Identifying multiple kinetic populations of DNA binding proteins in live cells using single-molecule fluorescence imaging

Han N. Ho, Daniel Zalami, Jürgen Köhler, Antoine M. van Oijen* and Harshad Ghodke*

* *Co-corresponding authors*

ABSTRACT

Understanding how multi-protein complexes function in cells requires detailed quantitative understanding of their association and dissociation kinetics. Analysis of the heterogeneity of binding lifetimes enables interrogation of the various intermediate states formed during the reaction. Single-molecule fluorescence imaging permits the measurement of reaction kinetics inside living organisms with minimal perturbation. However, poor photo-physical properties of fluorescent probes limit the dynamic range and accuracy of measurements of off rates in live cells. Interval imaging coupled to single-molecule fluorescence imaging can partially overcome the effect of photo-bleaching, however, limitations of this technique remain uncharacterized. Here, we present a structured analysis of which timescales are most accessible using the interval imaging approach and explore uncertainties in determining kinetic sub-populations. We demonstrate the effect of shot noise on the precision of the measurements, as well as the resolution and dynamic range limits that are inherent to the method. Our work provides a convenient implementation to determine theoretical errors from measurements and to support interpretation of experimental data.

INTRODUCTION

Understanding fundamental processes of life requires characterization of the kinetics of interactions between biological molecules. At single-molecule levels, these systems often exhibit kinetic heterogeneity that is inherent to the presence of multiple intermediate states (1-16). Advances in single-molecule imaging have enabled the detection and characterization of heterogeneous sub-populations in reactions conducted *in vitro* as well as, *in vivo*. Ultimately, these investigations enable the construction of detailed molecular mechanisms to explain how various biomolecular interactions proceed.

Compared to *in vitro* studies, live-cell investigations offer the key advantage of studying biochemical reactions at physiological conditions that can be difficult to reconstitute. Single-molecule live-cell imaging commonly relies on fluorescent proteins that are genetically fused to the protein of interest (Fig. 1A) (17-20). Tracking the fluorescence signal of thousands of molecules, one molecule at a time, enables the building of physical models, from which physical parameters such as diffusion constants and detachment rates from DNA can be determined. Where detachment rates are concerned, the trajectory lengths of thousands of molecules are aligned to obtain a cumulative residence time distribution (CRTD). At the single-molecule level, the dissociation of a protein from its substrate is a stochastic process. This phenomenon can be adequately described as a two-state kinetic model with the interconversion of populations being modelled as a Poisson process. The resulting CRTD can be fit with exponential functions to obtain decay rates. In the case of a fluorescently tagged protein where loss of fluorescence is attributable to either dissociation, or photobleaching of the chromophore, the decay rate represents a combination of dissociation rates and photobleaching rate (Fig. 1B-C) (2).

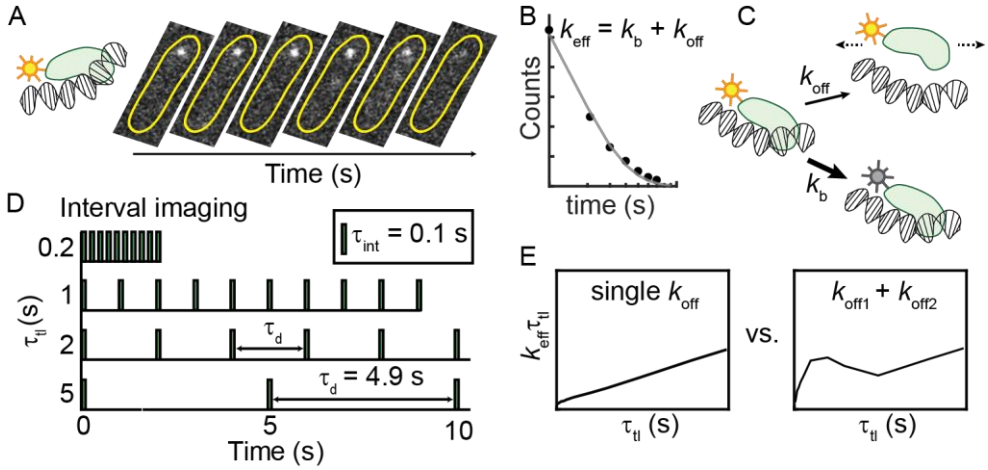

FIGURE 1 Experimental approach for characterizing kinetic heterogeneity of protein binding in live cells using single-molecule fluorescence imaging. (A) The protein of interest is tagged with a fluorescent protein. When the protein binds to DNA substrate, its fluorescence signal appears as a diffraction-limited focus that can be tracked in real time. Subsequent dissociation results in the disappearance of the focus and a redistribution of fluorescence signal throughout the cell. Yellow outlines illustrate the bacterial cell membrane. (B) The loss of fluorescence is attributable to either dissociation, or photobleaching of the chromophore. (C) Cumulative residence time distribution (CRTD) constructed from binding durations of thousands of molecules. Fitting the exponential function (Eq. 1) to CRTD yields an effective rate k_{eff}, which is the sum of off rate (k_{off}) of the protein of interest and photobleaching rate (k_b) of the fluorescent probe. (D) To deconvolute k_b and k_{off}, excitation and integration durations (τ_{int}) can be spaced with various dark intervals (τ_d). (E) Through exponential analyses, CRTDs obtained at various intervals result in $k_{eff}\tau_{tl}$ plots which are indicators of kinetic heterogeneity. A single kinetic population yields a straight line whereas deviations from linear fits indicate the presence of a second kinetic sub-population. For a single kinetic population, the slope is the off rate and y-intercept is proportional to the photobleaching rate.

Photobleaching, a result of fluorescent proteins being damaged upon exposure to excitation sources, leads to the loss of fluorescence signal (21). Under excitation conditions that guarantee good signal-to-background ratios, fluorescent proteins can only stay 'on' for a few frames during acquisitions, a problem referred as 'photon budget' (22). Thus, when photobleaching occurs faster than the dissociation process, lifetime measurements are limited by the photobleaching rate. To overcome this problem and extend the observation time, the observation time window can be expanded by temporally spacing the photon budget using stroboscopic imaging (23). In this method, a dark interval (τ_d) is inserted between integration time (τ_{int}), effectively scaling the observation time with a factor of τ_{tl}/τ_{int} ($\tau_{tl} = \tau_{int} + \tau_d$). Instead

of using one dark interval, Gebhardt and co-workers (2013) developed an interval imaging approach in which fluorescence acquisitions are collected at a series of intervals (Fig. 1D) (24). Briefly, the approach works as follows: First, several movies (each with a unique dark interval) are collected while keeping the photon budget constant (in practice this is achieved by keeping the number of frames constant across all the movies). In cases where the copy number of the tagged protein is high and single molecule imaging conditions may be difficult to attain, the cellular fluorescence is first photobleached such that only single molecule fluorescence is observable. Subsequently, using particle tracking algorithms that enable measurements of lifetimes of bound molecules within a specified localization radius, a cumulative residence time distribution can be compiled. Fitting the CRTDs to effective rates (k_{eff}), one can obtain the so-called $k_{eff}\tau_{tl}$ plot which is linear for mono-exponential distributions (Fig. 1E). This plot represents the total number of events that correspond to focus loss either due to protein dissociation or photobleaching of the tag. In this case, since the photobleaching rate is maintained constant across all conditions, it can be read off from the intercept on the Y-axis. A population of molecules dissociating with a finite and measurable off rate manifests as a straight line, where the slope reports on the off rate of the dissociation kinetics. A mixed population manifests as a deviation from the straight fit (Fig. 1E) (24). Fitting the experimental data to a model of mixed populations can then be used to extract the relative amplitudes and rates of the various populations. This power to deconvolute the photobleaching rate from multiple off rates has been successfully harnessed to dissect the kinetic heterogeneity of various DNA binding proteins including transcription factors and DNA replication and repair proteins in live cells (24-27).

However, limitations arising from the practical implementation of this elegant method remain uncharacterized. In particular, we address the following questions: 1) What is the minimum

number of observations needed to determine the time constant with specified confidence? 2) For a given experimental setup, what is the dynamic range in time constants that can be detected? 3) How many populations can be resolved and 4) What limits the ability to reliably resolve multiple populations? This study serves to provide a practical guide to realize the power as well as limitations of practical implementations of the interval imaging approach to measure intracellular binding kinetics of fluorescently tagged proteins.

METHODS

Rationale and model

Consider a system containing '*A*' number of fluorescently tagged DNA-bound proteins, wherein the proteins dissociate from DNA with a single off rate (k_{off}). Upon exposure to excitation photon sources, the fluorescent proteins exhibit photobleaching with a rate k_{b}, resulting in the loss of fluorescence signal. Additionally, dissociation contributes to the loss of fluorescent foci as protein molecules move out of the localization radius. Since dissociation and photobleaching are independent, and both are Poisson processes, the loss of observations as a function of time *t* can be described as:

$$f_1(t) = A\exp(-(k_{\mathrm{b}} + k_{\mathrm{off}})t) \tag{1}$$

Observation times of genetically expressible fluorescent proteins are severely limited to the duration of a few acquisition frames due to photobleaching, limiting measurements of long-lived binding events (28). To extend observation times, the frame rate can be reduced by inserting a dark interval (τ_{d}) after a short integration time (τ_{int}). Scaling the photobleaching rate appropriately, Eq. 1 then becomes:

$$f_2(t) = A\exp(-(k_\mathrm{b}\tau_\mathrm{int}/\tau_\mathrm{tl} + k_\mathrm{off})t) \tag{2}$$

where the time-lapse time τ_tl is the sum of τ_int and τ_d. The sum of two decay rates k_b and k_off can be approximated with an effective decay rate (k_eff):

$$k_\mathrm{eff} = k_\mathrm{b}\tau_\mathrm{int}/\tau_\mathrm{tl} + k_\mathrm{off} \tag{3}$$

Rearrangement of Eq. 3 yields:

$$k_\mathrm{eff}\tau_\mathrm{tl} = k_\mathrm{b}\tau_\mathrm{int} + k_\mathrm{off}\,\tau_\mathrm{tl} \tag{4}$$

As $k_\mathrm{b}\tau_\mathrm{int}$ is maintained constant at a certain imaging condition, $k_\mathrm{eff}\tau_\mathrm{tl}$ increases linearly with τ_tl, with the coefficient (slope) k_off.

In system with two sub-populations each dissociating at different rates k_off1 and k_off2, Eq. 2 then becomes:

$$\begin{aligned} f_3(t) = A\big(&B\exp(-(k_\mathrm{b}\tau_\mathrm{int}/\tau_\mathrm{tl} + k_\mathrm{off1})t) \\ &+ (1-B)\exp(-(k_\mathrm{b}\tau_\mathrm{int}/\tau_\mathrm{tl} + k_\mathrm{off2})t)\big) \end{aligned} \tag{5}$$

where B ($0 < B < 1$) and ($1 - B$) are the amplitudes of k_off1 and k_off2 sub-populations respectively.

Similarly, a system with three kinetic sub-populations can be described by:

$$\begin{aligned} f_4(t) = A\big(&B_1\exp(-(k_\mathrm{b}\tau_\mathrm{int}/\tau_\mathrm{tl} + k_\mathrm{off1})t) \\ &+ B_2\exp(-(k_\mathrm{b}\tau_\mathrm{int}/\tau_\mathrm{tl} + k_\mathrm{off2})t) \\ &+ (1-B_1-B_2)\exp(-(k_\mathrm{b}\tau_\mathrm{int}/\tau_\mathrm{tl} + k_\mathrm{off3})t)\big) \end{aligned} \tag{6}$$

where B_1, B_2 ($0 < B_1, B_2 < 1$ and $B_1 + B_2 < 1$) and ($1 - B_1 - B_2$) represent the amplitudes of k_off1,

k_{off2} and k_{off3} sub-populations respectively.

Simulating simultaneous dissociation and photobleaching

In order to maintain full control of the kinetic variables, we chose to perform simulations of the experiment. Simulations of exponential distributions and curve-fitting were performed with custom-written program in MATLAB (The MathWorks, Natick, MA). We simulated exponential distributions (Eq. 2,5,6) using the *exprnd* function in MATLAB. Unless otherwise stated, $k_b\tau_{int}$ was fixed at 0.7 and 11 values of τ_{tl} were used (0.1, 0.2, 0.3, 0.4, 0.6, 1, 2, 3, 5, 8, 10 s). The limit of 10 s applied on τ_{tl} is due to the practical challenge of tracking binding events on long timescales. The difficulties manifest in long timescales due to cellular growth and motions of the chromosome, and the higher probability of counting rebinding events in the same localization volume.

To mimic cumulative residence time distribution (CRTD) from experiments, the simulated data were binned in ten bins with edges corresponding to frame times (multiples of τ_{tl}). The number of observations (n), corresponding to the height of the first bin in the CRTD, was used as the stopping condition of the simulation. Practically, binding events are defined as peaks detected on at least two frames to account for slowly moving molecules that only stay on one frame. Hence, we rejected observations in the first bin (0 to τ_{tl}) and only carried observations from τ_{tl} to $10\tau_{tl}$ to the next step.

To simulate uncertainty in each simulation sample, ten rounds of bootstrapping were performed, each involved randomly sampling 80% of the simulated population. Next, fitting was performed on each bootstrapped CRTDs (henceforth referred simply as CRTDs). First, the CRTD at each τ_{tl} was fit to a mono-exponential model to obtain k_{eff} (Eq. 2, 3). These values for k_{eff}, corresponding to the number of τ_{tl}, were then used to construct the $k_{eff}\tau_{tl}$ plot. Error bands

in these plots represent standard deviations from ten bootstrapped samples.

Second, the CRTDs for all τ_{tl} were fit to multi-objective functions based on Eq. 2, 5 and 6 (global fitting). For fitting to mono-exponential, the global parameters k_b and k_{off} were solved by non-linear least square minimization using the trust-region-reflective algorithm in the *lsqnonlin* function in MATLAB. The full list of parameters, initialization, bound conditions and stopping criteria for fitting are present in Table S1. Solving for k_b, k_{off1}, k_{off2} and the amplitudes in bi-exponential model was performed in the same manner. Following minimization, the smaller off rate and its corresponding amplitude were assigned as k_{off1} and B respectively.

Minimization to the tri-exponential model (Eq. 6) required three additional constraints. The first constraint was the sum of amplitudes from the first two sub-population needs to be less than 1 ($B_1 + B_2 < 1$). To avoid solutions where two out of the three off rates are so closely spaced that they can be considered the same, we further restricted k_{off2} to be larger than $2k_{off1}$, and k_{off3} to be larger than $2k_{off2}$. Minimization with these additional constraints was handled using the trust-region-reflective algorithm in the *fmincon* function in MATLAB.

For each simulation, time constants and amplitudes are reported as the mean of CRTDs. The standard deviations of the time constant (σ_τ) from a hundred simulations using the same conditions was calculated according to Eq. 7.

$$\sigma_\tau = \left(\sum_{i=1}^{100} (\tau_i - \langle\tau\rangle)^2 \Big/ 100 \right)^{1/2} \tag{7}$$

where $<\tau>$ denotes the true time constant, which is calculated by $1/<k_{off}>$.

RESULTS

Uncertainties in time constants from mono-exponential distributions

We first explored the relationship between the number of observations (n) and uncertainties in estimates of time constants from mono-exponential distributions. To this end, we simulated a population of molecules dissociating at k_{off} of 0.1 s^{-1}, corresponding to a time constant $<\tau>$ of 10 s, and photobleaching at k_b equal to 7 s^{-1}. While τ_{int} was constant at 0.1 s, τ_{tl} varied from 0.1 s to 10 s. The theoretical $k_{eff}\tau_{tl}$ plot at infinite n is shown as the dashed line (Fig. 2A). At n = 10^3 observations, the $k_{eff}\tau_{tl}$ plot deviates noticeably from the theoretical line (Fig. 2A, purple curve). However, as n increases, the error bands reduce, and the plots closely resemble straight lines (purple curves, Fig. 2B-D). At 10^5 observations, linearly fitting the $k_{eff}\tau_{tl}$ plot (Fig. 2D) yielded a slope of 0.1 and y-intercept of 0.6992, reflecting the true k_{off} and $k_b\tau_{int}$. As expected, mono-exponential distributions with the same $k_b\tau_{int}$ but smaller off rate (k_{off} = 0.01 s^{-1}) or without off rate (k_{off} = 0 s^{-1}) yielded lines with smaller slope (Fig. 2A-D, green curves) or essentially flat lines (Fig. 2A-D, black curves).

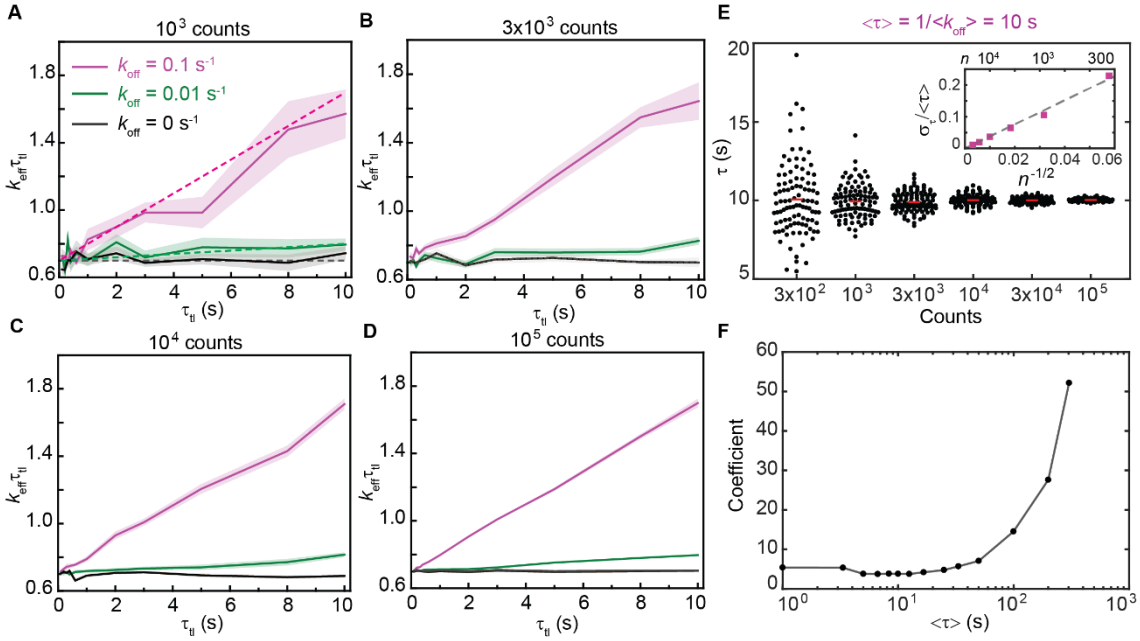

FIGURE 2 Determination of time constants from mono-exponential distributions. (A-D) $k_{eff}\tau_{tl}$ plots of mono-exponential distributions with $k_b\tau_{int}$ of 0.7 and k_{off} of 0.1 s^{-1} (purple curves), 0.01 s^{-1} (green curves) or 0 s^{-1} (black curves). The simulation stopped when the pre-set number of observations (n) of (A) 10^3, (B) $3x10^3$, (C) 10^4 or (D) 10^5 was reached. (A) Dash lines correspond to theoretical $k_{eff}\tau_{tl}$ plots (at infinite counts). Shaded error bars are standard deviations from 10 bootstrapped samples. (E) Scatter plots show distributions of τ obtained using global fitting from 100 simulated samples for each n value. Red bars represent the averages. (Inset) The relative error in determining τ (σ_τ /<τ>) reduces with the inverse square root of n ($n^{-1/2}$). Dash line is the linear fit to six data points. (F) Coefficient in function of σ_τ /<τ> versus $n^{-1/2}$ at various <τ>. The sharp increase in coefficients for <τ> larger than 50 s indicates larger uncertainties in measuring slow processes.

To characterize the uncertainty in the estimate of the time constant (σ_τ), we repeated the simulation a hundred times for each value of n and determined τ using global fitting (Methods, Fig. 2E). As expected for shot noise, the relative error $\sigma_\tau/\langle\tau\rangle$ is proportional to the inverse of the square root of n with a coefficient of 3.8 (Fig. 2E, inset). Importantly, the coefficient fluctuates between 3.7 and 5.7 for $\langle\tau\rangle \leq 50$ s, but rises sharply for $\langle\tau\rangle$ greater than 50 s (Fig. 2F). This result demonstrates that the uncertainty in estimating lifetime of long-lived molecules becomes arbitrarily large when the extended lifetime of the fluorophore (by introduction of τ_d) becomes comparable to the binding lifetime. In this circumstance, very large increases in the number of observations are required to effectively reduce the uncertainty in the lifetime measurement. Therefore, in practice for a bright fluorescent protein such as YPet, we propose that accurate measurements of binding lifetimes greater than 50 s require significant increases in the number of observations (n). In systems where a trade-off must be met between τ_{tl} and n, a lifetime of 50 s reflects a practical resolution limit for such measurements.

We anticipated that photobleaching rate also contributes to σ_τ as faster photobleaching reduces observation times. To examine the effect of $k_b\tau_{int}$, we performed a comprehensive set of simulations with $k_b\tau_{int}$ varying from 0.007 to 2.1 (k_b from 0.07 to 70 s^{-1} and τ_{int} from 0.01 to 0.1 s). We obtained the relationship between σ_τ/τ, n and $k_b\tau_{int}$ as in Eq. 8.

$$\frac{\sigma_\tau}{\tau} = \frac{(2.7379 k_b \tau_{int})^2}{n^{1/2}} \tag{8}$$

This formula describes the lower bound of errors as other sources of practical errors, such as localization uncertainties and experimental variations, have not been considered. The minimum number of observations required to determine τ ($\langle\tau\rangle \leq 50$ s) with a given uncertainty is therefore:

$$n = (\tau/\sigma_\tau)^2 \times (2.7379 k_b \tau_{\text{int}})^4 \quad (9)$$

As an example, when $k_b\tau_{int}$ is 0.7, the number of observations required to achieve relative error of 10% in the estimate of τ (where $<\tau> \leq 50$ s) is about 1350 (see Fig. 2E).

Uncertainties in time constants and amplitudes from bi-exponential distributions

Next, we examined the situation where a second kinetic sub-population is present in the system. A second population with a faster off rate yields $k_{eff}\tau_{tl}$ plots that deviate from straight lines (24). However, deviations can also be a result of shot noise at low n (see Fig. 2A-D). To identify the minimum n at which one can determine with a specified confidence that a bi-exponential model is appropriate, we simulated CRTDs using Eq. 5. First, we performed simulations with off rates that are an order of magnitude apart: $k_{off1} = 0.1$ s^{-1} (intermediate rate) and $k_{off2} = 1$ s^{-1} (fast rate). The amplitude B of the intermediate dissociating population was varied from 10% to 90%.

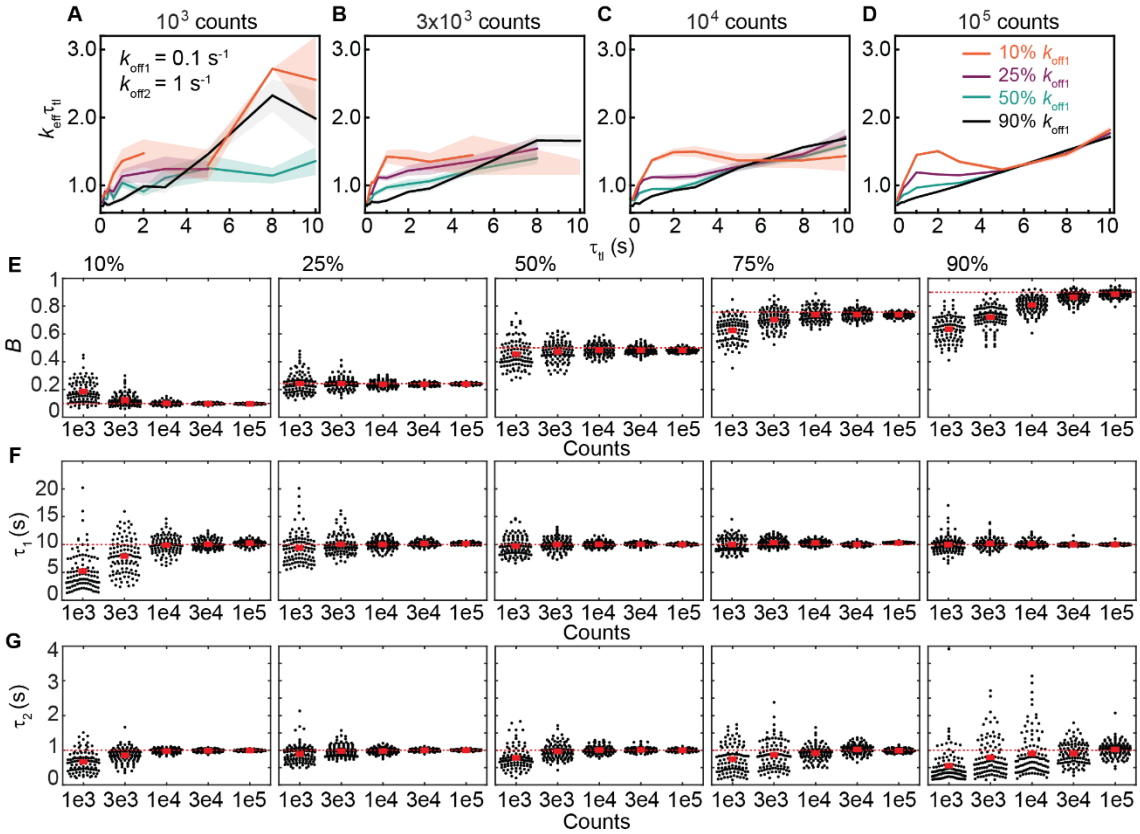

FIGURE 3 Determination of time constants and amplitudes from bi-exponential distributions with an intermediate rate (k_{off1}) and a fast rate (k_{off2} = 10 k_{off1}). (A-D) $k_{eff}\tau_{tl}$ plots of bi-exponential distributions with $k_b\tau_{int}$ of 0.7, k_{off1} and k_{off2} of 0.1 and 1.0 s^{-1} respectively, with (A) 10^3, (B) $3x10^3$, (C) 10^4 or (D) 10^5 observations. The amplitude of k_{off1} (B) is 10% (orange), 25% (purple), 50% (green) or 90% (black). Shaded error bars are standard deviations from 10 bootstrapped samples. (E-G) Scatter plots show distributions of *B*, τ_1 and τ_2 obtained using global fitting from 100 simulated distributions for each n value. Each panel corresponds to a pre-set *B*, which increases from 10%, 25%, 50%, 75% to 90% from left to right. In each panel, *n* increases from 10^3 (1e3) to 10^5 (1e5). Dashed lines and red bars represent the true values and the average respectively. To enhance visibility, outliers (less than 5% when present) were omitted from scatter plots.

When the majority of the population dissociates with the intermediate rate k_{off1} (B = 90%), the $k_{eff}\tau_{tl}$ plots resemble those of mono-exponential distribution with the single k_{off} of 0.1 s^{-1} (compare Fig. 3A-D, black curves and Fig. 2A-D). As before, increasing the number of observations significantly improved the quality of the $k_{eff}\tau_{tl}$ plots (Fig. 3A-D). These simulations reveal that a short-lived second sub-population does not manifest as a visible feature in the $k_{eff}\tau_{tl}$ plots when it is present only to the extent of 10% in the observations. To examine if the two populations could be resolved with global fitting using the bi-exponential model, we determined time constants and amplitudes from 100 simulations (Fig. 3E-G, right-most panel). Unsurprisingly, we found that the accuracies and precisions of determining B, τ_1 and τ_2 increase with n. While estimation of τ_1 is robust (relative error of 6% at 10^4 counts) (Fig. 3F), global fitting of CRTDs to the bi-exponential model at low counts suffers from a bias towards the fast dissociating sub-population, with its amplitude being overestimated and τ_2 being underestimated (Fig. 3E, G). This bias is observed to a lesser extent when k_{off1} is present at 75% or 50% (Fig. 3E-G).

As the amplitude of the fast dissociating sub-population increased (B equal to 25% or 10%), fewer observations were found at long intervals. Insufficient counts resulted in missing data points at these τ_{tl} ($\tau_{tl} \geq 5$ s) in $k_{eff}\tau_{tl}$ plots at low counts (10^3 and $3x10^3$, Fig. 3A-B). However, the $k_{eff}\tau_{tl}$ plots extended to the full τ_{tl} range of 10 s when n increases to 10^4 and 10^5 (Fig. 3C-D). As expected, deviations from straight lines were found in the 0-5 s regime, reflecting the presence of the fast dissociating sub-population. Since contributions from the fast dissociating sub-population drop sharply at long timescales, the $k_{eff}\tau_{tl}$ plots converge to the straight line exhibited by mono-exponential distributions with k_{off1} (Fig. 3C-D). Further analysis by integrating the area under the peaks in the 0-to-5 s region shows the area increases exponentially with the amplitude of the fast dissociating sub-population (Fig. S1).

When the fast dissociating sub-population represents the majority, the accuracy and precision in determining B, τ_1 and τ_2 also increase with n (Fig. 3E-G). Further, estimates of B are more precise compared to when the intermediate sub-population is the majority (Fig. 3E). This is due to the fast dissociating sub-population being sufficiently represented across τ_{tl} when B is small. As an example, when B is 10% or 25%, the errors in estimating B (σ_B) at 10^4 counts reduce to 1.4% and 2.9% respectively (Fig. 3E, two left-most panels). Along these amplitudes, τ_1 can be estimated with a 10% error at $3x10^4$ counts for 10% k_{off1} and 10^4 counts for 25% k_{off1} (Fig. 3F, two left-most panels). Further, estimations of τ_2 are more precise relative to τ_1, consistent with the abundance of fast dissociating sub-population (Fig. 3F-G).

These trends were also observed for bi-exponential distributions with rates that are further spaced (0.01 s^{-1} and 1 s^{-1}, Fig. S2). In general, we propose that deviations from straight lines in $k_{eff}\tau_{tl}$ plot can be used as reliable indicators for bi-exponential distributions when n is above 10^4.

Resolution limit in determining time constants and amplitudes from bi-exponential distributions

Due to the resolution limit that is inherent to exponential analysis (28), we anticipated the ability to resolve rates that are closely spaced would reduce. To test this hypothesis, we simulated bi-exponential distributions with rates that are only three-fold apart: an intermediate rate k_{off1} of 0.1 s^{-1} and a fast rate k_{off2} of 0.3 s^{-1}. Under conditions that yield sufficient observations at long intervals ($n \geq 10^4$), examination of the $k_{eff}\tau_{tl}$ plots often fails to identify the presence of multiple sub-populations in the form of deviation from straight lines (Fig. 4A-D). Only when the fast rate is present at 90%, can deviations be observed in the form of a broad convex spanning from 0 to 10 s (orange curves, Fig. 4C-D). Fitting to Eq. 5 yields unreliable results for B and τ_2 (Fig. 4E, G). While estimates of τ_1 are more precise, the accuracy in determining τ_1 requires 10^5 observations or $<B>$ to be larger than 50% (Fig. 4F). Fitting CRTDs at low counts ($n \leq 10^4$) to the bi-exponential model should be avoided as one often obtains two kinetics sub-populations with artificially enhanced rate separation and substantial amplitudes, regardless of the true amplitudes (Fig. 4E-G).

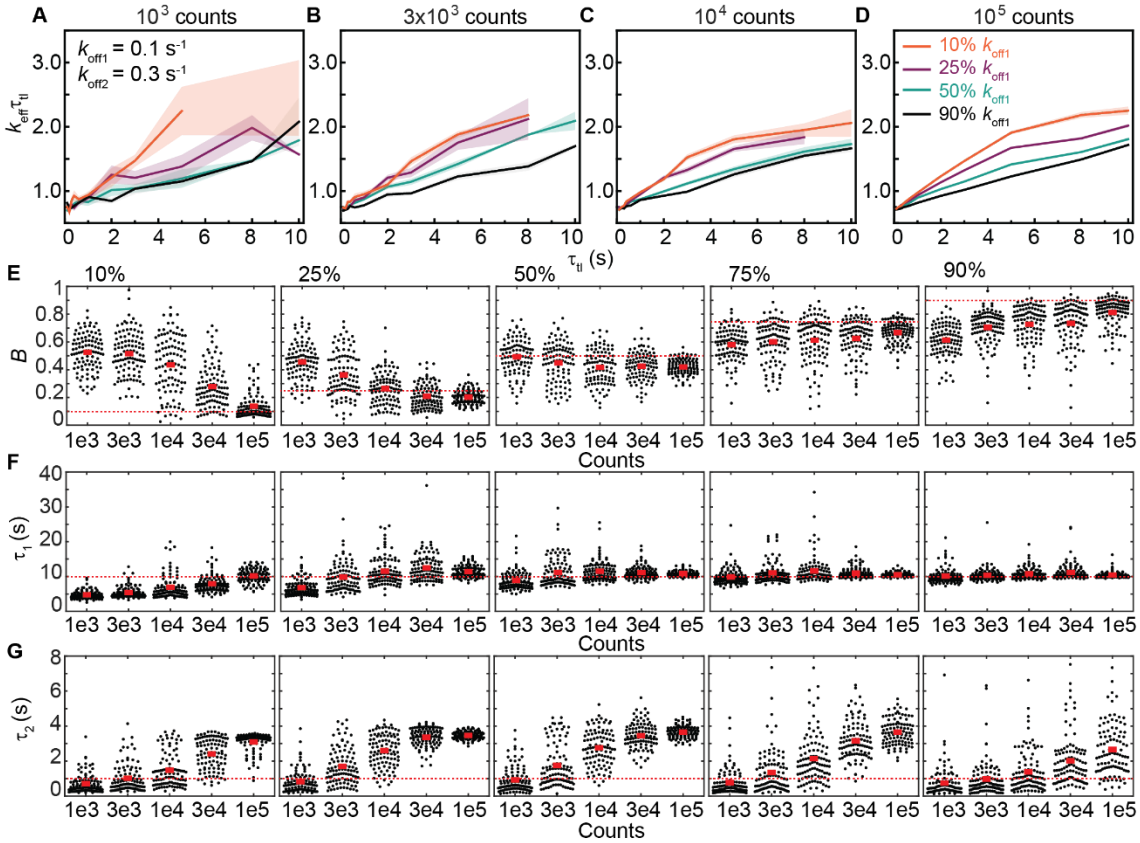

FIGURE 4 Determination of time constants and amplitudes from bi-exponential distributions with closely spaced rates ($k_{off2} = 3k_{off1}$). (A-D) $k_{eff}\tau_{tl}$ plots of bi-exponential distributions with $k_b\tau_{int}$ of 0.7, k_{off1} and k_{off2} of 0.1 and 0.3 s^{-1} respectively, with (A) 10^3, (B) $3x10^3$, (C) 10^4 or (D) 10^5 observations. The amplitude of k_{off1} (B) is 10% (orange), 25% (purple), 50% (green) or 90% (black). Shaded error bars are standard deviations from 10 bootstrapped samples. (E-G) Scatter plots show distributions of *B*, τ_1 and τ_2 obtained from fitting of 100 simulated distributions for each *n* value. Each panel corresponds to a pre-set *B*, which increases from 10%, 25%, 50%, 75% to 90% from left to right. In each panel, *n* increases from 10^3 (1e3) to 10^5 (1e5). Dashed lines and red bars represent the true values and the average respectively. To enhance visibility, outliers (less than 5% when present) were omitted from scatter plots.

Dynamic range limit in determining time constants and amplitudes from bi-exponential distributions

As we demonstrated in Fig. 2F, time constant measurements of slow processes ($\tau \geq 50$ s) entail larger uncertainties due to the limited observation time around one minute. To examine the implication of this dynamic range limit on resolving two kinetic sub-populations, we simulated bi-exponential distributions with a slow rate (k_{off1} of 0.01 s^{-1}; $<\tau>$ = 100 s) and an intermediate rate (k_{off2} of 0.1 s^{-1}; $<\tau>$ = 10 s). Unexpectedly, the $k_{eff}\tau_{tl}$ plots resemble straight lines (Fig. 5A-C), unlike those comprised of an intermediate rate ($<\tau>$ = 1 s) and a slow rate ($<\tau>$ = 10 s) (Fig. 3). The slopes of these lines increase with the amplitude of k_{off2}, reflecting the relative contribution of this sub-population. Only after $5<\tau_2>$ or 50 s, the contribution from the intermediate population becomes negligible and the $k_{eff}\tau_{tl}$ plots converge to that of mono-exponential distribution with k_{off1} (Fig. S3). Fitting the CRTDs at 10^5 counts to mono-exponential model (Eq. 2) results in apparent time constants (τ^*) that are between $<\tau_1>$ and $<\tau_2>$ (Fig. 5D). Fitting mean τ^* vs. B to exponential function yields Equation 10:

$$\tau^* = \langle\tau_2\rangle e^{log(\langle\tau_1\rangle/\langle\tau_2\rangle)B} \tag{10}$$

Thus, B can be derived from τ^* where $<\tau_1>$ and $<\tau_2>$ are known.

Fitting the CRTDs with n less than 10^6 to the bi-exponential model yields unreliable results (Fig. 5E-G). Across various amplitudes of k_{off1}, the slow time constant, τ_1, is often underestimated and corresponds to τ^* at that amplitude (compare Fig. 5F to Fig. 5D). Similarly, τ_2 is also underestimated, but eventually approaches $<\tau_2>$ of 10 s when n reached 10^6 counts and the amplitude of k_{off2} sub-population is more than 25% (Fig. 5G, three left-most panels).

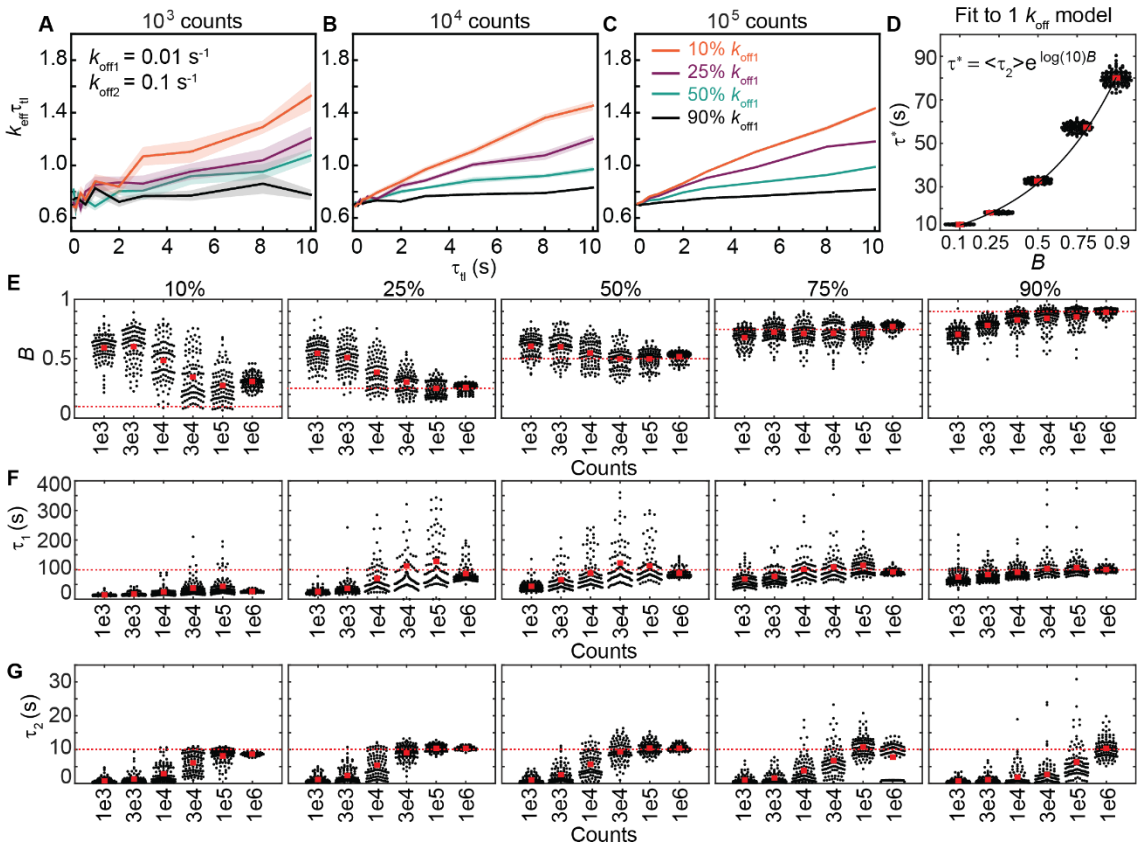

FIGURE 5 Determination of time constants and amplitudes from bi-exponential distributions with a slow rate (k_{off1}) and an intermediate rate (k_{off2} = $10k_{off1}$). (A-C) $k_{eff}\tau_{tl}$ plots of bi-exponential distributions with $k_b\tau_{int}$ of 0.7, k_{off1} and k_{off2} of 0.01 and 0.1 s^{-1} respectively, with (A) 10^3, (B) 10^4 or (C) 10^5 observations. The amplitude of k_{off1} (B) is 10% (orange), 25% (purple), 50% (green) or 90% (black). Shaded error bars are standard deviations from ten bootstrapped samples. (D) Scatter plots show distribution of apparent τ (τ^*) obtained from fitting of 100 simulated bi-exponential distributions at a specified *B* and 10^5 counts to mono-exponential model. Line is exponential fit between the average of τ^* (red bars) and *B*. (E-G) Scatter plots show distributions of *B*, τ_1 and τ_2 obtained from fitting of 100 simulated distributions to bi-exponential model. Each panel corresponds to a pre-set amplitude of k_{off1}, which increases from 10%, 25%, 50%, 75% to 90% from left to right. In each panel, *n* increases from 10^3 (1e3) to 10^6 (1e6). Dashed lines and red bars represent the true values and the average respectively. To enhance visibility, outliers (less than 5% when present) were omitted from scatter plots.

Resolution of three off rates

The resolution limit as well as dynamic range limit that we demonstrated above raise the question if tri-exponential distributions can be faithfully resolved under the specified experimental condition (τ_{tl} range, range of n). To address this issue, we simulated tri-exponential distributions (Eq. 6), with off rates spanning two orders of magnitude (0.01, 0.1 and 1 s^{-1}). The diversity in $k_{eff}\tau_{tl}$ plots obtained by varying B_1 and B_2 is illustrated in Fig. 6A. We further characterized uncertainties in amplitudes and time constants obtained using global fitting to the tri-exponential model (Methods). Importantly, the effects of increasing n on accuracy of the time constants and amplitudes are no longer easily predictable: improved accuracy in one parameter may come at an expense of the others (Fig. 6B-F). Overall, we found τ_3 to be an average of $<\tau_2>$ and $<\tau_3>$, τ_2 to be an average of $<\tau_1>$ and $<\tau_2>$ while τ_1 is commonly underestimated. As in the case of the bi-exponential simulations, we observed consistent biases towards faster time constants.

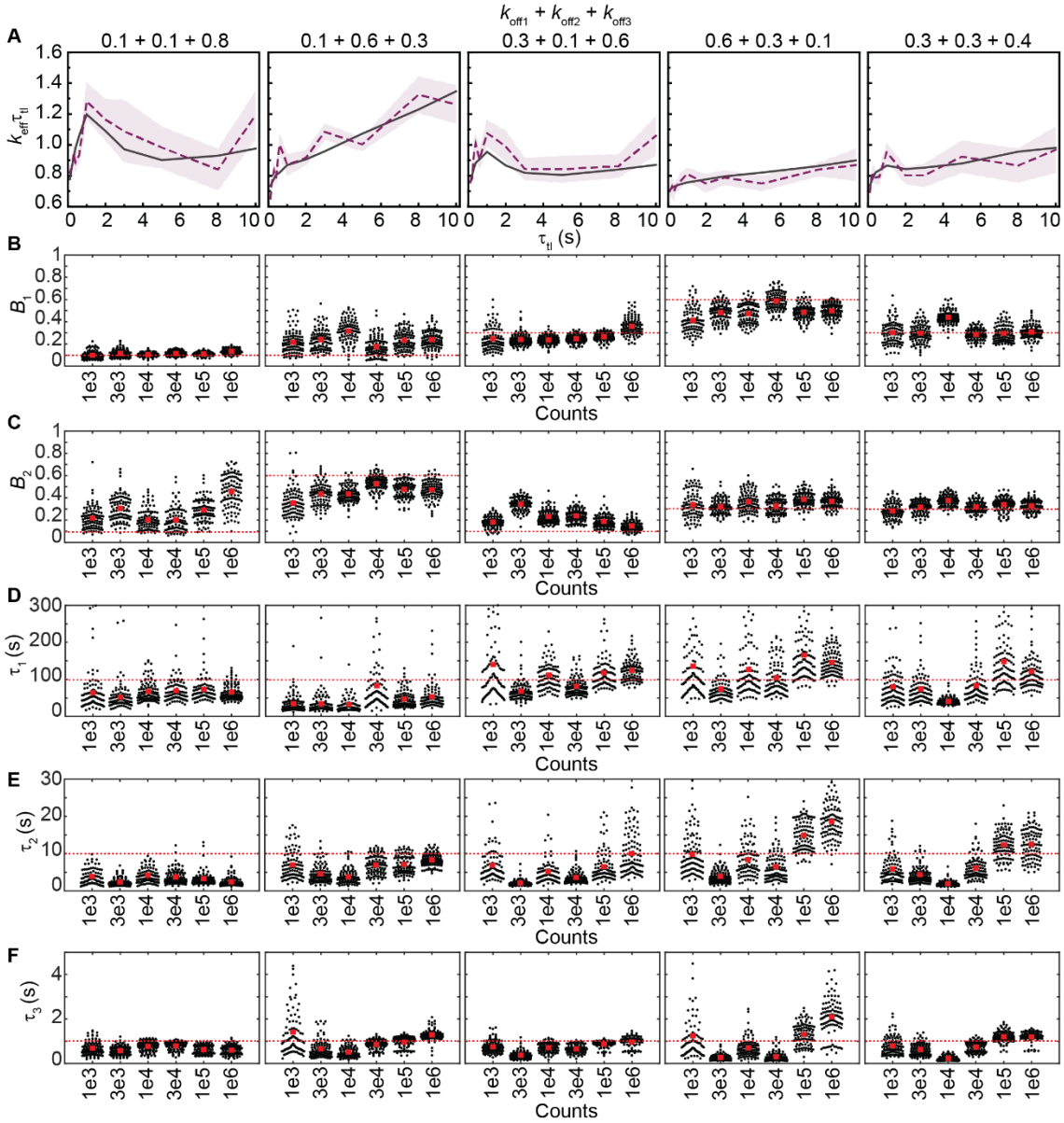

FIGURE 6 Determination of time constants and amplitudes from tri-exponential distributions with a slow rate (k_{off1}), an intermediate rate ($k_{off2} = 10k_{off1}$) and a fast rate ($k_{off3} = 10k_{off2}$). From left to right, five panels in each row correspond to different amplitudes of each sub-population (displayed on top). (A) $k_{eff}\tau_{tl}$ plots of tri-exponential distributions with $k_b\tau_{int}$ of 0.7, k_{off1}, k_{off2} and k_{off3} of 0.01, 0.1 and 1 s^{-1} respectively, with 10^3 (dash purple curves) and 10^6 (grey curves) observations. Shaded error bars are standard deviations from 10 bootstrapped samples. (B-F) Scatter plots show distributions of amplitudes (B_1 and B_2), τ_1, τ_2 and τ_3 obtained using global fitting from 100 simulated samples. In each panel, n increases from 10^3 (1e3) to 10^6 (1e6). Dashed lines and red bars represent the true values and the averages respectively. To enhance visibility, outliers (less than 5% when present) were omitted from scatter plots.

The choice of τ_{tl}

Given a finite amount of experimental time, should one sample with more τ_{tl} (increase $N_{interval}$) or obtain more observations (increase n) with a minimal τ_{tl} set? We defined the minimal τ_{tl} set to contain τ_{tl} values that are three-fold apart in the 0-to-10 s regime. Thus, this set contains 5 τ_{tl} values (N_{min} = 5): 0.1, 0.3, 1, 3 and 10 s. Extending Eq. 8 to the total number of observations ($nN_{interval}$), one could expect errors of estimates to reduce by a factor of $\sqrt{(N_{interval}/N_{min})}$ for the same n. Our simulations up to this point were performed with $N_{interval}$ equal to 11. Hence, error ratios (σ_{11}/σ_5) smaller than $1/\sqrt{(11/5)}$ or 0.67 indicate the benefit of increasing $N_{interval}$ outweighs the benefit of increasing n with the minimal τ_{tl} set whereas error ratios larger than 0.67 represent redundancy in τ_{tl}.

To identify the optimum choice of τ_{tl}, we simulated bi-exponential distributions with an intermediate rate (k_{off1} = 0.1 s^{-1}) and a fast rate (k_{off2} = 1 s^{-1}) using the minimal τ_{tl} set. These simulations yielded $k_{eff}\tau_{tl}$ plots that closely resemble those in Fig. 3 (see Fig. 7A-D) and similarly, deviations from straight lines are also reliable indicators of kinetic heterogeneity when B is less than 90%. As expected, estimates of B, τ_1 and τ_2 are more accurate with larger n (Fig. 7E-G). Further, at the same n, errors of estimates are almost always smaller in simulated distributions with 11 τ_{tl} ($\sigma_{11}/\sigma_5 < 1$, Fig. 7H). Redundancy in τ_{tl} were observed in some cases when the intermediate dissociating sub-population is the majority ($\sigma_{11}/\sigma_5 > 0.67$, B from 50% to 90%) (Fig. 7H). However, when the majority dissociates with the fast rate (B equal to 10% or 25%), the benefit of sampling with more τ_{tl} is clear ($\sigma_{11}/\sigma_5 < 0.67$), especially with $n \geq 10^4$. Thus, we concluded the net benefit of increasing $N_{interval}$ is greater than increasing n with the minimal τ_{tl} set.

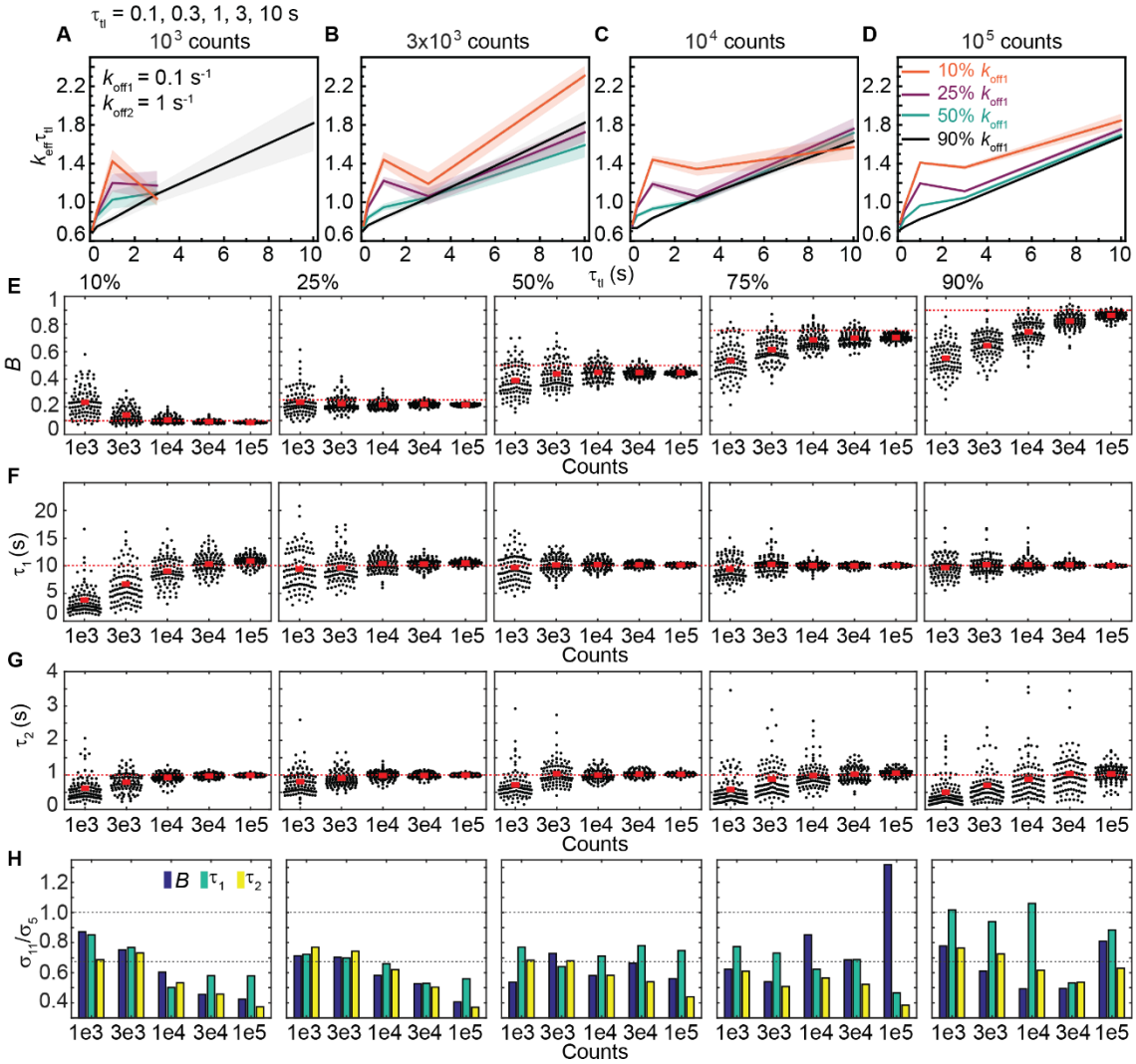

FIGURE 7 Determination of time constants and amplitudes from bi-exponential distributions using a minimal τ_{tl} set (five τ_{tl} values), an intermediate rate (k_{off1} = 0.1 s^{-1}) and a fast rate (k_{off2} = 1 s^{-1}). (A-D) $k_{eff}\tau_{tl}$ plots of bi-exponential distributions with (A) 10^3, (B) $3x10^3$, (C) 10^4 or (D) 10^5 observations. The amplitude of k_{off1} (*B*) is 10% (orange), 25% (purple), 50% (green) or 90% (black). Shaded error bands are standard deviations from 10 bootstrapped samples. (E-G) Scatter plots show distributions of *B*, τ_1 and τ_2 obtained using global fitting from 100 simulated distributions for each *n* value. Each panel corresponds to a pre-set *B*, which increases from 10%, 25%, 50%, 75% to 90% from left to right. In each panel, *n* increases from 10^3 (1e3) to 10^5 (1e5). Dashed lines and red bars represent the true values and averages respectively. To enhance visibility, outliers (less than 5% when present) were omitted from scatter plots. (H) Bar plots show ratios of error estimates obtained from simulations with 11 and 5 τ_{tl} values at the same *n*. Blue: *B*, green: τ_1, yellow: τ_2.

DISCUSSION

In this work, we used simulated data to explore the influence of shot noise, resolution limit and dynamic range limit on resolving multiple kinetic sub-populations in single-molecule interval imaging experiments (Fig. 8). Within the dynamic range and resolution limit, determination of time constants and amplitudes in mono-exponential and bi-exponential distributions are reliable in general, especially with at least 10^4 counts. Fitting to tri-exponential models, however, yields unreliable results and hence fitting outcomes should be interpreted with caution.

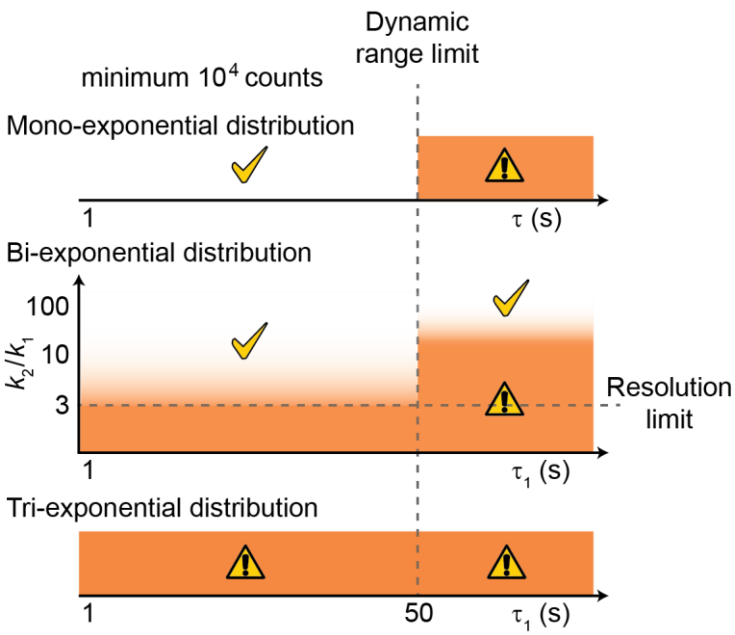

FIGURE 8 Dynamic range and resolution limits in resolving multiple populations using the interval imaging technique with photobleaching-prone fluorescent probes. Orange zones indicate conditions where errors in estimates of τ and the amplitude are high.

As showed in Eq. 8, the relative error in τ determination scales with the square of $k_b\tau_{int}$ and the inverse square root of n. This emphasizes the importance of choosing imaging conditions to minimize $k_b\tau_{int}$ as a two-fold increase in $k_b\tau_{int}$ needs to be compensated by a 16-fold increase in n. A balance has to be struck here to ensure good signal-to-background ratio, a prerequisite for reliable particle tracking. These findings also highlight the importance of developing and

using fluorophores with higher photo-stability and brightness for live-cell applications as these would greatly reduce uncertainties in measurements. In practice, the choice of fluorescent protein should be made with great care, as fluorescent proteins often exhibit undesirable properties that limit their utility (29-33).

Errors obtained from repeating the experiments can be an underestimation compared to inherent errors conferred by shot noise when fitting is ill-conditioned (34), which is often the case when minimizing using multi-objective functions (35). Therefore, reports of time constant measurements using these interval imaging approaches should clearly state $k_b\tau_{int}$ from fitting and n from experimental data. This would enable a theoretical error estimation of τ and avoid over-interpretation of experimental results.

We found $k_{eff}\tau_{tl}$ plots useful for guiding the fitting model when the number of counts is sufficiently large (more than 10^4) as deviations from straight lines faithfully reflect heterogeneity in binding kinetics. The reverse is not necessarily true. Good linear fits, seen at large n values, can reflect one of the following three scenarios: the absence of multiple populations, sub-populations with off rates that are within the resolution limit, or sub-populations where the off rate of one population lies beyond the dynamic range of the experiment. When the mono-exponential model is used to fit those data, an apparent time constant τ^*, whose value lies between the two true time constants, is obtained. While sub-optimal, τ^* depends on the proportion of molecules in each kinetic sub-population: a larger presence of the fast dissociating sub-population yields smaller τ^*. This in turn can report on change in binding kinetics when the biology is manipulable – for instance with binding partners or drugs.

Can statistical information such as reduced χ^2 be used to decide the model that best describes

the data? Computing these criteria requires the degree of freedom to be known, which is not the case when fitting to our non-linear models (36). Instead of using statistical criteria, the selection of the fitting model using $k_{eff}\tau_{tl}$ plot can be complemented with experimental design. For example, in case where a bi-exponential model is invoked, it might be tempting to attribute sub-populations to molecules performing certain activities such as binding of DNA repair proteins to a damaged or non-damaged substrate. These hypotheses can be tested using structure-function mutants in which one or few catalytic activities are inhibited, hence, yielding predictable changes in $k_{eff}\tau_{tl}$ plots and fitting results. Finally, where possible, we recommend approaches that utilize multiple experimental designs to reproducibly observe or enrich the hypothesized populations.

SOFTWARE

Our algorithms are freely available as open source MATLAB codes from https://github.com/hanngocho/off-rate-simulation.

SUPPORTING INFORMATION

Three supplementary figures are available following publication.

AUTHOR CONTRIBUTIONS

H.N.H., H.G. and A.M.v.O. designed research. H.N.H. wrote the codes, conducted the simulations and wrote the first draft. D.Z. contributed to the codes, with support from J.K. H.G. and A.M.v.O revised the manuscript and supervised research.

ACKOWLEDGEMENTS

This work was supported by Australian Research Council Grant DP150100956, Australian Laureate Fellowship FL140100027 (to A.M.v.O.). We thank Dr. Joris M. H. Goudsmits for assisting with the simulation codes in MATLAB. D.Z. gratefully acknowledges support from the Elite Network of Bavaria (ENB) program “Macromolecular Science”.

REFERENCES

1. Lu, H.P., L. Xun, and X.S. Xie, 1998 Single-molecule enzymatic dynamics. *Science*. **282**(5395): p. 1877-82.
2. Xie, X.S., 2002 Single-molecule approach to dispersed kinetics and dynamic disorder: Probing conformational fluctuation and enzymatic dynamics. *Journal of Chemical Physics*. **117**(24): p. 11024-11032.
3. Larson, M.H., R. Landick, and S.M. Block, 2011 Single-molecule studies of RNA polymerase: one singular sensation, every little step it takes. *Mol Cell*. **41**(3): p. 249-62.
4. Stratmann, S.A. and A.M. van Oijen, 2014 DNA replication at the single-molecule level. *Chem Soc Rev*. **43**(4): p. 1201-20.
5. Gahlmann, A. and W.E. Moerner, 2014 Exploring bacterial cell biology with single-molecule tracking and super-resolution imaging. *Nat Rev Microbiol*. **12**(1): p. 9-22.
6. Bell, J.C. and S.C. Kowalczykowski, 2016 Mechanics and Single-Molecule Interrogation of DNA Recombination. *Annu Rev Biochem*. **85**: p. 193-226.
7. Kaniecki, K., L. De Tullio, and E.C. Greene, 2018 A change of view: homologous recombination at single-molecule resolution. *Nature Reviews Genetics*. **19**(4): p. 191-+.
8. Kong, M., et al., 2017 Single-Molecule Methods for Nucleotide Excision Repair: Building a System to Watch Repair in Real Time. *Methods Enzymol*. **592**: p. 213-257.
9. Floyd, D.L., et al., 2008 Single-particle kinetics of influenza virus membrane fusion. *Proc Natl Acad Sci U S A*. **105**(40): p. 15382-7.
10. Yan, X., et al., 2016 Dynamics of Translation of Single mRNA Molecules In Vivo. *Cell*. **165**(4): p. 976-89.
11. Ghodke, H., et al., 2014 Single-molecule analysis reveals human UV-damaged DNA-binding protein (UV-DDB) dimerizes on DNA via multiple kinetic intermediates. *Proc Natl Acad Sci U S A*. **111**(18): p. E1862-71.
12. Volkov, I.L., et al., 2018 tRNA tracking for direct measurements of protein synthesis kinetics in live cells. *Nat Chem Biol*. **14**(6): p. 618-626.
13. Leake, M.C., et al., 2008 Variable stoichiometry of the TatA component of the twin-arginine protein transport system observed by in vivo single-molecule imaging. *Proc Natl Acad Sci U S A*. **105**(40): p. 15376-81.
14. Sako, Y., S. Minoghchi, and T. Yanagida, 2000 Single-molecule imaging of EGFR signalling on the surface of living cells. *Nat Cell Biol*. **2**(3): p. 168-72.
15. Dixit, R., et al., 2008 Differential regulation of dynein and kinesin motor proteins by tau. *Science*. **319**(5866): p. 1086-9.
16. Liu, J., et al., 2016 Cascading MutS and MutL sliding clamps control DNA diffusion to activate mismatch repair. *Nature*. **539**(7630): p. 583-587.

17. Li, Y., et al., 2018 Visualizing bacterial DNA replication and repair with molecular resolution. *Curr Opin Microbiol*. **43**: p. 38-45.

18. Ghodke, H., H. Ho, and A.M. van Oijen, 2018 Single-molecule live-cell imaging of bacterial DNA repair and damage tolerance. *Biochem Soc Trans*. **46**(1): p. 23-35.

19. Kapanidis, A.N., S. Uphoff, and M. Stracy, 2018 Understanding Protein Mobility in Bacteria by Tracking Single Molecules. *J Mol Biol*. **430**(22): p. 4443-4455.

20. Elf, J. and I. Barkefors, 2018 Single-Molecule Kinetics in Living Cells. *Annu Rev Biochem*.

21. Shaner, N.C., P.A. Steinbach, and R.Y. Tsien, 2005 A guide to choosing fluorescent proteins. *Nat Meth*. **2**(12): p. 905-909.

22. Liu, Z., L.D. Lavis, and E. Betzig, 2015 Imaging live-cell dynamics and structure at the single-molecule level. *Mol Cell*. **58**(4): p. 644-59.

23. Elf, J., G.W. Li, and X.S. Xie, 2007 Probing transcription factor dynamics at the single-molecule level in a living cell. *Science*. **316**(5828): p. 1191-4.

24. Gebhardt, J.C., et al., 2013 Single-molecule imaging of transcription factor binding to DNA in live mammalian cells. *Nat Methods*. **10**(5): p. 421-6.

25. Liao, Y., et al., 2016 Single-Molecule DNA Polymerase Dynamics at a Bacterial Replisome in Live Cells. *Biophys J*. **111**(12): p. 2562-2569.

26. Agarwal, H., et al., 2017 Direct Observation of Cell-Cycle-Dependent Interactions between CTCF and Chromatin. *Biophys J*.

27. Clauss, K., et al., 2017 DNA residence time is a regulatory factor of transcription repression. *Nucleic Acids Res*. **45**(19): p. 11121-11130.

28. Istratov, A.A. and O.F. Vyvenko, 1999 Exponential analysis in physical phenomena. *Review of Scientific Instruments*. **70**(2): p. 1233-1257.

29. Swulius, M.T. and G.J. Jensen, 2012 The helical MreB cytoskeleton in Escherichia coli MC1000/pLE7 is an artifact of the N-Terminal yellow fluorescent protein tag. *J Bacteriol*. **194**(23): p. 6382-6.

30. Landgraf, D., et al., 2012 Segregation of molecules at cell division reveals native protein localization. *Nature Methods*. **9**(5): p. 480-U98.

31. Wang, S., et al., 2014 Characterization and development of photoactivatable fluorescent proteins for single-molecule-based superresolution imaging. *Proc Natl Acad Sci U S A*. **111**(23): p. 8452-7.

32. Tuson, H.H., et al., 2016 Addressing the Requirements of High-Sensitivity Single-Molecule Imaging of Low-Copy-Number Proteins in Bacteria. *Chemphyschem*. **17**(10): p. 1435-40.

33. Ghodke, H., et al., 2016 Single-Molecule Specific Mislocalization of Red Fluorescent Proteins in Live Escherichia coli. *Biophys J*. **111**(1): p. 25-7.

34. Shrager, R.I. and R.W. Hendler, 1998 Some pitfalls in curve-fitting and how to avoid them: a case in point. *J Biochem Biophys Methods*. **36**(2-3): p. 157-73.

35. Kaiser, M.J., 2001 Nonlinear multiobjective optimization. *Interfaces*. **31**(3): p. 122-123.

36. Andrae, R., Schulze-Hartung, T. & Melchior, P., 2010 Dos and don'ts of reduced chi-squared. *arXiv:*. **1012.3754**.