Simulating Freely-diffusing Single-molecule FRET Data with Consideration of Protein Conformational Dynamics

James Losey,[†] Michael Jauch,[‡] Axel Cortes-Cubero,[¶] Haoxuan Wu,[§] Roberto Rivera,[¶] David S. Matteson,[§] and Mahmoud Moradi[*,†]

† *Department of Chemistry and Biochemistry, University of Arkansas, Fayetteville, AR 72701, U.S.A.*

‡ *Center for Applied Mathematics, Cornell University, Ithaca, NY 14850, U.S.A.*

¶ *Department of Mathematical Sciences, University of Puerto Rico, Mayaguez, Puerto Rico 00681*

§ *Department of Statistics and Data Science, Cornell University, Ithaca, NY 14850, U.S.A.*

E-mail: moradi@uark.edu

Abstract

Single molecule Förster resonance energy transfer experiments have added a great deal to the understanding of conformational states of biologically important molecules. While great progress has been made in studying structural dynamics of biomolecular systems, much is still unknown for systems with conformational heterogeneity particularly those with high flexibility. For instance, with currently available techniques, it is difficult to work with intrinsically disordered proteins, particularly when freely diffusing smFRET experiments are used. Simulated smFRET data allows for the control of the underlying process that generates the data to examine if a given smFRET data analysis technique can detect these underlying differences.

Here, we extend the PyBroMo software that simulates freely diffusing smFRET data to include a distribution of inter-dye distances generated using Langevin dynamics in order to model proteins with conformational flexibility within a given state. We compare standard analysis techniques for smFRET data to validate the new module relative to the base PyBroMo software and observe qualitative agreement in the results of standard analysis for the two timestamp generation methods. The Langevin dynamics module provides a framework for generating timestamp data with an known underlying heterogeneity of inter-dye distances that will be necessary for the development of new analysis techniques that study flexible proteins or other biomolecular systems.

1 Introduction

Structure and dynamics of proteins and other biomolecules are fundamental to their function[1]. Static structural data from high-resolution structure determination techniques such as x-ray crystallography and cryogenic electron microscopy can provide a detailed picture of these systems but they lack information on the transitions between the states. Other experimental techniques may allow for quantitative characterization of the transitions between states[2], however, with a much less spatial resolution as compared to One such technique is single-molecule Förster resonance energy transfer (FRET) spectroscopy or smFRET[2].

FRET is the non-radiative transfer of energy initially absorbed by a "donor" chromophore dye to a nearby "acceptor" dye[3,4]. The energy transferred between a donor and acceptor dye is dependent on the distance between the dyes and can be used to provide information on this distance. Therefore, FRET is often considered as a "spectroscopic ruler".[5] Ensemble FRET experiments, with simultaneous excitation of multiple donors at the same time, contain distance information but they suffer from bulk averaging that can obscure the protein conformational dynamics underlying the process. Through clever experimental design, valuable conformational information can still be gleaned[6–8].

The advent of single molecule spectroscopic techniques transformed biophysics into a source

of dynamic data on molecular structure as well as function[9]. smFRET experiments avoid ensemble averaging by taking advantage of exciting the donors and detecting the donor and acceptor signals at a single molecule level[10,11]. These techniques have become a popular source of spatio-temporal information on the conformational landscape of a molecule and have been applied in studies of a variety of systems from DNA[12], and RNA[13–15], to protein folding[16,17].

The two broad varieties of smFRET experiments are distinguished by how the labeled molecule is isolated from other FRET signals when it is excited. First, surface immobilized experiments fix the labeled molecule to a substrate, expose it to laser light to excite the donor dye, and collect the resulting photon timestamp data. This experimental procedure uses long exposure times to collect data on slower conformational dynamics, greater than 1 ms[18]. Despite experimental difficulties arising from surface impacts on dynamics and signal issues from photo-bleaching or other noise sources, surface immobilized experiments have been a fruitful area of study.

Second, freely diffusing smFRET methods record photon emissions from labeled molecules as they diffuse through a solution with a confocal laser focused inside the solution. Periodically, the path of a molecule will cross the focal region of the laser, where the probability of photon absorption and emission are high. The diffusion rates and concentrations of the molecules in solution as well as the size of the focal region are selected so that the observation of simultaneous excitations of more than one molecule is vanishingly rare within a particular observation time window. Photon detectors, tuned for the wavelengths of the donor and acceptor dyes, record timestamp data for each photon detected. The photon signal occurs in bursts as molecules diffuse into and out of the focal beam of the confocal laser. Freely diffusing experiments can capture dynamics occurring on faster scales[2] and avoid the potential impacts of the surface on conformational dynamics[19–21], but the short bursts of data provide a challenge for analysis.

While sophisticated statistical methodology is essential to the analysis of any smFRET experiment, the literature on this topic has primarily focused on surface-immobilized smFRET[22]. These techniques include histograms, Gaussian mixture models[23], hidden Markov models (HMM)[12,24–26], and Bayesian non-parametric approaches[27]. The freely diffusing smFRET technique is gaining in

popularity due to its simpler experimental methodology with no need for surface immobilization[28]. To further advance the developing fields of smFRET analysis, the ability to realistically simulate the underlying molecular processes in a systematic, controlled, and repeatable manner is a necessity.

Simulated smFRET data has been used in other studies[29,30] though it frequently focuses only on the generation of just the binned photon data. PyBroMo[31], an open source smFRET timestamp simulation software suite, uses a physical model of a diffusion smFRET experiment that combines a Brownian motion simulation to model the molecular diffusion in a solution, a numerical point spread function (PSF) to model the laser, and Poisson background noise to model background photon rates for each channel. These features provide a framework to generate timestamps for multiple populations of freely diffusing molecules with distinct diffusion constants and FRET efficiencies that have a single efficiency state or exhibit dynamic efficiency state switching. As an open source project, researchers can also extend the code to include other features not currently included in the software. For instance, PyBroMo uses a fixed efficiency for each population throughout the duration of the simulation. We propose an extension of PyBroMo to include heterogeneous efficiency states by modeling the underlying distances between the dyes as a dynamic process.

In reality, the distances between the dyes on a labeled molecule (dye-dye distance) are dynamic due to the thermal fluctuations of the molecule. A fixed efficiency assumes that the heterogeneity of dye-dye distances from molecular motion in a freely diffusing molecule is negligible compared to the other parts of the simulation. This simplification may be justifiable for highly structured molecules or at low temperatures. Reductions in molecular structures and greater fluctuations, like those observed in disordered proteins[32,33], will invalidate the assumption. This is especially true for disordered proteins with reduced secondary and tertiary structure to stabilize the conformations. The flexibility of the molecule leads to a heterogeneous conformational ensemble that poses further challenges to the analysis of experimental data. Biologically important systems often contain large heterogeneity of conformational states[34].

To more accurately model the conformational heterogeneity of dye-dye distances of a flex-

ible molecule during an smFRET simulation, an overdamped Langevin method of simulation was added to PyBroMo's existing software to model the internal conformational dynamics of the molecule. The Langevin dynamics will produce a trajectory of dye-dye distances for each molecule that conform to an underlying ground truth related to the potential energy used in the Langevin dynamics. This addition provides a more realistic smFRET simulation, particularly important for unstructured proteins or those associated with intrinsic disorder. This added realism will be necessary in the development of new analysis techniques that account for confomational heterogeneity.

The remaining sections of this paper will provide a more detailed description of PyBroMo, followed by a description of overdamped Langevin dynamics used to generate the distribution of dye-dye distances. Then, two example simulations are described to generate simulated data for molecules in a single state (Example 1) and for molecules that interconvert between two states (Example 2), in section 2. Section 3 shows the results of typical analysis methods applied to the example simulations. A standard analysis for smFRET data using thresholds and Gaussian mixture models was applied to the timestamp data for Example 1 using the base PyBroMo software (non-Langevin) and the extended PyBroMo using the added Langevin dye-dye distances (Langevin). Then, an analysis of the dynamic state model using the non-Langevin and Langevin timestamp data in section 2.4 uses a skew Gaussian mixture model, as well as changepoint analysis and hidden Markov models (HMMs) to assess the dynamics states. Section 4 provides a discussion of the results presented. Finally, section 5 presents conclusions based on the comparison of the analysis for the two simulated data sets.

2 Simulation methods

2.1 PyBroMo

PyBroMo[31] was developed by Ingargiola *et al.* to simulate photon emission from fluorescent dye pairs attached to freely diffusing molecules while recording the timestamps from those emissions,

similar to experimental FRET data. This software was designed to generate realistic FRET data by handling multiple populations of molecules with their own diffusion coefficients and FRET efficiencies, as well as generating background photons with separate emission rates for the donor and acceptor channels.

The first step in generating FRET timestamps is defining the basic elements of the simulation in the form of a Python script. In the script, the molecules are defined by a population number and diffusion coefficient, D_B. The simulation is defined by providing box dimensions, L_x, L_y, and L_z, as well as conditions for how to handle molecule interactions with the boundary. If a molecule's position is advanced across the box boundary, the position is either wrapped across the opposite boundary, or reflected back across the same boundary that was crossed. A point spread function (PSF) is defined to model the laser focal beam inside the simulation box. The PSF represents the emission probability of a molecule at any position within the simulation box. A Gaussian PSF is available where the emission probability in all dimensions is defined by

$$f(x) = \frac{1}{\sigma_x\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu_x}{\sigma_x}\right)^2} \tag{1}$$

where μ_x is the mean coordinate for the center of the function and σ_x is the standard deviation. Eq. (1) can be extended to include other Cartesian coordinates y and z. PyBroMo is also capable of importing custom PSF functions from tools like PSFLab[35] that can generate a custom numerical PSF that includes factors like light polarization. PyBroMo includes a default numeric PSF for use without the user having to create their own.

Next, the simulation inputs are passed to the Brownian motion simulation module along with a timestep (δt) and a maximum time to advance the molecules through the simulation box. The Brownian motion is a stochastic process where the position in each dimension are advanced from the current position by a random number drawn from a normal distribution,

$$x(t+\delta t) = \quad x(t) + \xi \tag{2}$$

where $\xi \sim N(0, 2D_B\delta t)$ for the white noise contribution. The Brownian motion simulation then repeatedly advances each molecule's position in three dimensions by the δt until the maximum time is reached. At each time step, the PSF calculates the normalized emission probability for every molecules position in a trajectory vector, $\mathscr{P}$. Molecules in regions of high emission probability, near the center of the PSF, emit more photons up to the maximum emission rate.

Finally, the timestamp generation module creates the number of photon emissions events, κ, through a discrete random Poisson process

$$f(\kappa, \lambda) = \frac{\lambda^{\kappa} e^{-\lambda}}{\kappa!} \tag{3}$$

where λ is the expected number of emissions. The values needed to calculate the λ values for every time step are a maximum total emission rate, ε_T, efficiency, E, for each population, and the emission probabilities, $\mathscr{P}$ from the Brownian motion simulation. Emission rates for the acceptor, ε_{Acc}, and donor , ε_{Don}, channels are then calculated

$$\varepsilon_{Acc} = \quad \varepsilon_T E \tag{4}$$

$$\varepsilon_{Don} = \quad \varepsilon_T (1 - E) \tag{5}$$

The efficiency, E, is constant for all timesteps. Separate expected counts for the acceptor, λ_{Acc} and donor, λ_{Don} are then calculated

$$\lambda_{Acc} = \mathscr{P}\varepsilon_{Acc}\delta t \tag{6}$$

$$\lambda_{Don} = \mathscr{P}\varepsilon_{Don}\delta t \tag{7}$$

and used to randomly draw emission events at every time step. Similarly, background emissions rates are also determined for the acceptor and donor detector channels by randomly drawn numbers from a Poisson distribution with expected values, $\lambda_{BGAcc}, \lambda_{BGDon}$, supplied as a simulation parameter.

The timestamps are merged and sorted into a single trajectory for output. A vector of labels is also generated to label the timestamp as being from the acceptor or donor channel. Other values of interest that may be included are the molecule ID that generated the photon emission or the position of the molecule in the PSF.

2.2 Overdamped Langevin Dynamics

The use of a static efficiency in the base PyBroMo software implies an underlying static relationship between the two fluorescent dyes labeling the molecule. Fluctuations in molecular structure, particularly in unstructured proteins, could impact how smFRET data is interpreted. To extend the PyBroMo software beyond the static efficiency assumptions, an overdamped Langevin dynamics module is added to simulate realistic dye-dye distance fluctuations over the simulation time as a one dimensional diffusion process within a potential energy field.

The Langevin trajectories are calculated according to the Euler-Maruyama method[36], where at each time step, the dye-dye distance is updated by calculating the contributions from the distance derivative of the potential energy function, $V(r)$ and a stochastic random contribution. This step update is defined as

$$r(t+\delta t) = \quad r(t) - \beta D_L \frac{dV(r)}{dr}\delta t + \xi_L \tag{8}$$

where D_L is the diffusion coefficient, $\xi_L \sim N(0, 2D_L\delta t)$, and $\beta = \frac{1}{k_B T}$ with k_B being the Boltzmann constant, and T is the system temperature. The diffusion coefficient for the dye-dye distance, D_L, is unique from the Brownian motion diffusion coefficient. The user defined potential energy field acts on the molecules as the white noise element perturbs the molecule.

A FRET efficiency model converts the dye-dye distance trajectories to efficiency trajectories. Two different efficiency models are used for the two example scenarios described in greater detail in sections 2.3 and 2.4. However, a constant that is common in efficiency models is the Förster radius, R_0, defined as the distance from the donor dye at which FRET efficiency is 0.5. This R_0

value is specific to the fluorescent dyes used in a smFRET experiment and based on the quantum yield of the donor dye and the spectral overlap of the two dyes.

Eq. (4) and (5) then generate vectors for the acceptor and donor emission rate ε_A and ε_D and Eq. (6) and (7) calculate the expected values λ_{Acc} and λ_{Don}. As with the base PyBroMo, random numbers are drawn from a Poisson distribution defined in Eq. (3) for each timestep. The background timestamp generation is unaffected by the Langevin dynamics module and contributes to the Poisson distributed background timesteps as before. Finally, the timestamps from acceptor, donor, and background are merged, as before, into a single trajectory with channel labels for each photon detected.

Next, we describe the two example simulations to demonstrate the ability of the Langevin dynamics module to generate timestamps.

2.3 Example 1: Molecules in a Single State

To demonstrate the generation of timestamps using the Langevin dynamics module, a simple example system of molecules in a harmonic potential is simulated for three independent simulations with all parameters held constant. The harmonic potential energy, V_H is defined as

$$V_H(r) = \frac{k_H}{2}(r - r_c)^2 \tag{9}$$

where k_H is the harmonic force constant, and r_c is the center of the potential function. 100 molecules are contained in a simulation box with lengths $L_x = L_y = 8$ μm, $L_z = 12$ μm. The Brownian diffusion coefficient, D_B, is set to 30 μm^2/s for all molecules. The Gaussian PSF is centered in the simulation box with a $\sigma_x = \sigma_y = 0.3$ μm, and $\sigma_z = 0.5$ μm. Three independent simulations are run for 10s each with a time step of 50 *ns*. For timestamp generation, a maximum emission rate of 200,000 counts per second (CPS) is used in all the simulations, as well as a background rate of 1,200 CPS for the acceptor channel and 1,800 CPS for the donor channel. The CPS values will be kept consistent for all simulations used in this work.

For the Langevin dynamics parameters, the thermodynamic coefficient β is 1.339 $(\text{kcal/mol})^{-1}$ which corresponds to a relatively high temperature of 378 K for large thermal fluctuations. The Langevin diffusion coefficient, D_L, is 13 Å^2/ms. The harmonic potential is defined by Eq. (9) with the coefficient k_H set at 0.025 (kcal/(mol Å^2)) with the center of the harmonic potential at 40 Å for 50 of the molecules, and at 65 Å for the remaining 50 molecules. Eq. (11) is used to convert the distances to efficiencies. In efficiency conversions, an R_0 of 56 Å is used.

A short trajectory of Langevin dye-dye distances is shown in Figure 1. The molecules in

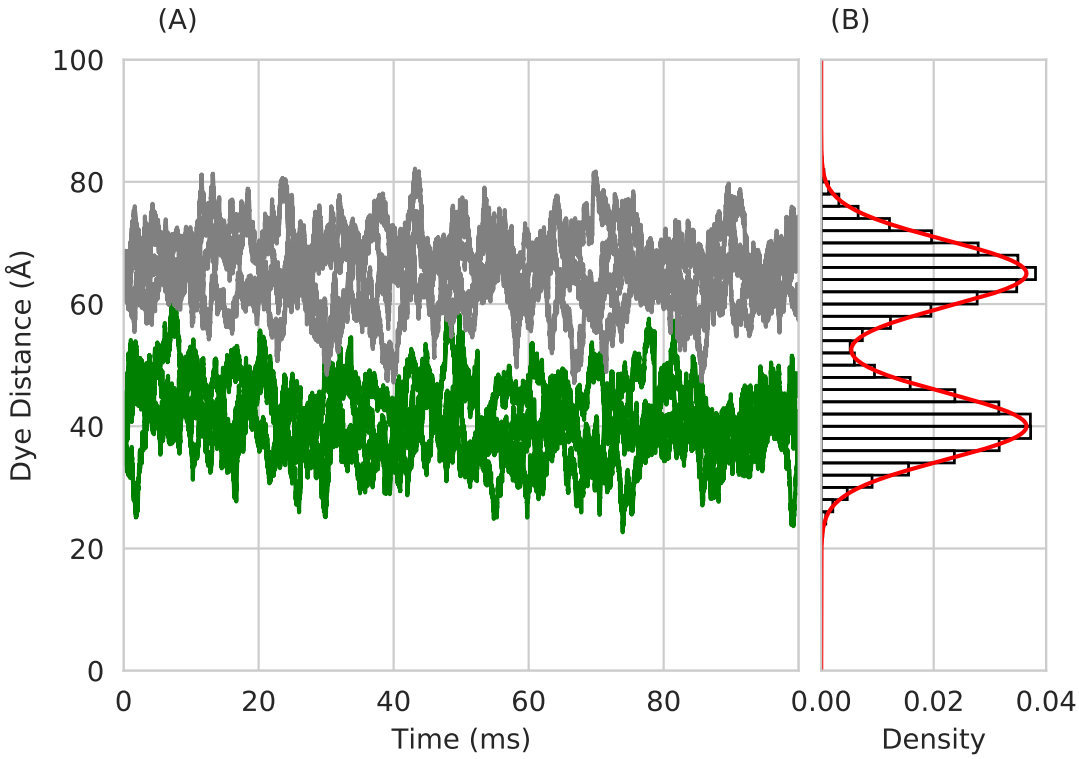

Figure 1: (A) A portion of a trajectory of Langevin dynamics for the dye-dye distance of 4 molecules in a harmonic potential centered at 40 Å (colored green) and 4 molecules in a harmonic potential centered at 65 Å (colored gray). (B) A histogram of the dye-dye distances for the combined populations along with the analytic solution for the distribution in red.

each population oscillate in the harmonic potential over time, with a probability of some dye-dye distance, $P(r)$, following the relation

$$P(r) = \frac{1}{2\sqrt{\frac{2\pi}{\beta k}}}\left(e^{-\beta V_{H1}(r)} + e^{-\beta V_{H2}(r)}\right) \tag{10}$$

where V_{H1} and V_{H2} are the harmonic potentials applied to the two molecule populations in the Langevin dynamics simulation.

For the harmonic simulations, an efficiency model developed for conformationally heterogeneous proteins[33] relating the dye-dye distances to FRET efficiency is,

$$E = \frac{1}{1 + 0.975\left(\frac{r}{R_0}\right)^{2.65}}, \tag{11}$$

where r is the dye-dye distance and R_0 was 56 Å. The FRET efficiencies used for photon generation are 0.41 and 0.71, which corresponded to Eq. (11) applied to distances of 40 Å and 65 Å, respectively. The distances matched the centers of the harmonic potentials used in the Langevin dynamics simulations. The other photon generation parameters for maximum emission rate and background noise were held constant.

To compare the results of the new Langevin dye-dye distance module with the base PyBroMo, three sets of simulated timestamps were generated with the base (non-Langevin) PyBroMo. These simulations used the same number of molecules, and other Brownian motion parameters for diffusion coefficient, simulation box, PSF, and background photons as described above. 50 molecules had an efficiency of $E = 0.71$ while the other 50 had an efficiency of $E = 0.41$. These efficiency values correspond to Eq. (11) applied to the harmonic centers from the Langevin dynamics, 40 Å and 65 Å respectively. The results of this comparison are provided in section 3.1.

2.4 Example 2: Molecules with Inter-conversion Between Two States

The harmonic Langevin simulations described above approximate a system where the dye-dye distance fluctuates around a single state for the duration of the simulation. However, biophysical intuition as well as experimental smFRET data suggest that many biomolecular systems correspond to two or more interconverting conformational states at equilibrium[37].

To simulate a system that dynamically moves between different states, a bistable potential energy with two symmetric wells are applied to a system of molecules in the Langevin dynamics

module. 90 molecules were simulated using the Langevin dye-dye distance module using the same Brownian diffusion coefficient, simulation box, and PSF as previously defined in Section 2.3. This bistable potential, $V_B(r)$, is defined as

$$V_B(r) = \frac{k_B}{4}\left((r - r_C)^2 - W^2\right)^2, \tag{12}$$

where k_B is the bistable force constant set at 10^{-4} (kcal/(mol Å^2)). The location of the center of the potential, r_C, is set in this example at 50 Å, and W is the the offset from the center where the potential wells were located, set at 15 Å. The locations of the potential energy minima is at $r_c \pm W$, or 35 and 65 Å. Using the bistable potential, a Langevin molecule will explore a local potential energy well until a large enough energetic contribution from the white noise in the Langevin dynamics gives the molecule the energy to overcome the energy barrier and explore the other well. A Langevin diffusion coefficient of $D_L = 40$ Å^2/ms is used.

FRET efficiency is modeled using the commonly used relation

$$E = \frac{1}{1 + \left(\frac{r}{R_0}\right)^6}, \tag{13}$$

where r is the dye-dye distance and $R_0 = 56$ Å, as before. The efficiency model in Eq. (13) is based on approximations of FRET theory and widely used in the smFRET literature. In order to gather a sufficient amount of data for analysis, a total of approximately 20 minutes of smFRET data is generated.

A short dye-dye distance trajectory using the bistable potential is shown in Figure 2. We see the dye-dye distances oscillate inside one of the potential wells for some period of time before eventually overcoming the energy barrier between the two wells and switching states. The distribution of dye-dye distances for the bistable Langevin simulation follows the relation

$$P(r) = \frac{1}{Z} e^{-\beta V_B(r)} \tag{14}$$

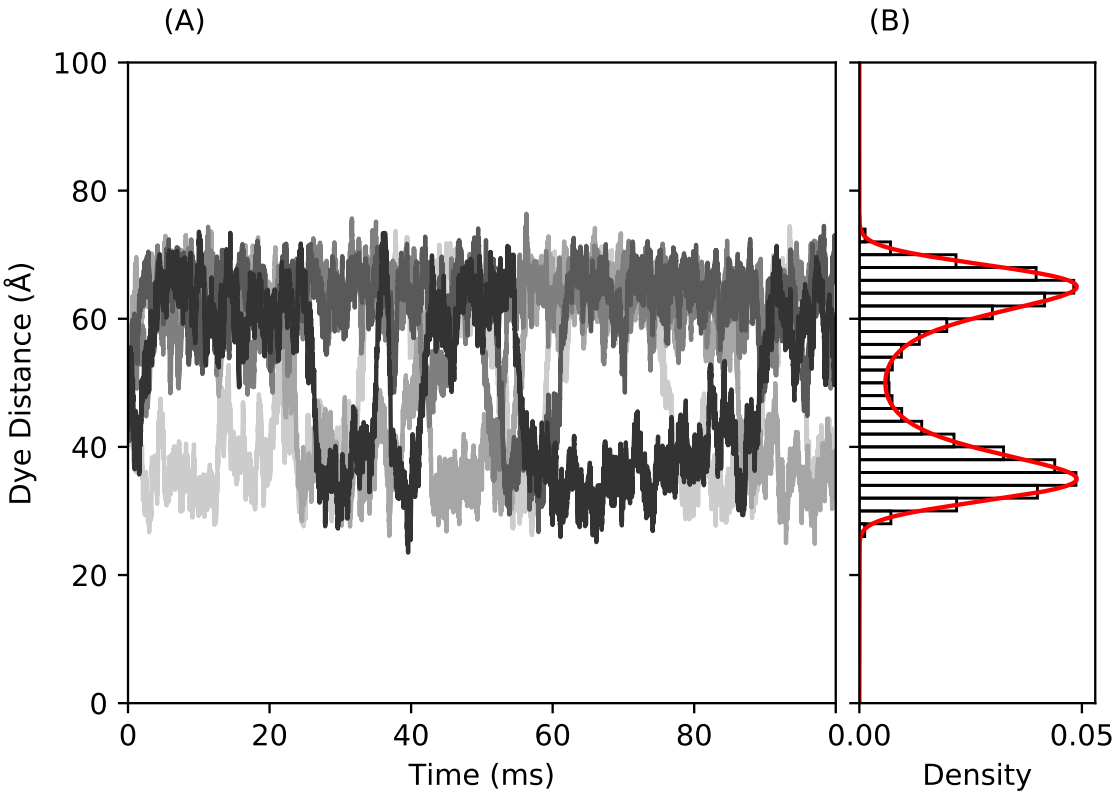

Figure 2: (A) A portion of Langevin dynamics trajectories for the dye-dye distance of 5 molecules moving in a bistable potential centered at 50 Å with potential energy minima at 65 Å and 35 Å. (B) A histogram of the the dye-dye distances is shown with the theoretical probability shown as a red line.

where the partition function for the bistable potential, $Z = \int_0^\infty e^{-\beta V_B(r)} dr$, normalizes the probability density to 1. A lower temperature, $T = 300$K, is used as compared to Example 1, with $\beta = 1.679$ (kcal/mol)$^{-1}$. The lower temperatures decrease the magnitude of thermal fluctuations for each timestep so the molecule will explore the local well long enough to emit sufficient photons for the state to be identifiable.

The analytical transition matrix of the bistable Langevin simulation, $T^{(0)}$, between different states is related to the transition rate matrix, Q, by

$$T^{(0)} = exp(\tau Q) \tag{15}$$

$$Q := \begin{bmatrix} Q_{1,1} & Q_{1,2} \\ Q_{2,1} & Q_{2,2} \end{bmatrix} \tag{16}$$

where τ is the lag time between state determination measurements. The entry $Q_{i,j}$ represents the transition rate from state i to state j. The transition rate between two non-identical states (here reactant, R and product, P) is calculated using relations from Berezhkovskii and Szabo,[38]

$$Q_{R\to P} = \frac{1}{\left(\int_{-\infty}^{x^*} e^{-\beta V(x)} dx\right)\left(\int_R^P e^{\beta V(x)} dx / D(x)\right)}, \tag{17}$$

where the integration limit x^* is the peak of the barrier at 50 Å, $V(x)$ is the potential energy, $D(x)$ is the position dependent diffusion coefficient, and $\beta = \frac{1}{k_B T}$. Substituting the bistable potential, $V(x) = V_B(x)$ and the constant Langevin diffusion coefficient, $D(x) = D_L$, the transition matrix can be computed theoretically as,

$$T^{(0)} = \begin{pmatrix} 0.968 & 0.032 \\ 0.032 & 0.968 \end{pmatrix}. \tag{18}$$

In addition to the 90 molecules in the bistable Langevin simulation, 10 molecules were kept in a constant "donor only" state of $E = 0$. Donor only states are present in experimental data and

represent molecules where only the donor dye is attached, with no FRET possible. The donor only population adds further realism to the analysis of dynamic state simulations as this is a source of error encountered by experimenters.

To provide a comparison with the bistable Langevin timestamps, non-Langevin timestamps were generated that simulated dynamic state switching. This is done by generating two timestamp traces of approximately 20 minutes in length, using the same parameters for Brownian motion as the bistable Langevin data. One set of timestamps used a fixed high efficiency state of $E = 0.944$, while the other used a fixed low efficiency state of $E = 0.290$. The efficiencies correspond to Eq. (13) using the locations of the well minima, $r_C = 35$ Å for the high efficiency state and $r_C = 65$ Å for the low efficiency state Also, a Förster radius of $R_0 = 56$ Å was applied in all the efficiency calculations. Again, the Brownian motion simulations parameters of Brownian diffusion constant, simulation box size, PSF, and background photons were the same for the non-Langevin PyBroMo as with the Langevin dye-dye distance module simulations above.

Transitions between states were simulated by drawing residence times from an exponential distribution with an average residence time of 31.126 ms. The trajectory of an efficiency state evolves like a step function alternating between the two states. This residence time leads to a transition matrix for the non-Langevin data that closely matches the transition rate matrix generated from the bistable potential. Using these residence times, a set of timestamps is created that switched between the two efficiency states, also 20 minutes in overall length.

The results from three analysis methods performed on the dynamic state model simulation timestamps are contained in section 3.2.

3 Results

Techniques for simulating freely diffusing smFRET experiments are valuable, in large part, because they allow researchers to evaluate statistical methods using realistic data with a known ground truth. With this in mind, we present a standard analysis of the timestamp data produced

from the parameters described in sections 2.3 and 2.4.

3.1 Analysis of Example 1

The first experiment was simulated with the base PyBroMo software described in Section 2.1, while the second experiment was simulated with the proposed Langevin dynamics module. The timestamp data generated by both non-Langevin and Langevin simulations was in the form of a column of ordered timestamps when a photon was detected. Additional columns label the channel that detected the photon (donor or acceptor), and a label to identify the molecule that emitted the photon. This molecule identifier would not be available in experimental data, but is information that is available in the simulation.

Data analyses of freely diffusing smFRET experiments typically begin by binning and thresholding the raw photon time stamp data[22,39]. The time bin size needs to be long enough to collect sufficient data such that the signal from the fluorescent dyes can be distinguished from the noise contributions. Conversely, the bin size needs to be small enough so that the FRET signal is only from one molecule. The specific choice of time bin length will be dependent on background noise rates, molecule diffusion rates, and confocal beam size, on the order of 1 ms.[40] In our analyses, we use a typical experimental bin width of one millisecond. For a given experiment, let I_t^{D} and I_t^{A} denote the photon counts in the donor and acceptor channels during time bin t and define the combined count $I_t^{\mathrm{C}} = I_t^{\mathrm{D}} + I_t^{\mathrm{A}}$. We restrict our analyses to those time bins with combined count exceeding 40 photons. Based on the simulation parameters that are used, a combined photon count at or above this magnitude indicates that the signal is very likely from a molecule diffusing across the focal beam and thus the proportion of photons in the acceptor channel reflects the molecule's conformational state. Thesholding also ensures that our estimates of the efficiencies within each time bin are not excessively variable due to low counts. No single method to determine photon thresholds has been universally accepted[41]. In the literature, there are a number of heuristics for choosing the threshold and many alternative approaches to identifying the diffusion of a molecule across the focal beam[42–44].

Central to our analysis are the estimates of efficiencies within each bin, which we refer to as *apparent efficiencies*. The apparent efficiency within bin t is defined as the proportion of the total photon count from that bin which was detected in the acceptor channel:

$$\hat{E}_t = \frac{I_t^{\mathrm{A}}}{I_t^{\mathrm{A}} + I_t^{\mathrm{D}}}. \tag{19}$$

When analyzing real smFRET experiments, estimation of efficiencies should also take into account the so-called γ factor, which accounts for the difference in quantum yields of the donor and acceptor dyes as well as the difference in photon detection efficiencies of the donor and acceptor channels.[45,46] This adjustment is not necessary for our analysis because the smFRET simulations in this article were run with equivalent quantum yields and equivalent detection efficiencies.

We analyze the simulated smFRET experiments using a simple histogram of the apparent efficiencies as well as a Gaussian mixture model fit to the apparent efficiencies. The histogram approximates the marginal distribution of efficiencies. It provides an idea of the relative amount of time a molecule spends at each efficiency and whether there exist easily-distinguished conformational states. In comparison to a histogram-based analysis, the analysis based on a Gaussian mixture model provides more quantitative information related to hypothesized latent conformational states. We suppose that there is a latent conformational state $s_t \in \{1, ..., K\}$ associated with each time bin t and that these latent conformational states are independent and identically distributed with probabilities $\pi_1, ..., \pi_K$. Given that $s_t = k$, we suppose that the apparent efficiency $\hat{E}_t$ follows a Gaussian distribution with mean μ_k and variance σ_k^2. The smFRET simulations were run with $K = 2$ conformational states, and we take this as given. We compute the maximum likelihood estimates of the unknown parameters via an expectation-maximization algorithm[47] as implemented in the `mixtools` package[48] in R[49].

Figure 3 compares the non-Langevin and Langevin simulations in terms of apparent efficiencies and the corresponding dye-dye distances. Figure 3 (A), based on the non-Langevin simulation, shows the estimated two-component Gaussian mixture density (in solid black) on top of a

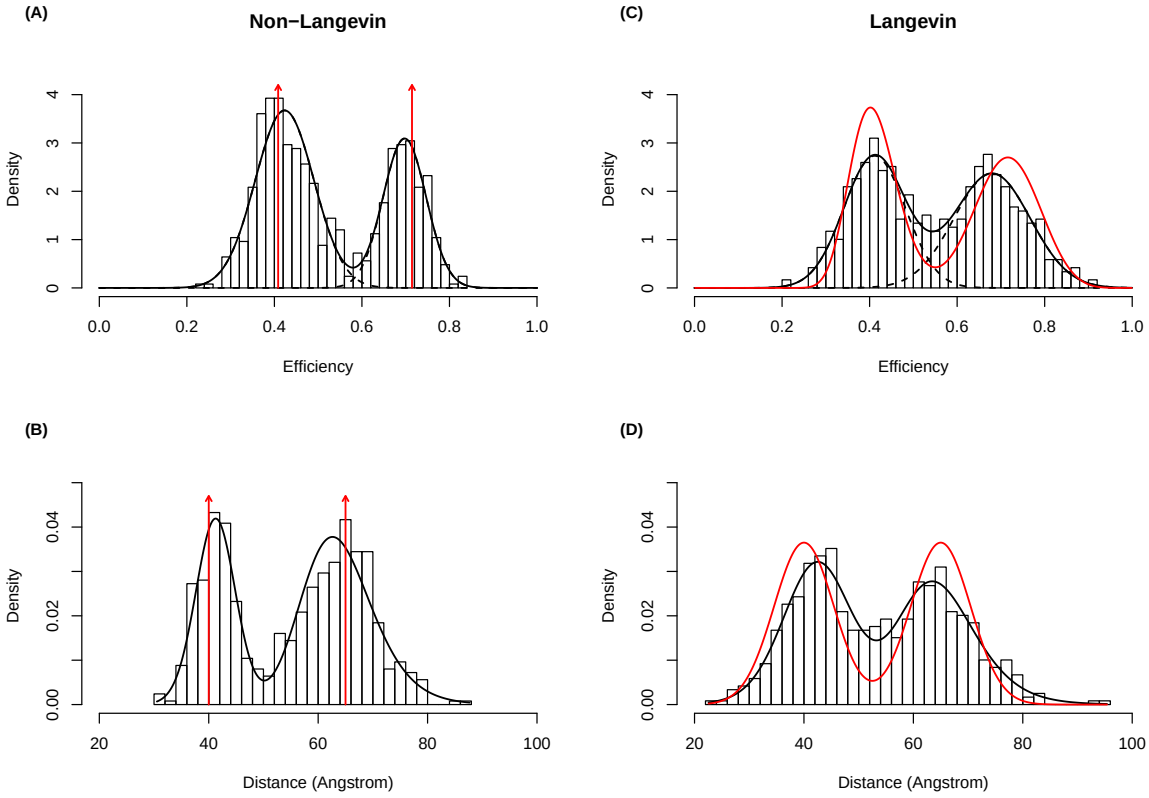

Figure 3: (A) The estimated Gaussian mixture density (solid black line) from the non-Langevin simulation on top of a histogram of the apparent efficiencies along with the two true efficiencies (red vertical arrows). (B) The corresponding plot in the distance space. (C) The analogous efficiency plot for the Langevin simulation. (D) The analogous distance plot for the Langevin simulation.

histogram of the apparent efficiencies. The dashed lines represent the (weighted) densities of the estimated component distributions. The low efficiency component has a mean of 0.42, a standard deviation of 0.07, and a mixture weight of 0.62. The high efficiency component has a mean of 0.70, a standard deviation of 0.05, and a mixture weight of 0.38. The vertical red arrows are placed at the true efficiency values used in the simulation. Figure 3 (B) shows the corresponding histogram, densities, and arrows after a transformation to the distance space. The probability distribution of distances is converted to a probability distribution of efficiencies through a change of variable based on the efficiency model in Eq. (11)

Figure 3 (C) and Figure 3 (D), in the right half of the figure, are analogues of Figure 3 (A) and Figure 3 (B) based on the Langevin simulation. The most substantial difference is that, instead of vertical red arrows at two true efficiencies (or distances), we have densities representing the true, non-degenerate theoretical distribution of efficiencies (or distances). In the distance space, the theoretical distribution is the two component Gaussian mixture specified by Eq. (10). The theoretical distribution in the efficiency space is again obtained through a change of variables from efficiency to distance. In Figure 3 (C), the low efficiency component has a mean of 0.41, a standard deviation of 0.07, and a mixture weight of 0.48, while the high efficiency component has a mean of 0.68, a standard deviation of 0.09, and a mixture weight of 0.52. Figure 3 (D) shows the corresponding histogram and densities after a transformation to the distance space, as done with Figure 3 (B), and the underlying Langevin dye-dye distance distribution shown as a red line. The distinct peaks observed in the non-Langevin timestamp analysis showed less overlap in distribution of the two populations compared with the Langevin simulation timestamp analysis which had a wider distributions with greater overlap. This small difference is reasonable due to the the overlap between the underlying distance distributions of the Langevin dynamics for the two populations. Overall, the analysis demonstrates that the addition of overdamped Langevin dynamics in a simple scenario produces timestamps that contain valuable information from the underlying distance distribution, like the location of efficiency peaks.

3.2 Analysis of Example 2

Next, we describe two more sophisticated analyses that account for additional realistic features included in the simulation, like donor-only particles and dynamic state changes, described in section 2.4. A histogram based analysis as well as analyses to infer state dynamics were performed. Again, the non-Langevin and Langevin timestamps generated using the simulation parameters described in Section 2.4 contained information consistent with the simulation parameters that was detectable by the analyses.

3.2.1 Skew Gaussian Mixture Model

We again analyze the non-Langevin and Langevin timestamps through mixture models. This time, we fit three component skew-Gaussian mixture models to the timestamps generated from Example 2. Adding a third component is necessary because these simulations include donor-only molecules, leading to a low FRET peak. The skew-Gaussian distribution has density

$$\frac{2}{\omega}\phi\left(\frac{x-\xi}{\omega}\right)\Phi\left[\alpha\left(\frac{x-\xi}{\omega}\right)\right] \tag{20}$$

where ϕ and Φ are the density and distribution functions of a standard Gaussian random variable, ξ is a location parameter, ω is a scale parameter, and α is a shape parameter[50,51]. This more flexible parametric family allows us to adequately model skewed distributions. Apparent efficiency distributions which lie near the boundary of the unit interval, including the low FRET peak, typically exhibit strong skewness. We compute the maximum likelihood estimates of the unknown parameters via an expectation-maximization algorithm as implemented in the `mixsmsn` package[52].

The results appear in Figure 4, which compares the non-Langevin and Langevin simulations in terms of apparent efficiencies and the corresponding dye-dye distances. Figure 4 is analogous to Figure 3, except here they depict the results of the skew Gaussian mixture model. The skew Guassian mixture analysis was able to recover the location of efficiency peaks from the timestamp data reasonably well for both the non-Langevin and Langevin data, as well as the donor-only peak.

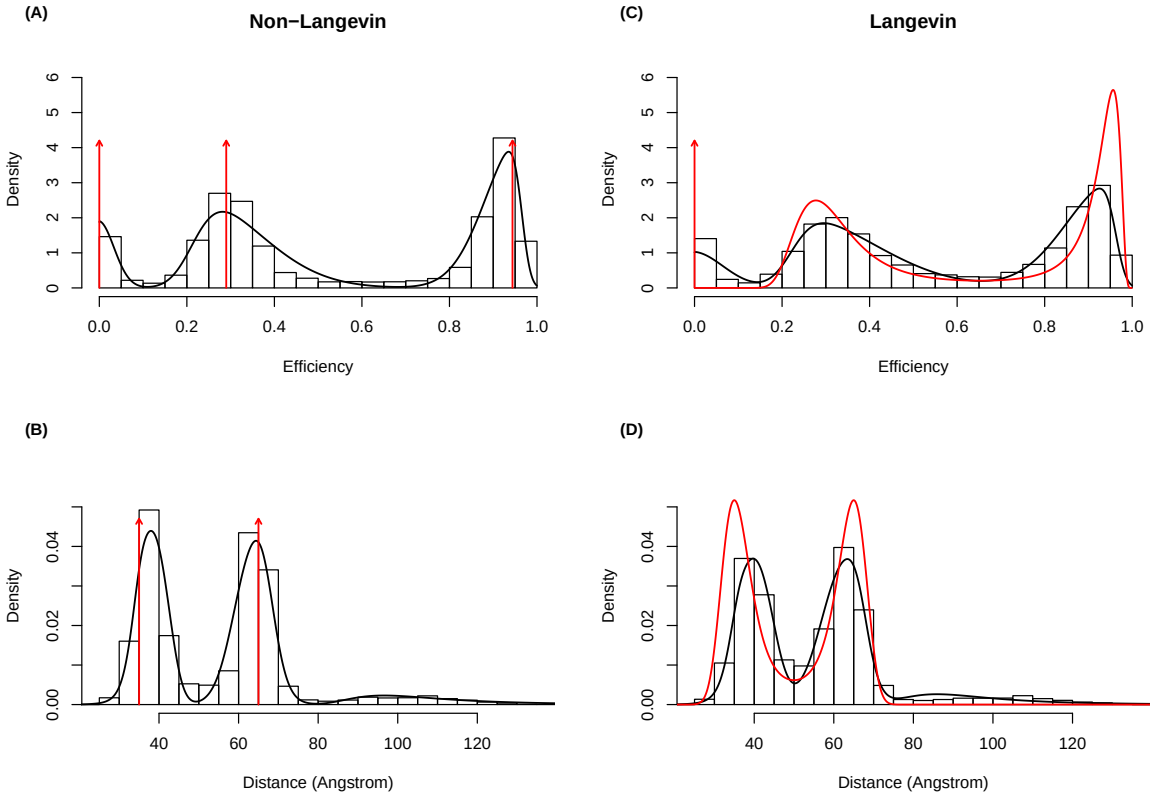

Figure 4: (A) The estimated skew Gaussian mixture density (solid black line) from the non-Langevin simulation on top of a histogram of the apparent efficiencies along with the true efficiencies (red vertical arrows). (B) The corresponding plot in the distance space. (C) The analogous efficiency plot for the Langevin simulation. (D) The analogous distance plot for the Langevin simulation.

Again, the efficiency states for the non-Langevin simulation timestamps showed higher, more well defined peaks with less overlap than the Langevin simulation timestamps, consistent with the point mass distribution in the distance . This method aggregates all the timestamp information over time into a histogram, losing temporal information about switches between states. The next two analysis methods will explore the state switching in the timestamp data with more depth.

3.2.2 HMM Analysis

We analyze the Example 2 timestamp data using a hidden Markov model (HMM)[53]. Specifically, we consider only the time-bins which are above a threshold (where the total photon count is above 40). In contrast to surface immobilized smFRET, in freely diffusing smFRET experiments the molecule is only sometimes in front of the focal beam[26,41]. We define a burst as a set of consecutive time bins such that for each of them, the total photon count is above the threshold. We then evaluate the sequence of apparent efficiencies for each burst. To perform dynamical analysis and detect transitions between the different FRET states, we treat the sequence of apparent efficiencies from each burst as an independent time-series to be modeled with the HMM[53,54], where the HMM parameters are constant for all the independent time-series. We fit the apparent efficiencies using two hidden states, and assume they are normally distributed conditionally on each state. Python's hmmlearn package was deployed to fit the HMM.

For the data generated using Langevin dynamics, the average photon burst duration is 2.18 bins of $1ms$. We fit the HMM using a total of 30053 such bursts and obtain a transition matrix estimate

$$T^{\mathrm{L}} = \begin{pmatrix} 0.960 & 0.040 \\ 0.056 & 0.944 \end{pmatrix}, \tag{21}$$

corresponding to two Gaussian states, for which we estimate means, $\mu_1 = 0.321$, $\mu_2 = 0.883$, and variances $\sigma_1^2 = 0.029$, $\sigma_2^2 = 0.004$, respectively.

For comparison, we analyze the data generated using non-Langevin dynamics, where the average photon burst duration is 2.20 bins of $1ms$. We fit the HMM using a total of 31354 such bursts.

Fitting the data results in a Transition matrix:

$$T^{\mathrm{NL}} = \begin{pmatrix} 0.956 & 0.044 \\ 0.061 & 0.939 \end{pmatrix}, \tag{22}$$

corresponding to two Gaussian states, with means, $\mu_1 = 0.291$, $\mu_2 = 0.910$, and variances $\sigma_1^2 = 0.025$, $\sigma_2^2 = 0.002$, respectively. The analytical transition matrix is the same as for the Langevin case.

Qualitatively, the measured transition matrices for both the Langevin and non-Langevin case look reasonably similar to the analytical transition matrix in Eq. (15). We observed marginally closer Gaussian state estimates for the non-Langevin transition matrix, while the error estimation in the transition matrix elements marginally favored the Langevin data. We present a more quantitative analysis of the error between the known and measured transition matrices for both cases in the Supporting Information Section 2 where our analysis finds smaller measures of error for the Langevin data, compared to the non-Langevin case.

A visualization of the transitions using changepoint analysis is presented in the Supporting Information, Figure S2, and shows reasonable qualitative agreement between Langevin and non-Langevin simulations. From these results we can infer that the Langevin dynamics module produces timestamps that include dynamic state changes in a controlled and realistic manner.

4 Discussion

The new Langevin module within the PyBroMo software allows for generating more realistic smFRET data consistent with what one expects to observe from freely diffusing smFRET experiments of molecules with flexible conformational states, where a fixed FRET efficiency or dye-dye distance does not provide a reasonable approximation. The comparison between the Langevin and non-Langevin models here was not to show the superiority of the Langevin method over the non-Langevin method as the Langevin method is considered an improvement simply because it is more

realistic. Instead, the comparison was made to show the newly added Langevin model can be recovered from the data using typical data analysis methods at least as accurately as the original non-Langevin model and is this compatible with the PyBroMo software.

In the results presented above, the data from two example simulations using the Langevin and the non-Langevin methods were analyzed using some typical methods applied to experimental sm-FRET data. Example 1 used a simple model for a flexible molecule where the dye-dye distances evolve dynamically using a Langevin simulation method in a harmonic potential, to give a distribution of distances and FRET efficiencies in a physically justifiable way. Example 2 used the same Langevin simulation method to evolve dye-dye distances in a bistable potential to model a system that inter-coverts between two states. Both examples are compared with simulated data generated with non-Langevin methods for single and bistable states with other parameters set to match the Langevin simulations as closely as possible. This is done as a validation exercise to identify any unintended artifacts from the new module when compared with the base PyBroMo software using standard analysis methods including applying photon count thresholds, binning data over 1 ms, creating histograms, and fitting HMMs for Example 2.

Our results demonstrate both, agreement between the Langevin and non-Langevin results as well as reasonable accuracy in reproducing some of the major parameters of the underlying simulation. For instance, the histogram analyses reproduced the locations of efficiency peaks used as Langevin simulation parameters, in approximately equal proportions for the dye-dye distance distributions. Additionally, the HMM estimated similar transition matrices for the Langevin and non-Langevin timestamp data. Importantly, the estimated transition matrices were reasonably accurate to the ground truth transition matrix.

Qualitative differences were observed between the Langevin and non-Langevin timestamp data in the histogram analysis. The histograms of the Langevin timestamp data showed broader distributions of the efficiency states, in general. The comparatively narrow distributions of efficiencies from the non-Langevin timestamp data were due solely to the Brownian motion of the molecule through the PSF, but the underlying efficiency distributions are point masses. Both Langevin and

non-Langevin simulation methods contained the same Brownian motion and PSF parameters so any broadening of the efficiency distribution for the Langevin timestamp data can be attributed to the ensemble of dye-dye distances from the Langevin module.

It is of note that the conversion between efficiency and distance, as done in the histogram analysis, is generally non-linear. Qualitative observations, like relative peak heights, can change after conversion. This is most obvious in Figure 4, where the two FRET states have different efficiency peak heights but the peaks of distance histograms (and underlying distribution for the Langevin simulation) are the same height. The two efficiency models used in this paper have qualitative similarities but each model required its own conversion. FRET is most accurate near the R_0 value for the dye pair, with efficiency data becoming more distorted as it approaches zero or one. Accurate conversion of efficiency histogram states into distance is required to infer the underlying state information.

Beyond validation, the qualitative similarity in results implies the need for more sophisticated analysis methods. Despite the stark differences in the ground truth of dye-dye distances, it would be difficult to identify the Langevin results from the non-Langevin results. Some identifiers of the underlying ground truth are present, like the wider spread of apparent efficiencies, but that is only visible with a direct comparison and could be missed if viewed alone.

The conventional analysis methods we applied to the timestamp data used time bins to collect the individual detected photons into an aggregate signal. An aggregate signal is necessary to collect enough FRET signal to overcome the background noise. For the Langevin simulation method, the time bins contain photons with an underlying ensemble of dye-dye distances and efficiencies, but the ensemble becomes averaged over the time of each bin. This is especially true when the underlying dynamics are significantly faster than the bin size. Reducing the size of time bins may reduce the averaging of conformations but also increases the proportion of background noise relative to the smFRET signal. A balance between time bin length and background noise limits how short the time bins can be while containing significant photon counts.

Using the new Langevin module added to the existing PyBroMo software, researchers will have

the ability to repeatedly generate large amounts of data with a known ground truth of heterogeneous dye-dye distances. Different simulation parameters can easily be changed to generate timestamps and test assumptions based on experimental diffusing smFRET data of flexible molecules with heterogeneous states. New analysis methods beyond the standard time bin methods can then be developed and tested against the simulated data with a known ground truth to assess the effectiveness of such approaches with the ultimate goal of extracting more information from diffusing smFRET experiments of flexible molecules.

5 Conclusion

In this work, we have shown that the addition of a Langevin dynamics module to the base PyBroMo software is capable of generating freely diffusing smFRET timestamp data with more realistic heterogeneity of dye-dye distance dynamics and distribution. The implementation of the Langevin dynamics provides a flexible approach for defining the underlying dynamics of the molecule with full knowledge of the ground truth. Simulated data with known ground truth of realistic heterogeneous dye-dye distances will play an important role in developing new techniques for the analysis of freely diffusing smFRET data for flexible molecules.

Acknowledgement

This research is supported by the National Science Foundation under Awards 1940188, 1945465, 1934985, 1940124, and 1940179. This research is also supported by the Arkansas High Performance Computing Center which is funded through multiple National Science Foundation grants and the Arkansas Economic Development Commission.

Supporting Information Available

Additional figures and analysis of simulated and experimental data are presented in the Supporting Information.

References

(1) Koshland, D. E. Correlation of structure and function in enzyme action. Science **1963**, 142, 1533–1541.

(2) Eitan Lerner,; Cordes, T.; Ingargiola, A.; Alhadid, Y.; Chung, S.; Michalet, X.; Weiss, S. Toward dynamic structural biology: Two decades of single-molecule Förster resonance energy transfer. Science **2018**, 359.

(3) Perrin, J. La fluorescence. Annales de Physique. 1918; p 133159.

(4) Förster, T. Zwischenmolekulare Energiewanderung und Fluoreszenz. Annalen der Physik **1948**, 437, 5575.

(5) Stryer, L.; Haugland, R. P. Energy transfer: a spectroscopic ruler. Proceedings of the National Academy of Sciences of the United States of America **1967**, 58, 719.

(6) Haas, E. Intrinsically Disordered Protein Analysis; Springer, 2012; pp 467–498.

(7) Lerner, E.; Orevi, T.; Ben Ishay, E.; Amir, D.; Haas, E. Kinetics of fast changing intramolecular distance distributions obtained by combined analysis of FRET efficiency kinetics and time-resolved FRET equilibrium measurements. Biophysical journal **2014**, 106, 667676.

(8) Rahamim, G.; Chemerovski-Glikman, M.; Rahimipour, S.; Amir, D.; Haas, E. Resolution of Two Sub-Populations of Conformers and Their Individual Dynamics by Time Resolved Ensemble Level FRET Measurements. PLOS ONE **2015**, 10, 121.

(9) Miller, H.; Zhou, Z.; Shepherd, J.; Wollman, A. J.; Leake, M. C. Single-molecule techniques in biophysics: a review of the progress in methods and applications. Reports on Progress in Physics **2017**, 81, 024601.

(10) Sharonda J. LeBlanc, K. R. W., Prakash Kulkarni Probing the Interaction between Two Single Molecules: Fluorescence Resonance Energy Transfer between a Single Donor and a Single Acceptor. PNAS **1996**, 93, 6264 – 6268.

(11) Ha, T.; Ting, A. Y.; Liang, J.; Caldwell, W. B.; Deniz, A. A.; Chemla, D. S.; Schultz, P. G.; Weiss, S. Single-molecule fluorescence spectroscopy of enzyme conformational dynamics and cleavage mechanism. Proceedings of the National Academy of Sciences **1999**, 96, 893–898.

(12) Beausang, J. F.; Zurla, C.; Manzo, C.; Dunlap, D.; Finzi, L.; Nelson, P. C. DNA looping kinetics analyzed using diffusive hidden Markov model. Biophysical journal **2007**, 92, L64–L66.

(13) Zhuang, X.; Bartley, L. E.; Babcock, H. P.; Russell, R.; Ha, T.; Herschlag, D.; Chu, S. A single-molecule study of RNA catalysis and folding. Science **2000**, 288, 20482051.

(14) Nierth, A.; Kobitski, A. Y.; Nienhaus, G. U.; Jäschke, A. AnthraceneBODIPY Dyads as Fluorescent Sensors for Biocatalytic DielsAlder Reactions. Journal of the American Chemical Society **2010**, 132, 26462654, PMID: 20131767.

(15) Keller, B. G.; Kobitski, A.; Jäschke, A.; Nienhaus, G. U.; Noé, F. Complex RNA Folding Kinetics Revealed by Single-Molecule FRET and Hidden Markov Models. Journal of the American Chemical Society **2014**, 136, 45344543, PMID: 24568646.

(16) Choi, U. B.; Weninger, K. R.; Bowen, M. E. Immobilization of proteins for single-molecule fluorescence resonance energy transfer measurements of conformation and dynamics. Methods in molecular biology (Clifton, N.J.) **2012**, 896, 320.

(17) Schuler, B.; Eaton, W. A. Protein folding studied by single-molecule FRET. Current Opinion in Structural Biology **2008**, 18, 1626, Folding and Binding / Protein-nucleic acid interactions.

(18) Roy, R.; Hohng, S.; Ha, T. A practical guide to single-molecule FRET. Nature methods **2008**, 5, 507516.

(19) Rhoades, E.; Gussakovsky, E.; Haran, G. Watching proteins fold one molecule at a time. Proceedings of the National Academy of Sciences **2003**, 100, 3197–3202.

(20) Cohen, A. E.; Moerner, W. Controlling Brownian motion of single protein molecules and single fluorophores in aqueous buffer. Optics express **2008**, 16, 6941–6956.

(21) Selvin, P. R.; Ha, T. Single-molecule techniques; Cold Spring Harbor Laboratory Press, 2008.

(22) Chen, Y.; Shen, K.; Shan, S.-O.; Kou, S. C. Analyzing Single-Molecule Protein Transportation Experiments via Hierarchical Hidden Markov Models. Journal of the American Statistical Association **2016**, 111, 951–966.

(23) Okamoto, K.; Terazima, M. Distribution Analysis for Single Molecule FRET Measurement. The Journal of Physical Chemistry B **2008**, 112, 73087314, PMID: 18491936.

(24) McKinney, S. A.; Joo, C.; Ha, T. Analysis of Single-Molecule FRET Trajectories Using Hidden Markov Modeling. Biophysical Journal **2006**, 91, 19411951.

(25) Okamoto, K.; Sako, Y. Variational Bayes analysis of a photon-based hidden Markov model for single-molecule FRET trajectories. Biophysical journal **2012**, 103, 13151324.

(26) Pirchi, M.; Tsukanov, R.; Khamis, R.; Tomov, T. E.; Berger, Y.; Khara, D. C.; Volkov, H.; Haran, G.; Nir, E. Photon-by-photon hidden Markov model analysis for microsecond single-molecule FRET kinetics. The Journal of Physical Chemistry B **2016**, 120, 13065–13075.

(27) Sgouralis, I.; Madaan, S.; Djutanta, F.; Kha, R.; Hariadi, R. F.; Presse, S. A Bayesian Nonparametric Approach to Single Molecule Forster Resonance Energy Transfer. The Journal of Physical Chemistry B **2019**, 123, 675–688.

(28) Jeong, C.; Cho, W.-K.; Song, K.-M.; Cook, C.; Yoon, T.-Y.; Ban, C.; Fishel, R.; Lee, J.-B. MutS switches between two fundamentally distinct clamps during mismatch repair. Nature structural & molecular biology **2011**, 18, 379.

(29) Gopich, I. V.; Szabo, A. Single-Molecule FRET with Diffusion and Conformational Dynamics. The Journal of Physical Chemistry B **2007**, 111, 12925–12932, PMID: 17929964.

(30) Gopich, I. V.; Szabo, A. Decoding the Pattern of Photon Colors in Single-Molecule FRET. The Journal of Physical Chemistry B **2009**, 113, 1096510973, doi: 10.1021/jp903671p.

(31) Ingargiola, A.; Laurence, T.; Boutelle, R.; Weiss, S.; Michalet, X. Open Computational Tools for Freely Diffusing Single-Molecule Fluorescence Analysis. Biophysical Journal **2016**, 110, 634a.

(32) Gomes, G.-N. W.; Krzeminski, M.; Namini, A.; Martin, E. W.; Mittag, T.; Head-Gordon, T.; Forman-Kay, J. D.; Gradinaru, C. C. Conformational ensembles of an intrinsically disordered protein consistent with nmr, saxs, and single-molecule fret. Journal of the American Chemical Society **2020**, 142, 15697–15710.

(33) Kuzmenkina, E. V.; Heyes, C. D.; Ulrich Nienhaus, G. Single-molecule FRET Study of Denaturant Induced Unfolding of RNase H. Journal of Molecular Biology **2006**, 357, 313324.

(34) Mazal, H.; Haran, G. Single-molecule FRET methods to study the dynamics of proteins at work. Current Opinion in Biomedical Engineering **2019**, 12, 8–17, Molecular & Cellular Engineering: single molecule technology Neural Engineering: High Resolution Cell Imaging.

(35) Nasse, M. J.; Woehl, J. C. Realistic modeling of the illumination point spread function in confocal scanning optical microscopy. J. Opt. Soc. Am. A **2010**, 27, 295–302.

(36) Maruyama, G. Continuous Markov processes and stochastic equations. Rendiconti del Circolo Matematico di Palermo **1955**, 4, 48.

(37) Gopich, I. V.; Szabo, A. FRET efficiency distributions of multistate single molecules. The Journal of Physical Chemistry B **2010**, 114, 15221–15226.

(38) Berezhkovskii, A. M.; Szabo, A. Committors, first-passage times, fluxes, Markov states, milestones, and all that. The Journal of Chemical Physics **2019**, 150, 054106.

(39) Schuler, B. Single-molecule fluorescence spectroscopy of protein folding. ChemPhysChem **2005**, 6, 1206–1220.

(40) Eggeling, C.; Berger, S.; Brand, L.; Fries, J.; Schaffer, J.; Volkmer, A.; Seidel, C. Data registration and selective single-molecule analysis using multi-parameter fluorescence detection. Journal of Biotechnology **2001**, 86, 163–180, DNA-Sequencing at the Single Molecule Level.

(41) Ingargiola, A.; Lerner, E.; Chung, S.; Weiss, S.; Michalet, X. FRETBursts: an open source toolkit for analysis of freely-diffusing single-molecule FRET. PloS one **2016**, 11, e0160716.

(42) Dahan, M.; Deniz, A. A.; Ha, T.; Chemla, D. S.; Schultz, P. G.; Weiss, S. Ratiometric measurement and identification of single diffusing molecules. Chemical Physics **1999**, 247, 85–106.

(43) Eggeling, C.; Fries, J. R.; Brand, L.; Günther, R.; Seidel, C. A. M. Monitoring conformational dynamics of a single molecule by selective fluorescence spectroscopy. Proceedings of the National Academy of Sciences **1998**, 95, 1556–1561.

(44) Michalet, X.; Colyer, R.; Scalia, G.; Ingargiola, A.; Lin, R.; Millaud, J.; Weiss, S.; Siegmund, O. H.; Tremsin, A. S.; Vallerga, J. V., et al. Development of new photon-counting detectors for single-molecule fluorescence microscopy. Philosophical Transactions of the Royal Society B: Biological Sciences **2013**, 368, 20120035.

(45) Lee, N. K.; Kapanidis, A. N.; Wang, Y.; Michalet, X.; Mukhopadhyay, J.; Ebright, R. H.; Weiss, S. Accurate FRET Measurements within Single Diffusing Biomolecules Using Alternating-Laser Excitation. Biophysical Journal **2005**, 88, 2939–2953.

(46) Deniz, A. A.; Dahan, M.; Grunwell, J. R.; Ha, T.; Faulhaber, A. E.; Chemla, D. S.; Weiss, S.; Schultz, P. G. Single-pair fluorescence resonance energy transfer on freely diffusing molecules: Observation of Förster distance dependence and subpopulations. Proceedings of the National Academy of Sciences **1999**, 96, 3670–3675.

(47) Dempster, A. P.; Laird, N. M.; Rubin, D. B. Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society. Series B (Methodological) **1977**, 39, 1–38.

(48) Benaglia, T.; Chauveau, D.; Hunter, D. R.; Young, D. mixtools: An R Package for Analyzing Finite Mixture Models. Journal of Statistical Software **2009**, 32, 1–29.

(49) R Core Team, R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing: Vienna, Austria, 2020.

(50) O'Hagan, A.; Leonard, T. Bayes estimation subject to uncertainty about parameter constraints. Biometrika **1976**, 63, 201–203.

(51) Azzalini, A.; Capitanio, A. The skew-normal and related families; Cambridge University Press, 2014.

(52) Prates, M. O.; Cabral, C. R. B.; Lachos, V. H. mixsmsn: Fitting Finite Mixture of Scale Mixture of Skew-Normal Distributions. Journal of Statistical Software **2013**, 54, 1–20.

(53) Zucchini, W.; MacDonald, I. L.; Langrock, R. Hidden Markov models for time series: an introduction using R; CRC press, 2017.

(54) Altman, R. M. Mixed hidden Markov models: an extension of the hidden Markov model to the longitudinal data setting. Journal of the American Statistical Association **2007**, 102, 201–210.

Graphical TOC Entry