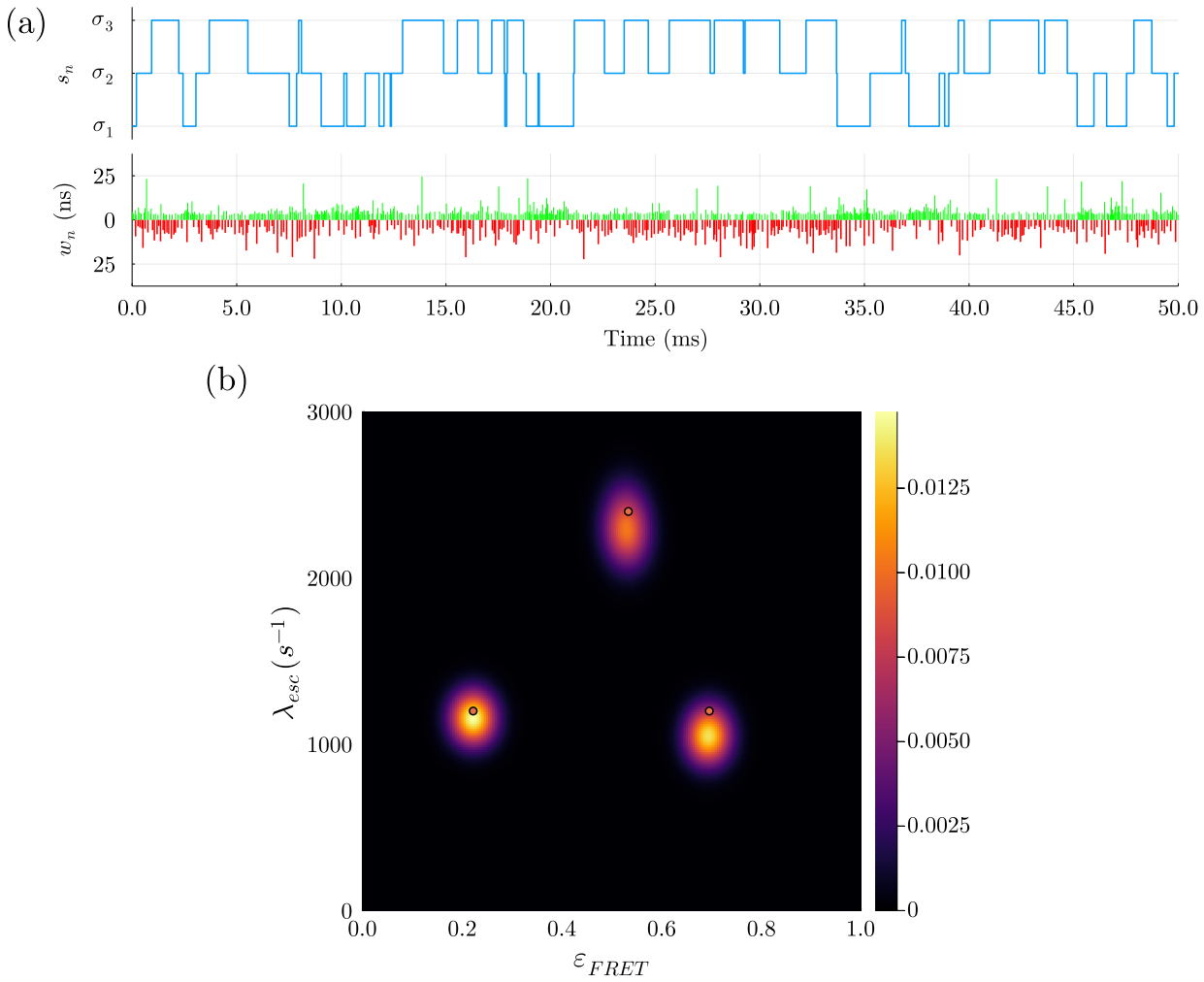


Figure 3: **Analysis on synthetic data for three system states**. In panel (a) we have a section of synthetic data produced with the values from Supplementary Information Table S3. The system state trajectory is seen in blue. Below this, the arrival times of donor and acceptor photons $\mu_n^d$ and $\mu_n^a$ are shown in green and red, respectively. In panel (b), we have the bivariate posterior for the system transition rates $\lambda_{esc}$ and FRET efficiencies $\varepsilon_{FRET}$. The ground truth values remain within two standard deviations of each posterior peak.

To showcase the critical role played by BNPs, we also consider the more difficult case of a system with three system states and faster system state dynamics ranging over 1200-2500 $s^{-1}$. To do so, we simulate traces of photons in both donor and acceptor channels over a period of ∼150 ms. An example of the synthetic data over 50 ms is depicted in Fig 3a.

Using direct photon arrivals from the generated trace of photons, our method predicts the correct number of system states, as shown in Fig 3b, while inferring all other parameters. Furthermore, our method learns the system transition rates where the ground truth fall within one standard deviation of the resulting distributions. Our method does so while

rigorously propagating uncertainty from all existing noise sources, such as background, rather than eliminating noise in the preprocessing steps [15, 19, 53]. The results for remaining parameters are provided in the Supplementary Information Section S7.

## 4.2 Experimental Data Analysis: Holliday Junction

In this section, we benchmark our method over a wide range of kinetic rates employing experimental data acquired using HJ with different kinetic rates arising from varying the concentration of $MgCl_2$ in buffer [52, 54]. The HJ kinetic rates have been extensively studied using both fluorescence lifetime correlation spectroscopy (FLCS) [54] and HMM analysis [55] on diffusing HJs assuming *a priori* a pair of high and low FRET system states. These previous studies show kinetic rates decreasing with increasing concentrations of $MgCl_2$ [42, 51]. This occurs due to the inability of the compact stacked X-structure to transition through the fully extended form of HJ at higher $Mg^{2+}$ concentration [42].

The observed rapid kinetics at low concentrations of $Mg^{2+}$ necessitates methods free of averaging and binning that can resolve such short timescales as well as transient conformations of HJs. As such, our nonparametric method and its implementation BNP-FRET eliminates all averaging and binning by contrast to FLCS or HMM analysis.

Here, we apply the BNP-FRET to data acquired from HJs at 1, 3, 5, and 10 mM $MgCl_2$ concentrations, and sample the distribution over the number of system states and rates. The acquired posterior distributions over the FRET efficiencies and escape rates are presented in Fig. 4. Moreover, estimates for the other parameters can be found in the SI Section S7. We note that our results are obtained on a single molecule basis with a photon budget of $10^4 - 10^5$ photons.

Our nonparametric sampler estimates the number of system states to be two, while this was given as an input to the other analysis methods [54, 55]. Moreover, the escape rates found for all predicted system states are of the orders $1400\,s^{-1}$, $140\,s^{-1}$, $72\,s^{-1}$, and $41\,s^{-1}$ for the four concentrations (see Fig. 4), respectively. These escape rates are in close agreement with values reported by FLCS and H2MM methods [54, 55] which lie well within the bounds of our posteriors shown in Fig. 4 while simultaneously, and self-consistently, learning state numbers.

## 4.3 Experimental Data Acquisition

In this section, we describe the protocol to prepare the surface immobilized HJ sample labeled with a FRET pair and the experimental procedure to record smFRET traces from individual immobilized molecules. The sample preparation method and the recording of experimental data follow previous work [56].

**Sample preparation:** The HJ used in this work consists of four DNA strands whose sequences are as follows:

**R-strand:** 5'-CGA TGA GCA CCG CTC GGC TCA ACT GGC AGT CG-3'
**H-strand:** 5'-CAT CTT AGT AGC AGC GCG AGC GGT GCT CAT CG-3'
**X-strand:** 5'-biotin-TCTTT CGA CTG CCA GTT GAG CGC TTG CTA GGA GGA GC-3'

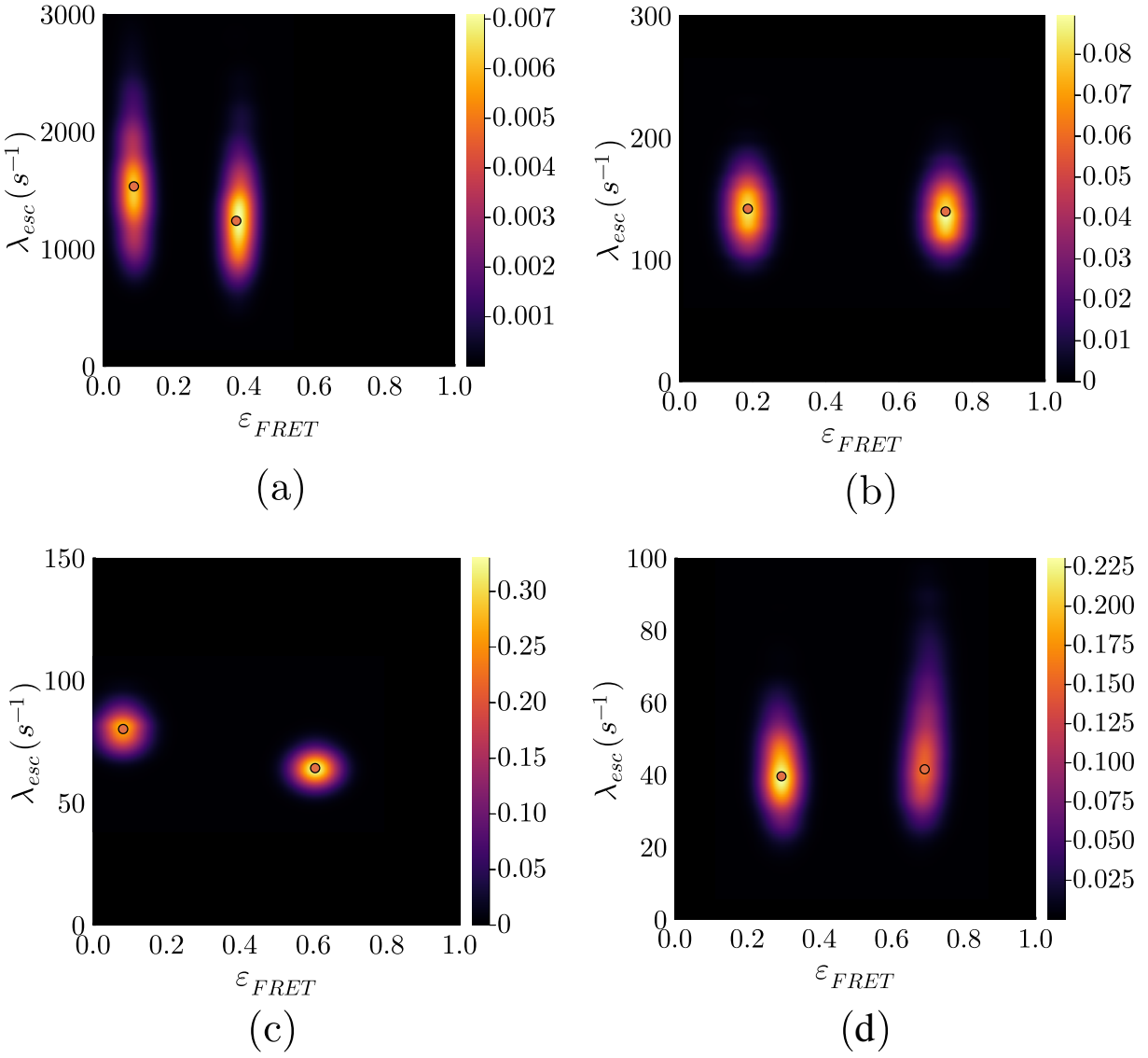


Figure 4: **The bivariate posterior for the conformational transition rates $\lambda_{esc}$ and FRET efficiencies $\varepsilon_{FRET}$.** The MAP estimates are shown with red dots. In panel (a), we show the posterior for a sample with 1 mM $MgCl_2$. We report escape rates of $1533.3 \pm 432.4\,s^{-1}$ and $1239.7 \pm 353.5\,s^{-1}$ in this case. The posterior for a sample with 3 mM$MgCl_2$ is shown in panel (b). We report escape rates of $139.4 \pm 22.2\,s^{-1}$ and $141.7 \pm 22.9\,s^{-1}$ for this case. In panel (c), we show our posterior for a sample with 5 mM$MgCl_2$. Here, we report escape rates of $64.1 \pm 9.9\,s^{-1}$ and $80.1 \pm 6.4\,s^{-1}$. The posterior in panel (d) is for a sample with 10 mM$MgCl_2$. We report escape rates of $41.5 \pm 15.6\,s^{-1}$ and $39.5 \pm 9.4\,s^{-1}$.

**B-strand:** 5'-GCT CCT CCT AGC AAG CCG CTG CTA CTA AGA TG-3'.

For surface immobilization, the X-strand was labeled with biotin at the 5'-end. For FRET

measurements, the donor (ATTO-532) and acceptor (ATTO-647N) dyes were introduced into the H- and B-strands, respectively. In both cases, the dyes were labeled to thymine nucleotide at the 6th position from the 5'-ends of respective strands (shown as T). All DNA samples (labeled or unlabeled) were purchased from JBioS (Japan) in the HPLC purified form and were used without any further purification.

The HJ complex was prepared by mixing 1 mM solutions of R-, H-, B-, and X-strands in TN buffer (10 mM Tris-HCl with 50 mM NaCl, pH 8.0) at 3:2:3:3 molar ratio, annealing the mixture at 94 °C for 4 minutes, and gradually cooling it down (2-3 °C $min^{-1}$) to room temperature (25 °C).

For smFRET measurements, we used a sample chamber (Grace Bio-Labs SecureSeal, GBL621502) with biotin-PEG-SVA (biotin-poly(ethylene glycol)-succinimidyl valerate) coated coverslip. The chamber was first incubated with streptavidin (0.1 mg $mL^{-1}$ in TN buffer) for 20 min. This was followed by washing the chamber with TN buffer (3 times) and injection of 1 nM HJ solution (with respect to its H-strand) for 3-10 seconds. After this incubation period, the chamber was rinsed with TN buffer (3 times) to remove unbound DNA and it was filled with TN buffer containing 1 mM (or 5 mM) $MgCl_2$ and 2 mM Trolox for smFRET measurements.

**smFRET measurements:** The smFRET traces from individual HJs were recorded using a custom built confocal microscope (Nikon Eclipse Ti) equipped with the Perfect Focus System (PFS), a sample scanning piezo stage (Nano control B16-055), and a time correlated single photon counting (TCSPC) module (Becker and Hickl SPC-130EM).

The broadband light generated by a supercontinuum laser operating at 40 MHz (Fianium SC-400-4) was filtered with a bandpass filter (Semrock FF01-525/30) for exciting the donor dye, ATTO-532. This excitation light was introduced to the microscope using a single-mode optical fiber (Thorlabs P5-460B-PCAPC-1), and directed onto the sample using a dichroic mirror (Chroma ZT532/640rpc) and a water immersion objective lens (Nikon Plan Apo IR 60x, numerical aperture = 1.27).

The excitation light was focused onto the top surface of the coverslip and, during measurements, the focusing condition was maintained using the PFS. The fluorescence signals were collected by the same objective, passed through the dichroic mirror, and guided to the detection assembly (Thorlabs DFM1/M) using a multimode fiber (Thorlabs M50L02S-A). Note that this multimode fiber (core diameter: 50 µm) also acts as the confocal pinhole. In the detection assembly, the fluorescence signals from the donor and acceptor dyes were separated using a dichroic mirror (Chroma Technology ZT633rdc), filtered using bandpass filters (Chroma ET585/65m for donor, and Semrock FF02-685/40 for acceptor), and detected using separate hybrid detectors (Becker & Hickl HPM-100-40-C).

For each detected photon, its macrotime (absolute arrival time from the start of the measurement) was recorded with 25.2 ns resolution and its microtime (relative delay from the excitation pulse) was recorded with 6.1 ps resolution using the TCSPC module operating in time-tagging mode. A router (Becker and Hickl HRT-41) was used to process the signals from the donor and acceptor detectors.

For recording smFRET traces from individual HJs, we first imaged a 10 µm×3 µm area of the sample using the piezo stage by scanning it linearly at a speed of 1 µm $s^{-1}$ in the X-direction and with an increment of 0.1 µm in the Y-direction. Individual HJs appeared as isolated bright spots in the image.

Next, we fitted the obtained donor and acceptor intensity images with multiple 2D Gaussian functions to determine the precise locations of individual HJs. Note that, during this image acquisition, the laser excitation power was kept to a minimum ($\sim$1 µW at the back aperture of the objective lens) to avoid photobleaching of the dyes. In addition, we also employed an electronic shutter (Suruga Seiki, Japan) in the laser excitation path to control the sample excitation as required.

Using the obtained precise locations of individual HJs, we recorded 30 s long smFRET traces for each molecule by moving them to the center of the excitation beam using the piezo stage. For each trace, the laser excitation was blocked (using the shutter) for the first 5 seconds and was allowed to excite the sample for the remaining 25 seconds. Note that the smFRET traces were recorded using 40 µW laser excitation (at the back aperture of the objective lens) to maximize the fluorescence photons emitted from the dyes. We automated the process of acquiring smFRET traces from different molecules sequentially and executed it using a program written in-house on Igor Pro (Wavemetrics).

# 5 Discussion

The sensitivity of smFRET has been exploited to investigate many different molecular interactions and geometries [8–11, 57]. However, quantitative interpretation of smFRET data faces serious challenges including unknown number of system states and robust propagation of uncertainty from noise sources such as detectors and background. These challenges ultimately mitigate our ability to determine full distributions over all relevant unknowns and, traditionally, have resulted in data pre- or post-processing compromising the information that is otherwise encoded in the rawest form of data: single photon arrivals.

Here, we provide a general BNP framework for smFRET data analysis starting from single photon arrivals under a pulsed illumination setting. We simultaneously enumerate the number of system states as well as determine rates by incorporating existing sources of uncertainty such as background and crosstalk.

We benchmark our method using both experimental and simulated data. That is, we first show that our method correctly learns parameters for the simplest case with two system states and slow system transition rates. Moreover, we test our method on more challenging cases with more than two states using synthetic data and obtain correct estimations for the number of states along with the remaining parameters of interest. To further assess our method's performance, we analyzed experimental data from HJs suspended in solutions with a range of $MgCl_2$ concentrations. These data were previously processed using other techniques assuming a fixed number of system states by binning photon arrival times [54].

Despite multiple advantages mentioned above for BNP-FRET, BNPs always come with an added computational cost as they take full advantage of information from single photon arrival times and all existing sources of uncertainty. For this version of our general BNP-FRET method simplified for pulsed illumination, we further reduced the computational complexity by grouping empty pulses together. Therefore, the computational complexity increased only linearly with the number of input photons as the photons are treated independently.

The method described in this paper assumes a Gaussian IRF. However, the developed framework is not limited to a specific form for the IRF and can be used for data collected

using any type of IRF by modifying Eq. 3. Furthermore, the framework is flexible in accommodating different illumination techniques such as alternating color pulses, typically used to directly excite the acceptor fluorophores. This can be achieved by simple modification of the generator matrix $\mathbf{G}^{non}$ in Eq. 3. A future extension of this method could relax the assumption of a static sample by adding spatial dependence to the excitation rate as we explored in previous work [33, 48, 58]. This would allow our method to learn the dynamics of diffusing molecules, as well as their photophysical and system state transition rates.

# 6 Acknowledgments

We thank Dr. Zeliha Kilic, Weiqing Xu, and J Shephard Bryan IV for their contributions and insight into the project. We also thank Dr Douglas Shepherd for providing insight into the workings of detectors and other experimental equipment and Irina Gopich for discussions on FRET likelihoods. S. P. acknowledges support from the NIH NIGMS (R01GM130745) for supporting early efforts in nonparametrics and NIH NIGMS (R01GM134426) for supporting single photon efforts.

# Bibliography

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

[2] Jennifer Lippincott-Schwartz, Erik Snapp, and Anne Kenworthy. Studying protein dynamics in living cells. *Nature reviews Molecular cell biology*, 2(6):444–456, 2001.

[3] Bo Huang, Mark Bates, and Xiaowei Zhuang. Super-resolution fluorescence microscopy. *Annual review of biochemistry*, 78:993–1016, 2009.

[4] Mickaël Lelek, Melina T Gyparaki, Gerti Beliu, Florian Schueder, Juliette Griffié, Suliana Manley, Ralf Jungmann, Markus Sauer, Melike Lakadamyali, and Christophe Zimmer. Single-molecule localization microscopy. *Nature Reviews Methods Primers*, 1(1):1–27, 2021.

[5] Mohamadreza Fazel and Michael J Wester. Analysis of super-resolution single molecule localization microscopy data: A tutorial. *AIP Advances*, 12(1):010701, 2022.

[6] Rupsa Datta, Tiffany M Heaster, Joe T Sharick, Amani A Gillette, and Melissa C Skala. Fluorescence lifetime imaging microscopy: fundamentals and advances in instrumentation, analysis, and applications. *Journal of biomedical optics*, 25(7):071203, 2020.

[7] Yuval Garini, Ian T Young, and George McNamara. Spectral imaging: principles and applications. *Cytometry Part A: The Journal of the International Society for Analytical Cytology*, 69(8):735–747, 2006.

[8] Rahul Roy, Sungchul Hohng, and Taekjip Ha. A practical guide to single-molecule FRET. *Nature Methods*, 5(6):507–516, June 2008.

[9] Hisham Mazal and Gilad Haran. Single-molecule FRET methods to study the dynamics of proteins at work. *Current Opinion in Biomedical Engineering*, 12:8–17, 2019.

[10] Benjamin Schuler. Single-molecule FRET of protein structure and dynamics - a primer. *Journal of Nanobiotechnology*, 11(1):S2, December 2013.

[11] Maolin Lu, Xiaochu Ma, Luis R. Castillo-Menendez, Jason Gorman, Nirmin Alsahafi, Utz Ermel, Daniel S. Terry, Michael Chambers, Dongjun Peng, Baoshan Zhang, Tongqing Zhou, Nick Reichard, Kevin Wang, Jonathan R. Grover, Brennan P. Carman, Matthew R. Gardner, Ivana Nikić-Spiegel, Akihiro Sugawara, James Arthos, Edward A. Lemke, Amos B. Smith, Michael Farzan, Cameron Abrams, James B. Munro, Adrian B. McDermott, Andrés Finzi, Peter D. Kwong, Scott C. Blanchard, Joseph G. Sodroski, and Walther Mothes. Associating HIV-1 envelope glycoprotein structures with states on the virus observed by smFRET. *Nature*, 568(7752):415–419, April 2019.

[12] Steven M. Mooney, Ruoyi Qiu, John J. Kim, Elizabeth J. Sacho, Krithika Rajagopalan, Dorhyun Johng, Takumi Shiraishi, Prakash Kulkarni, and Keith R. Weninger. Cancer/testis antigen PAGE4, a regulator of c-Jun transactivation, is phosphorylated by

homeodomain-interacting protein kinase 1, a component of the stress-response pathway. *Biochemistry*, 53(10):1670–1679, March 2014.

[13] Irina V. Gopich and Attila Szabo. Single-macromolecule fluorescence resonance energy transfer and free-energy profiles. *The Journal of Physical Chemistry B*, 107(21):5058–5063, May 2003.

[14] Irina V. Gopich and Attila Szabo. Theory of the energy transfer efficiency and fluorescence lifetime distribution in single-molecule FRET. *Proceedings of the National Academy of Sciences*, 109(20):7747–7752, May 2012.

[15] Hoi Sung Chung, John M. Louis, and Irina 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*, 120(4):680–699, February 2016.

[16] Peter Kapusta, Michael Wahl, Aleš Benda, Martin Hof, and Jörg Enderlein. Fluorescence lifetime correlation spectroscopy. *Journal of Fluorescence*, 17(1):43–48, January 2007.

[17] Kunihiko Ishii and Tahei Tahara. Two-dimensional fluorescence lifetime correlation spectroscopy. 1. principle. *The Journal of Physical Chemistry B*, 117(39):11414–11422, October 2013.

[18] Takuhiro Otosu, Kunihiko Ishii, and Tahei Tahara. Microsecond protein dynamics observed at the single-molecule level. *Nature Communications*, 6(1), July 2015.

[19] Janghyun Yoo, Jae-Yeol Kim, John M. Louis, Irina V. Gopich, and Hoi Sung Chung. Fast three-color single-molecule FRET using statistical inference. *Nature Communications*, 11(1), July 2020.

[20] Sean A. McKinney, Chirlmin Joo, and Taekjip Ha. Analysis of single-molecule FRET trajectories using hidden Markov modeling. *Biophysical Journal*, 91(5):1941–1951, September 2006.

[21] Jonathan E. Bronson, Jingyi Fei, Jake M. Hofman, Ruben L. Gonzalez, Jr., and Chris H. Wiggins. Learning rates and states from biophysical time series: A Bayesian approach to model selection and single-molecule FRET data. *Biophysical Journal*, 97(12):3196–3205, December 2009.

[22] Ayush Saurabh, Matthew Safar, Ioannis Sgouralis, Mohamadreza Fazel, and Steve Pressé. Single photon smFRET. I. theory and conceptual basis. *In preparation.*

[23] Ioannis Sgouralis and Steve Pressé. An Introduction to Infinite HMMs for Single-Molecule Data Analysis. *Biophysical Journal*, 112(10):2021–2029, 2017.

[24] Ioannis Sgouralis, Shreya Madaan, Franky Djutanta, Rachael Kha, Rizal F. Hariadi, and Steve Pressé. A Bayesian nonparametric approach to single molecule Förster resonance energy transfer. *The Journal of Physical Chemistry. B*, 123(3):675–688, January 2019.

[25] Emily B. Fox, Erik B. Sudderth, Michael I. Jordan, and Alan S. Willsky. A sticky HDP-HMM with application to speaker diarization. *The Annals of Applied Statistics*, 5(2A):1020 – 1056, 2011.

[26] Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[27] Jayaram Sethuraman. A constructive definition of dirichlet priors. *Statistica Sinica*, 4(2):639–650, 1994.

[28] JIM PITMAN. Poisson–dirichlet and GEM invariant distributions for split-and-merge transformations of an interval partition. *Combinatorics, Probability and Computing*, 11(5):501–514, 2002.

[29] Thomas S. Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230, 1973. Publisher: Institute of Mathematical Statistics.

[30] Samuel J. Gershman and David M. Blei. A tutorial on bayesian nonparametric models. *Journal of Mathematical Psychology*, 56(1):1–12, 2012.

[31] Ioannis Sgouralis, Miles Whitmore, Lisa Lapidus, Matthew J. Comstock, and Steve Pressé. Single molecule force spectroscopy at high data acquisition: A Bayesian nonparametric analysis. *The Journal of Chemical Physics*, 148(12):123320, March 2018.

[32] Meysam Tavakoli, J. Nicholas Taylor, Chun-Biu Li, Tamiki Komatsuzaki, and Steve Pressé. Single molecule data analysis: An introduction. *arXiv:1606.00403 [physics, q-bio]*, November 2016.

[33] Meysam Tavakoli, 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(1):011021, January 2020. Publisher: American Physical Society.

[34] Meysam Tavakoli, Sina Jazani, Ioannis Sgouralis, Wooseok Heo, Kunihiko Ishii, Tahei Tahara, and Steve Pressé. Direct photon-by-photon analysis of time-resolved pulsed excitation data using Bayesian nonparametrics. *Cell Reports Physical Science*, 1(11):100234, November 2020.

[35] J. Shepard Bryan IV, Ioannis Sgouralis, and Steve Pressé. Diffraction-limited molecular cluster quantification with Bayesian nonparametrics. *Nature Computational Science*, 2(2):102–111, February 2022. Number: 2 Publisher: Nature Publishing Group.

[36] Mohamadreza Fazel, Sina Jazani, Lorenzo Scipioni, Alexander Vallmitjana, Enrico Gratton, Michelle A. Digman, and Steve Pressé. High resolution fluorescence lifetime maps from minimal photon counts. *ACS Photonics*, 9(3):1015–1025, March 2022.

[37] Ayush Saurabh, Matthew Safar, and Steve Pressé. BNP-FRET: A software suite to analyze smFRET data using bayesian nonparametrics. *In preparation. Github link to appear.*

[38] Ammasi Periasamy, Nirmal Mazumder, Yuansheng Sun, Kathryn G. Christopher, and Richard N. Day. *FRET Microscopy: Basics, Issues and Advantages of FLIM-FRET Imaging*, pages 249–276. Springer International Publishing, Cham, 2015.

[39] M. D. Eisaman, J. Fan, A. Migdall, and S. V. Polyakov. Invited review article: Single-photon sources and detectors. *Review of Scientific Instruments*, 82(7):071101, 2011.

[40] Kenji Okamoto and Yasushi Sako. State transition analysis of spontaneous branch migration of the Holliday junction by photon-based single-molecule fluorescence resonance energy transfer. *Biophysical Chemistry*, 209:21–27, February 2016.

[41] Sungchul Hohng, Chirlmin Joo, and Taekjip Ha. Single-molecule three-color FRET. *Biophysical Journal*, 87(2):1328–1337, August 2004.

[42] Sean A. McKinney, Anne-Cécile Déclais, David M.J. Lilley, and Taekjip Ha. Structural dynamics of individual Holliday junctions. *Nature Structural Biology*, 10(2):93–97, February 2003.

[43] S. A. McKinney, A. D. J. Freeman, D. M. J. Lilley, and T. Ha. Observing spontaneous branch migration of Holliday junctions one step at a time. *Proceedings of the National Academy of Sciences*, 102(16):5715–5720, April 2005.

[44] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.

[45] W Keith Hastings. Monte carlo sampling methods using markov chains and their applications. 1970.

[46] J Shepard Bryan IV, Ioannis Sgouralis, and Steve Pressé. Diffraction-limited molecular cluster quantification with bayesian nonparametrics. *Nature Computational Science*, 2(2):102–111, 2022.

[47] Mohamadreza Fazel, Michael J Wester, Hanieh Mazloom-Farsibaf, Marjolein Meddens, Alexandra S Eklund, Thomas Schlichthaerle, Florian Schueder, Ralf Jungmann, and Keith A Lidke. Bayesian multiple emitter fitting using reversible jump markov chain monte carlo. *Scientific reports*, 9(1):1–10, 2019.

[48] Sina Jazani, Ioannis Sgouralis, Omer M Shafraz, Marcia Levitus, Sanjeevi Sivasankar, and Steve Pressé. An alternative framework for fluorescence correlation spectroscopy. *Nature communications*, 10(1):1–10, 2019.

[49] L.R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[50] Daniel T. Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics*, 22(4):403–434, 1976.

[51] I.G. Panyutin, I. Biswas, and P. Hsieh. A pivotal role for the structure of the holliday junction in dna branch migration. *The EMBO Journal*, 14(8):1819–1826, 1995.

[52] Zeliha Kilic, Ioannis Sgouralis, Wooseok Heo, Kunihiko Ishii, Tahei Tahara, and Steve Pressé. Extraction of rapid kinetics from smFRET measurements using integrative detectors. *Cell Reports Physical Science*, 2(5):100409, May 2021.

[53] Irina V. Gopich and Attila Szabo. Decoding the Pattern of Photon Colors in Single-Molecule FRET. *The Journal of Physical Chemistry B*, 113(31):10965–10973, August 2009.

[54] Wooseok Heo, Kazuto Hasegawa, Kenji Okamoto, Yasushi Sako, Kunihiko Ishii, and Tahei Tahara. Scanning two-dimensional fluorescence lifetime correlation spectroscopy: Conformational dynamics of DNA Holliday junction from microsecond to subsecond. *The Journal of Physical Chemistry Letters*, 13(5):1249–1257, February 2022.

[55] Menahem Pirchi, Roman Tsukanov, Rashid Khamis, Toma E. Tomov, Yaron Berger, Dinesh C. Khara, Hadas Volkov, Gilad Haran, and Eyal Nir. Photon-by-Photon Hidden Markov Model Analysis for Microsecond Single-Molecule FRET Kinetics. *The Journal of Physical Chemistry B*, 120(51):13065–13075, December 2016.

[56] Zeliha Kilic, Ioannis Sgouralis, and Steve Pressé. Generalizing HMMs to continuous time for fast kinetics: Hidden markov jump processes. *Biophysical Journal*, 120(3):409–423, 2021.

[57] Judith S Sebolt-Leopold and Jessie M English. Mechanisms of drug inhibition of signalling molecules. *Nature*, 441(7092):457–462, 2006.

[58] Sina Jazani, Ioannis Sgouralis, and Steve Pressé. A method for single molecule tracking using a conventional single-focus confocal setup. *The Journal of chemical physics*, 150(11):114108, 2019.