Supporting Information

Statistical Analysis of Scanning Fluorescence Correlation Spectroscopy Data Differentiates Free from Hindered Diffusion

Falk Schneider[1], Dominic Waithe[2,3], B. Christoffer Lagerholm[2], Dilip Shrestha[1], Erdinc Sezgin[1], Christian Eggeling[1,2,4], Marco Fritzsche[1,5*]

[1] MRC Human Immunology Unit, [2] Wolfson Imaging Centre Oxford and [3] MRC Centre for Computational Biology, Weatherall Institute of Molecular Medicine, University of Oxford, Headley Way, OX39DS Oxford, United Kingdom.

[4] Institute of Applied Optics Friedrich-Schiller-University and Leibniz Institute of Photonic Technology, Helmholtzweg 4, 07743 Jena, Germany.

[5] Kennedy Institute for Rheumatology, Roosevelt Drive, University of Oxford, Oxford OX37LF Oxford, United Kingdom.

*Correspondence to: marco.fritzsche@rdm.ox.ac.uk or +44 1865 2223555

S1 Analysis of sFCS data

Pre-processing and fitting of the sFCS data

Using the empirically determined scanning-frequencies and the pixel dwell times, the sFCS data were loaded and correlated using the FoCuS-scan software.[20] Initial bleaching was removed by cropping off the first 10 seconds of all measurements. To further reduce the bias due to bleaching the photobleaching correction by local averaging (16 seconds interval) was applied as described before.[20] Following temporal cropping, the correlation curves were then fitted using the described software package. The data for free and trapped diffusion simulations, as well as all experimental data were fitted to a single component 2D-diffusion model as follows.

$$G(\tau) = \frac{1}{N} \cdot \frac{1}{(1 + {}^{\tau}/_{\tau_D})} + O_f \tag{1}$$

τ denotes the correlation time not to be confused with the absolute measurement time t. Offset O_f, amplitude (given as inverse average number of particles in the observation volume) ${}^{1}/_{N}$ and transit time τ_D were fitting parameters. Fitting was performed in the range of 0.5 ms to 4000 ms. To obtain robust fits and an adequate error measure the data were bootstrapped 20 times. Data were exported and saved as Excel-sheets containing all fitted parameters and additionally the standard deviations and signal-to-noise values (based on $S/N = G(\tau)/var(G(\tau))^{0.5}$).[20]

The sFCS data for hindered hop diffusion comprising two processes with two different correlation times were fitted with a two-component model following as previously used in camera based FCS analysis:[11]

$$G(\tau) = \frac{1}{N} \cdot \left(A_1 \cdot \frac{1}{(1 + {}^{\tau}/_{\tau_{D,1}})} + A_2 \cdot \frac{1}{(1 + {}^{\tau}/_{\tau_{D,2}})} \right) + O_f, \tag{2}$$

where A_1 and A_2 denote the amplitudes of the two process with characteristic correlation times of $\tau_{D,1}$ and $\tau_{D,2}$.

Statistical analysis of transit time histograms

Statistical analysis of the sFCS data was performed on the basis of the empirical finding that all histograms of transit times from freely diffusing molecules were lognormally distributed as previously outlined.[20] The lognormal nature of the sFCS data was a consequence of an inherent bias in the data correlation resulting in a larger error for larger transit times. All quantitative analysis was executed using custom-written scripts in MATLAB (MathWorks, USA).

Lognormal distributed data are described by the Lognormal function that exists in three analytical forms in its linear, cumulative, and logarithmic representation. The linear single lognormal function $f_{sLogn,lin}(X_{lin}|\mu, \sigma)$ is given by:

$$f_{sLogn,lin}(X_{lin}|\mu, \sigma) = \frac{1}{X_{lin}\sigma\sqrt{(2\pi)}} \exp -\frac{(\ln(X_{lin}) - \mu)^2}{2\sigma^2}, \tag{3}$$

with two characteristic parameters μ and σ.

The single cumulative Lognormal function $f_{sLogn,cum}(X_{cum}|\mu,\sigma)$ is defined as:

$$f_{sLogn,cum}(X_{cum}|\mu,\sigma) = \frac{1}{2} + \frac{1}{2}\text{erf}\frac{(\ln(X_{cum}) - \mu)}{\sqrt{2}\sigma}, \tag{4}$$

where erf denotes the error function.

The final third presentation of the single Lognormal function $f_{sLogn,log}(X_{log}|\tilde{\mu},\tilde{\sigma})$:

$$f_{sLogn,log}(X_{log}|\tilde{\mu},\tilde{\sigma}) = \frac{1}{\tilde{\sigma}\sqrt{2\pi}} \exp\frac{-(X_{log} - \tilde{\mu})^2}{2\tilde{\sigma}^2} \tag{5}$$

Notably, the third presentation originates from the property that data X_{lin} were lognormally distributed with the logarithm $X_{log} = \ln(X_{lin})$ being normally distributed with $\tilde{\mu} = \mu + \frac{1}{2}(\sigma^2 - \sigma^4)$ and $\tilde{\sigma} = \sigma^2$.

Our analysis pipeline exploited these three representations of the Lognormal function. First, the data were read into MATLAB and pre-processed by applying a cutoff value to remove very large transit time values, accounting for bleaching and noise. To compute the sFCS histograms, the data sets were binned into 50 equally distributed bins, and then normalized to probability distribution functions (pdfs) for the linear and logarithmic representations, and normalized to cumulative distribution functions (cdfs) in the case of the cumulative representation, respectively. For the histogram fitting, the midpoints of the histogram data bars for all bins were extracted (*e.g.* for the linear representation the i-th data point would be $X_{lin,i}, y_{data_{lin,i}}$).
To increase fitting accuracy, the transit time histograms were fitted in all three representations. First, the cumulative histogram, then the histogram of the logarithmic data and finally the linear histogram were fitted to the respective form of the lognormal distribution. The fitting strategy is exemplified in Figure S2b for free diffusion and in Figure S3c for hindered trapped diffusion. After obtaining the first values from the cumulative only 20 % variation for μ and σ is allowed when fitting the other representations. Ultimately for free diffusion, we have three fits to determine just two parameters.
The μ value is closely related to the transit time. As a measure for evaluation we defined value recovery as $\frac{\tau_{D,in} - \tau_{D,out}}{\tau_{D,in}}$ where $\tau_{D,in}$ is the input transit time (input value for the simulation) and $\tau_{D,out}$ is the recovered value (either by averaging or fitting).

For hindered trapped diffusion, a combination of weighted two lognormal functions was fitted to the transit time histograms. The three distributions are given as:

$$\begin{aligned} f_{dLogn,cum}(X_{cum}|\mu_1,\sigma_1,\mu_2,\sigma_2,A) =& A \cdot (\frac{1}{2} + \frac{1}{2} \cdot \text{erf}\frac{(\ln(X_{cum}) - \mu_1)}{\sqrt{2}\sigma_1}) + \\ & (1-A) \cdot (\frac{1}{2} + \frac{1}{2} \cdot \text{erf}\frac{(\ln(X_{cum}) - \mu_2)}{\sqrt{2}\sigma_2}) \end{aligned} \tag{6}$$

$$\begin{aligned} f_{dLogn,log}(X_{log}|\tilde{\mu}_1, \tilde{\sigma}_1, \tilde{\mu}_2, \tilde{\sigma}_2, A) =& A \cdot (\frac{1}{\tilde{\sigma}_1\sqrt{2\pi}} \exp \frac{-(X_{log}-\tilde{\mu}_1)^2}{2\tilde{\sigma}_1^2})+ \\ & (1-A) \cdot (\frac{1}{\tilde{\sigma}_2\sqrt{2\pi}} \exp \frac{-(X_{log}-\tilde{\mu}_2)^2}{2\tilde{\sigma}_2^2}) \end{aligned} \tag{7}$$

$$\begin{aligned} f_{dLogn,lin}(X_{lin}|\mu_1, \sigma_1, \mu_2, \sigma_2, A) =& A \cdot (\frac{1}{X_{lin}\sigma_1\sqrt{(2\pi)}} \exp -\frac{(\ln(X_{lin})-\mu_1)^2}{2\sigma_1^2})+ \\ & (1-A) \cdot (\frac{1}{X_{lin}\sigma_2\sqrt{(2\pi)}} \exp -\frac{(\ln(X_{lin})-\mu_2)^2}{2\sigma_2^2}), \end{aligned} \tag{8}$$

where A accounts for the weighting factor of the two Lognormal distributions represented by μ_1, σ_1 and μ_2, σ_2, respectively.

In the case of hindered hop diffusion, we exploited a similar strategy but adapting the pre- processing of the sFCS data to a two-component FCS fitting model. The criterion to accept a two-component fit as justified was a contribution of > 10 % of a second process to the fit (meaning $0.1 < A_1 < 0.9$). Transit times below 1 ms and larger 300 ms at the resolution limit of sFCS with 2081 Hz were neglected in the statistical analysis. Subsequently the data were histogrammed, and the histograms were fitted to a combination of fast exponential and a slower lognormal probability distribution function starting with the cumulative representation, over the logarithmic ending with the linear form. The three weighted distributions are given as:

$$\begin{aligned} f_{LognExp,cum}(X_{cum}|\mu_1, \sigma_1, \mu_2, A) =& A \cdot (\frac{1}{2} + \frac{1}{2} \cdot \mathrm{erf} \frac{(\ln(X_{cum})-\mu_1)}{\sqrt{2}\sigma_1})+ \\ & (1-A) \cdot (1 - \exp \frac{-X_{cum}}{\mu_2}) \end{aligned} \tag{9}$$

$$\begin{aligned} f_{LognExp,log}(X_{log}|\tilde{\mu}_1, \tilde{\sigma}_1, \tilde{\mu}_2, A, C) =& A \cdot (\frac{1}{\tilde{\sigma}_1\sqrt{2\pi}} \exp \frac{-(X_{log}-\tilde{\mu}_1)^2}{2\tilde{\sigma}_1^2})+ \\ & (1-A) \cdot (\frac{-1}{\mu_2} \cdot X_{log} + \ln(\frac{1}{\mu_2})) + C \end{aligned} \tag{10}$$

$$\begin{aligned} f_{LognExp,lin}(X_{lin}|\mu_1, \sigma_1, \mu_2, A) =& A \cdot (\frac{1}{X_{lin}\sigma_1\sqrt{(2\pi)}} \exp -\frac{(\ln(X_{lin})-\mu_1)^2}{2\sigma_1^2})+ \\ & (1-A) \cdot (\frac{1}{\mu_2} \cdot \exp \frac{-X_{lin}}{\mu_2}), \end{aligned} \tag{11}$$

where A was a weighting factor for the two distributions with μ_1, σ_1 for the Lognormal and μ_2 for the exponential, respectively. C is an offset.

Fitting quality and model selection

The quality of the histogram fitting process was evaluated with two different strategies for both the single and double Lognormal function. First for visual inspection of the fitting quality, the weighted residuals were computed for each individual fit.They were calculated according to:

$$Residuals = \frac{(y_{data,i} - y_{fit,i})}{\sqrt{(y_{fit,i})}}, \tag{12}$$

with $y_{data,i}$ as the i-th y-values extracted from the histogram at the i-th bin and $y_{fit,i}$ as the i-th recalculated y-value using μ and σ from the fits (for example with $X_{lin,i}$).

Second, the goodness of fit value was also used to evaluate improvement in fitting when varying parameters, or applying a more complex model. Notably, the single Lognormal model over a simple Gaussian model, or the double Lognormal model over the single-Lognormal model produces significant changes in *Goodness of Fit* (compare for example Figure S3c). The goodness of the fit was calculated according to:

$$Goodness\ of\ Fit = \sum_{i=1}^{n} \frac{(y_{data,i} - y_{fit,i})^2}{y_{fit,i}} \tag{13}$$

For the selection of a model function, we utilized maximum likelihood estimation (MLE) to fit the linear representation of the histograms (compare Figure S4). The obtained estimators $\hat{\mu}_{1,2}$ and $\hat{\sigma}_{1,2}$ should be similar to the values obtained from our fitting strategy ($\mu_{1,2}$ and $\sigma_{1,2}$).

From the MLE we also obtained the optimized logarithmic likelihood vector for the Bayesian Information Criterion (BIC) analysis. The BIC value provided a statistical measure to evaluate which model represented the data best, applies penalties for the introduction of more fitting parameters, and was hence used to decide for a fitting model (free or hindered diffusion). In this way, the transit times were evaluated for example being Gaussian, single- or double-Lognormal distributed. The log-likelihood function for a single Gaussian is given by:

$$\begin{aligned} \ln \mathcal{L}_{Gauss}(\mu, \sigma^2) =& \sum_{i=1}^{n} \ln f_{Gauss}(x_i|\, \mu, \sigma^2) = \\ & -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 \end{aligned} \tag{14}$$

The log-likelihood function for a single-lognormal function is given by:

$$\begin{aligned} \ln \mathcal{L}_{sLogn}(\mu, \sigma^2) =& \sum_{i=1}^{n} \ln f_{sLogn,lin}(x_i|\, \mu, \sigma^2) = \\ & -\frac{n}{2}\ln(2\pi\sigma^2) - \sum_{i=1}^{n}\ln(x_i) - \frac{\sum_{i=1}^{n}\ln(x_i)^2}{2\sigma^2} + \frac{\sum_{i=1}^{n}\ln(x_i)\mu}{\sigma^2} - \frac{n\mu^2}{2\sigma^2} \end{aligned} \tag{15}$$

Accordingly, the log-likelihood function for a double-lognormal function is given by:

$$\begin{aligned}\ln\mathcal{L}_{dLogn}(\mu_1,\sigma_1^2,\mu_2,\sigma_2^2,A) &= \sum_{i=1}^{n}\ln f_{dLogn,lin}(x_i|\mu_1,\sigma_1^2,\mu_2,\sigma_2^2,A) = \\ -\frac{n}{2}\ln(2\pi)+\sum_{i=1}^{n}(-\ln(x_i))+\sum_{i=1}^{n}\ln(A\sigma_1^{-1}\exp(-&\frac{(\ln(x_i)-\mu_1)^2}{2\sigma_1^2})+ \\ (1-A)\sigma_2^{-1}\exp(-&\frac{(\ln(x_i)-\mu_1)^2}{2\sigma_1^2}))\end{aligned} \tag{16}$$

For numerical stability of the double-Lognorm function, we used in practice the alternative equation below incorporating a second weighting factor B which is not depended on A and we constrained the fit to a contribution of a second component. Notably, for the BIC calculation we assumed six fitting parameters.

$$\begin{aligned}\ln\mathcal{L}_{dLogn_{AB}}(\mu_1,\sigma_1^2,\mu_2,\sigma_2^2,A,B) &= \sum_{i=1}^{n}\ln f_{dLogn,lin_AB}(x_i|\mu_1,\sigma_1^2,\mu_2,\sigma_2^2,A,B) = \\ -\frac{1}{2}\ln(A+B)-&\frac{n}{2}\ln(2\pi)+\sum_{i=1}^{n}(-\ln(x_i))+ \\ \sum_{i=1}^{n}\ln(A\sigma_1^{-1}\exp(-\frac{(\ln(x_i)-\mu_1)^2}{2\sigma_1^2})+&B\sigma_2^{-1}\exp(-\frac{(\ln(x_i)-\mu_1)^2}{2\sigma_1^2}))\end{aligned} \tag{17}$$

Instead of maximizing the log-likelihood functions the negative values were minimized using MATLAB's solver fmincon. The optimized likelihood value was fed into MATLAB's built in BIC function aicbic along with the number of sFCS transit times as number of observations and the number of parameters of 2 or 6 for Gaussian and single-Lognormal or for a double-Lognormal, respectively.

Finally, we expanded our analysis to the case of hindered hop diffusion and also included a maximum likelihood estimation for a combination of lognormal and exponential probability distribution function to be able to make use of the BIC for model selection. The log-likelihood function for a weighted combination of Lognormal and Exponential is given by:

$$\begin{aligned}\ln\mathcal{L}_{LognExp}(\mu_1,\sigma_1^2,\mu_2,A) &= \sum_{i=1}^{n}\ln f_{LognExp,lin}(x_i|\mu_1,\sigma_1^2,\mu_2,A) = \\ \sum_{i=1}^{n}\ln(A\cdot x_i^{-1}(\sigma_1^2 2\pi)^{-1/2}\exp(-&\frac{(\ln(x_i)-\mu_1)^2}{2\sigma_1^2})+(1-A)\cdot\mu_2^{-1}\exp(-\frac{x_i}{\mu_2}))\end{aligned} \tag{18}$$

The model with the lowest BIC value represents the data best. To statistically compare the likelihoods of the different models, we used the relative likelihood (RL) values. They were calculated as follows:

$$RL_{\mathrm{i}} = \exp\left((BIC_{\min}-BIC_{\mathrm{i}})/2\right), \tag{19}$$

where RL_{i} represents the relative likelihood value for the i-th model with the the i-th BIC value (BIC_{i}) compared to the most likely model with the lowest BIC ($BIC_{\min}$).[33–35]

Together with the *Residuals* and the *Goodness of Fit* value it was possible to objectively decide whether a probe was undergoing free Brownian diffusion or hindered diffusion in the cellular membrane.

S2 Supporting Figures

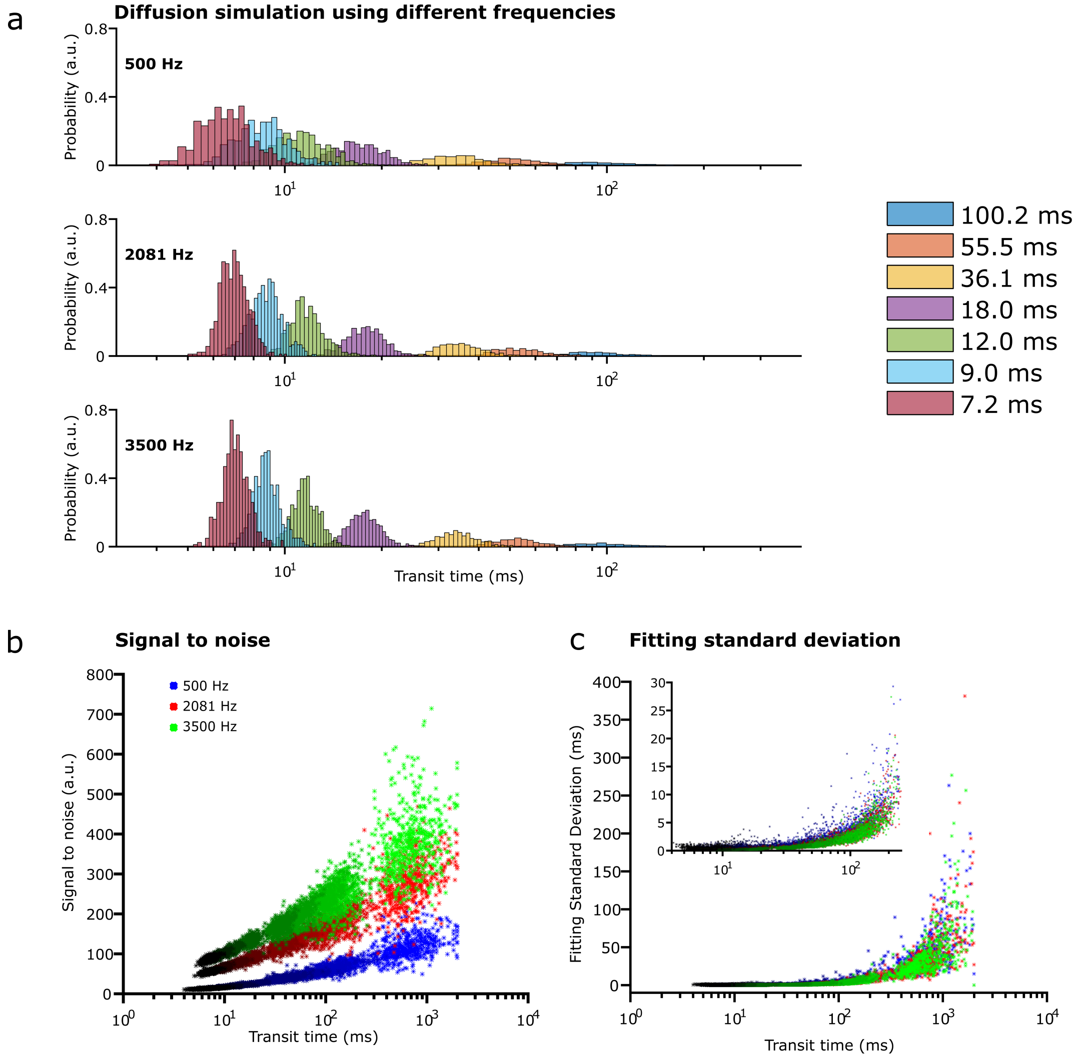

Figure S1: Spread of transit time values from individual FCS data generated from computer-simulated sFCS experiments of free diffusion with different implemented transit times and for different line scanning frequencies. (**a**) Transit time histograms with implemented average transit times and line scanning frequencies given in ms and Hz, respectively, revealing that slow diffusing molecules experience larger errors than faster particles, also depending on the scanning frequency. (**b**) Transit times versus signal-to-noise ratios and (**c**) transit times versus fitting standard deviation (as obtained from bootstrapping in the fitting program, see for details[20]) with different line scanning frequencies as marked and different implemented transit times indicated by dark (fast, low values) to bright (slow, high values) tones, indicating the dependency of the broadness of the distributions and the dis-symmetry towards longer transit times of the distribution on the magnitude of the measured transit time values, which results from the fact that the error of the measurements increases exponentially with the transit times accompanied by respective changes in the signal-to-noise (see [20] for details).

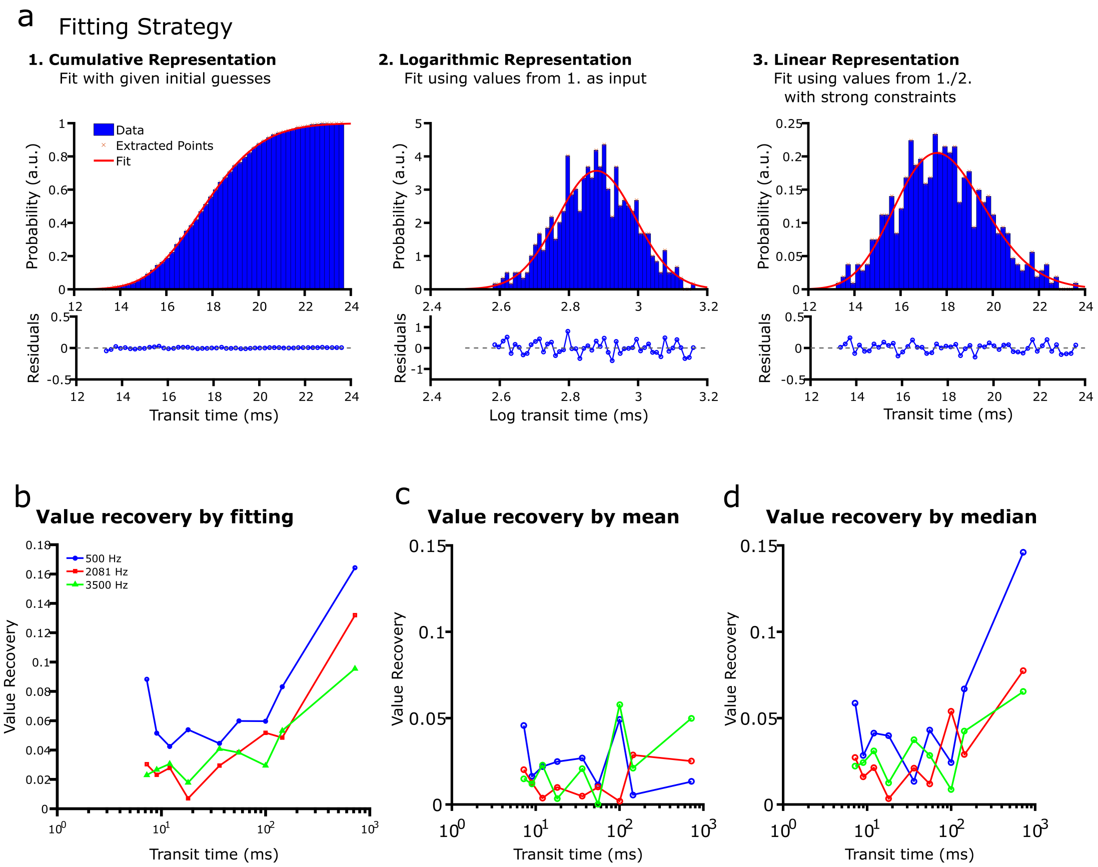

Figure S2: Fitting strategy and value recovery of computer-simulated diffusion data. (**a**) Fitting strategy: step1, LogNorm fit against cumulative transit time histogram (left, blue: data from computer-simulations of free diffusion; red: fit; lower panel: weighted residuals) with arbitrary start values of μ and σ; step 2, LogNorm fit against logarithmic transit time histogram with results of step 1 as input parameters; and step 3, LogNorm fit against linear transit time histogram with results of step 2 as input parameters and employing strong constraints on the fit (see **Supporting Information**). (**b-d**) Comparison of average transit time values as implemented in the computer-simulations and those resulting from the LogNorm fits to the transit time distributions (b, fitting), from the mean (c, mean) and median (d, median) of the transit time distributions for computer-simulated data of free diffusion for different transit times (x-axis) and line scanning frequencies (colored as labelled). Plotted are the value recovery = (transit time (fit) – transit time (implemented))/transit time (implemented). Usual errors are less than 5 %, and only up to 20 % for large transit times.

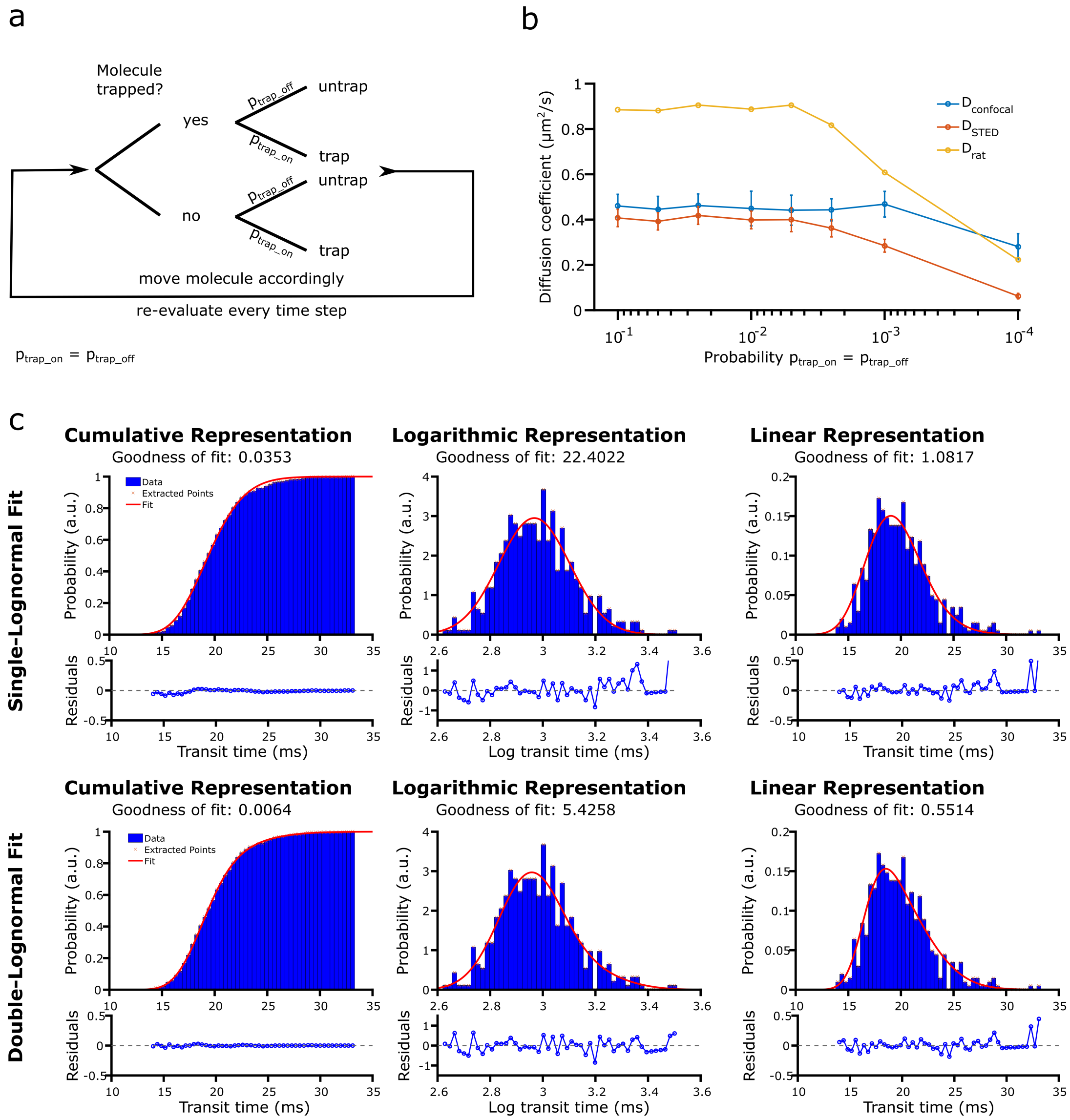

Figure S3: Algorithm and analysis of computer-simulated trapped diffusion data. (**a**) Principle of computer-simulation algorithm as described in the Methods section. (**b**) Ratio $D_{rat} = D_{STED}/D_{conf}$ of diffusion coefficients as obtained from the analysis of FCS computer-simulated trapped diffusion data with different trapping/untrapping probabilities p_{trap_on} and p_{trap_off} (x-axis); values of D_{STED} and D_{conf} were determined for observation spot sizes with FWHM = 80 nm (simulating super-resolved STED-FCS measurements) and FWHM = 200 nm (simulating confocal conditions), and only values $D_{rat} < 1$ indicate a significant and observable trapping extent. This allowed to identify accurate conditions for computer-simulations of trapped diffusion (we used $p_{trap\_on} = p_{trap\_off} = 0.001$ in our sFCS simulations (see **Methods**). (**c**) Cumulative (left), logarithmic (middle) and linear (right) transit time histograms of computer-simulated sFCS data of trapped diffusion (blue) with single- (upper panels) and double- (lower panels) LogNorm fits (red) with weighted residuals (respective lower panels) and Goodness-of-fit values, highlighting the more accurate description of the data using the double-LogNorm fits.

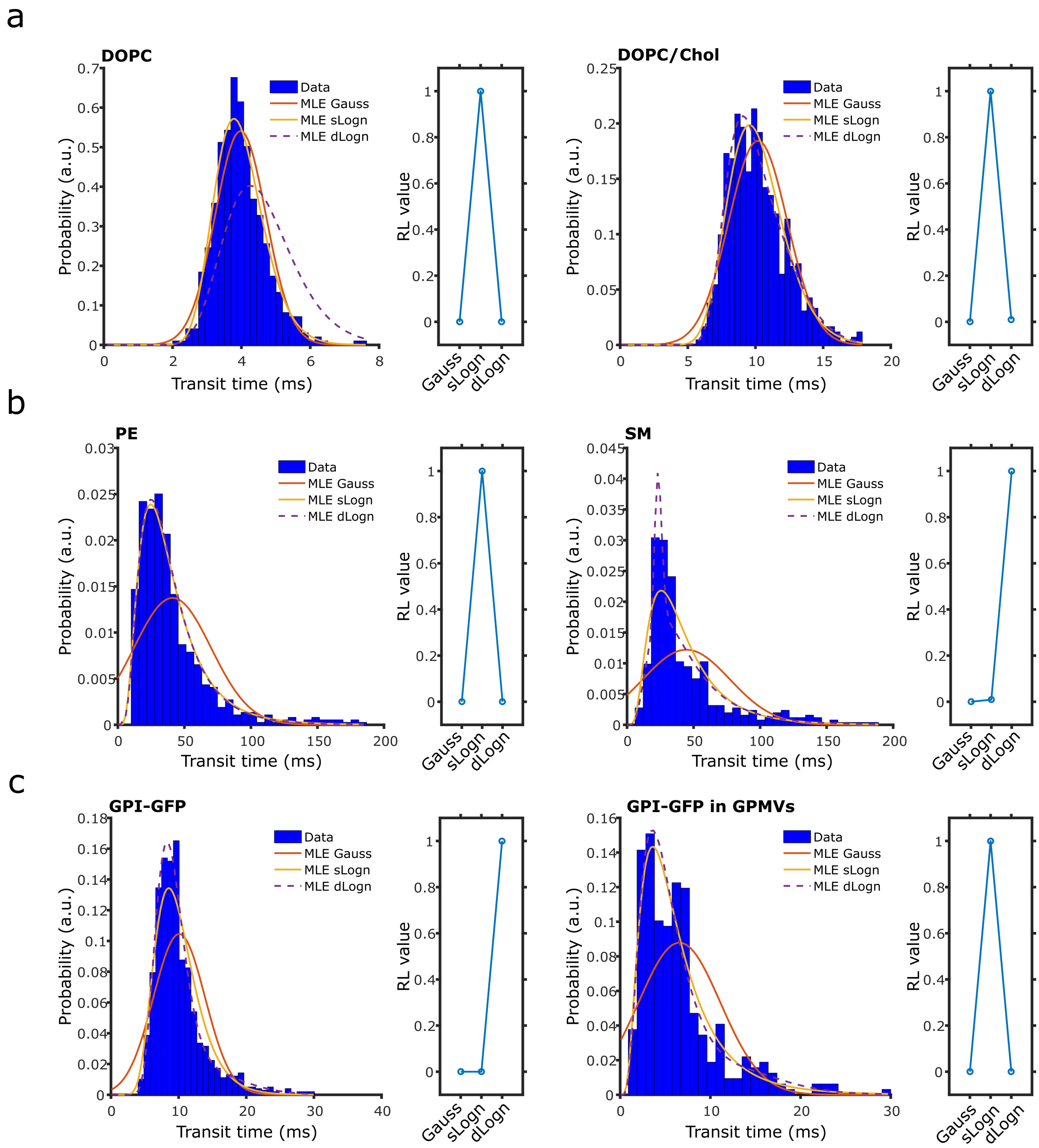

Figure S4: Maximum likelihood estimation of LogNorm fits to experimental data. Linear transit time histograms of sFCS data (blue) from (**a**) DPPE in DOPC (left) and DOPC/Chol (right) SLBs, (**b**) DPPE (left) and SM (right) in live PTK2 cells, and (**c**) GPI-GFP in live PTK2 cells (left) and GPMVs derived thereof (right). Fits to the data using a purely Gaussian (Gauss maximum likelihood estimation, orange lines), a single-LogNorm (sLogn maximum-likelihood-estimation, yellow lines), and a double-LogNorm (dLogn maximum likelihood estimation, purple dashed lines) are given, as well as RL values from the fits (right insets). The most likely model has a RL value of 1.

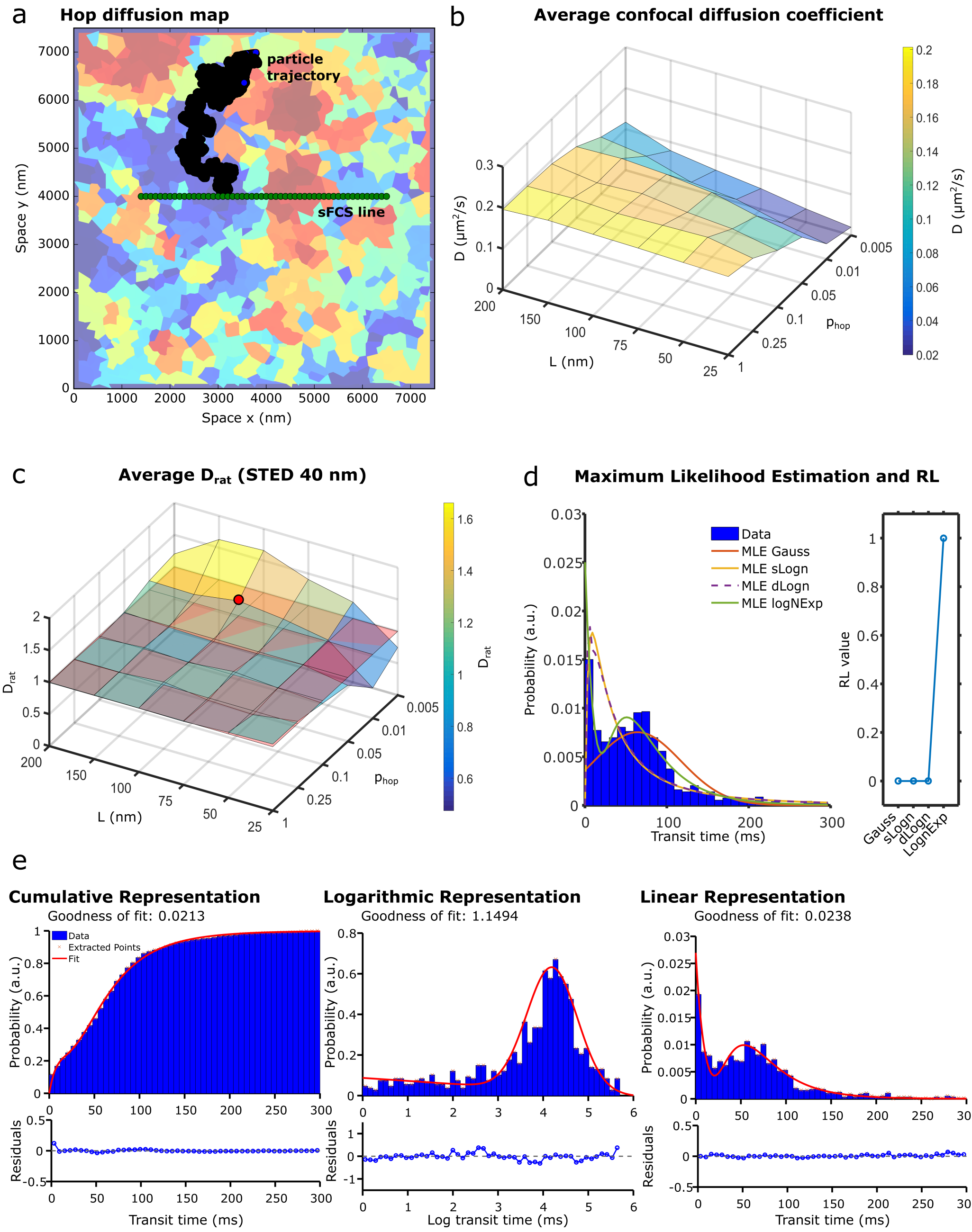

Figure S5: Computer-simulations for hindered hop sFCS diffusion data. (**a**) Exemplification of the hop diffusion simulations with coloured map of the meshwork causing the hindrance, a short particle trajectory (black) and the line for the sFCS measurement (centre of each pixel drawn as a green circle). (**b,c**) Average diffusion coefficient in confocal (**b**) and D_{rat} value (as in **Fig. S3 b**) (**c**) for a parameter estimation over the simulation inputs L and phop (using 30 second simulations, 4 carpets per condition). The red dot at L = 100 nm and $p_{hop} = 0.01$ indicates a region of reasonable hopping ($D_{rat} > 1$) used for the more dedicated simulations (45 seconds acquisition time, 10 carpets; the D_{rat} value for 80 nm STED observation spot diameter was $D_{rat} = 1.12$) analysed by maximum likelihood estimation along with BIC/RL evaluation in (**d**) and statistical analysis of the transit time histograms fitting combination of Lognorm and Exponential probability distribution function together with respective weighted residuals in (**e**). The red plane in (**c**) indicates a D_{rat} value of 1 representing free diffusion.

S3 Supporting Tables

Table S1: Summary of the fitted parameters for the data shown in Figure 1. NA refers to non-applicable.

	slogN		dlogN				
Data set	μ (exp μ)	σ	μ_1 (exp μ_1)	σ_1	μ_2 (exp μ_2)	σ_1	A
Simulation (free diffusion)	2.18 (8.81)	0.19	NA	NA	NA	NA	NA
Simulation (trapped diffusion)	NA	NA	2.93 (18.69)	0.19	3.13 (22.98)	0.21	0.73
SLB DOPC	1.44 (4.23)	0.41	NA	NA	NA	NA	NA
SLB DOPC/Chol	2.27 (9.72)	0.30	NA	NA	NA	NA	NA

Table S2: Summary of the fitted parameters for the data shown in Figure 2. NA refers to non-applicable.

	slogN		dlogN				
Data set	μ (exp μ)	σ	μ_1 (exp μ_1)	σ_1	μ_2 (exp μ_2)	σ_1	A
DPPE	3.40 (30.08)	0.68	NA	NA	NA	NA	NA
SM	NA	NA	4.10 (60.30)	0.69	3.24 (25.61)	0.42	0.40
GPI-GFP (Cells)	NA	NA	2.44 (11.42)	0.53	2.15 (8.54)	0.32	0.32
GPI-GFP (Cells)	1.59 (4.92)	0.74	NA	NA	NA	NA	NA