Simulating freely-diffusing single-molecule FRET data with consideration of protein conformational dynamics

James Losey,[1] Michael Jauch,[2] David S. Matteson,[3] and Mahmoud Moradi[1,*]

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

[2]*Center for Applied Mathematics, Cornell University, Ithaca, NY 14850, U.S.A.*

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

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, much is still unknown in systems that are highly flexible such as intrinsically disordered proteins because of the high degeneracy of distance states, particularly when freely diffusing smFRET experiments are used. Simulated smFRET data allows for the control of underlying process that generates the data to examine if analytic techniques can detect these underlying differences. We have extended the PyBroMo software that simulates the freely diffusing smFRET data to include a distribution of inter-dye distances generated using Langevin dynamics in order to model proteins with greater flexibility or disorder in structure. Standard analysis techniques for smFRET data compared highlighted the differences observed between data generated with the base software and data that included the distribution of inter-dye distance.

* moradi@uark.edu

INTRODUCTION

Förster resonance energy transfer (FRET) is the non-radiative transfer of energy between two chromophore dyes [1, 2]. The energy transferred between donor and acceptor dyes is dependent on the distance between the dyes, making this technique a "spectroscopic ruler" [3] for intra- and inter-molecular distances. Ensemble FRET experiments can suffer from bulk averaging of the conformation changes on the individual changes that occur during photon emission, though experimental design can overcome some of these challenges [4–6] . The advent of single molecule spectroscopic techniques transformed biophysics as a source of statistical and dynamic data on molecular structure as well as function [7]. Single molecule FRET (smFRET) experiments have become a popular source of spatiotemporal information on the conformational landscape of a molecule without ensemble averaging by taking advantage of FRET and the ability to label specific regions of a molecule with fluorescent dyes [8, 9]. These techniques have been applied in studies of systems like DNA [10], RNA [11–13], protein folding [14, 15].

Two broad divisions of smFRET experiments are surface immobilized and freely diffusing molecules. Surface immobilized experiments fix a molecule of interest that has been labeled with fluorescent dyes to a substrate, expose it to laser light to excite the donor, and collect the photon timestamp data. This experimental procedure uses long exposure times to collect data on slower dynamics, greater than 1 ms [16]. 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.

Freely diffusing smFRET methods record photon emissions from labeled molecules as they diffuse through a solution with a confocal laser focused inside the solution. Photon detectors tuned for the wavelengths of the donor and acceptor dyes record time series data with a timestamp and channel label for each photon detected. The diffusion rates and concentrations of the molecules are determined so that simultaneous excitation of multiple molecules was very rare in a particular time bin. Freely diffusing experiments capture dynamics occurring on faster scales [3] and avoid the difficulties of immobilization [17–19]. The photon signal occurs in bursts as molecules diffuse into and out of the focal beam of the excitation laser. Analysis techniques of freely diffusing smFRET experiments are in active development, making the need for realistic simulated data important. This generation of simulated data for freely diffusing smFRET experiments

While sophisticated statistical methodology is essential to the analysis of smFRET experiments,

the literature on this topic has focused on surface-immobilized smFRET [20], compared to the freely diffusing smFRET technique, a much easier experimental technique with no need for surface immobilization [21]. These techniques include simpler thresholds and histograms, as well as more complex Gaussian fitting of sub-populations [22], hidden Markov models (HMM) [10, 23–25], and non-Bayesian approaches [26] have been developed. To further advance freely diffusing smFRET analysis, we require the ability to accurately model and simulate the underlying molecular processes in a systematic, controlled, and repeatable manner.

PyBroMo[27], an open source smFRET timestamp simulation software suite, uses a Brownian motion simulation, a numerical point spread function (PSF) to model the laser, and Poisson background noise to model smFRET timestamps for multiple populations of freely diffusing molecules. These features provided a framework to generate timestamps with static or more complex state switching. As an open source project, researchers can also extend the code to include other features not currently included in the project. For instance, PyBroMo uses a fixed efficiency for each population throughout the duration of the simulation. A fixed efficiency assumes that the distribution of dye-dye distances in a freely diffusing molecule are negligible compared to the other parts of the simulation. This would not be the case for molecules with less rigid structures and greater fluctuations, like disordered proteins [28, 29]. To more accurately model the conformational state changes of a less structured molecule, an overdamped Langevin dynamics simulation was added to PyBroMo's existing software to model the internal conformational dynamics and heterogeneity of different states of the modeled protein. This addition provides a more realistic smFRET data simulation model, particularly for flexible proteins or those associated with intrinsic disorder.

The remaining sections of this paper will provide a more detailed description of PyBroMo, and the overdamped Langevin dynamics that generated the dye-dye distance distribution, as well as the parameters used in generating simulated data for the analysis section Next, a standard analyses for smFRET data using thresholds and Gaussian fitting was done on the timestamp data using fixed efficiency and Langevin efficiency. Conclusions are drawn based on the comparison of the analysis for the two simulated data sets. Real experimental smFRET data are also used for a closer comparison between the simulation and experimental data.

METHODS

PyBroMo

PyBroMo [27] was developed by Ingargiol et al. to simulate photon emission from fluorescent dye pairs attached to molecules freely diffusing in three dimensions and generate timestamps from those emissions, similar to experimental FRET data. This software was designed to handle multiple populations of particles with their own diffusion coefficients and FRET efficiencies, as well as generate background photons with separate rates for the donor and acceptor channels.

The first part of the simulation was defining the basic elements of the simulation. The particles are defined by a population number and diffusion coefficient, D_B. Next, the simulation box was defined by providing box length dimensions, L_x, L_y, and L_z as well as how to handle particle interactions with the boundary. Reflection and periodic boundary conditions are available.

A point spread function (PSF) was chosen to model the laser focal beam inside the simulation box. The PSF defines the emission probability of a particle at any position within the simulation box. A Gaussian PSF is available where the emission probability in all dimension is defined by the Gaussian/normal distribution

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

were μ_x are the coordinates for the center of the function and σ_x was the standard deviation, σ_x. This function can be extended to other the cartesian coordinates y and z. PyBroMo is also capable of importing custom PSF functions from tools like PSFLab [30] 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.

The simulation inputs were then passed to the Brownian motion simulation module along with a timestep (δt) and a maximum time to advance the particles through the simulation box. The Brownian motion was a stochastic process where the position in each dimension was based on the current position plus a random number drawn from a normal distribution.

$$x(t+\delta t) = x(t) + N(0, 2D_B\delta t) \tag{2}$$

where N is a random number drawn from a normal distribution centered at 0 with a variance of $\sigma^2 = 2D_B\delta t$. The Brownian motion simulation then repeatedly advanced each particle's position in three dimension by the δt until the maximum time was reached. At each time step, the PSF

calculated the the normalized emission probability for every particles position in a vector, $\mathscr{P}$. Particles in regions of high emission probability, often near the center of the PSF, emit more photons.

The timestamp generation module generated photon emissions through a discrete random Poisson process where the number of emission events, κ, followed the distribution

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

where λ was the expectation interval for emissions. The values needed to calculate the λ values for every time step were a maximum total emission rate, ε_T, and 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 were calculated

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

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

where ε_A and ε_D are the emission rates for the acceptor and donor, respectively. The efficiency, E, was constant for all timesteps. Separate expectation intervals for the acceptor, λ_{Acc} and donor, λ_{Don} were 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 were also determined for the acceptor and donor detector channels by randomly drawn numbers from a Poisson distribution with expectation interval, $\lambda_{BGAcc}, \lambda_{BGDon}$, supplied as a simulation parameter.

Finally, the timestamps were merged and sorted into a single vector for output. A vector of labels was also generated to identify if the timestamp was from the acceptor or donor channel. Other values of interest that may be included are the particle ID that generated the photon emission or the position of the particle in the PSF.

Overdamped Langevin Dynamics

To extend the PyBroMo software, an overdamped Langevin dynamics module was added to simulate the dye-dye distance as diffusion in one dimension in a harmonic potential. The Langevin

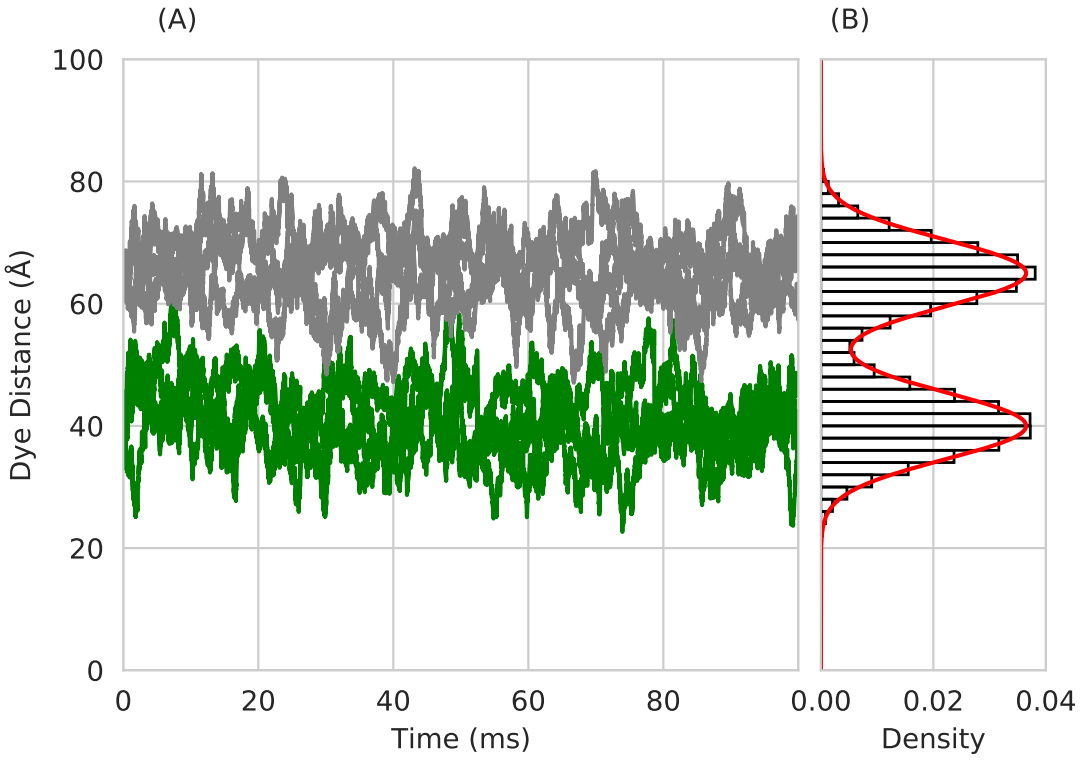

FIG. 1. (A) A portion of a trajectory of Langevin dynamics for the dye-dye distance of 4 particles in a high state (colored green) with a harmonic potential centered at 40 Å and 4 particles in a medium state (colored gray) with a harmonic potential centered at 65 Å. (B) A histogram of the dye-dye distances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

trajectories were calculated according to the Euler-Muryama method [31]. At each time step, the dye-dye distance was updated by calculating the contributions from the potential energy function and the stochastic random contribution:

$$r(t+\delta t) = r(t) - \beta D_L \frac{dV(r)}{dt}\delta t + N(0, 2D_L\delta t) \tag{8}$$

$$V(r) = \frac{1}{2}k(r-r_0)^2 \tag{9}$$

where D_L was the diffusion coefficient, and $\beta = \frac{1}{k_B T}$ and k_B is the Boltzmann constant. The diffusion coefficient for the dye-dye distance, D_L, is unique from the Brownian motion diffusion coefficient. A short trajectory of Langevin dye-dye distance is shown in figure 1.

With the Langevin distance trajectories, a dynamic efficiency for each timestep was calculated

using an efficiency model as a function of dye-dye distance.loped for less structured proteins [29],

$$E = \frac{1}{1 + 0.975\left(\frac{r}{R0}\right)^{2.65}} \tag{10}$$

where $R0$ is the distance that results in a 50% FRET efficiency.

Equations 4 and 5 then generate vectors for the acceptor and donor emission rate ε_A and ε_D and equations 6 and 7 calculate the expectation intervals λ_{Acc} and λ_{Don}. As with the base PyBroMo, random numbers were drawn from a Poisson distribution defined in equation 3 for each timestep. The background timestamp generation was unimpacted by the langevin dynamics, contributing Poisson distributed background timesteps as before. Finally, the timestamps from acceptor, donor, and background were merged.

Simulation Details

To compare the differences of timestamps with a static efficiency with timestamps from an efficiency distribution, 3 10s simulations were run with all other parameters held constant. 100 particles were contained in a simulation box with lengths $L_x = 8\mu m, L_y = 8\mu m, L_z = 12$ The Brownian diffusion coefficient D_B, was set to 30 $\mu m^2/s$ for all particles. A Gaussian PSF was used that was centered in the simulation box with a $\sigma_x = \sigma_y = 0.3\mu m$, and $\sigma_z = 0.5\mu m$. Three independent simulations were run for 10s each with a time step 50 ns. For timestamp generation, the maximum emission rate of 200,000 counts per second (CPS) was used in both simulations, as well as a background rate of 1200 CPS for the acceptor channel and 1800 CPS for the donor channel.

For the Langevin dynamics,the thermodynamic coefficient β was 1.33873220189633 $(\text{kcal/mol})^{-1}$ and the Langevin diffusion coefficient, D_L, was 12937.9931633482 Å^2/s. The harmonic coefficient, k, was 0.025 with the center of the harmonic potential was at 40 Å for 50 of the particles, and at 65 Å for the remaining 50 particles. Rel. 10 was used to convert the distances to efficiencies. In the static efficiency simulations, 50 particles had an efficiency of 0.76 while the other 50 had an efficiency of 0.41. These efficiency values corresponded to equation 10 applied to the harmonic centers from the Langevin dynamics, 40 Å and 65 Å respectively. An $R0$ of 56 Å was used in this efficiency conversion.

The analytic distribution for the Langevin simulation distances was defined by

$$P(r) = \frac{1}{2\sqrt{\frac{2\pi}{\beta k}}}\left(e^{-\beta V_1(r)} + e^{-\beta V_2(r)}\right) \tag{11}$$

where N is a normalization factor and V_1 and V_2 were the harmonic potentials used in the Langevin simulation. Converting the distance probability distribution to an efficiency probability distribution was done through the following transformation

$$P(E) = P(r(E))\frac{dr}{dE} \tag{12}$$

$$r(E) = R0\left(\frac{\frac{1}{E}-1}{0.975}\right)^{1/2.65} \tag{13}$$

$$\frac{dr}{dE} = \frac{R0}{2.65(E-1)E}\left(\frac{\frac{1}{E}-1}{0.975}\right)^{1/2.65} \tag{14}$$

where equation 13 was found by rearranging equation 10 to isolate r in terms of E and $\frac{dr}{dE}$ is the derivative of equation 13 with respect to E.

Data analysis methods

Techniques for simulating freely diffusing smFRET experiments are valuable, in large part, because they allow researchers to evaluate statistical methods using realistic data with known ground truth. With this in mind, we present a simple analysis of two simulated freely diffusing smFRET experiments. The first experiment was simulated with the base PyBroMo software described in Section , while the second experiment was simulated with the Langevin dynamics module discussed in Section . We give the details of our analysis here, and present the results in the Results section. We are most interested in ways in which the addition of Langevin dynamics changes our results.

Data analyses of freely diffusing smFRET experiments typically begin by binning and thresholding the raw photon time stamp data [20]. In our analyses, we use a 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. The hope is that a combined photon count of this magnitude indicates that a molecule is indeed 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. 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.

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:

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

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. 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 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 in this article were run with $K = 2$, and we take this as given. We compute the maximum likelihood estimates of the unknown parameters via an expectation-maximization algorithm as implemented in the `mixtools` package [32].

RESULTS AND DISCUSSION

We now summarize and discuss the results of simulations and the data analysis described in Section . We will make frequent references to Figures 2 - 4.

Figure 2 includes trace plots of the combined acceptor and donor counts for both the non-Langevin and Langevin simulations. The counts were computed using one millisecond bins. The simulated smFRET experiments were both 10 seconds long, but we plot just one second of each for visual clarity. Qualitatively, the plots look similar to each other and also to the corresponding

trace plot from a real smFRET experiment in Figure 4.

Figure 3 compares the non-Langevin and Langevin simulations in terms of apparent efficiencies and the corresponding dye-dye distances. Plot A, based on the non-Langevin simulation, shows the estimated two-component Gaussian mixture density (in solid black) on top of a 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 .42, a standard deviation of .07, and a mixture weight of .62. The high efficiency component has a mean of .70, a standard deviation of .05, and a mixture weight of .38. The vertical red arrows are placed at the true efficiency values used in the simulation. Plot B shows the corresponding histogram, densities, and arrows after a transformation to the distance space using Equation 10. Plots C and D in the right half of Figure 3 are analogues of Plots A and 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 Equation 11. The theoretical distribution in the efficiency space is obtained through the change of variables described in Equation 12. In Plot C, the low efficiency component has a mean of .41, a standard deviation of .07, and a mixture weight of .48, while the high efficiency component has a mean of .68, a standard deviation of .09, and a mixture weight of .52.

For reference, we also include a simple analysis of real freely diffusion smFRET data for an intrinsically disordered protein. The results appear in Figure 4. Plot A shows a trace of the combined acceptor and donor counts from a one second segment of the 10 second experiment. Plot B shows the estimated two-component Gaussian mixture density on top of a histogram of the apparent efficiencies. The low efficiency component has a mean of 0.14, a standard deviation of .05, and a mixture weight of .48, while the high efficiency component has a mean of .41, a standard deviation of .18, and a mixture weight of .52. Plot C shows the corresponding histogram and density after a transformation to the distance space.

CONCLUSION

In this work, we have added a module to PyBroMo software to provide a more accurate simulation tool to generate simulated freely diffusing smFRET data for flexible or disordered proteins with a high degree of flexibility. We showed that the simulated data with consideration of internal

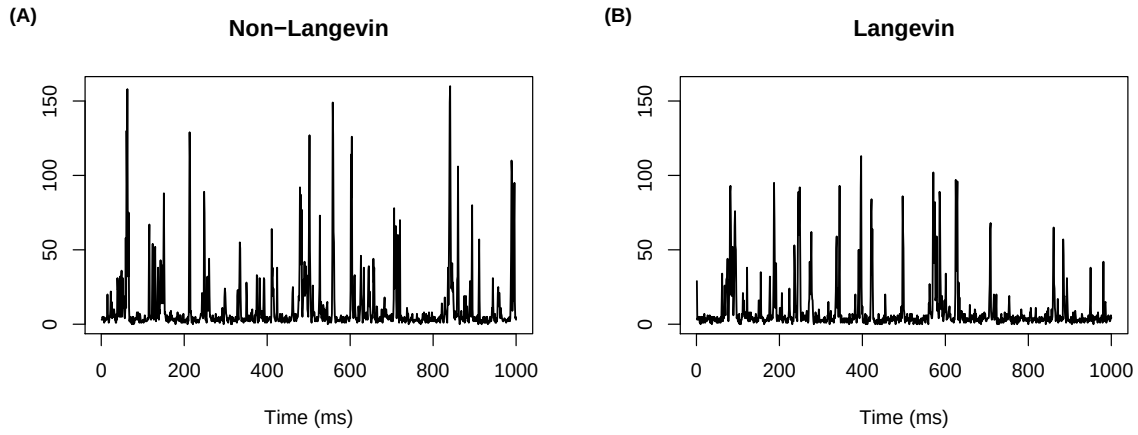

FIG. 2. Trace plots of the combined acceptor and donor counts from one second segments of the (A) non-Langevin and (B) Langevin simulations.

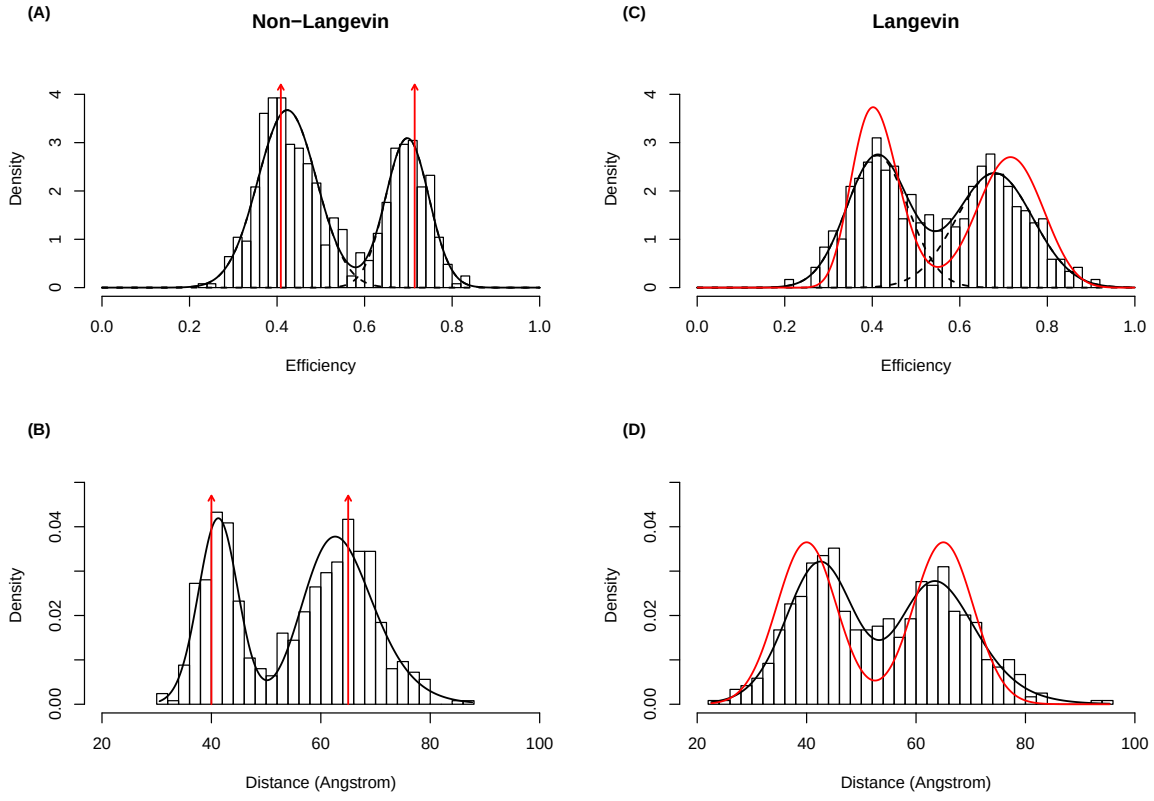

FIG. 3. (A) The estimated Gaussian mixture density (solid black line) from the non-Langevin simulation on top of a histogram of the apparent efficiencies along with the two true efficiencies (red vertical arrows). (B) The corresponding plot in the distance space. (C) The analogous efficiency plot for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

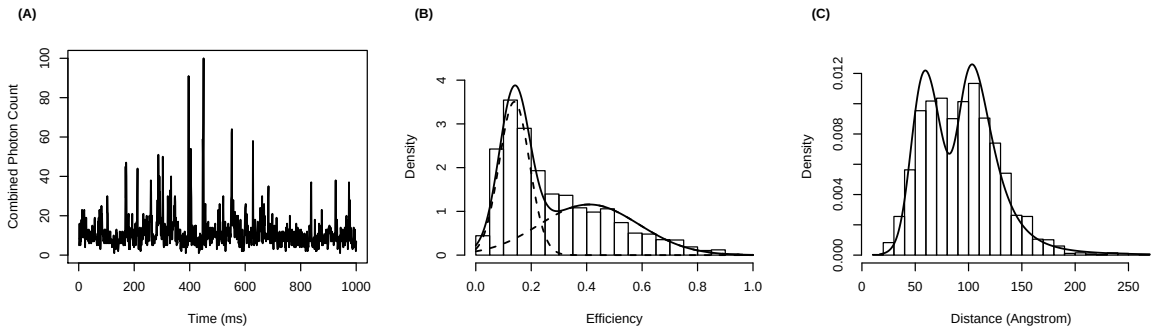

FIG. 4. (A) Trace plot of the combined acceptor and donor counts from a one second segment of the cAlb experiment. (B) The estimated Gaussian mixture density (solid black line) on top of a histogram of the apparent efficiencies. (C) The corresponding plot in the distance space.

protein dynamics more closely resembles experimental data and the Gaussian mixture model fits better with the Langevin-enabled simulation data than the conventional one. The relative value of state probabilities can be determined using standard Gaussian mixture models. The original single-point efficiencies provided within PyBroMo cannot be recovered since broad distributions are observed instead both in the efficiency and in the distance spaces. On the other hand, when the simulated data takes use of an overdamped Langevin dynamics, the estimated distributions and the theoretical distributions are quite close. In brief, the Langevin-based simulated data resembles the real experimental data better than the single-point efficiency model.

Acknowledgement

This research is supported by the National Science Foundation under Awards 1940188, 1945465, 1934985, and 1940124. 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.

[1] Jean Perrin. La fluorescence. In *Annales de Physique*, volume 9, page 133159, 1918.

[2] Th. Förster. Zwischenmolekulare energiewanderung und fluoreszenz. *Annalen der Physik*, 437(1-2):5575, 1948.

[3] Eitan Lerner, Thorben Cordes, Antonino Ingargiola, Yazan Alhadid, SangYoon Chung, Xavier Michalet, and Shimon Weiss. Toward dynamic structural biology: Two decades of single-molecule förster resonance energy transfer. *Science*, 359(6373), 2018.

[4] Elisha Haas. Ensemble fret methods in studies of intrinsically disordered proteins. In *Intrinsically Disordered Protein Analysis*, pages 467–498. Springer, 2012.

[5] E Lerner, T Orevi, E Ben Ishay, D Amir, and E Haas. Kinetics of fast changing intramolecular distance distributions obtained by combined analysis of fret efficiency kinetics and time-resolved fret equilibrium measurements. *Biophysical journal*, 106(3):667676, February 2014.

[6] Gil Rahamim, Marina Chemerovski-Glikman, Shai Rahimipour, Dan Amir, and Elisha Haas. Resolution of two sub-populations of conformers and their individual dynamics by time resolved ensemble level fret measurements. *PLOS ONE*, 10(12):121, 12 2015.

[7] Helen Miller, Zhaokun Zhou, Jack Shepherd, Adam JM Wollman, and Mark C Leake. Single-molecule techniques in biophysics: a review of the progress in methods and applications. *Reports on Progress in Physics*, 81(2):024601, 2017.

[8] Keith R. Weninger Sharonda J. LeBlanc, Prakash Kulkarni. Probing the interaction between two single molecules: Fluorescence resonance energy transfer between a single donor and a single acceptor. *PNAS*, 93(13):6264 – 6268, 1996.

[9] Taekjip Ha, Alice Y Ting, Joy Liang, W Brett Caldwell, Ashok A Deniz, Daniel S Chemla, Peter G Schultz, and Shimon Weiss. Single-molecule fluorescence spectroscopy of enzyme conformational dynamics and cleavage mechanism. *Proceedings of the National Academy of Sciences*, 96(3):893–898, 1999.

[10] John F Beausang, Chiara Zurla, Carlo Manzo, David Dunlap, Laura Finzi, and Philip C Nelson. Dna looping kinetics analyzed using diffusive hidden markov model. *Biophysical journal*, 92(8):L64–L66, 2007.

[11] Xiaowei Zhuang, Laura E Bartley, Hazen P Babcock, Rick Russell, Taekjip Ha, Daniel Herschlag, and Steven Chu. A single-molecule study of rna catalysis and folding. *Science*, 288(5473):20482051, 2000.

[12] Alexander Nierth, Andrei Yu. Kobitski, G. Ulrich Nienhaus, and Andres Jäschke. Anthracenebodipy dyads as fluorescent sensors for biocatalytic dielsalder reactions. *Journal of the American Chemical Society*, 132(8):26462654, 2010. PMID: 20131767.

[13] Bettina G. Keller, Andrei Kobitski, Andres Jäschke, G. Ulrich Nienhaus, and Frank Noé. Complex rna folding kinetics revealed by single-molecule fret and hidden markov models. *Journal of the American Chemical Society*, 136(12):45344543, 2014. PMID: 24568646.

[14] Ucheor B Choi, Keith R Weninger, and Mark E Bowen. Immobilization of proteins for single-molecule fluorescence resonance energy transfer measurements of conformation and dynamics. *Methods in molecular biology (Clifton, N.J.)*, 896:320, 2012.

[15] Benjamin Schuler and William A Eaton. Protein folding studied by single-molecule fret. *Current Opinion in Structural Biology*, 18(1):1626, 2008. Folding and Binding / Protein-nucleic acid interactions.

[16] Rahul Roy, Sungchul Hohng, and Taekjip Ha. A practical guide to single-molecule fret. *Nature methods*, 5(6):507516, June 2008.

[17] Elizabeth Rhoades, Eugene Gussakovsky, and Gilad Haran. Watching proteins fold one molecule at a time. *Proceedings of the National Academy of Sciences*, 100(6):3197–3202, 2003.

[18] Adam E Cohen and WE Moerner. Controlling brownian motion of single protein molecules and single fluorophores in aqueous buffer. *Optics express*, 16(10):6941–6956, 2008.

[19] Paul R Selvin and Taekjip Ha. *Single-molecule techniques*. Cold Spring Harbor Laboratory Press, 2008.

[20] Yang Chen, Kuang Shen, Shu-Ou Shan, and S. C. Kou. Analyzing Single-Molecule Protein Transportation Experiments via Hierarchical Hidden Markov Models. *Journal of the American Statistical Association*, 111(515):951–966, 2016.

[21] Cherlhyun Jeong, Won-Ki Cho, Kyung-Mi Song, Christopher Cook, Tae-Young Yoon, Changill Ban, Richard Fishel, and Jong-Bong Lee. Muts switches between two fundamentally distinct clamps during mismatch repair. *Nature structural & molecular biology*, 18(3):379, 2011.

[22] Kenji Okamoto and Masahide Terazima. Distribution analysis for single molecule fret measurement. *The Journal of Physical Chemistry B*, 112(24):73087314, 2008. PMID: 18491936.

[23] Sean A. McKinney, Chirlmin Joo, and Taekjip Ha. Analysis of single-molecule fret trajectories using hidden markov modeling. *Biophysical Journal*, 91(5):19411951, 2006.

[24] Kenji Okamoto and Yasushi Sako. Variational bayes analysis of a photon-based hidden markov model for single-molecule fret trajectories. *Biophysical journal*, 103(6):13151324, 2012.

[25] 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 mi-

crosecond single-molecule fret kinetics. *The Journal of Physical Chemistry B*, 120(51):13065–13075, 2016.

[26] Ioannis Sgouralis, Shreya Madaan, Franky Djutanta, Rachael Kha, Rizal F. Hariadi, and Steve Presse. A Bayesian Nonparametric Approach to Single Molecule Forster Resonance Energy Transfer. *The Journal of Physical Chemistry B*, 123(3):675–688, 2019.

[27] Antonino Ingargiola, Ted Laurence, Robert Boutelle, Shimon Weiss, and Xavier Michalet. Open computational tools for freely diffusing single-molecule fluorescence analysis. *Biophysical Journal*, 110(3):634a, 2016.

[28] Gregory-Neal W Gomes, Mickaël Krzeminski, Ashley Namini, Erik W Martin, Tanja Mittag, Teresa Head-Gordon, Julie D Forman-Kay, and Claudiu C Gradinaru. Conformational ensembles of an intrinsically disordered protein consistent with nmr, saxs, and single-molecule fret. *Journal of the American Chemical Society*, 142(37):15697–15710, 2020.

[29] Elza V. Kuzmenkina, Colin D. Heyes, and G. Ulrich Nienhaus. Single-molecule fret study of denaturant induced unfolding of rnase h. *Journal of Molecular Biology*, 357(1):313324, 2006.

[30] Michael J. Nasse and Jörg C. Woehl. Realistic modeling of the illumination point spread function in confocal scanning optical microscopy. *J. Opt. Soc. Am. A*, 27(2):295–302, Feb 2010.

[31] Gisiro Maruyama. Continuous markov processes and stochastic equations. *Rendiconti del Circolo Matematico di Palermo*, 4(1):48, 1955.

[32] Tatiana Benaglia, Didier Chauveau, David R. Hunter, and Derek Young. mixtools: An R package for analyzing finite mixture models. *Journal of Statistical Software*, 32(6):1–29, 2009.