Paving the Way for Structural Modelling by smFRET Measurements

kumulative Dissertation zur Erlangung des Doktorgrades Dr.rer.nat. im Fachbereich Physik der Fakultät für Naturwissenschaften der Universität Ulm

vorgelegt von

Tobias Eilert

aus Bottrop

2018
Abstract

The thesis at hand contributes to the field of structure determination of biomacromolecules. It focuses on the Förster resonance energy transfer (FRET), which allows to infer distances at the molecular level (≈ 2 – 10 nm) making FRET a promising tool for structure analysis. In the process of FRET, the energy of a photon, absorbed by a fluorophore, called donor, is transferred to a second dye, called acceptor. FRET experiments result in an average FRET efficiency which in the first place depends on the distance between donor and acceptor. If the two dyes are attached to specific sites of a macromolecule, we can infer intramolecular distances not only in vitro, but even in vivo. Furthermore, FRET can be observed on a single-molecule level in real time. Thus, different conformations and their dynamics of one macromolecule are observable and distinguishable by single-molecule FRET (smFRET) experiments. These features encouraged the development of the Nano-Positioning System (NPS). The basic idea of NPS is that we can localise an unknown position unambiguously, if we know its distances to at least four known positions. Given a network of dyes, NPS allows thus for the dependent localisation of unknown positions of fluorophores, such that the structure of macromolecular complexes can be illuminated by FRET measurements. As a Bayesian analysis tool, the theoretic basis of NPS is a probability distribution, the so-called posterior. It tells us how probable a spatial arrangement of dyes in the network is conditional on the experimental smFRET efficiencies measured between the dyes. Hence, the present thesis deals on the one hand with the development and implementation of a fast and adaptive sampling algorithm which extracts the structural information hidden within the posterior leading to the release of Fast-NPS. On the other hand, we promoted smFRET as a structure analysis method by providing a manual guide of how to perform smFRET measurements at a TIRF microscope and how to subsequently infer a structure in using the software Fast-NPS. This is not only presented in a manuscript, but also in a how-to video.

A major difficulty in the inference of structures by smFRET measurements is the transformation of the smFRET efficiency to a distance. An important factor here is how the fluorophores are described. The publications of the cumulative thesis at hand report the progressing development of dye models from one single conservative model to a set of primitive models incorporating different assumptions. For the application of the latter, we developed a consistency test such that we can test, if the experimental smFRET efficiency is still in accordance with the applied dye models. This sophisticated procedure led to a 4-fold enhancement of localization precision on average.

In a benchmark study, we applied these models to analyse smFRET data obtained from dsDNA and a protein-DNA complex at a TIRF microscope. We could show that Fast-NPS infers the correct position in both cases. However, we saw that some models are too strict in their assumptions, such that the
inferred structures become inconsistent with the smFRET data. A further concern is that we could also find model combinations that do not infer the correct positions, but still are consistent with the experimental data. The latter problem guided us to a more sophisticated theory of dye models. We understood that the FRET efficiency is inseparably determined by the geometry and kinetics of both dyes. This leads to a complete description of the smFRET efficiencies by expectation values. For dyes whose motions can be treated approximately as time-independent, we can compute the expected value of the smFRET efficiency by Monte Carlo integration. Completing the theory, we also established a stochastic simulation in order to calculate the expected value in the case of time-dependent motions. We accomplished this by explicitly simulating the rotational and translational diffusion of donor and acceptor simultaneously to the donor de-excitation. It is important to note that these simulations are guided by experimental parameters obtained from fluorescence lifetime and time-resolved anisotropy measurements.

In further utilizing the stochastic simulation, we developed a statistical method to classify the motions of donor and acceptor during a single de-excitation event into one of the so-called transfer regimes telling us how to average the motion. Applying this method, an investigation of dyes attached to dsDNA showed us that their kinetic assumptions, usually applied in structural inference, are not valid over the whole distance range inferred by FRET. This has a great impact on the future analysis of smFRET experiments.
Kurzbeschreibung

Die Hauptschwierigkeit bei der Inferenz von Strukturen mittels Einzelmolekül-FRET-Experimenten ist die Umwandlung der Einzelmolekül-FRET-Effizienzen zu einer Distanz. Ein wichtiger Faktor hierbei ist wie die Fluorophore beschrieben werden. Die Publikationen der vorliegenden kumulativen Thesis berichten die fortschreitende Entwicklung der Modelle zur Beschreibung der Farbstoffe von einem einzigen konservativem Model zu mehreren primitiven Modellen, die verschiedene Annahmen integrieren. Für die Anwendung der primitiven Modelle haben wir einen Konsistenz-Test entwickelt, der uns erlaubt zu testen ob die experimentellen FRET-Effizienzen mit den Farbstoff-


In einem weiteren Schritt nutzten wir die zeitabhängige Simulation um eine statistische Methode zu entwickeln mit der wir die Bewegungen von Donor und Akzeptor während des Deaktivierungsprozesses in eines der sogenannten Transfer Regime zu klassifizieren. Diese Regime sagen uns dann wie die Farbstoff-Bewegungen gemittelt werden müssen. Die Anwendung dieser Methode auf an dsDNA gebundene Farbstoffe zeigte, dass die für die Strukturanalyse üblicherweise angewandten Annahmen über die Kinetik der Farbstoffe nicht zutreffend sind, und zwar über den gesamten Distanz-Bereich den wir mit FRET inferrieren. Dies hat einen großen Einfluss auf die zukünftige Analyse von Einzelmolekül-FRET-Experimenten.
Contents

1 Overview of Publications 1

2 Introduction 3

3 Background 7
   3.1 Physical Background ........................................ 7
      3.1.1 Fluorescence ............................................ 7
      3.1.2 Förster Resonance Energy Transfer ...................... 7
      3.1.3 Time-resolved Anisotropy ............................... 11
      3.1.4 Measurement of single-molecule FRET Efficiency ...... 13
   3.2 Mathematical Background ................................. 14
      3.2.1 Stochastic Processes .................................. 14
      3.2.2 Fundamentals of Monte Carlo Methods ................. 16
      3.2.3 Markov Chains and MCMC Algorithms ................ 18

4 Publications 21
      Chain Monte Carlo-based analysis tool to obtain structural in-
      formation from single-molecule FRET measurements ....... 21
   4.2 J. Vis. Exp. 120, 2017: Structural Information from Single-
      molecule FRET Experiments Using the Fast Nano-positioning
      System ................................................................... 37
   4.3 J. Chem. Phys. 148, 2018: Precision and accuracy in smFRET
      based structural studies - a benchmark study of the Fast-Nano-
      Positioning System ............................................... 53

5 Summary and Discussion 87

6 Conclusion and Outlook 93

7 Appendix 97
   List of Figures .................................................. 101
   References ...................................................... 103
   List of Publications ........................................... 107
   List of Conferences ........................................... 109
   Curriculum Vitae ................................................ 111
   Danksagung ..................................................... 113
1 Overview of Publications

This cumulative thesis concerns the following publications:

[i] T. Eilert, M. Beckers, F. Drechsler and J. Michaelis.
Fast-NPS – A Markov Chain Monte Carlo-based analysis tool to obtain structural information from single-molecule FRET measurements.
Computer Physics Communication **219**, 2017

Structural information from single-molecule FRET experiments using the Fast-Nano-Positioning System.
Journal of Visual Experiments **120**, 2017

Precision and accuracy in smFRET based structural studies - a benchmark study of the Fast-Nano-Positioning System.
Journal of Chemical Physics **148**, 2018

Complete kinetic theory of FRET.

Single-molecule FRET experiments on PARP-1 binding to DNA single-strand breaks.
In preparation and discussed in Section 6
Physics is the natural science that tries to describe nature by relating experiments to theoretical models, which are able to predict the future behaviour of phenomenons of interest. A prerequisite here is that we can conduct experiments in order to observe the phenomenon directly or indirectly. The major problem in understanding life at the molecular level is simply that we cannot see it. The field of biophysics tries to visualise cells and their compartments, and even localize single molecules. The Nobel prize in chemistry in the year 2014 was awarded for the latter, i.e. for the development of super-resolved fluorescence microscopy. Much more challenging, though, is the elucidation of the structure of macromolecules themselves. A breakthrough was achieved by X-ray crystallography, which is able to resolve the structure up to an Ångström level. However, a major disadvantage of this technique are the necessary crystals of the macromolecules. Oftentimes it takes long or it is simply not possible to crystallise the macromolecule of interest, e.g. because of size and/or conformational instability.

In single-particle cryogenic electron microscopy (cryo-EM), macromolecules are rapidly frozen in a thin aqueous layer and imaged at cryogenic temperature in a transmission electron microscope. Thus, the sample preparation is not limited by macromolecular weight or other factors, such that even the ribosome can be resolved. However, the resolution was considerably behind the X-ray crystallography. Recent technical progress concerning direct electron detectors in 2012 brought cryo-EM to the focus of structural biologists again, such that its development was awarded by a Nobel prize in Chemistry in 2017, too.

Although the sample preparation of both techniques are different, they have common weaknesses. First, the macromolecules of interest are in an artificial environment or state. The crystal environment may alter the overall or local configuration. While the macromolecules are in an aqueous solution in cryo-EM, the rapid freezing may also introduce structural artifacts. Thus, it is possible that in both methods the obtained structure does not reflect the truth. Second, since both techniques average over many copies of the same molecule, which may occupy several or a continuum of configurations, flexible parts may be distorted or even not resolvable. Although, especially in cryo-EM, there evolved specialized algorithms to entangle conformational states over the past years, transient conformations of the macromolecular complexes and their dynamics during reactions cannot be observed directly in real time.

The so-called Förster resonance energy transfer (FRET) can help to overcome these shortcomings. FRET is a quantum mechanical phenomenon in which a photon is absorbed by a fluorophore, called donor, and its energy is transferred to a second fluorophore, called acceptor. The transfer process is based on a dipole-dipole near field interaction without occurrence of a photon. FRET experiments lead to the so-called FRET efficiency that in the first place depends on the distance between donor and acceptor. If the two fluorophores are attached to specific sites of a macromolecule, we can infer distances within macromolecules in vitro or even in vivo. This can be used to complement or
even expand structures obtained from X-ray crystallography or cryo-EM. Furthermore, FRET can be observed on a single-molecule level in real time. Since FRET is performed on a non-fixed substrate, i.e. the macromolecule of interest is neither in a crystal nor frozen, all – even transient – conformations are observable and distinguishable by single-molecule FRET (smFRET) experiments.\textsuperscript{10,11,12} However, smFRET is not limited to determining distances, but it can also provide information about the dynamics of conformational changes and thus is an excellent tool in order to unravel the mechanism of biomolecular systems.\textsuperscript{13}

All of these advantages make smFRET a promising experimental tool for analysing biomolecular complexes which are inherently flexible in structure. Over the last decade two different approaches were presented in order to utilise smFRET measurements for structure determination of biomacromolecules and their complexes: the Nano-Positioning System (NPS)\textsuperscript{14} and the FRET-restrained positioning and screening (FPS).\textsuperscript{15} In both approaches, several smFRET measurements are performed between a set of dyes within the complex of interest. A special feature here is that the fluorophores are attached to the macromolecule by flexible linkers. These introduce a lack of knowledge about the precise conversion of the measured FRET efficiency to the distance between the dyes. Although, the general principle of the two programs is similar, they differ strongly in their assumptions regarding this transformation. And it is intuitively clear that this transformation is crucial for the correct structural inference of biomolecular complexes.

Until the beginning of this thesis in February 2015, there was a limited amount of models describing the behaviour of the dyes involved in a FRET measurement. In the 1970s, there was a heated discussion about the rotational dynamics of the dyes attached to a macromolecule by flexible linkers and its influence on FRET.\textsuperscript{16,17} These two models assume the opposite speed limits of rotational motion: very slow and very fast. Only the fast – which was considered the more likely assumption – considers geometrical constraints on rotations. Only more than 40 years later – when numerical approaches for structural inference became feasible – the uncertainty in dye position became a focus of research.\textsuperscript{14,18} However, it was the common opinion that the dyes do not translate or at least very slow. As these assumptions are crucial for the success of structure determination by FRET, one of the main topics of the thesis at hand is the ongoing development of progressively more sophisticated dye models. Their detailed description and implementation can be found in the publications 4.1 and 4.4.

The Nano-Positioning System was invented in the laboratory of Jens Michaelis.\textsuperscript{14} It is the main issue of this thesis. Since it uses a Bayesian approach, the structural information implied by the FRET data is encoded in a probability distribution, the so-called posterior. This information is given in the form of expected values, marginal distributions and credible volumes of dye positions. These can be easily computed, if we have a set of realizations drawn from the posterior. This can be done by a so-called sampler. Since structural analysis may include testing several geometries and various dye models, the sampling
engine of NPS needs to be fast. Further, we require that the sampler is completely adaptive. The user should not need to worry about algorithmic details, because his/her only focus should lie on answering the biochemical questions. Hence, a major part of the thesis covers the development of a robust, fast and adaptive sampling engine what led to the release of Fast-NPS. The in-depth description of its development can be found in the publication 4.1.

Although, the user can treat the sampling engine as a black box, the structure determination of a biomolecular complex, beginning with the smFRET measurements and ending in the data analysis by Fast-NPS, is a complicated procedure. Scientists, who are new to Fast-NPS or even to smFRET, but are interested in the structural analysis of a particular biomolecule, might be scared off by the complexity of the experiment and software. The first fear of contact can be usually reduced by a step-by-step manual, which can be found in the publication 4.2.

Every analysis tool needs to be validated, before it is declared trustworthy to be used. This is usually done in a benchmark study. Here, a problem is synthetically created, such that we know its solution. Subsequently, the data is obtained experimentally and analysed by the software which is to be tested, as if the answer is not known. In our case we designed two test macromolecules, acquired the corresponding smFRET efficiencies and did a detailed analysis with Fast-NPS. We did not only test the correctness of the sampling engine, but also investigated the application of the developed dye models. The complete benchmark study can be found in the publication 4.3.

After we have motivated the cumulative thesis at hand, we will present the physical and mathematical background information which is necessary to understand the publications in their complete entity.
3 Background

The topics addressed in the cumulative thesis at hand range from the physics of spectroscopy over computer simulations to stochastic calculus and statistics. In order to understand the successive development of the presented theory, it is necessary to give a general overview of the used methods. This information is summarised in the following sections.

3.1 Physical Background

The determination of macromolecular structures by Fast-NPS is based on the analysis of fluorescence decays, time-resolved anisotropy decays and FRET measurements. The review of their underlying physics is unavoidable in order to understand their applications in the presented theory. It is important to mention that we review here the state of the FRET theory as it was at the beginning of this thesis (February 2015), because the included publications establish step by step an advanced and complete mathematical framework of the FRET process. Most of the presented information can be found in [19] or will be referenced on the spot.

3.1.1 Fluorescence

Fluorescence is the spontaneous emission of a photon due to the transition of a molecule from a higher excited electronic singlet state (e.g. $S_1$) to its electronic ground state ($S_0$). This process is depicted in a Jabłoński diagram (on the left of Figure 1). The fluorescence over time is usually an exponential decay which is completely described by a characteristic lifetime or rate, $\tau = k^{-1}$.

For the interpretation of such an experiment, it is important to understand that the rate of the fluorescence decay is the sum of the rates of all processes depopulating the excited state of the fluorophore, i.e. it serves as an indicator for all de-exciting processes, not only for emission.

3.1.2 Förster Resonance Energy Transfer

Förster resonance energy transfer (FRET) is the process of radiation-free energy transfer from one fluorophore, the so-called donor, to another, the acceptor. This electrodynamic phenomenon is due to a dipole-dipole interaction between the two fluorophores. An important factor for the probability to transfer a photon, the so-called FRET efficiency, is the relative orientation of the transition dipole moments of donor and acceptor, but most importantly their distance. The latter property is the reason for the majority of applications for FRET. Because distances distinguishable via FRET are in the range of the diameter of common macromolecular complexes, FRET became rapidly established as a standard method in life sciences and medical diagnostics.

A figurative description of FRET can be provided by a Jabłoński diagram (see Fig. 1). On the left side, it depicts the transition of the donor from the electronic ground state $S_0^D$ to $S_1^D$ due to absorption of a photon with frequency...
After vibrational relaxation, the donor emits a photon of lower energy with frequency $\nu_{\text{em}}^D$, de-excites to the ground state by non-radiative processes\(^a\) or transfers the energy to the acceptor dye which would be therefore excited. In the latter case the acceptor relaxes then to its vibrational ground state of $S_1^A$, and subsequently may de-excite via non-radiative processes (not shown) or by emission of a photon with frequency $\nu_{\text{em}}^A$. Because excitation by a photon or by non-radiative energy transfer typically elevates the molecules to higher vibrational energy levels, which relax much faster than the respective excited electronic states, the emission of the donor is spectrally separable from donor excitation and acceptor emission, meaning $\nu_{\text{exc}} > \nu_{\text{em}}^D > \nu_{\text{em}}^A$.

\[ k_T = k_{D0} \frac{R_0^6}{d^6}, \]  \hfill (1)

where $k_{D0}$ is the characteristic rate of the donor depopulation in the absence of the acceptor. The Förster radius, $R_0$, is the distance at which half of the number of donor de-excitations lead to transfer, i.e. $k_T = k_{D0}$\(^8\). The Förster radius depends on the spectral overlap $J$ of donor emission and acceptor absorption spectra, the quantum yield $\Phi_{D0}$ of the donor in the absence of the acceptor, the refractive index $n$ of the medium/environment in which FRET is taking

\(^a\) As non-radiative processes we summarise internal conversion, intersystem crossing and other radiation-free quenching phenomena.
place and the orientation factor $\kappa^2$, which is a measure for the alignment of
the transition dipole moments of the dyes, i.e. we have

$$R_0^b = 8.79 \cdot 10^{-28} \frac{\kappa^2 \Phi_{D0} J(\lambda)}{n^4}. \quad (2)$$

The spectral overlap can be expressed in terms of the spectral density of the
donor emission fluorescence intensity $F_D(\lambda)$ and the absorption spectrum of
the acceptor represented by its absorption coefficient $\epsilon(\lambda)$ both as a function
of the emission or absorption wavelength $\lambda$, respectively,

$$J(\lambda) = \frac{\int_0^\infty F_D(\lambda) \epsilon_A(\lambda) \lambda^4 d\lambda}{\int_0^\infty F_D(\lambda) d\lambda}. \quad (3)$$

For a FRET pair whose transition dipole moments are fixed, the orientation
factor $\kappa^2$ can be calculated by

$$\kappa^2 = (\cos \Psi_{DA} - 3 \cos \Theta_D \cos \Theta_A)^2 \in [0, 4]. \quad (4)$$

Figure 2 shows the relative geometry, also called configuration of donor emission
and acceptor absorption dipole moments and highlights the angles entering
the equation of the orientation factor. The angle $\Psi_{DA}$ is solely defined by the
transition dipole moments of donor and acceptor and is therefore independent
of their positions in space. The angles $\Theta_{D/A}$ though depend on one of the
transition dipole moments and the vector linking the two dye positions.

It is important to note that all three angles contributing to $\kappa^2$ depend on at
least one transition dipole moment. As a consequence, when both partners
of the FRET pair are unrestricted in rotation and reorient much faster than
the fluorescence lifetime of the donor in the presence of the acceptor, the
orientation factor adopts its unrestricted and isotropic average of 2/3. The
Förster radius under these conditions is called the isotropic Förster radius,
$R_0^{iso}$. Its relation to the Förster radius is given by

$$R_0 = R_0^{iso} \sqrt{\frac{3}{2}} \kappa^2. \quad (5)$$

The major advantage of this quantity is that it depends solely on the spec-
troscopic characteristics of the dyes and their molecular environments, but is
independent of the actual orientations of donor and acceptor. Common values
of the isotropic Förster radius for structure inference lie between 50 and 70 Å,
representing the common length scale at the macromolecular level.

In a FRET measurement we obtain a FRET efficiency, which depends on the
transfer rate – if we again assume a fixed configuration – as follows:

$$E = \frac{k_T}{k_{DA}} \in [0, 1], \quad (6)$$

b We use the term configuration also for the relative geometry of more than two dyes.
Figure 2: Relative geometry of the fluorophores’ transition dipole moments. The positions of the transition dipole moments of donor and acceptor are denoted by $p_D$ and $p_A$, respectively. The transition dipole moments themselves are indicated by arrows and lie in the coloured planes (green: donor, yellow: acceptor). This geometry defines the orientation factor $\kappa^2$ by three angles: $\Theta_{D/A}$ are the angles between transition dipole moment of donor or acceptor, respectively, and the vector connecting donor and acceptor positions. The angle between donor and acceptor transition dipole moments themselves is denoted by $\Psi_{DA}$.

where the rate of donor de-excitation in the presence of the acceptor is given by $k_{DA} = k_T + k_{D0}$. In using equations (1) and (5) we arrive at

$$E = \frac{1}{1 + \frac{2}{3\pi^2} \left( \frac{d}{R_0^6} \right)^6},$$

a formula which explicitly shows the dependence of the FRET efficiency on the distance between donor and acceptor and their relative orientation. For fast and isotropic rotation of both dyes (i.e. $\kappa^2 = 2/3$), a small efficiency means that the donor is far away from the acceptor and a large efficiency that donor and acceptor are close. For the inference of distances, a FRET efficiency around 50% is advantageous, because it is most sensitive to changes in the distance (see Fig. 3). However, the FRET efficiency should at least be in the range of 5 to 95%, since for too short distances Förster’s assumptions break down or the dyes may collide and for large distances the FRET efficiency becomes almost insensitive to a change in distance.

As we already stated and is obvious in equation (7), the FRET efficiency does not only depend on the distance between donor and acceptor, but also heavily on their relative orientation. In order to bind the error on the orientation factor for a given FRET pair, we perform time-resolved anisotropy experiments in addition to the FRET measurements. The next section describes how these
provide us with information about the confinement of the dye rotation and therefore also about the relative orientation factor.

### 3.1.3 Time-resolved Anisotropy

For the structural analysis by Fast-NPS, donor and acceptor need to be attached chemically to the sites of the macromolecular complex of interest. Often times the attachment of commonly hydrophobic dyes on proteins leads to destabilisation of their native conformation. In order to prevent this and avoid artificial results, the dyes are bound to the protein by flexible linkers. This produces a situation which is neither described by rotationally immobile fluorophores nor by fluorophores completely free to rotate, as discussed in the previous section. We assume that the dye is able to rotate, but is constrained. Since not only the linker, but also the macromolecular environment of the attachment site constrains the rotational motion, the degree of confinement can vary from sample to sample. In order to assess this effect experimentally, we perform *time-resolved anisotropy* measurements of each dye attached to the biomolecule. The fluorescence anisotropy \( r(t) \) at time \( t \) is a measure to describe the depolarisation of the emitted photons, when the fluorophores are excited with polarised light. It is defined by

\[
r(t) = \frac{I(t)_\parallel - I(t)_\perp}{I(t)_\parallel + 2I(t)_\perp} \in [0, 0.4],
\]  

(8)
where $I(t)\parallel$ stands for the intensity of the emitted photons parallel and $I(t)\perp$ perpendicular to the polarisation axis of the excitation pulse at time $t$. If we assume coupled rotation, i.e. not only the dye, but also the macromolecule itself rotates, we obtain the following fit formula for the anisotropy decay:

$$r(t) = \left( (r_0 - r_\infty)e^{-t/\rho_{\text{Dye}}} + r_\infty \right) e^{-t/\rho_{\text{Macro}}},$$

where $\rho_{\text{Dye}}$ and $\rho_{\text{Macro}}$ are the rotational correlation times of the dye and macromolecule, respectively. The absorption dipole moment may not coincide with the emission dipole moment due to redistribution of the electron density. This instantaneous depolarisation is measured by the fundamental anisotropy, $r_0 := \lim_{t \to 0} r(t)$. In contrast, the limiting anisotropy, $r_\infty := \lim_{t \to \infty} r(t)$, gives us information about the rotational constraint of the dye itself. If the fluorophore is not hindered in rotation, the dye’s fluorescence is completely depolarised, i.e. we have $r_\infty = 0$. In case we have a non-zero limiting anisotropy, we have evidence that the dye is constrained in rotation.

Dale and Eisinger developed a model that assumes that the fluorophores do not translate during FRET, i.e. we have just one distance, but reorient rapidly under the constraints of the macromolecule. The latter assumption requires that the fluorescence lifetime of the donor in the presence of the acceptor must be much longer than the rotational correlation time of the dye, i.e. $\tau_{DA} \gg \rho_{\text{Dye}}$. This limit leads to a dynamic average of the transfer rate,

$$\langle k_T \rangle = k_D \frac{\langle R_0^6 \rangle}{d^6},$$

such that the dynamic average FRET efficiency can be expressed by

$$E = \frac{1}{1 + d^6/\langle R_0^6 \rangle},$$

where $\sqrt[6]{\langle R_0^6 \rangle}$ is the dynamically averaged Förster radius. With regard to equation 5, $\langle R_0^6 \rangle$ can be calculated as

$$\langle R_0^6 \rangle = \langle R_0^{\text{iso}} \rangle^6 \frac{3}{2} \langle \kappa^2 \rangle.$$

This means if we know the dynamic average orientation factor $\langle \kappa^2 \rangle$ under the molecular constraints for donor and acceptor, we have a one-to-one relationship between the average FRET efficiency and the interdye distance. This model is the so-called classic dye model which is introduced and discussed in the publications 4.1 and 4.3 of this thesis.

The beauty of Dale’s and Eisinger’s approach is that they do not need to assume specific orientation distributions for the dyes. Their one and only requirement is that they are axially symmetric, i.e. their solution is valid for a complete family of distributions. Under this assumption, the dynamic average of the orientation factor only depends on the relative angles of the symmetry axes of the orientation distributions and the observed depolarisation factors of
donor and acceptor, $A_D^\infty$ and $A_A^\infty$, respectively,

$$
\langle \kappa^2 \rangle = \left( \cos \Psi_{AD} - 3 \cos \Theta_D \cos \Theta_A \right)^2 \frac{A_D^\infty A_A^\infty}{\sqrt{A_D^\infty A_A^\infty}} + \left( \frac{1}{3} + \cos^2 \Theta_D \sqrt{A_D^\infty} \right) \left( 1 - \sqrt{A_A^\infty} \right) + \left( \frac{1}{3} + \cos^2 \Theta_A \sqrt{A_A^\infty} \right) \left( 1 - \sqrt{A_D^\infty} \right).
$$

(13)

The observed depolarisation factors are given by the ratio of the residual fluorescence anisotropy and the limiting anisotropy of donor and acceptor, respectively,

$$
A_D^\infty = \frac{r_D^\infty}{r_D^0} \quad \text{and} \quad A_A^\infty = \frac{r_A^\infty}{r_A^0}.
$$

(14)

Although this approach is based on many assumptions that might be not true, it is an elegant way of inferring distances from experimentally measured FRET efficiencies. This conversion is - in a formal sense - a transformation, and its improvement is one of the main topics of this thesis.

### 3.1.4 Measurement of single-molecule FRET Efficiency

For the analysis of structures by Fast-NPS, we perform single-molecule FRET measurements. Experiments on single molecules, in contrast to ensembles, have the advantage that we can separate different conformational species and can even resolve their dynamic transitions. In order to detect the signal of single fluorophores, it is important to separate the molecules and to increase the signal to noise ratio, which is usually achieved by decreasing the detection volume. The two types of setups used in this thesis are the total internal reflection (TIRF) microscope and the confocal microscope. In a TIRF microscope, the fluorescently labelled molecules of interest are attached to the surface of a cover glass such that the single molecules are spatially separable. From the other side, a laser beam is directed onto the glass-water surface in an angle such that the incident light is totally internally reflected, i.e. the complete ray is reflected. However, beyond the boundary surface an evanescent field appears which decays exponentially into the liquid with a characteristic length of $\approx 100$ nm. This reduces the observation volume drastically such that FRET at a single-molecule level is observable.

The term confocal originates from the geometry of the microscopic setup. It tightly focuses the excitation beam by a microscope objective in a point with a detection volume of $\approx$ femtoliters. In a solution in the picomolar range, then, only one molecule is present in the focal spot at a time. The emission light from the excited molecules is collimated and focused on a pinhole, which blocks the out-of-focus light. As a result, we detect only the emitted photons from the molecule in focus increasing the signal to noise ratio.
3.2 Mathematical Background

In this section we will summarise the mathematical background necessary to follow the present publications and occasionally emphasise the topic’s importance by giving examples in the FRET context. We present various stochastic processes which are essential for the different aspects of the FRET process. We discuss some methods of Monte Carlo integration, because they are extensively used in this thesis. We will focus especially on Markov chains on general state spaces and summarise the theoretical requirements for an unbiased Markov chain Monte Carlo (MCMC) algorithm, because it is the basis of the sampling engine in Fast-NPS. The presented theory was collected from the books and publications [22, 23, 24, 25, 26, 27, 28].

In the following the triplet \((\Omega, \mathcal{B}(\Omega), P)\) represents a probability space, i.e. a sample space, \(\Omega \neq \emptyset\), is the set of all possible outcomes, a \(\sigma\)-algebra \(\mathcal{B}(\Omega)\) on \(\Omega\), that means a set of measurable subsets of \(\Omega\), called events, and a probability measure \(P\) on \((\Omega, \mathcal{B}(\Omega))\). A probability measure is a function \(P : \mathcal{B}(\Omega) \to [0,1]\), which is \(\sigma\)-additive and satisfies \(P(\Omega) = 1\). The tuple \((\mathcal{X}, \mathcal{B}(\mathcal{X}))\) is a nonempty set \(\mathcal{X}\), also called the state space, and a \(\sigma\)-algebra \(\mathcal{B}(\mathcal{X})\) on \(\mathcal{X}\). In the context of structure determination by smFRET measurements, the sample space is the configuration of dyes and the state space consists of the parametrised representations of these configurations, usually \(\mathbb{R}^n\) for some \(n \in \mathbb{N}\) and its Borel \(\sigma\)-algebra.

Further, a \((\mathcal{X}, \mathcal{B}(\mathcal{X}))\)-valued random variable is a measurable function \(X : \Omega \to \mathcal{X}\), that means, for every subset \(A \in \mathcal{B}(\mathcal{X})\), we have \(X^{-1}(A) = \{\omega \in \Omega : X(\omega) \in A\} \in \mathcal{B}(\Omega)\). If \(\mathcal{X} \subset \mathbb{R}^n\), then \(X\) is called a random vector. The probability distribution function of \(X\) is then defined by \(P_X(A) := P(X^{-1}(A))\) for any subset \(A \in \mathcal{B}(\mathcal{X})\). We further assume that \(P_X\) has a probability density function or probability mass function, \(f_X\), i.e. a measurable function with the property that \(P_X(A) = \int_{X^{-1}(A)} dP = \int_A dP_X = \int_A f_X(x) \mu(dx)\) for all \(A \in \mathcal{B}(\mathcal{X})\) with respect to a measure \(\mu\), usually the Lebesgue or the counting measure, respectively. For convenience, we write \(dx\) instead of \(\mu(dx)\) and skip in some parts the random variable in the index of a density, mass function or distribution.

3.2.1 Stochastic Processes

Stochastic processes are essential to this work. Not only the above mentioned MCMC algorithms are based on a stochastic process, but also all aspects of FRET on the single-molecule level, i.e. fluorescence, transfer and diffusion, are stochastic in nature.

In general, a stochastic process is a family of \((\mathcal{X}, \mathcal{B}(\mathcal{X}))\)-valued random variables or vectors, \((X_t)_{t \in \mathbb{T}}\), indexed by some set \(\mathbb{T}\) and defined on a common probability space \((\Omega, \mathcal{B}(\Omega), P)\). The index set \(\mathbb{T}\) is usually \(\mathbb{R}^+\) or an ordered subset, e.g. the natural numbers, and is interpreted as the time. If \(\mathbb{T}\) is at least countable, the corresponding process is called discrete, else continuous. In some processes the time may be the one or a random variable, too.
An important type of stochastic processes are stationary processes. Here, the distribution of $(X_{t+s})_{t \in \mathbb{T}}$ is identical for any shift $s \in \mathbb{T}$. Hence, the mean and variance, if they exist, do not change over time either. An essential example is the time-discrete white noise process, where the random variables are not only identically, but also independently distributed with zero mean and finite variance. As an example of a discrete-time white noise process in the field of FRET, we can name the data obtained by a TIRF microscope (see Figure 8A in publication 4.4). In a TIRF measurement we collect the photons of a single donor and acceptor in consecutive and disjoint time intervals usually of length $30 - 100$ ms. Here, the sequence of e.g. the donor intensities is called an intensity trace which can be modelled by a discrete-time white noise process with offset (see Section 7).

Most of the stochastic processes which are found in our work are Lévy processes. They are said to have stationary increments, i.e. the increase depends only on the length of the time interval, but not on the time where the interval is placed. The second condition is that the increments are independent, which means the growth in two disjoint intervals is independent. A special case is the Wiener process, whose increments are normally distributed. It is the mathematical model for the Brownian motion which describes the random behaviour of objects at the molecular level due to thermal energy. As a basis for the proposal kernel in our MCMC algorithm it plays a major role in the software Fast-NPS (see publication 4.1).

Another fundamental stochastic process is the renewal process. It models patterns of events occurring at random times. In a renewal process it is assumed that the random interarrival times between two events, say $(T_i)_{i \in \mathbb{N}}$, are independently and identically (iid) distributed. Their sum, the so-called jump or arrival time, determines the time for the $n$-th renewal denoted by $S_n = \sum_{i=1}^{n} T_i$ where $S_0 = 0$. The counting version of a renewal process is then represented by the sequence of random variables $(N_t)_{t \in \mathbb{R}^+}$, where $N_t = \sup\{n \in \mathbb{N}_0 | S_n \leq t\}$.

A special renewal process is the time-homogeneous Poisson counting process. It is a Lévy process and is characterised by the fact that the number of jumps in a finite time interval $\Delta t$ follows a Poisson distribution, $N_{\Delta t} \sim Pois(k \Delta t)$, where $k$ is the time-invariant rate of the process. The de-excitation of a single fluorophore, excited by a laser of constant power below saturation, is well described by a Poisson counting process (see Section 7). Two inherent features are essential: the interarrival times are independently and exponentially distributed with rate $k$ and, consequently, the arrival times within a finite time interval are uniformly distributed conditioned on the number of events. In exploiting these characteristics, we have the following two ways of simulating a Poisson process. If the time interval $\Delta t$ is known a-priori, we can draw a realisation $n$ from $Pois(k \Delta t)$ and distribute the $n$ events uniformly on $[0, \Delta t]$. In contrast, if we have only knowledge about the number of events $n$, but not $\Delta t$, we can produce a realisation of a Poisson counting process by summing up $n$ realisations of independent and exponential random variables with rate $k$.

Another simulation method exploits the infinitesimal definition of the Poisson process which amongst other states that the probability that an event occurs
in some short time \( h \) is \( P(N_{t+h} - N_t = 1) = h k + o(h) \) for \( h \to 0 \). Thus, by drawing a realisation from a uniform distribution on \([0, 1]\), we can decide if an event in \( h \) happens or not. We can progress until an event happens, start anew and continue for a predefined time interval. The resulting pattern will asymptotically be a realisation of a Poisson counting process with rate \( k \). The major advantage of this method is that it allows to simulate a time-inhomogeneous Poisson counting process. Here, the rate is a function of time, \( k(t) \), that means the process loses its property of stationary increments. The infinitesimal method is extensively used in this thesis in order to simulate the FRET process, which is a Cox or also called doubly stochastic Poisson process, i.e. the rate is a stochastic process as well (see publication 4.4).

3.2.2 Fundamentals of Monte Carlo Methods

In most of the dye models developed in 4.1 and 4.4 we face the task to compute the expected value of the transfer rate or the FRET efficiency, which have in most cases no analytic solution. Additionally, their domain of integration is usually multidimensional, such that deterministic numerical integration methods do not perform well. Here, Monte Carlo methods provide a stochastic approach in order to solve these integrals efficiently.

Formally, the expected value of a measurable real-valued function \( g : \mathcal{X} \to \mathbb{R} \) is defined by

\[
\mathbb{E}(g(X)) = \int_{\mathcal{X}} g(x) \, f_X(x) \, dx.
\]  

(15)

For the following sections, we assume that \( \mathbb{E}(g(X)) \) and \( \mathbb{E}(g(X)^2) \) exist and are finite. In the publications, we use the physical notation, i.e. \( \langle g \rangle := \mathbb{E}(g(X)) \). In the case the density of \( X \) is uniform over a set \( A \in \mathcal{B}(\mathcal{X}) \), the expected value reduces to the integral

\[
\mathbb{E}(g(X)) = \int_A g(x) \, dx/|A|,
\]  

(16)

where \(|A| := \int_A dx < \infty \) is the Lebesgue measure of \( A \). In the classic Monte Carlo integration we place a sequence of \( N \) independent samples \( \{x_i\}_{i \leq N} \) uniformly in the set \( A \). If \( A \) is a simple union of hyperrectangles we can easily produce a set of uniform samples by a pseudorandom number generator. The average of the function values at the samples allows us to formulate an estimate for the expected value by

\[
\mathbb{E}(g(X)) \approx \frac{1}{N} \sum_{i=1}^{N} g(x_i).
\]  

(17)

The strong law of large numbers states that the sample average converges almost surely to the expected value. Further, the central limit theorem tells us that the error in the Monte Carlo estimate decreases with \( N^{-1/2} \) regardless of the number of dimensions.

If it is, however, not possible to sample uniformly from \( A \) alone, we choose the smallest superset of \( A \), say \( B \), in which we can place uniform samples. In
rejecting any sample which is in $B \setminus A$, we get uniform samples in $A$. This is the simplest version of the acceptance-rejection method, also called rejection sampling. Its efficiency is given by the ratio $|A| \setminus |B|$.

For a univariate random variable $X$ with a non-uniform density, we can use the inverse transform sampling. It is based on the probability integral transform which states that if $X$ has a continuous cumulative distribution function $F_X$, the random variable $Y = F_X(X)$ is uniformly distributed on $[0, 1]$. If we can compute an inverse $F_X^{-1}$, then $X = F_X^{-1}(Y)$ is distributed according to $F_X$ for $Y \sim U(0, 1)$. We can now estimate the expected value of any function as given by equation (15). In this manner we simulated, for example, the photoselection of fluorophores in the simulation of time-resolved anisotropy measurements.

The dyes with an angle $\theta$ to the axis of polarised laser light get excited with probability density $f_{\Theta}(\theta) = 3 \cos^2(\theta) \sin(\theta)$. Its inverse cumulative distribution function is $F_{\Theta}^{-1}(y) = \arccos(\sqrt{1-y})$, such that we can efficiently draw a set $\{\theta_i\}$ iid from $F_{\Theta}$ by simply generating pseudorandom numbers $\{y_i\}$ on $[0, 1]$. Figure 4 shows the analytic functions and the histogram of $10^6$ sampled angles.

![Figure 4](image_url)

Figure 4: Example of the inverse transform method for the photoselection of fluorophores by polarised light. The density of the photoselected fluorophores and its cumulative distribution are shown in red and green, respectively. The quantile function is plotted in yellow and the estimated density is shown in blue. We drew $10^6$ uniform samples and binned the transformed samples into intervals of length $0.01 \text{ rad}$.

In case we cannot find the inverse of the cumulative distribution function, we need to apply a more advanced version of rejection sampling. Let us assume we have a random variable $Y$ with density $f_Y$ from which we can sample independently. If we can find an $M$ such that $f_X(x) \leq M f_Y(x)$ for all $x \in \mathcal{X}$,
we can draw from $f_X$ by drawing from $f_Y$ and accepting the proposal $x$ with probability $\frac{f_X(x)}{M f_Y(x)}$. The efficiency of this sampling method is $M^{-1}$, because $f_Y$ and $f_X$ are normalised. This method is not limited to the univariate case.

3.2.3 Markov Chains on General State Spaces and MCMC Algorithms

The importance of Markov chains for this thesis derives from the fact that Fast-NPS uses a Bayesian approach in order to infer macromolecular structures. In Bayesian statistical inference the knowledge about quantities is expressed in probabilities. The likelihood function, say $L(\{E_i\}|x,M)$, is the density of the data, in our case the FRET efficiencies $\{E_i\}$, given the unknown parameters, $x \in \mathcal{X}$, i.e. a configuration of dyes, and a statistical model, $M$. The knowledge about the dye configurations we have before any measurement is represented in the prior density, $\Pi$. Then the density of the dye configurations given the data and model, the so-called posterior, is proportional to the product of likelihood and prior,

$$\pi(x|\{E_i\},M) \propto L(\{E_i\}|x,M)\Pi(x|M). \quad (18)$$

We denote its distribution by $\mathcal{P}$. The posterior provides the structural information in the form of expected values, marginal distributions and their credible volumes. However, all of these include an integration over a high-dimensional space of the dye configurations. The domain of integration is usually much larger than the area of major posterior mass and we have no well-founded knowledge of any dominating function from which we can sample iid. This makes rejection sampling infeasible.

The Markov Chain Monte Carlo solution to this problem is to construct a Markov chain on the state space that produces a set of dependent dye configurations drawn from the posterior, with which we can estimate the characteristics of the posterior. Essential for a Markov chain is its Markov kernel, i.e. a rule for the transition from one point to another region in state space. Given a general state space $\mathcal{X}$, i.e. a measurable space $(\mathcal{X},\mathcal{B}(\mathcal{X}))$, a Markov or stochastic kernel is a function $\kappa: \mathcal{X} \times \mathcal{B}(\mathcal{X}) \rightarrow [0,1]$ such that

(i) $\kappa(x, \cdot)$ is a probability measure on $(\mathcal{X},\mathcal{B}(\mathcal{X}))$ for all $x \in \mathcal{X}$ and

(ii) $\kappa(\cdot,A)$ is a non-negative measurable function for every $A \in \mathcal{B}(\mathcal{X})$.

Then, a Markov chain $(X_n)_{n \in \mathbb{N}_0}$ on a general state space is a discrete-time-homogeneous Markov chain with Markov kernel $\kappa$, if

$$P(X_{n+1} \in A|X_0 = x_0, ..., X_n = x_n) = P(X_{n+1} \in A|X_n = x_n)$$
$$= \int_A \kappa(x_n, dx), \quad (19)$$

with $A \in \mathcal{B}(\mathcal{X})$ and $x_i \in \mathcal{X}$ for any $n \in \mathbb{N}_0$. Intuitively that means, the next step of a Markov chains depends only on the current configuration in state space. In order to obtain samples from our posterior, we need the Markov chain to have $\mathcal{P}$ as its stationary or invariant distribution, that means we
have \( P(A) = \int_X P(dx) \kappa(x, A) \) for all \( A \in \mathcal{B}(X) \). A sufficient criterion to assure this is reversibility. A Markov chain is reversible with respect to \( P \), if
\[
P(dx) \kappa(x, dy) = P(dy) \kappa(y, dx), \forall x, y \in X.
\] (20)

This condition is better known as detailed balance in the context of physics. It states that the net flow of some quantity between two regions of the state space is zero. Exploiting this, we can construct a stochastic kernel which produces a reversible Markov chain with \( P \) as stationary distribution. The main concept is to split the Markov kernel into the product of two functions, a proposal kernel, \( q \), and an acceptance probability, \( \alpha \). The first generates a proposal, \( y_{n+1} \), given a previous state, \( x_n \) and the second decides, if it is accepted, \( x_{n+1} = y_{n+1} \), or rejected, \( x_{n+1} = x_n \). We can construct the acceptance probability with help of the detailed balance equation, such that the Markov chain is stationary for \( P \) by
\[
\alpha(x, y) = \min \left\{ 1, \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)} \right\}.
\] (21)

If we proceed in that manner, we get a set of samples \( \{x_i\} \) drawn from the posterior. This scheme describes the Metropolis-Hastings algorithm, which is the basis for all other MCMC algorithms. It is noteworthy that all involved densities and kernels do not need to be normalised, since the normalising constants in (21) cancel each other out. This is one reason for the widespread use of MCMC methods in Bayesian statistics, where the normalising constant of the posterior, also called marginal likelihood or model evidence, is usually not known or even impossible to compute.

Especially on a complicated support, a Markov chain with stationary distribution \( P \) might not converge to stationarity from any starting point \( x_0 \in X \). In other words it might get stuck in some region, such that the characteristics of the posterior estimated by the collected samples is strongly biased. So we require from a Markov chain that for every starting point all parts of the state space can be reached. For chains on general state spaces this can be assured by the notion of \( \phi \)-irreducibility.

A Markov chain is called \( \phi \)-irreducible, if there exists a measure \( \phi \) on \( X \) such that, whenever \( \phi(A) > 0 \), we have \( L(x, A) > 0 \) for all \( x \in X \) and \( A \in \mathcal{B}(X) \). Here is \( L(x, A) = P_x(\tau_A < \infty) \) the probability that the chain started from \( x \in X \) ever returns to \( A \) with \( \tau_A = \min \{ n \geq 1 : X_n \in A \} \) the first return time to \( A \). It is possible to show that the Metropolis-Hastings algorithm for an everywhere finite \( d \)-dimensional target density \( \pi \) equipped with a positive and continuous proposal kernel on \( \mathbb{R}^d \times \mathbb{R}^d \), is \( \mathcal{P} \)-irreducible.

Although a Markov chain is \( \phi \)-irreducible, it might still not converge to its stationary distribution due to periodic behaviour. A Markov chain is called \( d \)-periodic if a set of \( d > 1 \) disjoint subsets \( A_1, \ldots, A_d \subset X \) exists with \( \kappa(x, A_j) = 1 \) for all \( x \in A_i \) and \( j = (i + 1) \mod d \), such that \( \mathcal{P}(A_i) > 0 \) for at least one \( i \). Else the chain is called aperiodic. Interestingly, with the requirements for a \( \mathcal{P} \)-irreducible Markov chain stated above, the Metropolis-Hastings algorithm produces an aperiodic chain.
Assume that $\mathcal{B}(\mathcal{X})$ is the Borel $\sigma$-algebra on $\mathcal{X} = \mathbb{R}^n$, as is the case in Fast-NPS. Then, for a $\phi$-irreducible and aperiodic Markov chain with stationary distribution $\mathcal{P}$, we have

$$\lim_{n \to \infty} \kappa^n(x, A) = \mathcal{P}(A)$$

(22)

for all $A \in \mathcal{B}(\mathcal{X})$ and for $\mathcal{P}$-almost all $x \in \mathcal{X}$. Here, the $n$-step transition kernel is inductively defined by $\kappa^n(x, A) = \int_{\mathcal{X}} \kappa(x, dy)\kappa^{n-1}(y, A)$ for $x \in \mathcal{X}$ and $A \in \mathcal{B}(\mathcal{X})$. Here, we set $\kappa^0(x, A) = \delta_x(A)$, where the latter is the Dirac measure. In words, equation (22) states that the constructed Markov chain converges to the posterior, no matter where we started. So the collected samples $\{x_i\}$ are drawn from the posterior.

More importantly for us though these assumptions also assure that a strong law of large numbers applies:

$$\mathbb{E}(g(X)) = \lim_{n \to \infty} n^{-1} \sum_{i=1}^{n} g(x_i),$$

(23)

almost surely for a measurable function $g : \mathcal{X} \to \mathbb{R}$ and $\mathbb{E}(g(X)) < \infty$. This allows us to estimate the characteristics of the posterior by a realisation of a Markov chain constructed by the Metropolis-Hastings algorithm.

In the next sections follow the publications which were motivated in the introduction. Afterwards we will give a comprehensive summary, discuss the major results and give a outlook on future improvements and goals.
4 Publications


---

Title: Fast-NPS – A Markov Chain Monte Carlo-based analysis tool to obtain structural information from single-molecule FRET measurements

Authors: Tobias Eilert, Maximilian Beckers, Florian Drechsler and Jens Michaelis

Journal: Computer Physics Communications

Copyright: Elsevier

The use of the article in a PhD thesis is kindly allowed by Elsevier, according to “Personal Use” under “Journal Author rights”.

https://www.elsevier.com/about/our-business/policies/copyright#
Author’s Contribution

The author of the thesis at hand structured and wrote the publication except for corrections and comments by the co-authors. Further, he provided the theoretical framework for most of the applied methods. He implemented and structured most parts of the algorithm. The stop criterion for the adaptation of the proposal kernel was derived, tested and adapted by the author. The complete adaptive scheme of the parallel tempering, i.e. the adaptation of the swap probabilities and the number of chains as well as the novel stop criteria are his work. Also, the efficient adaptation of the proposal kernels of the higher temperature chains was introduced by the author. Further, he corrected the implementation of the dye models and established the theory and code for the selection of individual dye models. The author completed the theoretical framework for and extended the consistency check to the marginal consistency check, which he applied to a real-case scenario. The remaining code was written by Florian Drechsler and Maximilian Beckers, but were corrected and extensively rewritten by the author.
Fast-NPS—A Markov Chain Monte Carlo-based analysis tool to obtain structural information from single-molecule FRET measurements

Tobias Eilert a, Maximilian Beckers b, Florian Drechsler a, Jens Michaelis a,*,

a University of Ulm, Institute for Biophysics, Albert-Einstein-Allee 11, 89081 Ulm, Germany
b European Molecular Biology Laboratory, Structural and Computational Biology Unit, Meyerhofstrasse 1, 69117 Heidelberg, Germany

ABSTRACT

The analysis tool and software package Fast-NPS can be used to analyse smFRET data to obtain quantitative structural information about macromolecules in their natural environment. In the algorithm a Bayesian model gives rise to a multivariate probability distribution describing the uncertainty of the structure determination. Since Fast-NPS aims to be an easy-to-use general-purpose analysis tool for a large variety of smFRET networks, we established an MCMC based sampling engine that approximates the target distribution and requires no parameter specification by the user at all. For an efficient local exploration we automatically adapt the multivariate proposal kernel according to the shape of the target distribution. In order to handle multimodality, the sampler is equipped with a parallel tempering scheme that is fully adaptive with respect to temperature spacing and number of chains. Since the molecular surrounding of a dye molecule affects its spatial mobility and thus the smFRET efficiency, we introduce dye models which can be selected for every dye molecule individually. These models allow the user to represent the smFRET network in great detail leading to an increased localisation precision. Finally, a tool to validate the chosen model combination is provided.

Programme summary

Programme Title: Fast-NPS
Programme Files doi: http://dx.doi.org/10.17632/7ztzj63r68.1
Licencing provisions: Apache-2.0
Programming language: GUI in MATLAB (The MathWorks) and the core sampling engine in C++
Nature of problem: Sampling of highly diverse multivariate probability distributions in order to solve for macromolecular structures from smFRET data.
Solution method: MCMC algorithm with fully adaptive proposal kernel and parallel tempering scheme.

© 2017 Elsevier B.V. All rights reserved.

1. Introduction

Structural biology is a key field of life sciences. It aims to give a picture of life at the molecular level, which is a prerequisite for obtaining a mechanistic molecular understanding of cellular processes. Modern techniques, such as X-ray crystallography or cryo-EM, can resolve the structure of macromolecules up to an Angstrom level. However, in order to understand nature it is important to not only resolve the static structure of macromolecules in artificial environments, but also to elucidate their dynamic structure or even transient dynamic complexes in an aqueous milieu. Here, an interesting tool is Förster resonance energy transfer (FRET) [1]. FRET has been termed a molecular ruler, since the distance range that can be measured is on the length scale of proteins [2]. Furthermore, single-molecule FRET (smFRET) has become a widely used technique for studying the dynamics of macromolecular complexes [3–5]. While distance information can be obtained, determining distances quantitatively remains a challenge [6–8]. Using the trilateration of several smFRET distances one can determine an unknown position with respect to a macromolecular complex [9,10]. Moreover, smFRET results have been combined with other techniques for structural biology approaches [11–13].

We present the Fast-Nano-Positioning System (Fast-NPS), an advanced software package that utilises smFRET measurements between dye molecules (further referred to simply as dyes) to gain quantitative structural information about macromolecular
complexes. It aims to localise the unknown dye positions, the so-called antennas, by means of dyes covalently bound to known positions on a macromolecular complex, the satellites, with a Bayesian model. In Bayesian statistics the degree of knowledge about an unknown quantity is expressed by a probability distribution conditional on the evidence obtained from experimental data, the so-called posterior [14].

In Fast-NPS the data constitutes the measured average smFRET efficiencies with experimental error, the determined isotropic Förster radii and the measured steady-state fluorescence anisotropies [15,16]. The unknown quantities of interest are the positions of the antennas. Since not only the positions, but also the orientations of the dyes have a tremendous effect on the measured smFRET efficiencies, their transition dipole moments (TDMs) are explicitly modelled in Fast-NPS.

In previous works the Nano-Positioning System (NPS) has already been used to study the position of the exiting RNA from the eukaryotic RNA polymerase II [10] and to investigate the influence of the transcription factor TFIIH on the position of the nascent RNA [11]. Further, the position of the non-template and upstream DNA in yeast Polymerase II transcription elongation complexes [17] and the architecture of a minimal Polymerase II open promoter complex [18] were analysed. Moreover, NPS was also applied to shed light on the archaeal initiation complex [19].

In Section 2 we recapitulate the Bayesian framework of NPS which forms the basis for the analysis of smFRET data in order to gain structural information from macromolecular complexes. In Section 3.1 we present a novel Markov Chain Monte Carlo (MCMC) algorithm, which forms the basis for Fast-NPS, to analyse the posterior in order to extract structural information from an arbitrarily large smFRET network using standard desktop computers in a reasonable amount of time. In Section 3.2 we establish the theory of individual dye models, accounting for the different spatial behaviour over time depending on the dye’s molecular environment. A method to assess the consistency of the prior information with the smFRET data is presented in Section 3.3. In Section 4 we apply the dye models to a real smFRET network guided by the help of our consistency check in order to maximise the localisation precision. In Section 5 we give a comprehensive discussion of Sections 3–4 and end with a conclusion and a future outlook in Section 6.

2. Theory

In Bayesian statistics the normalised posterior \( P \) is proportional to the product of likelihood \( L \) and prior \( \Pi \), i.e. \( P \propto L \cdot \Pi \). The likelihood is the probability distribution of the parameters, in our case the dye positions and orientations, given the experimental data and a statistical model. According to the central limit theorem the distribution of the average smFRET efficiency \( \langle E \rangle \) converges to a normal distribution centred around \( \langle E \rangle \), the expectation of the smFRET efficiency \( E \), when the number of data points becomes large. Assuming that there is a unique configuration of dyes giving rise to the data, \( \langle E \rangle \) is the associated smFRET efficiency. Thus, the likelihood function is defined by a normal distribution centred around the average smFRET efficiency with standard deviation \( \sigma \), where \( \sigma \) is the experimentally determined measurement error, i.e. we have \( L : = \mathcal{N}\langle E \rangle, \sigma^2 \rangle \) [15]. Further, the dependence of the average smFRET efficiency \( \langle E \rangle \) on the distance \( d \) between donor and acceptor and their Förster distance \( R \) is given by

\[
\langle E \rangle = \frac{1}{1 + (d/R)^6}.
\] (1)

For simplicity both random variables and their realisations will be denoted with the same symbols throughout this paper.

The Förster distance \( R \) is given by

\[
R = \frac{R_0}{\sqrt{2}}.
\] (2)

where the orientation factor \( k^2 \) is a function of the positions and TDMs of the dye molecules [1]. Depending on the relative orientation of donor and acceptor it can take values from 0–4 [20]. When both dye molecules are free to rotate and reorient faster than the fluorescence lifetime of the donor in the presence of the acceptor, \( k^2 \) adopts its isotropic value of 2/3, such that all orientation effects vanish. Under this condition the Förster distance is called the isotropic Förster distance \( R_{iso} \) [21]. The position of a dye \( i \) is parametrised by \( x_i, y_i, z_i \) and its TDM by an azimuthal angle \( \theta_i \) and a polar angle \( \varphi_i \). Substitution of Eq. (1) into the likelihood represents a transformation from the range of the average smFRET efficiency to the configuration space \( \Omega \) of both dyes. The parameter vector of the likelihood defined on \( \Omega \) is then \( x := (x_1, y_1, z_1, \theta_1, \varphi_1, x_2, y_2, z_2, \theta_2, \varphi_2)^T \).

Fast-NPS is a hybrid approach. The prior knowledge about the position of the dye molecule is gained from an accessible volume (AV) computation with respect to the structure of the molecule of interest obtained from X-ray crystallography, NMR studies or cryo-EM [11]. These volumes serve as a flat position prior for Fast-NPS. Finally, the prior for the TDMs is uniform over the unit hemisphere (see Section 3.1.3).

Since the unambiguous localisation of one or more antennas requires the measurement against several satellites, the joint likelihood is given by an uncorrelated (assuming independent measurements) multivariate normal distribution in the space of the average smFRET efficiencies. The joint prior consists then of the product of the uniform distributions on the individual AVs. Finally, according to Bayes’ law the joint posterior is proportional to the product of joint likelihood and joint prior [15]. The parameter vector is then given by \( x := (x_1, y_1, z_1, \theta_1, \varphi_1, \ldots, x_n, y_n, z_n, \theta_n, \varphi_n)^T \), where \( n \) denotes the number of dyes in the network.

In the following section we focus on the development of an adaptive sampling engine for the structural inference of a large variety of smFRET networks.

3. Methods

3.1. Algorithm

In Section 2 we have defined a probability distribution on the configuration space \( \Omega \) providing us with the information how likely a realisation \( x \) is. The structural information, which we can extract from the posterior, is given by a volume which specifies how likely it is that a certain dye position is found inside. Although the posterior can be written in a closed form, the marginalisation down to the position of one dye is analytically unfeasible. In order to solve this problem we chose to develop a sampling algorithm which produces a set of realisations \( \{x\} \), so-called samples, drawn from the posterior. Then, the marginalisation is reduced to the simple projection onto the positions of the dye of interest.

Since we want to express a complete lack of position knowledge about an antenna when starting the analysis, the search space covers usually more than 25 times the volume of the macromolecule. However, the major posterior mass which is displayed by a credible volume, i.e. the smallest volume that includes a certain probability, covers only a small fraction of this search space (Fig. 1A). In order to efficiently draw samples from the posterior, we have chosen an importance sampling algorithm, or more specifically, a Metropolis-within-Gibbs sampler [22].

Localisation geometry enforced by limitations of biochemical labelling strategies can also induce banana- or spherical shell-like
Fig. 1. Examples of the position prior and credible volumes for antennas analysed by Fast-NPS. (A) The transparent box excluding the macromolecule (here as an example, the archaeal RNA Polymerase) depicts the volume where an antenna may be located a priori. This search space is many times greater than the credible volume of an antenna (red and green shown at 68% credibility). (B) Example of a credible volume in green (at 68% credibility) exhibiting a banana-like shape. (C) Example of a credible volume in red (at 68% credibility) showing bimodality. The pictures were produced in UCSF Chimera [23]. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Fig. 2. General Workflow of a complete Fast-NPS analysis. The blue rectangles represent the input and output data. The yellow rectangle represents the GUI of Fast-NPS. The subsequent phases of the algorithm are rendered in grey. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

shaped credible volumes (Fig. 1C). This together with huge differences in size and shape of satellite and antenna position priors represents a major challenge for efficiently proposing adequate configurations. In order to overcome these obstacles the sampler is adjusted to the posterior (see Section 3.1.5). This adaption does not only include the shape of the likelihood, but also incorporates the demanding characteristics of the almost fractal-like prior.

The surface of biomacromolecules may exhibit a complex topology with numerous grooves, cavities and even enclosures. This can lead to a multimodal credible volume (Fig. 1B). In order to ensure that the sampler is able to visit every point in configuration space it is equipped with a parallel tempering scheme (see Section 3.1.6) [24].

The respective difficulties of all these effects can vary from network to network. Hence, we established a flexible, i.e. adaptive, and robust sampling engine which is presented now.

3.1.1. General workflow

Via the graphical user interface (GUI) of Fast-NPS the user provides the input for a new smFRET network. The input data comprises the experimental data, the a priori information, e.g. the PDB file of the macromolecule, and the default starting parameters for the sampling algorithm. The sampler is divided into three phases which will be discussed in detail below: burn-in, adaption and sampling. In the first phase the algorithm tracks the convergence to the posterior. Once stationarity appears to have been reached, the following samples are allowed to be used for the adaption of the sampler. In the adaption phase the algorithm learns the characteristics of the posterior such that the sampling is most efficient. The last phase constitutes the actual collection of samples. The general workflow of a complete Fast-NPS analysis is schematically depicted in Fig. 2.

3.1.2. Sampling scheme

The basis of the developed MCMC sampler is a Metropolis-within-Gibbs algorithm, i.e. each component within x is updated one after another with a Metropolis scheme until a completely new configuration has been reached [22]. We call a complete update an iteration. In a Metropolis update a proposed configuration y, the so-called proposal, is compared to the current one x. We accept the proposal with probability

\[
\min \left(1, \frac{P(y)}{P(x)} \right),
\]

and otherwise we reject it. This algorithm produces a realisation of a Markov chain \{x\}, i.e. a set of samples, drawn from the posterior \(P\).

Since the basic object of an smFRET network is the dye, we further apply a dye-wise blocking scheme, such that we update all five components of the ith dye, \(x_i := (x_i, y_i, z_i, \theta_i, \phi_i)^T\), at once.

In order to pick an unbiased starting point for the sampler we draw a configuration, the so-called random seed, from a uniform distribution over all allowed positions and orientations, i.e. from the joint prior.

3.1.3. Proposal kernel

A crucial ingredient for the efficiency of every importance sampler is its proposal kernel, i.e. the distribution from which a new configuration is drawn that is either accepted or rejected. As a basis of our proposal kernel we chose the normal distribution, because it not only concentrates the major part of the proposals near the current configuration, thus yielding a good acceptance rate, but also provides the possibility to perform jumps to distant configurations. Since we are using a dye-wise update (see Section 3.1.2), the complete proposal distribution consists of a set of multivariate
normal distributions \{q_i\} each defined on the five dimensions of dye \textit{i} with the current configuration \textbf{x}, as the mean vector and a covariance matrix \textbf{Σ_i} describing the width and correlation of the distribution. The adaption of the covariance matrices is described in Section 3.1.5 [25].

In order to represent the accessible volume of a dye, we approximate the configurational \textit{N} by a collection of boxes, the so-called box collection prior. If a proposed configuration \textbf{y} results in a position outside the current box, the box collection prior is searched. Since a box collection prior usually consists of several thousands of boxes the search would be computationally expensive. Hence, we have established a neighbourhood system such that every box knows all its box neighbours located within a cube centred around the box of interest. We found that a cube edge length of 40 Å is a good compromise to minimise the search time and to maximise the allowed volume for the proposal. In order to ensure that the proposal is within the neighbourhood of the current box we truncated each proposal distribution \textit{q_i} of dye \textit{i} to obtain

\[
q_i(\textbf{y} | \textbf{x}_\text{trunc}) = \begin{cases} q_i(\textbf{y} | \textbf{x}) & \text{if } \|\textbf{y}^{\text{pos}} - \textbf{x}_i^{\text{pos}}\|_1 \leq 20 \text{ Å} \\ 0 & \text{else}, \end{cases}
\]

where \textbf{x}_i^{\text{pos}} is the current position, \textbf{y}_i the proposed configuration, \textbf{y}_i^{\text{pos}} the proposed position of dye \textit{i} and \| \cdot \|_1 denotes the L\textit{1}-norm. This modification of the proposal kernel decreases the time duration for a proposal without crucially limiting its allowed volume. In order to obtain proposals from the truncated kernel, we draw a realisation from \textit{q_i} and accept it if it is within the proposal volume, and otherwise reject it, i.e. we use the acceptance–rejection method [26]. For convenience, we will denote the truncated proposal kernel by \textit{q_i} in the rest of this paper.

Besides the dependency on the position, the smFRET efficiency is highly sensitive to the relative orientation of the TDM of donor and acceptor, as expressed by the orientation factor \(s^{2}\). Since the TDMs are described by unit vectors, we need to sample the surface of a unit sphere. Further, it is easily seen that \(s^{2}\) is invariant under a flip in direction by 180° of one or both TDMs (see Appendix A). Hence, sampling of a unit hemisphere is sufficient. Consequently, we parameterised the TDM of dye \textit{i} by the polar angle \(\varphi_i \in (0, \pi]\) and azimuthal angle \(\theta_i \in (0, \pi]\). Since the sphere surface area grows with \(\cos(\varphi)\), we chose to work with the variable \(-\cos(\varphi) \in (-1, 1]\) instead. This parametrisation facilitates an unbiased proposal on the sphere. If the sampler attempts to cross the borders of the chosen hemisphere, the angular coordinates are remapped appropriately exploiting the spherical periodicity. Before we can turn to the adaption of the proposal kernel, we need to find a criterion that tells us when the Markov chain has converged to the posterior. This phase, called burn-in, is described in the next section.

3.1.4. Burn-In

The adaption of the proposal kernel requires that the Markov chain is stationary, i.e. we have no bias due to the initial random seed. In order to monitor the convergence to the posterior distribution, we use a criterion based on the cross entropy [27]. One can show that a stationary Markov chain leads to the convergence of the cross entropy \(H\) (see Appendix B), which can be approximated by its estimate \(\tilde{H}\):

\[
\tilde{H}(F; n, k) = -\frac{1}{n} \sum_{j=1}^{n} \log(F_j(\textbf{x}_k)),
\]

where \(n\) is the number of samples used to compute \(\tilde{H}\), \(k\) is the spacing between samples and \(\textbf{x}_k\) denotes the \(k\)th sample. Here, we use \(n = 10\) and \(k = 100\) in order to account for the inherent correlation of the sampler. As a criterion for the convergence of the cross entropy, we monitor the slope of its successive estimates. To this end, we perform a linear regression on 10 of these estimates and a \(t\)-test upon the linear regression coefficient \(b\) for the slope to a confidence level of 90%. If the null hypothesis \(H_0 : b = 0\) (i.e. the cross entropy does not change any more) can no longer be rejected against \(H_1 : b \neq 0\) (i.e. the cross entropy still decreases or increases), we assume that the Markov chain has reached stationarity. In the next section, we face the task to establish an adaptable proposal kernel in order to efficiently produce samples from the posterior.

3.1.5. Adaption of proposal kernel

In order to efficiently sample the posterior, the proposal distribution needs to adapt to the shape of the target distribution. As a consequence, we developed the following simple strategy. At first, we estimate the diagonal entries of the covariance matrix of the posterior, which we shall call global variances, and subsequently scale them to obtain an efficient proposal kernel.

In the first step we calculate the global variances from the samples taken at stationarity. The recursive algorithm used (see Appendix C) takes only every 200th sample in order to account for the inherent correlation. For the record, if the sample variances of the posterior \(\sigma^2\) for parameter \(j\) of dye \textit{i} are precise enough, we save the last 10 consecutive estimates (in a spacing of 1000 iterations) and calculate their mean and standard error. If the standard errors divided by their corresponding means, i.e. the relative standard errors \(\sigma_j(\text{pos})/\sigma_j\), for all variances are less than 5%, we stop the recursive approximation. With these variances we then construct a set of uncorrelated covariance matrices \(\{\Sigma_i\}\), that is \(\Sigma_i = \text{diag}(\sigma_1, \ldots, \sigma_d)\).

It can be shown that in order to sample the posterior most efficiently, the global variances need to be scaled such that we have an acceptance probability of 0.234 [28]. Thus, we adapt the proposal covariance matrices \(\{\Sigma_i\}\) according to \(\Sigma_i^\text{pos} = \exp(2c_i^2) \Sigma_i\), i.e. we keep the ratio of the variances for the proposal of dye \textit{i} constant (see Fig. 3A). In this way the proposal geometry adapted to the shape of the posterior is preserved. The variance scaling factor \(c_i^2\) is approximated by a Robbins–Monro type of stochastic optimisation algorithm [29], that is

\[
c_i^{\text{pos}+1} = \begin{cases} c_i^2 \frac{1}{\delta(n)} & \text{if } \text{ge}^2 > 0.234 \\ c_i^2 \frac{1}{\delta(n)} & \text{else}, \end{cases}
\]

in order to obtain an average acceptance rate of \(\text{ge}^2 + 1 \approx 0.234\). Here, \(n\) is the number of scaling steps and the step size is given by \(\delta(n) = \min(0.05, n^{-2/3})\), such that we have a limited and diminishing adaption [30]. We perform an update once every 200 iterations. The initial variance scaling factor should be set to \(c_i^2 = \frac{2}{\delta(n)}\) for each dye \textit{i}, where \(d = 5\) are the dimensions of the proposal parameter space of \textit{q_i} [31]. To terminate this adaption phase the final variance scaling factors are set to \(c_i^2 := c_i^2\) as soon as \(0.2 < \text{ge}^2 < 0.35\), and the proposal kernel is no longer adapted.

A proposal \(y_i\) from \(q_i\) is then given by \(y_i = x_i + \Sigma_i^{1/2} \textbf{u}\), where \(\textbf{u}\) is a vector of realisations of standard normally distributed random variables. It is noteworthy that we neglect the correlation terms of the posterior (see Fig. 3A), which precludes correlated proposals, but reduces the computation time drastically.

Since we apply an acceptance–rejection method to draw from the truncated proposal kernel (Section 3.1.3) and the scaled variances may grow large, we would need to draw many times from the untruncated kernel in order to get a valid proposal within the proposal volume. Thus, we truncate the variances if they exceed (20 Å)² in position and (2 Å)² in angle space, the latter because of periodic remapping (Section 3.1.3). With this we ensure that more than 68% of the draws fall into the proposal volume.

Having described the adaption scheme of the proposal kernel in order to explore the posterior on a local scale, in the next section we turn to the question of how to adapt the parallel tempering scheme in order to efficiently explore the posterior globally.
3.1.6. Parallel tempering

In position space the posterior distribution may show a steep slope, because of its dependency on the sixth power of the distance between dyes (see Eq. (1)). In case of multimodality this leads to vast regions of low-probability separating the different modes. Additionally, because of the complex surface structure of biomacromolecules, it is frequently observed that a region of high probability is split apart by the position prior, i.e. regions with zero probability. These areas constitute critical barriers for a Metropolis proposal. These are constituted of energy barriers originating from the standard Boltzmann dependence $\exp(-\frac{E}{kT})$, where $E$ is the energy on temperature. The combination of the temperature dependence $\exp(-\frac{E}{kT})$ and the prior $\Pi$, which originates from the standard Boltzmann dependence $\exp(-\frac{E}{kT})$ on energy, prevents the system from reaching the ground-state posterior.

Our parallel tempering scheme is implemented as follows. After every dye in every chain has been updated, the lowest-temperature chain attempts to swap its configuration with the “hottest” chain. A swap is accepted with probability

$$p^\text{swap} \propto \exp(-\frac{E}{kT}).$$

The parallel temperings scheme is most efficient if every replica spends the same amount of time at every temperature $T$, as well as to allow every point in configuration space to be accessible from any other point. With these two conditions satisfied it is then possible to obtain the total number of chains $N$ needed, leading to the so-called temperature ladder $\left\{ T_k \right\}_{k=1}^N$. The step size is given by

$$\gamma \approx 0.234.$$
prior, i.e., the configuration space $\Omega$ is finite, our observation is that the limit $\beta \to 0$ is not required. We find that the average swap probabilities stay constant at $\tilde{a}_{k,k+1} \approx 0.234$, but then (depending on the complexity of the smFRET network, of course) suddenly start to grow rapidly until it reaches slowly 100%, implying that we have approximately reached the uniform prior (see Fig. 3B). This led us to the following robust and simple heuristic to select the optimal number of chains $N$ needed: We let

$$N := \min \left( \{ k : \tilde{a}_{k,k+1} > 0.9 \} \right).$$

(10)

With this we define the optimal number of chains for which the average swap probability to $N + 1$ is greater than 90%, since this implies that $p_{\text{sw}}$ is approximately equal to $p_{\text{sw}+1}$ everywhere in state space. Thus, we set $p_{\text{max}} = 0.9$. Applying this criterion, the most complex smFRET network we have tested needs 30 chains. Thus, we start for every network with 35 temperatures and subsequently reduce the number of chains according to Eq. (10). Should a user encounter a more complex smFRET network, it is possible to set the maximal number of chains manually.

In order to pick an unbiased starting point for the parallel tempering scheme every chain has its own random seed. Further, in the phase of the proposal kernel adaption we need to adapt every chain. Here, we take for each chain the estimated diagonal covariance matrix of the ground-state posterior, but scale them independently. This approach follows the aim that the shape of the posterior, i.e., the ratio of the spread in different directions, due to tempering should stay nearly constant.

The next section describes the sequence of steps introduced before in order to yield a fully adaptive MCMC algorithm to infer structural information from arbitrary smFRET networks.

### 3.1.7. Sequential algorithm embedded within the software package Fast-NPS

There are three possible input parameters for the algorithm, i.e., the maximal number of chains $N$ (see Eq. (10)), the number of effective samples collected and the thinning interval, i.e. how many direct samples are produced until one effective samples is saved. Their default values ($35$, $10^4$ and $10$) are set such that for a large variety of smFRET networks no user input is required.

These default parameters together with the experimental data and the prior information (that is the box collection priors) are passed via the graphical user interface to the algorithm implemented in C++. The algorithm commences by initialising and pre-processing the smFRET network data and the prior information into appropriate C++ classes. After these initialisation steps, the algorithm proceeds to the core sampler, which is implemented using OpenMP, an API for multi-platform shared memory multiprocessing programming, for an efficient parallel execution of the tempered chain operations.

The core algorithm within this parallelised region is structured as follows: Initially, each chain is started from an independent random seed. Then, we monitor the convergence of the cross entropy, such that each chain has reached stationarity. Subsequently, we start the recursive variance calculation only in the first chain demanding a relative standard error $\sigma(\mu_{\text{var}})$ below 5% of every dye $i$ and parameter $j$ in order to proceed. In the next step, the diagonal covariance matrices ($\Sigma$) adapted to the ground-state posterior are applied to all chains and scaled independently in order to reach an optimal acceptance rate. Simultaneously, we adapt the temperatures in order to obtain the most efficient swapping scheme. We stop this procedure only if all Metropolis and swap acceptance rates have converged at the same time, because they depend on each other. In the next step, the algorithm checks if there are any high temperature chains which merely represent approximately the prior and are thus redundant. If this is true, the parallel region is restarted with the reduced number of chains $N$ (computed by Eq. (10)). Subsequently, the production of samples is started. Having reached the predefined number of effective samples, the algorithm stops and returns the samples to the GUI. Fig. 4 summarises the detailed algorithm embedded in the software package Fast-NPS in a flowchart.

### 3.2. Dye models

Recently we introduced several dye models accounting for different rotational and spatial behaviour of a dye due to its molecular environment [16,36]. The original and most conservative model was termed classic. Here, it is assumed that the dye resides in one particular location within its AV (static model), and the orientation of the TDM obeys an axially symmetric orientation distribution. The opening angle of this distribution is related to the respective fluorescence anisotropy [37]. In contrast, the iso model represents the situation when a dye is fixed at a certain position, but is free to rotate so that the orientation averages dynamically across the lifetime of the fluorescent state. Thus, its fluorescence anisotropy is set to zero. If it is additionally assumed that the dye may occupy every position within its AV with equal probability, i.e. static position averaging is fulfilled, we obtain the $\text{meanpos-iso}$ model. Note that in this model the dye mobility is low compared to the fluorescent lifetime, so that for each excitation a different position within the AV is probed, but during the lifetime of the excited state the position is constant. In the var-meanpos-iso model, we apply dynamic orientation and static position averaging, but allow the freedom that the centre of mass is no longer at the centre of the AV, i.e. some regions of the AV are more likely to be visited than others. Finally, if we use this position model and again restrict the dye rotation to an axially symmetric distribution, we obtain the var-meanpos model.

Since the local environments of a dye molecule are in general different, in an optimal analysis one should be able to select the model for each dye individually in order to reach the optimal precision while retaining consistency. Because we have 5 different models, there are 25 possible model combinations. Since it is irrelevant for a given model combination, which model the acceptor or donor follows, we obtain $15 = \binom{5}{2}$ unique model combinations (see Appendix D). For model combinations where at least one dye is averaged over several positions, the expected value of the smFRET efficiency $\bar{E}(\xi)$ is different from the smFRET efficiency $E$ calculated from the mean position distance $R_{\text{MP}}$ of both dyes by Eq. (1). For a given $R_{\text{MP}}$, the expected value of the smFRET efficiency is calculated by

$$E(\xi) = \int_\Psi E(\xi) \; p(\xi) \; d\xi,$$

(11)

where $p(\xi)$ stands for the corresponding weight function for a given configuration $\xi$. The parameter domain of integration is denoted by $\Psi$, which is defined by the restrictions of the specific model combination on the dye positions and orientations (see Fig. 5).

Prior to every network analysis we perform Monte Carlo simulations in order to solve Eq. (11). This is necessary for every unique measurement, because $\bar{E}(\xi)$ is sensitive to the isotropic Förster radius $R_0^{\text{iso}}$. We perform the Monte Carlo integrations for every $R_{\text{MP}} \in \{1 \text{ Å}, \ldots, 150 \text{ Å} \}$ and fit the 150 pairs of values of $R_{\text{MP}}$ and $\bar{E}(\xi)$ for each dye pair with a polynomial of degree 11. Depending on the complexity of the model combination, we obtain either none, 1, $n$ or $n^2$ fitting polynomials for each dye pair, where $n$ is the number of equally sized regions on the TDM orientation hemisphere (see Appendix D). In the Monte Carlo simulations we assume that the statically position averaged dyes are uniformly distributed within a sphere of radius 20 Å. This simplification is made because radially
asymmetric AVs would lead to 6 more degrees of freedom resulting in a tremendously increased complexity. The radius of 20 Å was chosen to represent a typical dye molecule of size and linker length as used when attaching dyes to proteins and DNA.

3.3. Consistency check

There is a great need to determine whether the sampled configurations are actually in agreement with the smFRET data, since not only the position prior of each dye introduces previous knowledge, but also the assignment of models to the dyes may significantly change the posterior in smFRET space. As stated in the introduction the joint likelihood is an uncorrelated multivariate normal distribution in smFRET space. Based on this, we established a method that indicates if the combined prior information clashes with the smFRET input data. If the configuration samples collected from the posterior agree with the data, the smFRET efficiencies recalculated from these configurations should be drawn from the likelihood. Thus, approximately 95% of the smFRET efficiencies of the sampled configurations should lie within the 95% confidence ellipsoid of the likelihood (see Fig. 6A). In case the applied prior information is not consistent with the smFRET data, the configuration samples will lead to a shift of the recalculated smFRET efficiencies, such that much fewer than 95% of the smFRET samples lie within the 95% confidence ellipsoid (see Fig. 6B). In order to check whether a sample lies within the confidence ellipsoid, we calculate the normalised Euclidean distance as inconsistent if the proportion of samples within the 95% confidence ellipsoid is less than 90%.

The consistency of a network may drop drastically if only one dye is selected in an inappropriate model. Although this method allows to assess the agreement of the samples with the smFRET data of the whole network, the information which measurement(s)
is/are causing the inconsistency is hidden. Thus, if the network is inconsistent, we look at the uncorrected marginal confidence intervals of each smFRET measurement (see Fig. 6B). If several measurements involving the same dye are marginally inconsistent, this is a strong hint that this dye causes the inconsistency.

Fig. 7 shows an example of the marginal consistency check, which is interpreted as follows: As the measurement between dye 20 and dye 2 is highly inconsistent, we can infer that the chosen prior information of dye 20 and/or dye 2 is untrustworthy. Since also the measurement between dye 12 and dye 2 is marginally inconsistent and the remaining measurements involving dyes 12 and 20 are marginally consistent, the marginal consistency check suggests that the model for dye 2 causes the inconsistency. The calculation should be repeated with a different, i.e. less restrictive, model for dye 2.

4. Results

The consistency check provides a measure that tells the user if the a priori information consisting of dye models is in accordance with the smFRET data. Further, the marginal consistency check helps to identify which dye has been described by an incorrect model. This method can be used to find the maximal precision while staying consistent. In order to test this procedure we used the smFRET network of an archaerial pre-initiation complex as a test system [19]. In the beginning we applied to all dyes within this network the var-meanpos-iso model. This showed to be highly inconsistent, so we identified the dyes which were involved in most of the inconsistent measurements and changed their models to classic. We continued this procedure until the overall consistency had risen to over 90%. As a result we see that it is necessary to use for 14 out of 20 dyes the classic model to reach the desired consistency threshold (see Fig. 8).

For this example this procedure results on average in a more than 4-fold decrease in size of the credible volumes. The highest observed improvement in precision for any antenna was 18-fold (data not shown).

5. Discussion

5.1. Algorithm

The major obstacle in the structural inference of smFRET data is the huge variety of smFRET networks. They vary in dimensions (from 3 up to over 100 parameters) and macromolecular localisation geometry, i.e. in the spatial arrangement of dyes and the architecture of the macromolecule itself. As a result we also expect very different shapes of the posterior. This fact prevents the optimisation of the algorithm’s parameters to a given network, since one needs to ensure that parameters will be reasonably well-adapted for every smFRET scenario. Further, Fast-NPS needs to be simple to use, since the scientific user is only interested in the structural information one can get from the given smFRET
In order to fulfill these requirements, we developed an algorithm which differs in many points from the current suggestions for MCMC samplers. At first, the adaption processes are usually applied simultaneously to the sampling in a diminishing manner [30,40]. On the contrary, we established stopping criteria such that we can separate the adaption processes into successive phases. This idea enables us to find an endpoint for the adaption and to subsequently start sampling without any side-processes which would slow down the computation [41]. A disadvantage of this approach is that if the sampler has not efficiently sampled every major probability region of the parameter space until the end of the adaption phase, the adaption is incomplete or, worse, counterproductive. However, since we monitor the convergence to stationarity and optimise the parallel tempering scheme for each network individually, this is unlikely.

In order to monitor the convergence to the posterior we compute estimates of the cross entropy (see Eq. (5)). The corresponding stopping criterion is realised by a t-test upon the linear regression coefficient b of these estimates (see Section 3.1.3). Here, a potential danger is the fact that the convergence might show an oscillatory behaviour, which could trigger the stopping criterion in a false-positive way. Using basic elements of information theory, we were able to show that the cross entropy of an ergodic Markov chain is monotonically decreasing (see Appendix B). Thus, the linear regression-based stopping criterion should not terminate prematurely.

In our implementation we use the sample diagonal covariance matrix of the posterior estimated from the posterior samples for proposing a new configuration. This has advantages and disadvantages. By doing so we cannot provide correlated proposals, where changes in one parameter would be connected to changes in other parameters. Neglecting correlations between parameters increases the correlation between samples and hence, decreases the efficiency of the sampler. However, if we would have chosen to take the correlation terms into account, drawing from the proposal distribution would be computationally much more expensive.

Instead, we collect only every 10th sample, thus efficiently removing correlations. This simpler approach leads to a faster sampling of the posterior.

Since we estimate the global variances of the posterior, we have a fixed proposal distribution no matter where we are in configuration space. Thus, strictly speaking, the proposal kernel might be suboptimal everywhere in state space. A solution to this problem would be a state-dependent proposal kernel. There exist a few MCMC methods that incorporate the information from the gradient and/or Hessian matrix at each point in state space, e.g. the Riemannian Manifold Hamiltonian Monte Carlo, but these are difficult to tune and the operations in such an approach (i.e. the computation of the gradient and the Hessian matrix) are highly time-consuming [42]. Further, e.g. the gradient of the posterior can lead to problems when the prior is discontinuous in a certain subset of the state space. In such a situation the gradient of the posterior might drive the sampler to the edge of the prior, where it is likely to get stuck or the sampling efficiency is at least tremendously reduced. In contrast, our approach includes the information of the prior, in other words, the sample variances of the likelihood are different from those of the posterior. Thus, we can account for abrupt borders enforced by the prior. This feature is highly important so that we are able to deal with different prior shapes and sizes.

However, in the case of multimodality, e.g. two narrow, but distant modes of the posterior, the global variances lead the proposal kernel wrongly to believe that the posterior has a large expansion. This leads to low acceptance rates, since the sample variances are large compared to the width of a single mode. In our implementation we overcome this problem by scaling the covariance matrices to regain efficient sampling (see Eq. (6)).

In order to minimise the time to search the box collection prior (see Section 3.1.3), we were forced to truncate the proposal volume. As already mentioned in Section 3.1.5, we further bounded the scaled variances in order to avoid unnecessary sampling steps. As a consequence, it is possible that at higher temperatures all variances reach these limits and, hence, the shape of the proposal kernel which has adapted to the posterior is lost. Here, we can argue that if a proposal standard deviation of above 20 Å in position space is not enough to lower the Metropolis acceptance rate to 0.234, a radially symmetric shape of proposal in position, which would be the outcome of our implementation, is optimal for the given proposal volume.

Atchadé et al. have derived the characteristics that the posterior needs to satisfy, such that a geometric spacing in the inverse temperatures leads to a uniform swap probability [33]. Following this criterion, formally the smFRET posterior is not a candidate for geometric spacing. However, extensive testing has shown that all known smFRET scenarios in all model combinations express a nearly uniform swap probability under geometric spacing. As seen in Fig. 3B, there are deviations from the optimal swap acceptance rate as expected, but the sampling efficiency is still near the optimum [33]. Although at this point it is impossible to foresee all smFRET networks that could be analysed by Fast-NPS, the inherent similarity of all network posterior is a strong hint that no extraordinary deviations in swap probability are to be expected. The major advantage of the geometric spacing is that the time to adapt the temperature distances is almost independent of the number of chains. Thus, we can start with a higher number of chains and reduce them subsequently without drastically increasing runtime.

The heuristic to reduce the number of chains given in Eq. (10) is a novel result for the parallel tempering scheme. It is important to note that this heuristic applies only if the underlying state space is bounded, for example by a proper uniform prior. The swapping probability threshold to reduce the number of chains was set to

Fig. 8. Marginal consistency procedure in order to find the maximal precision while staying consistent. On the x-axis the number of dyes within the network which are described by the var-meanpos-iso model is displayed. The connected blue circles show the overall consistency of the network with the given model combination.
90%, because the convergence to 100% occurs only very slowly. Thus, with e.g. a threshold of 99% we would have many chains that are all similar to the prior, but yield no sampling advantage and would slow down the sampler. We have tested several networks and saw that the last chain defined by a swapping probability threshold of 90% placed samples nearly uniformly everywhere in configuration space.

The advantage of the heuristic is that it is intuitive and easily implemented. Further, it frees the user from the trial and error approach to find the optimal number of chains. It is noteworthy that the rate of convergence from the optimal average swap probability to the border of 90% is high for low-dimensional and low for complex scenarios. Thus, the heuristic is more costly for complex scenarios, because we have chains that accept too many steps, i.e. are inefficient. The power of the heuristic is seen when comparing the number of chains found for different smFRET scenarios. We saw not only networks that needed up to 30 chains, but also such ones which required only 2 chains.

The default values for the sampler were determined empirically. The optimal number of temperature chains increases with the complexity of the network which depends predominantly on the dimensionality of the posterior, i.e. on the number of dyes and their models. We have investigated a complex scenario involving 20 dyes which yields 100 dimensions if for all dyes the classic model is applied. For this network we needed 30 chains. We therefore set the starting value to 35 chains. If the user meets a scenario where the number of chains is not reduced by the heuristic, larger number of chains should be tested.

The number of effective samples determines how precisely the posterior is approximated. Hence, the larger the credible volumes are, the more samples we need. Since the classic model results in the largest credible volumes, we again took the above mentioned all-classic network to test how many samples are needed. Here, one million of effective samples yielded a good sampling of the posterior and showed no excessive computation time. If the user experiences that credible volumes are fragmented into many regions, an increased number of samples is advisable.

The thinning interval determines how many direct samples are collected until one effective sample is saved. It is an established method to reduce correlation between samples. High correlations can lead to an increase in surface roughness of the displayed credible volumes. Since the above mentioned network has highly curved credible volumes (see Fig. 1B), the correlation between samples is expected to be high. Thus, this network is taken as a worst case scenario. Here, the combination of $10^6$ effective samples with a thinning interval of 10 yielded smooth credible volumes in a short inference time. As mentioned before if the user encounters a more difficult network these default values might need to be adapted manually.

With all these adaptive features the underlying algorithm of Fast-NPS needs no user intervention such that the scientist can focus on the experiments and on the interpretation of the inference results. Further, the adaptive proposal kernel decreases the correlation and the heuristic tremendously reduces the number of chains needed, such that small smFRET networks can be processed in under 5 min. We saw that on a common desktop PC the runtime varies from below 5 min to 5 h. This comparatively short computation time allows the user to investigate many different strategies, geometries and model combinations. One major advantage of the software package Fast-NPS is that the analysis does not need any additional pre- or post-processing steps through external software, e.g. MD simulations, which would require in-depth understanding and make its application again unfeasible for a large fraction of life-scientists.

5.2. Dyemodelsandconsistencycheck

Oftentimes in smFRET experiments, it is assumed that the dyes are free to rotate within their fluorescence lifetime, i.e. their TDM distribution is uniform on the unit sphere. This then results in an orientation factor of $k^2 \equiv 2/3$, increasing the localisation precision enormously. This assumption has proven to be consistent with MD simulations for dyes attached to DNA (with no other macromolecules around) [43]. This can be explained by the fact that the DNA is mainly hydrophilic on the outside, and the dyes – being hydrophobic – are not prone to interact. However, structural smFRET studies are rather rarely applied solely to DNA samples, but more often to DNA-protein or protein-only constructs [44]. In a recent smFRET study regarding the structure of Hsp90 a geometrically averaged steady-state anisotropy of donor and acceptor below 0.2 was used as a heuristic threshold for assuming that the isotropic averaging condition is fulfilled [45]. Interestingly, still an empirically determined error of 5% in the Förster distance was applied to account for potential errors caused by the isotropic averaging assumption. Although this heuristic has not been rigorously proven, its application within a self-consistent field optimisation showed a tremendous decrease in uncertainty.

Fast-NPS pursues a different strategy. No assumptions about the dye motion which are proven wrong by experiment are made, instead different models are used to account for the heterogeneity found in nature. Thus, we started off with the most conservative classic model described by the theory of Dale and Eisinger and added the models iso, meanpos-iso, var-meanpos-iso and var-meanpos [37]. All these can be combined within a single network, such that each dye is assigned the best suited model.

In the Monte Carlo simulations solving for the expected FRET efficiency the geometric averaging poses problems. Testing showed that the object of integration in Eq. (11) may have an effect on $\mathbb{E}(\mathbf{E})$. However, the consideration of asymmetric dye volumes would entail a huge computational cost prior to the actual network inference, because the relative orientation of asymmetric dye volumes would have an influence on the conversion polynomial (see Section 3.2). Thus, for computing the polynomials needed for static position averaging, we assume that each dye moves freely within a sphere of radius 20 Å. It is important to account for position averaging, but its impact on the posterior is rather small compared to the impact on the posterior when fixing the position at the centre of the AV or when assuming orientational dynamic averaging. Thus the error introduced here should be negligible.

Since it is difficult to decide in advance which dye might follow which model, we have implemented an approach to test if the configuration samples taken from the posterior, which relies on possibly invalid model assumptions, violate the smFRET input data. This consistency check enables the user to easily investigate whether the assumptions made actually reflect the data. Here, because of the increased speed of Fast-NPS compared to NPS [15], repeated inferences can be made to find the best model combination. This can lead to a tremendous increase of localisation precision while maintaining consistency of model and data.

The danger in structural modelling via smFRET data is that an ill-posed localisation geometry (e.g. approximately coplanar satellites) may lead to phantom regions of high probability additional to the true position. Thus, a model combination may favour a phantom region while underweighting the true position, and still be entirely consistent. This, in turn, leads to an artificial picture of the macromolecular complex. This inherent problem of model assumptions strongly supports our conservative strategy. Hence, the user is advised to be careful in the choice of dye model combinations.
5.3. Results

The procedure to follow the marginal consistencies allows to achieve the maximal precision while staying in accordance with the smFRET data. Its application to an smFRET network of an archaear pre-initiation complex showed that rather few dyes could be set to a more restrictive model (see Fig. 8) [19]. Nevertheless, the increase in precision is on average more than 4-fold and for individual dyes up to 18-fold.

6. Conclusion and outlook

The software package Fast-NPS constitutes a tool for the structural analysis of smFRET experiments. The specialised adaption scheme of the underlying sampling engine makes the inference robust, efficient and fast. Since Fast-NPS is completely self-contained and its application follows the biochemical logic, it is user-friendly for all molecular life scientists. The addition of more restrictive dye models allows the user to increase the precision of the analysis of smFRET experiments. As a test that the dye models are applied in accordance with the experimental data, Fast-NPS provides a consistency check. This tool guides and assists the user in the application of the models.

It is of major interest that future, more complex dye models do not reflect assumptions and human intuitions whose consistency is tested afterwards, but their derivation should be driven solely by experimental data, such as time-resolved anisotropy measurements. Fast-NPS is now capable of fast inference of unknown dye positions. However, in the future Fast-NPS needs to be extended to allow for a prediction of structures consisting of several known macromolecular structures. Both a rigid-body approach as well as flexible docking of macromolecular complexes should be feasible as shown for structure determination by NMR data [46].

Acknowledgements

We thank Adam Marek Muschielok and Julia Nagy for helpful discussions regarding the physics of dye molecules and regarding the conduction of smFRET experiments.

Funding

J.M. acknowledges funding by the ERC starting grant remodelling and by DFG project number MI 749/4-1.

Disclosure

The authors have nothing to disclose.

Appendix A. Invariance of the average orientation factor ($k^2$) under a flip in direction by 180° of one or both average TDMs

Under the assumption that the distribution of fluorophore orientations is axially symmetric and the dynamic orientational averaging limit is fulfilled, Dale and Eisssinger showed that the average value of the orientation factor ($k^2$) only depends on the relative angles of the average transition dipole moments and the average axial depolarisation factors ($d_{1l}^2$) and ($d_{1l}^2$) of donor ($D$) and acceptor ($A$),

$$\langle k^2 \rangle = \frac{1}{3} \left( cos \theta_D d_{1l}^2 + cos \theta_A d_{1l}^2 \right) \left( 1 - \langle d_{1l}^2 \rangle \right) \left( 1 - \langle d_{1l}^2 \rangle \right),$$

(A.1)

where $k^2 = \cos \Psi_{DA} - 3 \cos \theta_D \cos \theta_A$ is the axial orientation factor [37]. Here, $\Psi_{DA}$ is the angle between the donor and acceptor average TDMs and $\theta_D, \theta_A$ are the angles between the vector interconnecting the donor and acceptor and their average TDMs, respectively. Let further $I_{DA}$ be the normalised vector interconnecting the dyes and $m_D, m_A$ the average TDM of donor or acceptor in Cartesian coordinates. The axial orientation factor can then be expressed by

$$k^2 = (m_D \cdot m_A - 3 (I_{DA} \cdot m_D) (I_{DA} \cdot m_A))^2,$$

(A.2)

and similarly the average value of the orientation factor by

$$\langle k^2 \rangle = k^2 \left( \langle d_{1l}^2 \rangle \right) \left( 1 - \langle d_{1l}^2 \rangle \right)$$

(A.3)

where $\cdot$ denotes the standard scalar product. Now it is straightforward to see that $k^2$ and $\langle k^2 \rangle$ are invariant under a flip in direction by 180° of one or both average TDMs.

Appendix B. Convergence of a Markov chain to its unique stationary distribution leads to convergence of the cross entropy

Let $P_n$ be an arbitrary distribution on a state space $\Omega$ of a Markov chain at time $n$, and let $P$ be the unique stationary distribution. The cross entropy $H(P, P_n)$ between these two probability distributions is then defined by

$$H(P, P_n) = E_P(-\log(P_n)) = \int_{\Omega} P(x) \log(P_n(x)) \, dx,$$

(B.1)

where $E_P$ stands for the expectation with respect to $P$ [47]. In order to understand why the convergence of the cross entropy implies stationarity of a Markov chain, let us express the cross entropy in another way:

$$H(P, P_n) = H(P) + D(P||P_n).$$

(B.2)

Here, $H(P)$ denotes the entropy of $P$ and $D(P||P_n)$ the Kullback–Leibler divergence or relative entropy of $P_n$ from $P$ [47]. The latter is a measure for the similarity between the two distributions. Since the stationary distribution does not change, the entropy of $P$ is a constant. Further, it is a known result that the relative entropy between a distribution $P_n$ on $\Omega$ at time $n$ and the stationary distribution $P$ decreases with $n$, i.e. we have

$$D(P||P_n) \geq D(P||P_{n+1}).$$

(B.3)

Thus, we also have

$$H(P, P_n) \geq H(P, P_{n+1})$$

(B.4)

or in other words, the cross entropy of a Markovian process is a monotonically decreasing sequence [47]. Since the estimator for the cross entropy given in the main part is unbiased, we can assume that its sequence is (apart from statistical fluctuations) monotonically decreasing, too. Thus, convergence of the cross entropy is an indicator for convergence of the Markov chain (while not guaranteeing it).

Appendix C. Recursive computation of the variance

In the adaption phase we need to calculate the sample variance (of many parameters) repeatedly, in order to decide if the computed value is precise enough (see Section 3.1.5). The sample variance $\text{Var}(X_k)$ of a vector of parameter values $X_k = (x_1, \ldots, x_k)^T$ is given by

$$\text{Var}(X_k) = \frac{1}{k-1} \sum_{i=1}^{k} (x_i - \bar{X}_k)^2,$$

(C.1)

where $\bar{X}_k = \frac{1}{k} \sum_{i=1}^{k} x_i$ is the sample mean of $X_k$ and $k$ is the number of parameter values collected. Assume we now collect an
additional sample, $X_{k+1}$. For the calculation of the sample variance $\text{Var}(X_{k+1})$ by Eq. (C.1) we need $X_{k+1}$, i.e. we would need to store $X_{k}$ and run again through $X_{k}$. This is costly both in required memory and runtime.

The following recursive algorithm avoids this problem. Set $M_{0} = E(X_{0})$ and $Q_{0} = (k - 1)\text{Var}(X_{0})^{2}$. With $M_{0} = Q_{0} = 0$ and the recursive equations

$$M_{k} := M_{k-1} + \frac{1}{k}(X_{k} - M_{k-1}) \tag{C.2}$$

and

$$Q_{k} := Q_{k-1} + \frac{(k - 1)}{k}(X_{k} - M_{k-1})^{2} \tag{C.3}$$

we obtain the sample variance at every step $k$ by

$$\text{Var}(X_{k}) = \frac{Q_{k}}{k - 1}. \tag{C.4}$$

Here, one only needs to store the last sample $x_{k}$, $M_{k-1}$ and $Q_{k-1}$ to calculate the current sample variance $\text{Var}(X_{k})$. Further, the recursive algorithm uses each sample only once.

### Appendix D. Monte Carlo scheme for smFRET efficiency conversion

The smFRET efficiency between dyes with a given model combination, where at least one dye is averaged over several positions, differs from the smFRET efficiency calculated from the mean position distance $R_{\text{MP}}$ of both dyes by Eq. (1). Clearly, the conversion polynomial for the combination of a classic donor with an iso acceptor is the same as for an iso donor with a classic acceptor and so forth, since the smFRET efficiency remains unchanged under an exchange of donor and acceptor positions. Thus, we obtain 15 unique model combinations which are displayed in Table D.1.

These combinations can be grouped into the four following sets due to their averaging regime:

1. no dye shows static position averaging:
   - classic–classic
   - classic–iso
   - iso–iso

2. at least one dye shows static position averaging and
   
   (a) both dyes are isotropically averaged:
   - iso–mpi
   - iso–vmp
   - mpi–mpi
   - mpi–vmp
   - vmp–vmp
   
   (b) one dye is isotropically averaged
   - classic–mpi
   - classic–vmp
   - iso–vmp
   - mpi–vmp
   - vmp–vmp
   
   (c) no dye is isotropically averaged:

For set 1 we need no conversion of the smFRET efficiency, since we assume fixed positions for both dyes. In set 2 we collected the remaining model combinations, i.e. all measurements that include at least one dye which is statically averaged over all positions. In subset 2a both dyes are isotropically averaged over all orientations, hence the average orientation factor is $(\kappa^{2}) = 2/3$. This results in only one polynomial.

In 2b and 2c at least one dye has a fixed orientation. Since a different orientation leads to a different expected smFRET efficiency, the Monte Carlo simulation would be necessary for all angles of one or both dyes. Since this would slow down the inference tremendously, we partition the surface of the TDM orientation hemisphere into 25 regions of equal area and use the polar coordinates of the centre of each surface area for the calculation of the conversion polynomial. This polynomial is then used for all TDMs pointing to this area. Thus, if one dye is fixed and one is isotropically averaged, the number of polynomials increases to 25 (set 2b). If both dyes have a fixed orientation (set 2c), i.e. we have two partitioned hemispheres, we have 25 · 25 = 625 unique polynomials.

### Appendix E. Normalised Euclidean distance

The Mahalanobis distance $D_{\text{MD}}$ is a similarity measure between two points in a vector space weighted by a covariance matrix $\Sigma$ [38]. It deviates from the Euclidean distance such that it accounts for scaling and correlation under $\Sigma$. In our case the Mahalanobis distance between a given point in smFRET space, i.e. $E := (E_{1}, \ldots, E_{n})^{T}$, and the mean vector $(\langle E \rangle_{1}, \ldots, \langle E \rangle_{n})^{T}$ of a given network with $n$ measurements is given by

$$D_{\text{MD}}(E, \langle E \rangle) = \sqrt{(E - \langle E \rangle)^{T} \Sigma^{-1} (E - \langle E \rangle)}. \tag{E.1}$$

Since the smFRET efficiencies follow an uncorrelated multivariate normal distribution (see Section 2), we have a diagonal covariance matrix and the Mahalanobis distance reduces to the normalised Euclidean distance given by

$$D_{\text{MD}}(E, \langle E \rangle) = \sqrt{\sum_{i=1}^{n} (E_{i} - \langle E \rangle_{i})^{2}}, \tag{E.2}$$

where $\sigma_{i}$ is the standard error of the $i$th measurement [39]. Since the argument of the square root in Eq. (E.2) is the sum of squared realisations of independent, standard normal random variables, it follows the $\chi^{2}$-distribution with $n$ degrees of freedom. Given the threshold $\chi^{2}(0.95)$ of the quantile function of the $\chi^{2}$-distribution to the confidence level of 95%, a realisation $E$ lies within the $n$-dimensional confidence ellipsoid of the normal distribution if we have $D^{2}_{\text{MD}}(E, \langle E \rangle) \leq \chi^{2}(0.95)$.

### References

4.2 J. Vis. Exp. 120, 2017: Structural Information from Single-molecule FRET Experiments Using the Fast Nano-positioning System

Title: Structural Information from Single-molecule FRET Experiments Using the Fast Nano-positioning System

Authors: Thilo Dörfler, Tobias Eilert, Carlheinz Röcker, Julia Nagy and Jens Michaelis

Journal: Journal of Visualized Experiments

Copyright: Authors

The corresponding video may be found on https://www.jove.com/video/54782
Author’s Contribution

The author of the present thesis by himself wrote the part of the manuscript (chapter 10 - 16) explaining the handling of the Fast-NPS software from producing the position priors over defining the network geometry to calculation and visualization of the results. Here, the consistency check is shortly explained and the author shows how the marginal consistency check is applied. All these points are presented by the author in the accompanying video tutoring the user through the whole data analysis by Fast-NPS (from 6 min 46 sec to 10 min 06 sec).
Structural Information from Single-molecule FRET Experiments Using the Fast Nano-positioning System

Thilo Dörfler1, Tobias Eilert1, Carheinz Röcker1, Julia Nagy1, Jens Michaelis1
1Institute of Biophysics, Ulm University
*These authors contributed equally

Correspondence to: Jens Michaelis at jens.michaelis@uni-ulm.de

URL: https://www.jove.com/video/54782
DOI: doi:10.3791/54782

Keywords: Biochemistry, Issue 120, Nano-Positioning System, Fast-NPS, single-molecule fluorescence, single-molecule Förster Resonance Energy Transfer, structural biology

Date Published: 2/9/2017


Abstract

Single-molecule Förster Resonance Energy Transfer (smFRET) can be used to obtain structural information on biomolecular complexes in real-time. Thereby, multiple smFRET measurements are used to localize an unknown dye position inside a protein complex by means of trilateration. In order to obtain quantitative information, the Nano-Positioning System (NPS) uses probabilistic data analysis to combine structural information from X-ray crystallography with single-molecule fluorescence data to calculate not only the most probable position but the complete three-dimensional probability distribution, termed posterior, which indicates the experimental uncertainty. The concept was generalized for the analysis of smFRET networks containing numerous dye molecules. The latest version of NPS, Fast-NPS, features a new algorithm using Bayesian parameter estimation based on Markov Chain Monte Carlo sampling and parallel tempering that allows for the analysis of large smFRET networks in a comparably short time. Moreover, Fast-NPS allows the calculation of the posterior by choosing one of five different models for each dye, that account for the different spatial and orientational behavior exhibited by the dye molecules due to their local environment.

In order to study the dynamics of macromolecular complexes and the structure-function relationship in particular, single-molecule methodologies have provided useful information. Several new strategies were developed providing an orthogonal approach on acquiring structural and dynamic information. Examples are high speed AFM6, mechanical manipulation7, fluorescence localization microscopy8, as well as single-molecule Förster Resonance Energy Transfer (smFRET)9,10. Since quite early on FRET has been termed a molecular ruler, due to the distance dependence on the length scale of biomacromolecules10.

One particularly interesting application of smFRET is to use the distance information obtained from smFRET measurements to infer structural information11,12,13,14,15,16,17,18,19,20,21,22,23. Due to the high time resolution of smFRET, the position of mobile parts of a protein structure can be localized. However, in order to extract quantitative information from smFRET data important correction parameters about the dye molecules need to be determined during the measurement14,15. With these correction factors, the FRET efficiency $E_{\text{FRET}}$ can be calculated using the formula

$$E_{\text{FRET}} = \frac{I_a - \beta \cdot I_d}{I_a + \gamma \cdot I_d}.$$
where $I_D$ and $I_A$ are the fluorescence intensities of the donor and the acceptor molecule, respectively (see Figure 2). The β-factor accounts for cross-talk, the leakage of donor emission into the acceptor channel and is calculated by

$$\beta = \frac{I_A}{I_D}$$

where $I_D$ and $I_A$ are the fluorescence intensities of the donor and the acceptor molecule after photo bleaching of the acceptor molecule.

The γ-factor corrects the difference in the relative detection efficiencies in the two channels as well as the differences in the fluorescence quantum yield of the donor and the acceptor dye. It is calculated from every individual time trace by

$$\gamma = \frac{\Delta I_A}{\Delta I_D} = \frac{I_A - I_A'}{I_D - I_D'}$$

Note, that this description neglects direct excitation of the acceptor molecule, which sometimes becomes important and would need to be corrected for as well. For determining these correction factors it is useful to excite both the donor as well as the acceptor in an alternating scheme in order to differentiate between photo-physical changes and structural dynamics.

In order to not only obtain quantitative smFRET efficiencies but also quantitative structural information, the Nano-Positioning System (NPS) was introduced in 2008. The name was chosen based on its similarities to the satellite-based Global Positioning System (GPS). The NPS is a hybrid technique combining smFRET and X-ray crystallography data for the localization of unknown dye positions in biomacromolecular complexes. The crystal structure serves as a reference frame and the smFRET results are used to obtain distance information between an unknown fluorophore position (antenna) and a position known from the crystal structure (satellite). In consecutive experiments the distances between the antenna and several satellites are measured and the position of the antenna is determined by means of a statistically rigorous analysis scheme based on Bayesian parameter estimation. As a result, not only the likeliest position of the antenna is computed, but its complete 3D uncertainty distribution, the so-called posterior, visualized by credible volumes. Moreover, NPS was expanded to allow for the analysis of complete smFRET networks.

The NPS has been used to solve a number of important questions in eukaryotic transcription, namely the course of the upstream DNA, the non-template DNA and the nascent mRNA within the RNA Polymerase II elongation complex, also demonstrating the effect of transcription initiation factors and the dynamic architecture of an open-promoter complex. Moreover, the NPS was used to elucidate the structure of the archaeal RNA Polymerase open complex in particular the position of transcription initiation factor TFE, which binds competitively to the same site as transcription elongation factor Sp4/5.

Since then, a number of smFRET based structural approaches have been published. When comparing different smFRET based structural methods, it becomes clear that the apparent precision of the method is highly dependent on the particular choice of dye models. One should note that dye molecules may exhibit different spatial and orientational behavior depending on their local environment.

To this end, Fast-NPS was introduced. Fast-NPS uses an advanced sampling algorithm reducing the calculation times drastically. Furthermore, Fast-NPS allows one to perform a structural analysis and for each dye molecule the user can choose from a set of five different dye models which will be described next. The most conservative model, called classic, assumes that the dye occupies only one, but unknown, position. At this position, the fluorophore can rotate freely within a cone, whose size is determined from its respective (time-dependent) fluorescence anisotropy. The orientation of the cone is not known, which leads to large uncertainties when converting measured smFRET efficiencies into distances. In this respect, the model is conservative, since it will lead to the smallest precision compared to the other dye models. Only for very short distances should the assumptions made by the classic model lead to a noticeably incorrect position determination. For typical smFRET values, the correct position is always enclosed in the comparatively large credible volume.

However, since a higher precision is desirable, it is important to develop and test alternative dye models, which could help to improve the precision. If the dye rotates much faster than its inherent fluorescence lifetime, the so-called iso model can be applied. Here, the orientation factor $\kappa^2$ (needed for calculating the characteristic isotropic Förster radius $p^{iso}$) is set to 2/3. As a result, the computed credible volumes are almost two orders of magnitude smaller as compared to those in the classic model. In the case that the fluorophore is found in an environment that enables not only fast reorientation, but additionally fast motion all over its accessible volume, the meanpos-iso model should be used. In this model, the dye effectively occupies only one mean position, where the spatial averaging is accounted for by a polygonal distance conversion. This model applies if for the (commonly hydrophobic) dye is attached to a hydrophilic region, e.g., the DNA. Application of the meanpos-iso model leads to a further reduction in the size of the credible volumes by a factor of approximately two. However, a dye linked to a protein might bind reversibly to several hydrophobic patches in its sterically accessible volume (AV). A fluorophore that instantaneously switches between these regions, but within one region undergoes free rotation and fast localized motion is best described by the var-meanpos-iso model. For a similar situation in which the dye is not free to rotate the var-meanpos model applies. More details about these models can be found in our recent publication.

These models provide an extensive repertoire to specifically account for the various environments a dye might encounter and applying them wisely optimizes its localization precision. In Fast-NPS, every dye molecule attached to a specific position can be assigned to an individual model, such that FRET-partners are allowed to have different models. This enables limitless and close-to-nature modelling. However, it is important that one performs rigorous statistical tests to ensure that the result obtained by the final model combination is still in agreement with the experimental data. These tests are included in the Fast-NPS software.

In order to apply Fast-NPS to experimental data the measurement of (only) three input parameters is required. First, the dye-pair specific isotropic Förster radii ($p^{iso}$) have to be determined. Therefore, the quantum yield (QY) of the donor dye, the donor fluorescence emission spectra and the acceptor absorption spectra need to be measured. These measurements can be carried out in bulk, using a standard spectrometer.
and a standard fluorescence spectrometer. For each pair, the $R_c$ is then calculated using the freeware PhotochemCAD and can be used in the NPS analysis. Moreover, the (time-resolved) fluorescence anisotropies of the dye molecules need to be obtained using a polarization (and time) sensitive fluorescence spectrometer. However, the most important input parameters for Fast-NPS are the smFRET efficiencies measured on a single-molecule fluorescence microscopy setup, such as a total internal reflection fluorescence microscope (TIRFM).

Here, we present a step-by-step protocol for obtaining smFRET data and applying Fast-NPS (Figure 1).

### Protocol

#### 1. Prerequisites and Lab Equipment

1. **Cut quartz glass slides using a diamond drill (0.75 mm) at positions indicated in Figure 4.** The design of the quartz glass slides is asymmetric in order to differentiate between the two sides of each slide.

2. **Perform the smFRET measurements on a prism type total internal reflection fluorescence microscope (TIRFM) (Figure 6).**

3. **Mount the flow chamber into the sample holder in a way that the holes in the quartz glass slide match the threads in the sample holder.**

4. **Perform the absorption measurement used for determining the quantum yield on a UV-VIS spectrometer (see Materials and Methods).**

5. **Prepare sample chambers according to published procedures.**

6. **Cut silicone tubing (0.58 mm ID, 0.96 mm OD) into 20 cm long pieces.** Insert one of the pieces to the inlet and the outlet screw of the acrylic glass holder to fix the position of the flow chamber.

7. **Perform the smFRET measurements on a prism type total internal reflection fluorescence microscope (TIRFM) (Figure 6).**

#### 2. Mounting of Flow Chambers in a Custom Holder

1. **Pull silicone tubing (0.8 mm ID, 2.4 mm OD) into hollow tap screws (M4) and cut the tubing on both ends straight leaving an overhang of 1 cm on both sides by using a sharp razor blade. Adjust the overhang of the tubing to about 2 mm on one side of the tap screw.**

2. **Mount the flow chamber into the sample holder in a way that the holes in the quartz glass slide match the threads in the sample holder. Tighten inlet and outlet screws gently to ensure that the inlet and outlet of the sample chamber are still penetrable. Gently tighten the four screws of the acrylic glass holder to fix the position of the flow chamber.**

3. **Cut silicone tubing (0.58 mm ID, 0.96 mm OD) into 20 cm long pieces. Insert one of the pieces to the inlet and the outlet screw of the measurement chamber. Close the inlet and outlet tubing by using a clamp.**

#### 3. smFRET Measurement on the TIRF Microscope

1. **Use a syringe to wash the sample chamber with 500 µL of PBS. Prevent air bubbles from entering the sample chamber at all times by creating a droplet at the end of the inlet tubing before changing to a different buffer solution.**

2. **Flush the sample chamber with 100 µL neutravidin (0.5 mg/mL in PBS) solution and incubate 15 min at room temperature.**

3. **Wash out the neutravidin solution with 500 µL of PBS.**

4. **Screw the metal holder for the prism onto the sample chamber.**
5. Mount the sample chamber to the micrometer stage of the TIRF-microscope. Make sure to mount the sample chamber horizontally as straight as possible in front of the objective to avoid defocusing during scanning process.

6. Start the software controlling the EM-CCD camera and the software for controlling the piezo-motors of the stage.

7. Adjust the focus of the microscope objective by looking at the reflections of the IR laser.

8. Place the prism (PS991, n = 1.52) on top of the metal prism holder. Adjust the lateral position of the prism to make sure that the laser beams hit the prism then use an adhesive and incubate with UV light for 5 min.

NOTE: Mounted prism can be reused by cleaning.

9. In the camera control software click "Setup Acquisition" and define the following acquisition parameters: 100 ms integration time, 401 frames/movie (green camera), 400 frames/movie (red camera), electron multiplier gain 225, pre-amplifier gain 5x and readout rate 3 MHz at 14 bit.

10. Create a folder on the local hard drive for the measurement. Choose a desired name for the measurement files, e.g., Year-Month-Day. In the software settings go to the "Auto-Save" rider, enable "Auto save" and choose file format *.sif for movie acquisition. Select the folder on the hard disk. Use the folder name as filename.

11. Enable the "AutoIncrement" function (set start value to 1). Enable the attachment of an operator to the filename. Use "DON" and "ACC" for the donor and the acceptor channel, respectively. Choose "_" as separator.

12. In the camera control software click on "Video" to start the live image of the camera and bleach background fluorescence by scanning the sample chamber using maximum laser intensity of three lasers (combined >3,000 W/cm² for 10 s per field of view).

13. Switch off the blue laser. Decrease the intensity of the green laser to around 200 W/cm² and to around 40 mW/cm² for the red laser if alternating laser excitation (ALEX) is used.

14. Dilute the biotinylated fluorescent sample to a concentration of 50-100 pM. Load 100 µL of the solution. The sample is immobilized on the chamber surface upon binding.

NOTE: Make sure not to overload the chamber. Neighboring molecules must be separated from each other.

15. If necessary, load an additional 100 µL of a 2x more concentrated sample to the chamber.

16. Seal the inlet and outlet tubing of the measurement chamber using clamps after loading is completed.

17. Switch off all lasers and use the piezo-motors to move the flow chamber two fields of view further.

18. In the camera control software click on "Take signal" to start movie recording and switch on the laser at the same time. Ensure that more than 80% of the molecules are bleached by the end of the movie by adjusting the laser power.

19. Repeat steps 3.17 and 3.18 for the whole prebleached sample chamber region.

4. Acquisition of the Transformation Map ("beadmap")

1. Prepare a flow chamber as described in Sections 1.1, 1.2 and 2.

2. Use avidin-coated fluorescent multispectral beads which show fluorescence emission in the donor and the acceptor channel. Vortex the stock for 1 min, then dilute 50 µL of the stock in 50 µL ddH₂O. Vortex again for 1 min, sonicate 1-2 min, then vortex another 10 s.

3. Carry out steps described for smFRET measurements (Section 3.5-3.10).

4. Load 100 µL (1 chamber volume) of the 1.2 diluted fluorescent beads into the flow chamber. Wait 10 min for the fluorescent beads to bind to the surface.

5. Use the acquisition parameters in 3.9 but change movie length to 26 (green camera) and 25 (red camera) and the electron multiplier gain to 10.

6. Set the intensity of the green laser to a value of 20 W/cm².

7. Take one movie at a field of view with approximately 50-100 beads.

5. Processing and Analysis of smFRET Data

1. Use the custom written software SM FRET for the analysis of the beadmap (see Materials and Methods) and the acquired movies. Start the program viewPlot1.m.

2. Click on Analysis|Batch Analysis, uncheck the option "ALEX" if it has not been used. For best performance choose the threshold for peak finding "high". Press "OK".

3. Choose "NO" when asked if a beadmap has already been analyzed. Browse the folder containing the acquired beadmap and select the *.sif file (by double-clicking on it). In the next dialogue window press "OK".

NOTE: If a beadmap has already analyzed in a previous measurement, choose "YES" here and select the saved beadmap by browsing to the correct folder and double clicking on the beadmap *.map file. Continue with step 5.8.

4. Choose two single beads positioned in opposite corners of the field of view. The pixel intensities are color-coded from dark blue (low intensity) to dark red (high intensity).

5. Click on the center of the first bead. If the center of the molecule can be located clearly by the color-coding, choose "YES" or else click "NO" and choose a different molecule pair.

6. Position the crosshair on the pixel showing the maximum intensity and press "KEEP". Repeat the process with the second channel.

7. Click on the molecule in the opposite corner and repeat steps 5.5 and 5.6.

NOTE: The relative pixel shift of the two channels is displayed in the command window and the transformation map is automatically saved as *.map file into the folder containing the beadmap *.sif file.

8. To load the donor and acceptor movies (*.sif) for the "batch analysis", browse to the folder, select all movies that shall be analyzed and click on "OK". In the next dialogue window, press "OK".

NOTE: The batch analysis is finished when the last lane displayed in the command window starts with "Finished analyzing...". The detected molecules are displayed in a new window, which also indicates the relative shift of the donor and the acceptor channel determined from the transformation map.

9. To load the batched movie files click on File|Load. Uncheck the option "ALEX" if it has not been used. Set the smoothwidth to 10 and click "OK". Choose the folder containing the *.ttr files and click on "select all" and "OK" in the next context menu.
10. If the displayed trace features the characteristic smFRET phases (Figure 2) press on the toggle button "Not selected" and first select the time point of the beginning of the FRET event by moving the line with the mouse cursor and clicking the left mouse button. Next select the time point of the bleaching of the acceptor molecule and finally the time point of the bleaching of the donor molecule.

11. In the next window the FRET efficiency is plotted in blue. To select the trace press the "Yes" button otherwise choose "No". To re-access a time trace click the "Prev" button.

12. Repeat the procedure until the last molecule of the movie.

13. After analyzing the last molecule in a movie save the selected traces by clicking on "File|Save". Save the selected traces in the same folder as the *.sif files.

14. Repeat steps 5.10-5.13 for all acquired movies.

15. Execute the program combine_fret_results.m. Select the folder containing the *.res files and all the *FRETonly_trace files. Save the molecule-wise FRET and frame-wise FRET files as MW.dat and FRW.dat, respectively.

NOTE: The *.dat files are saved as ASCII files. The FRW.dat file contains six columns and one row for each FRET-frame. The sixth column contains the corrected frame-wise FRET efficiency. The MW.dat file contains 21 columns and one row per selected FRET molecule. The third column contains the molecule-wise FRET efficiency.

6. Displaying smFRET Data in Histograms

NOTE: In order to extract the mean smFRET efficiency of all recorded smFRET data the frame-wise data or the molecule-wise data are plotted in histograms and analyzed using Gaussian fits to (multiple) peaks. In the following, the protocol uses a commercial data analysis software (see Materials List). However, any other available software can be used instead.

1. Open a data analysis software (see Materials List). Click on File|Import|multiple ASCII. Select the folder containing the FRW.dat file. Select the file and press "OK". Accept the input option with "OK" without a change.

2. Select the third column C(Y) containing the corrected FRET efficiencies, right-click on the column and select Plot|Statistics|Histogram. In the histogram window double-click on the columns and deselect "automatic binning" and select a desired bin-size, e.g., 0.05. Also choose beginning and end values, e.g., -0.025 and 1.025.

3. Select the columns of the histogram by left-clicking on them. Then right-click and choose "go to Bin Worksheet". Select the "Counts" column by left-clicking on it and then right-click and select Plot|Column|Bar Pie|Column.

4. In the column bar plot go to Analysis|Fitting|Nonlinear curve fit|Open dialog. Choose "Gaussian" under Function then go to the "Parameter" tab. Select auto parameter initialization. Fix the offset value (y0) at 0. Click on "Fit".

NOTE: The fit function as well as the fitting details are now displayed in the column plot. The "xc" value gives the center of the fit function, i.e., the mean FRET efficiency that serves as the input parameter for the NPS software.

7. Measurement of the Quantum Yield

1. Perform quantum yield determination with the relative method similar to the procedure described by Würth et al.\(^{35}\), using Rhodamine 101 dissolved in ethanol (QY = 91.5%) as a standard.

2. Record absorption spectra at a UV-VIS spectrometer using an 80 µL volume in an absorption cuvette of 1 cm path length. The absorbance at the wavelength which will be used for fluorescence excitation has to be ≤ 0.05.

3. Record emission spectra on a lamp-calibrated spectrometer operated in photon-counting mode. Perform the measurements with Glan-Thompson polarizers in the excitation (0°) and emission (54.7°) path (magic angle conditions) using a spectral bandwidth of about 5 nm and 2.5 nm for the excitation and emission monochromator, respectively. Measure samples after transferring them to a fluorescence cuvette with 3 mm path length taking care that the count rate does not exceed 10^6 s\(^{-1}\).

1. Calculate quantum yield according to

\[ \Phi = \frac{n^2}{n_{\text{std}}^2} \cdot \frac{1}{\int_\lambda f(\lambda) d\lambda} \cdot \frac{\left[1 - 10^{-4f(\lambda)}(\lambda) \right]}{\left[1 - 10^{-4f_\text{std}(\lambda)} \right]} \cdot \Phi_{\text{std}} \]

where \( n \) and \( n_{\text{std}} \) are the refractive indices of the solvent of the sample and the standard, respectively. \( f(\lambda) \) and \( f_\text{std}(\lambda) \) are the fluorescence intensities of the sample and the standard at the wavelength \( \lambda \). \( A(\lambda_{\text{ex}}) \) and \( A_{\text{std}}(\lambda_{\text{ex}}) \) are the absorbance of the sample and reference at the excitation wavelength and \( \Phi_{\text{std}} \) is the quantum yield of the standard.

8. Calculation of the Isotropic Förster Radius

1. Calculate the isotropic Förster radius (R\( _{0}^{\text{iso}} \)) from the emission spectrum of the donor molecule, the absorption spectrum of the acceptor molecule, the quantum yield of the donor and the refractive index of the medium. Use the freeware PhotochemCAD for the calculation of R\( _{0}^{\text{iso}} \). However any other available software can be used instead.\(^{36}\)

9. Measurement of the Anisotropies

1. Determine steady-state fluorescence anisotropies from recordings of fluorescence spectra with various excitation/emission polarizer settings (V/V, V/H, H/V, H/H)\(^{36}\).
2. Calculate the G-factor, which corrects for polarization artefacts of the instrument, for each wavelength from the ratio

\[ G(\lambda) = \frac{I_{xy}(\lambda)}{I_{yy}(\lambda)} \]

and use it to calculate the anisotropy value for each wavelength:

\[ r(\lambda) = \frac{I_{xx}(\lambda) - G(\lambda) \cdot I_{yy}(\lambda)}{2G(\lambda) \cdot I_{yy}(\lambda)} \]

where \( I_{xy} \) indicates the intensity for excitation polarization \( x \) and emission polarization \( y \).

3. Average the values across the emission spectral range to calculate the steady-state fluorescence anisotropy.

10. Installation of Fast-NPS Software

1. Download UCSF Chimera from http://www.cgl.ucsf.edu/chimera and follow the installation guide.
2. Go to the website of the "Institute of Biophysics" at Ulm University: https://www.uni-ulm.de/en/nawi/institute-of-biophysics/software.html. Download the current version of Fast-NPS and extract it to a folder of choice. Open the subfolder "Redistributable" and install the Visual C++ Redistributable which is appropriate for the system.

11. Centering the pdb File

1. Open the pdb file(s) of interest in Chimera. Select all atoms of the macromolecular complex and calculate the coordinates of the centroid (Tools|Structure Analysis|Axes|Planes|Centroids|Define centroid…|Ok).
2. Open the Reply Log (Favorites|Reply Log) and the Transformation Tool (Tools|Movement|Transform Coordinates). Enter the coordinates of the centroid shown in the Reply Log into the textbox "Shift" of the Transform Coordinates window and change the sign of every coordinate. Press "Apply" and save the file with "Save PDB" (File|Save PDB).

12. Setting up the Position Priors

NOTE: All values are considered in angstrom.

1. Start the technical computing language and change the current folder to the local Fast-NPS folder. Enter in the command window: FastNPS.
2. Create a new jobfile in the Project Manager (Project|New).
3. Set up the position prior (Tools|Model dye prior).
4. In the panel "prior basics" define the spatial resolution of the position prior by entering its value (2 Å is recommended).
5. Exclude the interior of the macromolecule by activating the checkbox and clicking on the "load PDB" button. Select and load the centered pdb file as described in Section 11.
6. Specify the approximate diameter (13 Å is recommended, see discussion) of the dye by entering its value.
7. Enter a skeletonization distance, i.e., the distance the dye molecule may penetrate into the macromolecule (2 Å is recommended).
8. In the panel "maximum prior size" enter the minimal and maximal coordinates of the position prior (recommended: x in [-150,150], y in [-150,150] and z in [-150,150]).
9. When defining a satellite, activate the checkbox "attachment via flexible linker" in the panel "prior basics" and enter in the panel "linker" the coordinates of the atom (in the centered pdb file) at which the dye molecule is attached. Further, specify the length and the diameter of the linker by entering their values (13 Å and 4.5 Å are recommended, see discussion). In case of an antenna skip this point.
10. Press the button "calculate accessible volume".
11. Save the position prior and optionally export it for visualization purposes using software such as Chimera.

13. Defining the Network Geometry

1. Open the Define Measurement Window (Mode|Edit Geometry).
2. Create a new dye molecule by pressing the button "New" in the panel "Dyes".
3. Set its fluorescence anisotropy (Section 9) by entering a value and select a dye model within the dropdown menu "Dye model".
4. Press the button "Load", select the corresponding position prior and check the activate check box of the dye. Repeat this procedure for all dyes, i.e., for all antennas as well as for all satellites.
5. After creating all dyes, define the measurements. Create a new measurement by clicking "New" in the panel "Measurements".
6. Select its FRET partners in the dropdown menus "Dye1" and "Dye2" below.
7. Enter the smFRET efficiency with error and the isotropic Förster radius of this dye pair.
8. Finally, check the activate check box of the measurement. Repeat this procedure for all measurements.

NOTE: Oftentimes the network becomes increasingly complex, so that the user might get confused. In order to prevent mistakes, check the network visually by pressing the "Check Network" button. The figure displays the activated dyes and indicates measurements via lines interconnecting the FRET dyes.
14. Calculation

1. Open the Calculation Window (Mode|Calculation).
2. If every dye in the network has a specific model assigned, select "User defined" and start the calculation by pressing "Calculation". To have all dyes in the same model, select one of the five models (classic, iso, meanpos-iso, var-meanpos-iso and var-meanpos) and continue.

NOTE: The command window will indicate the progress of the calculation. Fast-NPS will do so with a pop-up message, when the calculation has been completed.

15. Visualization of Results

1. In order to export the credible volumes of the dyes, open the View Results Window (Model|View Results).
2. Export dye densities:
   1. Export dyes singly or all simultaneously. In order to export a single dye select it in the panel "Displayed Dyes" and press "Export Density". Enter a resolution (2 is recommended) and choose a file type for exportation. On the right the density is previewed and some of its mathematical characteristics are shown.
   2. In order to export all dyes simultaneously push "Batch Export".
3. Open the resulting density files in Chimera.

16. Consistency Check of Chosen Model Combination

1. Open the View Results Window (Model|View Results). If in the panel "Calculation Info" the text box "Consistency" displays a value lower than 90% the current model does not represent the measured smFRET efficiencies sufficiently and is thus inconsistent.
2. In case of inconsistency push the button "Detailed Consistency". Search for the measurements that have a value below 90%. If one or more dyes are predominantly involved in these measurements, their models are likely to cause the inconsistency. Consider different dye models for these dyes and rerun the Fast-NPS calculation.

Representative Results

Transcription is the first step in gene expression in all organisms. In Archaea, transcription is carried out by a single RNA polymerase (RNAP). Compared to eukaryotes, the archaeal RNAP bears a striking structural resemblance to their eukaryotic counterparts while having a simpler transcriptional machinery. Thus, Archaea can be used as a model system to study eukaryotic transcription initiation by RNA Polymerase II (Pol II). Recently, the complete architecture of the archaeal RNA polymerase open complex has been determined from single-molecule FRET and NPS. The data from NPS analysis was used to build a model of the complete archaeal open promoter complex, which provides useful insights into the mechanism of transcription initiation.

To elucidate this structure, smFRET efficiencies were measured between unknown antenna dye molecules located within the open promoter complex and several known satellite dye molecules that were incorporated at five reference sites in the RNAP, whose positions are known from crystallographic structures (pdb-ID: 2WAQ)

Figure 7 depicts the model of the complete archaeal open promoter complex built from the NPS analysis. It comprises the double stranded promoter DNA (light and dark blue), the RNA Polymerase (grey) and the transcription initiation factors TBP (purple), TFB (green) and TFE (yellow). The model is superimposed with the results from the NPS analysis, the credible volumes, which were calculated using the classic model (A), the iso model (B), the meanpos-iso model (C), the var-meanpos-iso model (D) and the var-meanpos model (E).
Figure 1: Workflow of the acquisition and processing of the parameters needed for the Fast-NPS calculation. Please click here to view a larger version of this figure.

Figure 2: Exemplary fluorescence intensity time trace of a smFRET event. The fluorescence intensities of the donor (green) and the acceptor molecule (red) showing the three characteristic phases, namely I: smFRET, II: donor fluorescence after acceptor photobleaching, III: background fluorescence after donor photobleaching. Please click here to view a larger version of this figure.
Figure 3: Schematic illustration of the flow chamber for smFRET experiments. The flow chamber is mounted onto a customized metal holder with acrylic glass holders. The sandwich-design of the flow chamber comprises a quartz glass (fused silica) slide with two holes for attaching inlet and outlet tubing, a sealing film and a coverslip that closes the flow chamber. The prism for TIRF illumination is mounted onto the lower half of the flow chamber. Hollow tab screws provide the inlet and outlets for the flow chamber. Please click here to view a larger version of this figure.

Quartz glass slide
76x26x1mm

Figure 4: Preparation of the quartz glass slide and the sealing film. Mechanical drawing of the quartz glass slide indicating the positions of the holes (given in millimeters). Please click here to view a larger version of this figure.
Figure 5: Mechanical drawing of the flow chamber. The measures for the aluminum prism holder, acryl glass holders and aluminum mounting frame are given in millimeters. Please click here to view a larger version of this figure.
Figure 6: Schematic illustration of the prism-type TIRF setup used for smFRET experiments. Abbreviations for optical components: A, aperture; DM, dichroic mirror; F, emission filter; L, lens; M, mirror; O, objective; P, prism; PSD, position sensitive photo-diode; S, sample; PS, positioning stage; T, telescope. Please click here to view a larger version of this figure.
Figure 7: Simulation results of the different model assumptions. All pictures show the archaeal RNA polymerase (pdb-ID: 2WAQ, top view) together with the model for promoter DNA (tDNA and ntDNA in blue and cyan, respectively), TBP (purple), TFB (green) and TFE (yellow) in the archaeal open complex. The credible volumes are superimposed for the NPS simulation results of (A) the classic model, (B) the iso model, (C) the meanpos-iso model, (D) the var-meanpos-iso model and (E) the var-meanpos model. All volumes are shown at 68% credibility. The classic and the var-meanpos networks are consistent with the smFRET data. In contrast, networks where for all dyes the iso, meanpos-iso or the var-meanpos-iso model is chosen are inconsistent with the measured data. Please click here to view a larger version of this figure.

Discussion

We present the setup and experimental procedure to accurately determine FRET efficiencies between dyes attached via flexible linkers to biomacromolecules, i.e., nucleic acids and/or proteins.

In order to ensure precise smFRET measurements (Section 3), it is crucial to exclude air from the flow chamber at any time during the measurement. Furthermore, make sure to not overload the flow chamber with fluorophores. The fluorophores must be clearly separated to ensure correct analysis. As smFRET pairs, which do not show bleaching of the donor have to be excluded from the analysis, make sure that >80% of the molecules in the field of view are bleached at the end of the movie. To account for inhomogeneties in the sample the β-factor and
the γ-factor, correcting the cross-talk and relative detection efficiencies of the donor and acceptor channel, respectively, are calculated for each FRET pair individually.

The camera settings (integration time, electron multiplier gain, pre-amplifier gain and readout rate described in Section 3.9) should be set to values giving the best tradeoff between signal to noise ratio, dynamic range and time-resolution. They need to be re-adjusted for different experiments or if different hardware is used. The numbers of frames need to be high enough to ensure that most of the donor molecules bleach within the observation time.

For the measurements on the fluorescence spectrometer (Sections 7 to 9) a good compromise between the signal intensity and the spectral resolution of the recorded data has to be found. To this end the slits in the excitation and emission pathway of the fluorescence spectrometer have to be adjusted dependent on the instrument used and the sample concentration.

Moreover, we present the Fast-NPS analysis method to obtain structural information of transient or dynamic macromolecular complexes. NPS has been applied to reveal the path of the non-template DNA strand and the position of transcription initiation factors in the archaeal RNA polymerase open complex. Using the network of more than 60 different distance measurements, we showed that Fast-NPS, equipped with a newly implemented sampling engine (Eilert, T., Beckers, M., Drechsler, F., & Michaelis, J. in preparation), reduces the time needed for the analysis of this complex smFRET network by ≈2 orders of magnitude, as compared to the original global NPS method27. The algorithm’s robustness is rooted in a Metropolis-within-Gibbs sampler combined with a parallel tempering scheme. Fast-NPS shows exact reproducibility of network results and is consistent with results published earlier27.

Several different methods have been published aiming to infer structural information from smFRET measurements11,12,13,14,15,16,17,18. All of these approaches provide only one specific dye model. Thus, dyes, that do not fulfill the assumptions made by the respective model, cannot be used or lead to false structural information. Fast-NPS, on the contrary, allows to select for each dye molecule a different model. This helps to account for different conformational behavior of both, the dye molecule itself, as well as the linker used for its attachment. The local molecular surroundings of the dye molecule, as well as its physical properties will determine which model is most appropriate.

For the analyzed smFRET network of the archaeal initiation complex, an isotropic assumption for all dye molecules leads to a drastic decrease in the size of the credible volumes as compared to the classic model. In combination with a dynamic position averaging for all dye molecules the median of all credible volume sizes (at 95%) reduces to less than 0.5 nm. However, these dye molecule posteriors are no longer consistent with their smFRET measurements, indicating that the assumptions made lead to false structural information. In contrast, the posteriors determined in the classic model are consistent with the determined smFRET efficiencies.

As the assumption of isotropic and/or dynamic position averaging for all dye leads to inconsistencies, Fast-NPS enables dye molecule priors in which each dye can be assigned one of the five models. Each model uses the same accessible volume. The algorithm for the calculation of the dye AVLs makes several assumptions. At first, the fluorophore’s spatial shape is approximated by a sphere. Thus, a diameter taking into account the fluorophore’s width, height and thickness should be used (Section 12). Further, the linker’s shape is approximated by a flexible rod. The values presented in Section 12 were computed for the dye Alexa 647 attached via a 12-C linker. To date, it is not possible to accurately determine a priori which model is most suited, given an experimental geometry, and thus all models should be tested. In general, one will choose the model which gives the smallest posterior size, while still being consistent with the data. To test whether a choice of models is consistent with the smFRET data, we calculate both the posterior and the likelihood. Consistency means that more than 90% of the samples collected from the posterior are within the 95% confidence interval of the likelihood.

While it is true that the lower the anisotropy, the smaller the distance uncertainty, in a smFRET network geometric arrangements of the dye molecules also have to be taken into account. Thus, while representing dye molecules with a low fluorescence anisotropy with an iso model is a typical first choice, the consistency test provides a more direct means for selecting the correct dye model. The optimal choice of dye models can lead to a drastic increase in localization precision and at the same time retain the network’s consistency with its FRET data.

To summarize, Fast-NPS allows to gain structural and dynamic information of large macromolecular complexes. In contrast to common structural methods such as x-ray crystallography or cryo electron microscopy this allows for monitoring highly flexible or transient complexes, thus greatly widening our mechanistic understanding of complex biological processes.

Disclosures
The authors have nothing to disclose.

Acknowledgements
The authors thank B. Gruchmann for the mechanical drawings of the flow chamber. Further, we want to express our gratitude to Max Beckers and Florian Drechsler for insightful comments and discussions regarding NPS and the underlying sampling engine.

References


Title: Precision and accuracy in smFRET based structural studies - a benchmark study of the Fast-Nano-Positioning System

Authors: Julia Nagy, Tobias Eilert and Jens Michaelis

Journal: Journal of Chemical Physics

Copyright: Authors
Author’s Contribution

The author of the thesis at hand partially wrote the paper, amongst others the text about the dye assumptions and models. He conducted the Fast-NPS calculations and wrote the corresponding section (II C). The statistical test on different means described in Section II D was designed and applied by the author. He constructed and conducted the methods to assess the precision and accuracy of both benchmark studies (Sections II E and F). Further, the author was strongly involved in the interpretation of the analysis results.
Precision and accuracy in smFRET based structural studies—A benchmark study of the Fast-Nano-Positioning System

Julia Nagy, Tobias Eilert, and Jens Michaelis
Ulm University, Institute of Biophysics, Albert-Einstein-Allee 11, Ulm 89069, Germany
(Received 25 September 2017; accepted 28 November 2017; published online 18 December 2017)

Modern hybrid structural analysis methods have opened new possibilities to analyze and resolve flexible protein complexes where conventional crystallographic methods have reached their limits. Here, the Fast-Nano-Positioning System (Fast-NPS), a Bayesian parameter estimation-based analysis method and software, is an interesting method since it allows for the localization of unknown fluorescent dye molecules attached to macromolecular complexes based on single-molecule Förster resonance energy transfer (smFRET) measurements. However, the precision, accuracy, and reliability of structural models derived from results based on such complex calculation schemes are oftentimes difficult to evaluate. Therefore, we present two proof-of-principle benchmark studies where we use smFRET data to localize supposedly unknown positions on a DNA as well as on a protein-nucleic acid complex. Since we use complexes where structural information is available, we can compare Fast-NPS localization to the existing structural data. In particular, we compare different dye models and discuss how both accuracy and precision can be optimized. Published by AIP Publishing.

https://doi.org/10.1063/1.5006477

I. INTRODUCTION

The understanding of fundamental biological questions oftentimes relies on our structural knowledge of their molecular constituents as well as on the in-depth analysis of the dynamic changes underlying their biological function. Here, single-molecule Förster resonance energy transfer (smFRET) provides an ideal method to analyze flexible domains and conformational changes which remain undetected by traditional biochemical or structural methods. In particular, smFRET allows us to gain real-time distance information between a pair of dye molecules attached to a complex (Ha et al., 1996; Hohlbein et al., 2013; and Weiss, 1999).

In the past years, a growing number of approaches have been developed to use smFRET data in order to obtain structural information (Rasnik et al., 2004; Andrecka et al., 2008; Brunger et al., 2011; Kalinin et al., 2012; and Svensson et al., 2014). In particular, to achieve accurate smFRET-based localization of unknown positions within macromolecular complexes, a probability-based analysis method, termed the Nano-Positioning System (NPS), was developed for the analysis of a single position (Muschiolok et al., 2008) or a network of distances (Muschiolok and Michaelis, 2011). The NPS combines data from smFRET measurements with existing structural information and Bayesian parameter estimation analysis. As a result, three-dimensional probability density functions for dye molecules attached to positions in unknown, flexible regions of the complex of interest can be calculated. This method has been used in investigations of the eukaryotic RNA polymerase II (Pol II) transcription elongation complex to study the position of the exiting RNA (Muschiolok et al., 2008) and the position of non-template and upstream DNA (Andrecka et al., 2009). Moreover, the architectures of a minimal Pol II open complex (Treutlein et al., 2012) and of the complete archaeal RNA polymerase open complex (Nagy et al., 2015) have been determined. Also, the conformational states of Rep helicase undergoing ATP hydrolysis have been analyzed using the NPS (Balci et al., 2011).

Recently, the Fast-Nano-Positioning System (Fast-NPS) has been introduced for the analysis of large smFRET networks, yielding a reduction of the computation times by two orders of magnitude as compared to the original NPS (Beckers et al., 2015 and Eilert et al., 2017). The novel feature of the Fast-NPS is to account for specific aspects in the local environment of the dye molecule by being able to choose a specific model for each dye position. In all the five different models, we assume that the rotation of the transition dipole moment of the dye is dynamic (i.e., the rate of rotation $k_{R}$ is much greater than the transfer rate $k_{T}$, $k_{R} \ll k_{T}$) and the translation of the dye is static (i.e., the rate of translation $k_{Trans}$ is much smaller than the transfer rate, $k_{T} \gg k_{Trans}$).

The most conservative model is termed classic (Eilert et al., 2017). Here, it is assumed that the dye resides in one particular, but unknown, location within its sterically accessible volume (AV), and the orientation of the transition dipole moment obeys an axially symmetric orientation distribution. The opening angle (or restriction) of this distribution is related to the respective fluorescence anisotropy, which can be therefore used to limit the range of possible values of the orientation factor (Dale et al., 1979). In the iso model, the dye is still fixed at a certain position, but in contrast to the classic model, an unrestricted dye rotation is assumed, i.e., the rotational
distribution is uniform. In the \textit{meanpos}-\textit{iso} model, we further assume that the dye undergoes translational diffusion and that the whole AV is probed \textit{uniformly}. In other words, the dye can switch between different positions; for each excitation of the dye, a different position is probed and all positions are sampled with equal probability. In the \textit{var-meanpos-iso} model, the mean of the dye molecule position is no longer fixed to the center of the AV, but variable, i.e., some regions of the AV are more likely to be visited than others due to attractive or repulsive interactions. Finally, if we use this position model and again restrict the dye rotation to an axially symmetric distribution, we obtain the \textit{var-meanpos} model. The assumptions made in these five different models are summarized in Table I.

Here we present two proof-of-principle benchmark studies where the result of the Fast-NPS localization of a dye molecule is compared to available structural information. In these studies, one can compare the results for different dye models in achievable precision and accuracy. For the first case study, a double-stranded DNA (dsDNA) molecule was chosen where the unknown dye position ("antenna") was located at the center of the DNA, surrounded by five approximately evenly distributed dyes, used as FRET partners with known positions ("satellites"). In the second study, the archaeal transcription factors TATA-binding protein (TBP) and transcription factor B (TFB) from \textit{M. jannaschii} were studied in complex with a DNA carrying a promoter element.

II. MATERIALS AND METHODS

A. DNA and recombinant protein preparation, labeling, and assembly

Labeled and unlabeled DNA single-strands were purchased from IBA (Göttingen, Germany) and annealed as described previously (Andrecka et al., 2008). The ntDNA strand was labeled with the donor dye Cy3B at either position (+12), (+7), (−18), (−30), or (−37) or with 6-TAMRA at position (−24). For the dsDNA benchmark study, the tDNA strand was labeled with the acceptor dye Alexa647 (A647) at position (−9). For surface immobilization, the ntDNA strand had biotin attached at the 5′-end via a C6-amino linker. The transcription factors TBP and TFB from the hyperthermophilic archaeal model system \textit{M. jannaschii} were expressed in recombinant form in \textit{E. coli} (Werner and Weinzierl, 2005). TFB was mutated site specifically yielding a single cysteine at position 71 and labeled with an A647 fluorophore via a cysteine-maleimide coupling strategy as described recently (Nagy et al., 2015).

In the case of the dsDNA benchmark study, the double-labeled nucleic acid scaffold (2 pM) was loaded directly into the sample chamber of the total internal reflection fluorescence (TIRF) microscope. In the case of the DNA-protein study, the complexes were assembled freshly before each smFRET experiment by adding 1 µL each of nucleic acid scaffold (2 µM), TBP (19 µM), and TFB (11 µM) to 97 µL HNME buffer (40 mM HEPES [pH 7.3], 250 mM NaCl, 2.5 mM MgCl₂, 0.02 mM EDTA, 1% glycerol, and 2 mM DTT). The mixture was incubated at 60 °C for 10 min. The complexes were then diluted 10⁶-fold in HNME buffer and loaded into the sample chamber. All smFRET experiments were performed on a custom-built prism-based total internal reflection fluorescence microscope (TIRFM) (Grohmann et al., 2011 and Dörfler et al., 2017).

B. Determination of isotropic Förster radii and anisotropies

For each donor-acceptor pair, the isotropic Förster radius \( R_0^{iso} \) was determined using standard procedures (Vámosi et al., 1996). First, the quantum yield of the donor samples was determined using cresyl violet dissolved in ethanol as a standard (\( QY = 58\% \)) (Rurack and Spieles, 2011) (Table S1 of the supplementary material) with an experimental error of 5%. Second, overlap integrals (\( J \)) were calculated from recorded donor emission spectra (528–700 nm with an excitation wavelength of 523 nm) and acceptor absorption spectra (400–700 nm) using published extinction coefficients. Due to the uncertainty of the extinction coefficient, we estimate the uncertainty of the overlap integral to 10%. To account for the local environment of the fluorophores, an index of refraction of \( n = 1.35 \) was used together with a 5% error. The isotropic Förster radii \( R_0^{iso} \) were determined for all the different donor-acceptor pairs (Table S1 of the supplementary material). We used normal error propagation to determine the uncertainty of \( R_0^{iso} \). Due to the functional dependence of \( R_0^{iso} \) on \( QY, J, \) and \( n, R_0^{iso} \propto \frac{QY}{Jn^2} \), the relative error can be computed as

\[
\frac{\Delta R_0^{iso}}{R_0^{iso}} = \frac{1}{6} \left( \frac{\Delta QY}{QY} \right)^2 + \left( \frac{\Delta J}{J} \right)^2 + \left( \frac{4 \Delta n}{n} \right)^2 \approx 4\%. \tag{1}
\]
To account for this uncertainty in our analysis, we extended the recently published Fast-NPS (Eilert et al., 2017) by defining $R_{\text{iso}}^{\text{post}}$ as a normally distributed variable with a standard error of 4%.

Additionally, to account for uncertainties in the Förster distance due to orientation effects, we measured the steady-state fluorescence anisotropies of the donor and acceptor dyes for all attachment sites using a fluorescence spectrometer (Fluorolog II, SPEX) (Table S1 of the supplementary material). Both isotropic Förster distances and fluorescence anisotropies were used as prior information in the Fast-NPS analysis (Eilert et al., 2017).

### C. Fast-NPS calculations

The NPS is an analysis tool based on Bayesian parameter estimation. The aim of a Bayesian analysis is to determine the posterior distribution that says how probable a certain set of parameters is in light of the given data and a statistical model. For the case of the Fast-NPS, this results in the probability distribution of the configurations, i.e., the positions and orientations of all dyes. The posterior is obtained by the product of likelihood and prior. The likelihood is the probability distribution of the data, given the configuration. In our case, we know that the smFRET efficiencies follow a multivariate normal distribution centered at the experimental smFRET efficiencies with standard measurement errors. A transformation of the FRET efficiencies to the configuration space yields the likelihood. The prior represents the knowledge we have beforehand and the applied assumptions, i.e., the AVs of the dyes and the dye models.

The Fast-NPS computes the AV by taking several parameters into account: the structure of the macromolecule to which the dye is attached, the attachment point, the dye diameter, linker length and linker thickness. The dye and the linker are modeled on a lattice such that no clash with the macromolecule occurs. The lattice points satisfying this condition give rise to the AV.

In the Fast-NPS, an advanced and highly adaptive Markov chain Monte Carlo algorithm samples the posterior, such that we can approximate the credible volumes (which are 3D analogs of a credible interval) (Eilert et al., 2017). These are cumulative distribution functions obtained from the sampled positions sampled by decreasing probability. We typically display the credible volumes at 68% (using a 10 sigma interval) or 95% (2σ).

Since the prior includes assumptions, e.g., the dye models, it is possible that the experimental smFRET efficiencies cannot be mapped to sampled configurations, i.e., there is no conformation possible that satisfies all distance constraints determined from the smFRET data given the assumptions. Therefore at the end of the Fast-NPS calculations, a consistency check is performed. Since the likelihood is a multivariate normal distribution, we can determine confidence intervals. To this end, we compute smFRET efficiencies from the sampled configurations and test how many of these lie within the 95% confidence interval. For a consistent model, we expect to find 95% of the samples within this confidence interval. If this is true for less than 95% of the samples, we can deduce that some assumptions are not consistent with the data. In order to account for statistic fluctuations, we define that a network is inconsistent if less than 90% of the samples lie within the 95% confidence interval. In this case, we compute the marginalized consistencies. These can then reveal which measurement(s) is (are) leading to the inconsistency.

### D. Statistical test on different means

The mean position of the antenna calculated from the DNA structure is enclosed by the credible volume only at a confidence level of 95%. Thus, we wanted to test if this prior mean position significantly differs from the mean of our posterior. Since the posterior in position space is bimodal and elongated along an arch, the expected value of the position is not representative. Thus, we transform the prior mean position distances to prior mean smFRET efficiencies. This mean vector, $\bar{x}_{\text{prior}}$, is then tested against the mean smFRET efficiency vector, $\bar{x}_{\text{post}}$, of the posterior, i.e., the measured data. We perform Hotelling’s $t$-squared test (the multivariate generalization of the $t$-test) and test the null hypothesis $H_0 : \mu_{\text{prior}} = \mu_{\text{post}}$. The variance-covariance matrix, $\Sigma_{\text{post}}$, of the posterior mean vector consists of the measurement errors at the diagonal entries. Although the mean position distances are known exactly, the isotropic Förster radii, which are needed to recalculate $\bar{x}_{\text{prior}}$, have a measurement error. Thus, we conducted a normal error propagation of $E_{\text{FRET}}$. Given the functional dependence $E_{\text{FRET}} = \frac{1}{1 + \sigma_{\text{FRET}}^2}$, the relative error of $E_{\text{FRET}}$ is given by

$$\frac{\Delta E_{\text{FRET}}}{E_{\text{FRET}}} = \frac{6}{1 + \left(\frac{d}{R_{\text{iso}}^{\text{post}}} \right)^6} \frac{\Delta R_{\text{iso}}^{\text{post}}}{R_{\text{iso}}^{\text{post}}} \frac{1}{6}$$

These errors are then used as the diagonal entries in $\Sigma_{\text{prior}}$, i.e., the variance-covariance matrix of the prior mean vector. The test statistic is given by $T = (\bar{x}_{\text{prior}} - \bar{x}_{\text{post}})^T (\Sigma_{\text{prior}} + \Sigma_{\text{post}})^{-1} (\bar{x}_{\text{prior}} - \bar{x}_{\text{post}})$. The test statistic follows a $X^2$-distribution, and we obtain a $p$-value of 0.0729.

### E. Computation of precision and accuracy for the dsDNA benchmark study

To obtain the precision of the Fast-NPS analysis along the direction of the DNA helix, we projected the posterior position samples of the antenna onto the DNA axis. Using these projected samples, we calculated the standard deviation of the posterior samples as a measure of precision in 1D. To compute the 3D localization precision, we determine the cube root of the credible volume of the antenna at 68% credibility.

In order to access the accuracy in 1D for the dsDNA benchmark study, i.e., the accuracy of localization along the direction of the DNA, we computed the projection of the AV derived mean position of the antenna onto the axis of the DNA and set this position at the origin. To obtain the 1D accuracy, we compared this position to the mean of the projected posterior position samples of the antenna. Note that an accuracy can only be determined if a meanpos-iso model is assumed for the antenna; for all other dye models, the actual position of the antenna dye is unknown and thus no accuracy can be
determined. To determine the 3D accuracy for the dsDNA benchmark study, we computed the distance between the mean position of the antenna and the mean of the posterior position samples of the antenna.

F. Computation of precision and accuracy for the protein-DNA complex benchmark study

The determined posterior distribution is excessively extended along one dimension, because of localization geometry, but shorter in the other two directions. Therefore, we performed the following analysis in order to obtain a measure of the precision along these directions. First, we conducted a principal component analysis (PCA) of the posterior samples in order to get the orthonormal basis of the space in which the posterior density, which resembles a crescent arch, lies. If one described the arch as a segment of a ring, we divided this ring into 50 angular segments of uniform length along the direction of the ring, with the length measured from the ring center (Fig. S1 of the supplementary material). The samples of each segment were projected onto the bisectrix of each angular segment. A subsequent PCA gives us the eigenvalues (or variances) of the angular segment along the principal directions of intermediate and maximal precision and their square root is equal to the respective standard deviations. As a quantitative measure, we give the geometric mean over these standard deviations along the directions of intermediate and maximal precision. Segments that contain less than 10% samples compared to those of the segment with maximal number of samples were neglected. To give a measure for the 3D precision, we again computed the cube root of the values (or variances) of the angular segment along the principal component axes.

III. RESULTS

Fast-NPS analysis of macromolecular complexes can be divided into four parts: obtaining the required background information (prior), the actual smFRET measurements (likelihood), the computation of the 3D probability densities for the position of the antenna dye molecules (posterior), and the visualization of the analysis results (Fig. 1 and Sec. II).

In the following, we will present two benchmark studies for the Fast-NPS. We will use the Fast-NPS to determine the position of a dye molecule attached to a DNA molecule and of a dye molecule attached to a protein in a DNA-protein complex.

A. DNA benchmark study

1. Design, data analysis, and Fast-NPS calculation

The dsDNA molecules used in this benchmark study consist of a 66-nucleotide template (t) and non-template (nt) strand. The sequence [Fig. 2(a)] represents a strong viral SSV T6 promoter DNA which has been used in previous transcription studies (Werner and Weinzierl, 2002). The numbering originates from the definition of the transcription start site at +1. Due to the reason that the same DNA sequence was also used in the second benchmark study presented here, namely, the promoter specific binding of TBP and TFB, the same base numbering was chosen throughout both experiments for simplicity. To obtain a structural model for the DNA, modeling was performed using the 3D-DART server (van Dijk and Bonvin, 2009).

For the dsDNA benchmark study, the FRET acceptor dye molecule was attached to position (−9) on the tDNA strand, representing the unknown dye called “antenna,” which was localized with the Fast-NPS software (Eilert et al., 2017). The nt strand was labeled with a donor dye at either position (+12), (+7), (+18), (−24), or (−30), representing five reference dyes known from the structure called “satellites” [Fig. 2(a)]. The fluorescent dyes were attached to the DNA strands via flexible linkers. While in an NPS study the attachment point of the antenna molecule is known from the respective PDB structure (in this case the model for helical dsDNA), the precise location of the dye molecule is not. Therefore, the AV needs to be computed using the size of the dye molecule, the linker length, and the point of attachment (Muschielok et al., 2008) [Fig. 2(b) and Table S2 of the supplementary material and Sec. II].

For the determination of the position of the dye attached to position t(−9), the acceptor labeled tDNA strand is annealed with each one of the five donor labeled ntDNA strands, respectively. For each double-labeled dsDNA construct, three independent smFRET measurements were performed on a custom-built prism-based total internal reflection fluorescence microscope (Nagy et al., 2015). From the measured fluorescence intensities, the smFRET efficiencies were calculated using a molecule-wise gamma correction (Lewis et al., 2008). The computed FRET efficiencies from all FRET molecules were plotted in histograms [Figs. 3(a)–3(e)], and the peaks were fitted with a Gaussian function to extract the mean smFRET efficiencies. For each histogram, the number of molecules, the computed mean smFRET efficiency, and the width of the Gaussian function are summarized in Table II.

FIG. 1. Workflow of a structural analysis with smFRET and Fast-NPS.
FIG. 2. Benchmark study design: overview of the labeling positions. (a) Cartoon depicting the sequence of the promoter DNA (tDNA strand in blue, ntDNA strand in cyan), which is used throughout this study. Labeling sites on the DNA are marked with stars. (b) Benchmark study on dsDNA. The accessible volumes (AV) of the fluorescent dyes attached to the five reference sites on the dsDNA via flexible linkers (Muschielok et al., 2008) are displayed: nt(+12) (yellow), nt(+7) (gray), nt(−18) (light blue), nt(−24) (green), and nt(−30) (violet). To compute these position priors, a helical structure of the dsDNA was built with the DNA structure modeling server 3D-DART (van Dijk and Bonvin, 2009). (c) Benchmark study on a protein-DNA complex. The computed AVs of the satellite positions for nt(−18) (light blue), nt(−24) (green), nt(−37) (dark red), and TFB-G262 (orange) are shown together with the crystal structure of the TBP/TFB/DNA sub-complex from Pyrococcus woesei [PDB file: 1D3U, (Littlefield et al., 1999)] including DNA template strand (blue), non-template strand (cyan), and transcription factor TFB (green). The DNA was extended in both directions using the structural modeling tools of Chimera (Pettersen et al., 2004). For better visualization of the labeling sites, the structure of TBP is not displayed.

Moreover, for each dye pair, the steady-state fluorescence anisotropies and the isotropic Förster radii $R_{iso}^0$ were determined (Table S1 of the supplementary material and Sec. II). One should note that the determination of $R_{iso}^0$ requires knowledge about the index of refraction, the overlap integral, and the quantum yield of the donor. We used Gaussian error

FIG. 3. Experimental smFRET data used for the Fast-NPS benchmark localization on the dsDNA. [(a)–(e)] Frame-wise smFRET histograms used in the Fast-NPS localization of t(−9). Shown is the smFRET data from measurements between t(−9) and nt(−12) (a), nt(+7) (b), nt(−18) (c), nt(−24) (d), and nt(−30) (e), respectively. The histograms were fitted with a single Gaussian distribution indicated by the gray lines and the extracted mean smFRET efficiencies (displayed in Table II) were used in the Fast-NPS analysis.
TABLE II. Overview of the experimental data used for the positioning of \( t(9) \).

<table>
<thead>
<tr>
<th>Satellite</th>
<th>Antenna</th>
<th>( E_{\text{FRET}} ) (%)</th>
<th>Number of molecules</th>
<th>Width (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>nt(+12)_Cy3B</td>
<td>t(9)_A647</td>
<td>43</td>
<td>291</td>
<td>14</td>
</tr>
<tr>
<td>nt(+7)_Cy3B</td>
<td>t(9)_A647</td>
<td>66</td>
<td>567</td>
<td>12</td>
</tr>
<tr>
<td>nt(+18)_Cy3B</td>
<td>t(9)_A647</td>
<td>89</td>
<td>458</td>
<td>9</td>
</tr>
<tr>
<td>nt(+24)_TMR</td>
<td>t(9)_A647</td>
<td>74</td>
<td>295</td>
<td>12</td>
</tr>
<tr>
<td>nt(+30)_Cy3B</td>
<td>t(9)_A647</td>
<td>46</td>
<td>276</td>
<td>15</td>
</tr>
</tbody>
</table>

propagation resulting in an overall error of 4% (Sec. II). In the updated Fast-NPS algorithm used for this study, we now included these errors as prior information. The Fast-NPS then computed the three-dimensional probability distribution of the position of the fluorescent dye attached to the unknown position \( t(9) \). The resulting credible volume was displayed in comparison with the structure of the dsDNA and the calculated mean position within the AV (Fig. 4) or the accessible volume (Fig. S2 of the supplementary material). Unless otherwise stated, the size of all credible volumes presented in this study corresponds to 68% credibility, representing the smallest volume which encloses a probability of 68% given the experimental uncertainties and the respective choice of dye models.

2. The classic model makes conservative assumptions, yet yielding a precision of \( \approx 2 \) basepairs (bp) in the direction of maximal precision

A novel sophisticated feature of the Fast-NPS analysis software is the possibility to describe the time dependent behavior due to the chemical environment for every dye molecule individually by choosing the most suitable model from a set of five different dye models (Eilert et al., 2017). To evaluate and compare the analysis using different dye models, the precision, a measure of statistical variability reflecting the closeness of agreement among the generated samples, was analyzed. Due to the very anisotropic arrangement of dye

FIG. 4. Fast-NPS results for position \( t(9) \). (a) The credible volume of \( t(9) \) (yellow) resulting from the Fast-NPS calculation using the classic model for all dyes is drawn at 68% and is displayed together with the dsDNA structure and the calculated mean position (red sphere) within the AV. Localization along the DNA axis is rather defined (6 Å), while localization perpendicular to the DNA axis is rather imprecise (3D precision 36 Å) (Table III). [(b) and (c)] The credible volume of \( t(9) \) (brown) resulting from the Fast-NPS calculation using the meanpos-iso model for all dyes, drawn at 68% (b) and 95% (c) credibility. The precision has increased dramatically in 1D (1.7 Å) and in 3D (13 Å) (Table III). The distance between the NPS credible volume mean position and the true mean position is around 6 Å in 3D and 0.4 Å when projected onto the DNA axis (Table IV).
molecules, we computed the precision both in 1D as a projection of the samples onto the DNA axis and in 3D (Sec. II).

In the classic model, the dye occupies only one (unknown) position within the AV. At this position, the dye molecule can rotate albeit not freely. Rotation is restricted to an axially symmetric orientational distribution whose size can be determined by the respective fluorescence anisotropy measurements (Dale et al., 1979). The orientation of the average axis of the orientational distribution is not known, leading to comparatively large uncertainties in the conversion from measured smFRET efficiencies into distances. Due to these rather conservative assumptions, the classic model has the smallest precision of all tested models though the correct position is typically enclosed in the relatively large credible volume. Analysis of the DNA (t=9) position using the classic model resulted in a credible volume which encircles the DNA in a shape of a crescent arch with a comparatively large width [Fig. 4(a)]. This shape is expected due to the almost one-dimensional arrangement of the satellite dyes along the long axis of the dsDNA. Given the fact that each individual satellite has a relatively large AV caused by the flexible dye linker, the uncertainty about the satellite positions increases the size of the credible volume.

Therefore, the position of the dye along the direction of the DNA axis is well defined, while its localization perpendicular to the DNA axis is rather imprecise. The computed 1D precision along the axis of the DNA (Sec. II) is $\pm 6$ Å, or $\pm 2$ bps, thus yielding a high localization precision (Table III). By contrast, the almost linear arrangement of satellites in space leads to a decrease in precision in directions perpendicular to the DNA axis resulting in a 3D precision (Sec. II) of $\approx 36$ Å. One should note that this comparatively low precision is not a limit of the technique but rather reflects the fact that for optimal performance, satellite positions should be spaced around the antenna position such that rays connecting the antenna position with satellite positions are isotropic.

In the framework of a classic model, the accuracy cannot be determined since only the position of the attachment of the linker is known, but the position of the actual antenna dye can be anywhere within the AV. However, a previous study used a comparison of smFRET data and molecular dynamics simulations to show that the complicated movement of the dye molecule throughout the AV can be approximated by its mean position (Sindbert et al., 2011). If one compares the center of the posterior of the classic model to the mean position within the AV, one finds an accuracy of $\approx 1.4$ Å in 1D (Table IV). Note that computing a 3D accuracy is not meaningful if the surface of the credible is not convex as it is the case here.

### TABLE III. Overview of the precision of the different dye models for the DNA only study at a relative R² error of 0.04.

<table>
<thead>
<tr>
<th>Precision</th>
<th>classic</th>
<th>iso</th>
<th>vmpi</th>
<th>mpi</th>
<th>vmp</th>
</tr>
</thead>
<tbody>
<tr>
<td>1D (projection)</td>
<td>stdev (Å)</td>
<td>6.02</td>
<td>6.16</td>
<td>3.85</td>
<td>1.70</td>
</tr>
<tr>
<td></td>
<td>stdev (base)</td>
<td>1.77</td>
<td>1.81</td>
<td>1.13</td>
<td>0.50</td>
</tr>
<tr>
<td>3D</td>
<td>stdev (Å)</td>
<td>36.32</td>
<td>33.12</td>
<td>26.89</td>
<td>12.51</td>
</tr>
</tbody>
</table>

### TABLE IV. Overview of the accuracy for the DNA only study.

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>classic</th>
<th>iso</th>
<th>vmpi</th>
<th>mpi</th>
<th>vmp</th>
</tr>
</thead>
<tbody>
<tr>
<td>1D (projection)</td>
<td>Distance (Å)</td>
<td>1.38</td>
<td>0.26</td>
<td>0.52</td>
<td>0.43</td>
</tr>
<tr>
<td></td>
<td>Distance (base)</td>
<td>0.41</td>
<td>0.08</td>
<td>0.15</td>
<td>0.13</td>
</tr>
<tr>
<td>3D</td>
<td>Distance (Å)</td>
<td>5.78</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Distance (base)</td>
<td>1.70</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3. The precision can be increased by applying more restrictive models

For biological applications, it is oftentimes important to maximize the precision of the structural information. When comparing the classic model to more informative models, it is clear that the classic model is conservative both with respect to the position of the dye molecule and with respect to the calibration of the smFRET ruler, i.e., the knowledge about the precise Förster distance. It is therefore intuitive to use more informative models of these constraints and observe the effect on localization precision and accuracy.

Therefore, we first wanted to investigate the effect of the uncertainty in the Förster distance due to the uncertainty about the value of $\kappa^2$ on the localization accuracy. In contrast to the classic model, in the iso model, it is assumed that dye rotation is not restricted, i.e., the dye molecules show a sufficiently small steady-state fluorescence anisotropy, such that we can assume $\kappa^2 = 2/3$. Thus, in the iso model, the Förster distance reduces to the isotropic Förster distance; however, there remains the uncertainty of $R_0$ (Sec. II). The computed credible volume for the iso model is comparable in shape and size to that for the classic model [Fig. S3(A) of the supplementary material]; however, the volume is less symmetric indicating certain statistically forbidden regions. We obtain a precision of $\approx 6$ Å in 1D and $\approx 33$ Å in 3D (Table III), again comparable to the values obtained using the classic model. The accuracy of this model in 1D is $\approx 0.3$ Å (Table IV). However, this surprisingly high accuracy has to be put into perspective by the value of the precision which is significantly larger. One should note that again computation of the 3D accuracy is not meaningful due to the complex shape of the credible volume.

Next, we wanted to test the effect of the uncertainty in the dye positions on the resulting posterior. We therefore repeated the Fast-NPS analysis using the var-meanpos model, where the dye molecule samples a new conformation within the AV for each new excitation; however, the positions do not need to be visited uniformly. Again, the computed credible volume resembles that of the classic model [Fig. S3(B) of the supplementary material], but the precision in 1D has improved by 20% while staying comparable in 3D (less than 10% difference) and the accuracy has improved by 40% (Tables III and IV). To further improve precision, several restrictions need to be applied at the same time. We therefore next tested the var-meanpos-iso model combining the restrictions from the iso model and the var-meanpos model. The computed credible volume is now visibly smaller than for the first three models.
In the most restrictive model, termed *meanpos-*iso, we assume that all possible positions within the AV are uniformly visited by the dye molecule (Sindbert et al., 2011), resulting in a mean position at the center of the AV. Fast-NPS computation using the *meanpos-*iso model resulted in a credible volume whose size is reduced dramatically [Fig. 4(b)], and as a result, the precision is increased to 1.7 Å in 1D and 13 Å in 3D (Table III). In the *meanpos-*iso model, the accuracy, i.e., the comparison of the computed mean dye position of the samples to the mean dye position calculated for the model of DNA structure, is also strictly defined now since an isotropic averaging over all AV positions is assumed in this model. Since the DNA structure represents an object where one dimension (the length of the molecule) is much larger than the other (its diameter), it is again insightful to compute the accuracy both in 3D and in 1D, i.e., along the projection of the posterior onto the long axis of the DNA (Sec. II). The distance between the NPS credible volume mean position and the true mean position is ≈6 Å in 3D and 0.4 Å when projected onto the DNA axis (Table IV). Due to the geometrical constraint of the satellites being arranged linearly along the quasi one-dimensional dsDNA, the accuracy in 1D is much better than that in 3D.

Despite these fairly good values for precision and accuracy for the meanpos-*iso* model, the actual expected position is not enclosed in the credible volume when displayed at 68% credibility [Fig. 4(b)]. Only when increasing the credibility to 95%, one finds that the expected position is enclosed in the credible volume [Fig. 4(c)]. A multivariate t-test on equal mean smFRET efficiencies of the corresponding prior mean positions and the experimental data yields a low p-value of ≈7% (Sec. II). Thus, the assumptions made in the model are most likely not correct, i.e., either the dye is not sampling all possible positions within the AV with equal probability (*meanpos*), the dye is not completely free to rotate (*iso*), or the DNA structure used as prior information is not correct possibly due to dynamic movements of the DNA molecule. The observed steady-state anisotropies of all the dyes are around r = 0.2, which indicates that the $x^2 = 2/3$ assumption might not be valid. Empirical studies have shown that these anisotropies lead to an additional distance uncertainty of 5% (Hellenkamp et al., 2017). These three error sources show us again that unjustified and possibly wrong prior information may lead to a result that is in accordance with the data but does not reflect the correct solution in position space.

**B. DNA-protein benchmark study**

1. **Design, data analysis, and Fast-NPS calculation**

While structural studies on nucleic acids are important, there is a much wider interest in using these methodologies to investigate large protein or protein-nucleic acid complexes. Therefore, we also performed a benchmark study for a protein-nucleic acid complex. To this end, we chose a minimal complex of the transcription initiation machinery from the hyperthermophilic archaean *M. jannaschii*, consisting of a dsDNA, the TATA-binding protein (TBP), and transcription factor B (TFB) (Sec. II). The DNA sequence, which is identical to that of the DNA benchmark study, has been investigated previously and can be used for promoter-dependent *in vitro* transcription (Werner and Weinzierl, 2002). In order to obtain a structural model for the ternary complex of TFB, TBP, and DNA, we used the published model for the TFB/TFB/DNA complex from *Pyrococcuswoesei* [PDB file: 1D3U, (Littlefield et al., 1999)] and extended both dsDNA ends by 5 bp of B-DNA using the structure modeling tool in Chimera (Pettersen et al., 2004). Extending the DNA is necessary for the dye prior to calculation since these critically depend on the available free space.

For the DNA-protein benchmark study, we localized residue S71 of TBP. To this end, a single cysteine mutant of TBP was labeled with a fluorescent acceptor dye at position S71C (Sec. II). As reference sites, we chose three positions on the ntDNA strand, namely, nt(−18), nt(−24), and nt(−37), and one position on TFB, namely, TFB-G262, which were labeled with a donor dye [Fig. 2(c) and Sec. II]. Next, we calculated the accessible volumes of the reference sites [Fig. 2(c) and Table S3 of the supplementary material]. For this purpose, we removed the residues of TBP from the structural model in order to mimic a crystal structure where we have no knowledge about the structure, binding site, and orientation of the protein in question. Moreover, residues P208 to A300 of TFB were removed for computing the AV for nt(−37) since the position of the AV is coinciding with the binding of TFB and one cannot rule out that the dye (slightly) alters the local conformation of the transcription factor.

For the determination of the position of the dye molecule attached to TBP-S71, we assembled the acceptor labeled protein with one of the four donor labeled components, namely, nt(−18), nt(−24), nt(−37), or TFB-G262. In the case of a donor labeled nt strand, unlabeled TFB was used. In case TFB was donor labeled, unlabeled dsDNA was added to the labeled proteins (Sec. II). For each TFB/TFB/DNA construct, we performed at least three independent smFRET measurements; the data are summarized in smFRET efficiency histograms [Figs. 5(a)–5(d)]. The mean smFRET efficiencies required for Fast-NPS analysis were determined from the histograms by fitting Gaussians to the data (Table V). Moreover, steady-state fluorescence anisotropies where measured for all complexes and the isotropic Förster radii were determined for each dye pair (Table S1 of the supplementary material and Sec. II). For the calculation of the three-dimensional probability distribution of the position of the dye attached to TBP-S71, we used a modified structural model with the information of TBP removed as prior information in the Fast-NPS analysis, mimicking a realistic situation where no structural information is known about the protein of interest.

2. **The classic model makes conservative assumptions, yet yielding a precision of 1.5 bp in the direction of maximal precision**

In the classic model, the resulting credible volume of TBP-S71 is crescent-shaped due to the rather planar
FIG. 5. Experimental smFRET data used for the Fast-NPS benchmark localization on the protein-DNA complex. [(a)–(d)] Framewise smFRET histograms used in the Fast-NPS localization of TBP-S71. Shown are the smFRET data from measurements between TBP-S71 and nt(−18) (a), nt(−24) (b), nt(−37) (c), and TFB-G262 (d), respectively. The histograms were fitted with a single Gaussian distribution indicated by the gray lines and the extracted mean smFRET efficiencies (displayed in Table V) were used in the Fast-NPS analysis.

arrangement of the satellite dyes [Fig. 6(a)]. To estimate the 3D precision of the localization using the classic model, we again used the third root of the accessible volume (Sec. II) and obtained a value of ≈37 Å (Table VI). The crescent arch-like profile of the posterior reflects the geometric arrangements of the satellites, similar to that in the DNA benchmark studies. Therefore, the localization of the dye position is fairly precise in two directions and rather imprecise along the direction of the crescent arch. In order to determine the precision in the two better resolved directions, we performed a principal component analysis (Sec. II and Fig. S1 of the supplementary material) to obtain a precision of ≈4.9 Å and ≈8.9 Å (Table VI) in the two principal directions perpendicular to the tangent of the arch.

Next, we wanted to estimate whether the localization using the classic model also provides an accurate measure about the position of the antenna dye attached to TBP-S71. Therefore, we calculated the AV of the antenna dye using the coordinates from the structural model of the complete complex. The resulting AV of the antenna dye overlaps with the NPS credible volume [Fig. 6(a)]. It should be noted that a perfect overlap is not expected given the nature of the model based on prior and the NPS posterior. The AV does not reflect the positions which the dye really visits (in fact the classic model assumes a fixed dye position), but it encompasses the complete volume the dye molecule could be residing in, given its size, the length and width of the flexible linker, and the position of the anchor point defined by the structural model. The NPS credible volume, in contrast, represents the probability for the dye position, and its size is a measure of the uncertainty after the measurement, i.e., its precision. As long as there is overlap between prior and posterior, the model is in accordance with the data and the accuracy is at least as high as the precision.

3. The precision can be increased by applying more restrictive models

While the classic model is in agreement with the position determined from the crystal structure, the size of the posterior is comparatively large, reflecting the conservative assumption that went into the model. Therefore, we wanted to also test different, more restrictive models. When all dye molecules are described by the var-meanpos-iso model, the computed credible volume is reduced, yielding a 3D precision of ≈25 Å (Table VI). Again, the posterior is consistent with the prior of TBP-S71 [Fig. S4(A) of the supplementary material], and the results are also consistent with the data. By contrast, when the meanpos-iso model is used for all dye molecules [Fig. S4(B)], the analysis result is no longer consistent with the data, i.e., this model is not a valid description of the experiment. This could be caused by one or more dyes that do not (uniformly) visit the complete AV. Another reason for this discrepancy could be that the rotation of (at least) one dye is restricted, i.e., we have an orientation factor different than 2/3, and thus the assumed Förster distance is incorrect. As expected, the precision of the meanpos-iso analysis in 3D is improved dramatically, but since Fast-NPS can show that this model is not consistent with the data, this increase in precision is meaningless. In contrast to the dsDNA study, where all the dyes were attached to the DNA, the situation might be more complicated when dyes are attached

<table>
<thead>
<tr>
<th>Satellite</th>
<th>Antenna</th>
<th>FRET (%)</th>
<th>Number of molecules</th>
<th>Width (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>nt(−18)_Cy3B</td>
<td>TBP-S71_A647</td>
<td>79</td>
<td>97</td>
<td>20</td>
</tr>
<tr>
<td>nt(−24)_TMR</td>
<td>TBP-S71_A647</td>
<td>89</td>
<td>113</td>
<td>17</td>
</tr>
<tr>
<td>nt(−37)_Cy3B</td>
<td>TBP-S71_A647</td>
<td>84</td>
<td>82</td>
<td>25</td>
</tr>
<tr>
<td>TFB-G262_DL550</td>
<td>TBP-S71_A647</td>
<td>54</td>
<td>98</td>
<td>35</td>
</tr>
</tbody>
</table>
to proteins because of interactions with protein surface regions.

In the analysis of large smFRET networks, Fast-NPS allows for a good assessment of dye molecules which are inconsistent with the rest of the network (Eilert et al., 2017). Moreover, it allows one to exchange the model on a molecule by molecule basis. The consistency check of the network with all dyes in the meanpos-iso model identified the

measurement between TBP-S71 and nt(−24) to have less than 80% consistency (Fig. S5 of the supplementary material). A possible explanation is that nt(−24) interacts with the protein, and thus either the sampling of the AV is incomplete or the rotation of the dye molecule is no longer unrestricted. Therefore, we tested a hybrid model where only the dye molecule attached to nt(−24) as well as the antenna dye is described by the var-meanpos-iso model and all other dyes are described by the meanpos-iso model [Fig. 6(b)]. In contrast to all the meanpos-iso models, this hybrid model is consistent with the data. Again, the 3D precision has improved compared to the classic and to the var-meanpos-iso model to ≈12 Å, and the precision along the two directions perpendicular to that of the largest uncertainty has also been reduced to ≈1.0 Å and ≈3.5 Å. While the credible volume drawn at 68% did not yet have any overlap with the AV, overlap starts at a credible volume of 75% and increases when drawn at 95% [Fig. 6(c)]. Therefore, for this model, the accuracy is comparable to the precision. Since the computed credible volume is located close to the surface of TBP, this model predicts that the dye is interacting with the protein surface. In order to test whether all the

TABLE VI. Overview of the precision of the different dye models for the protein-DNA study at a relative Riso error of 0.04.

<table>
<thead>
<tr>
<th>Precision</th>
<th>classic</th>
<th>nt-24 vmpi/rest mpi</th>
<th>vmpi</th>
<th>mpi</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1D (projection)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>In the direction of intermediate precision (Å)</td>
<td>8.90</td>
<td>3.54</td>
<td>5.69</td>
<td>2.88</td>
</tr>
<tr>
<td>In the direction of maximal precision (Å)</td>
<td>4.86</td>
<td>1.04</td>
<td>3.54</td>
<td>1.70</td>
</tr>
<tr>
<td><strong>3D</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>stdev (Å)</td>
<td>36.62</td>
<td>12.54</td>
<td>25.40</td>
<td>13.18</td>
</tr>
</tbody>
</table>

FIG. 6. Fast-NPS results for position TBP-S71. The credible volumes of TBP-S71 resulting from the Fast-NPS calculation using the classic model (a) or the hybrid model (b) and (c) for all dyes are shown together with the accessible volume (AV) of TBP-S71 (orange, transparent), and the extended crystal structure of the TBP/TFB/DNA subcomplex from *Pyrococcus woesei* [PDB file: 1D3U, (Littlefield et al., 1999)] including DNA template strand (blue), non-template strand (cyan), TATA-binding protein TBP (gray, transparent), and transcription factor TFB (green). For the Fast-NPS calculation, the protein structure of TBP was deleted to mimic a realistic situation where no structural information is known of the protein. (a) The credible volume resulting from the Fast-NPS calculation using the classic model (yellow) drawn at 68% overlaps with the AV of TBP (b) and (c) The credible volume resulting from the Fast-NPS calculation using the hybrid model [purple, nt(−24) described by the var-meanpos-iso model and all other dyes by the meanpos-iso model], drawn at 68% (b) and 95% (c) credibility.
assumptions made in the hybrid model are correct, additional measurements would need to be performed, thus improving the precision. If the more precise model stays consistent, it is a good indication that the assumptions made are correct.

In a real case scenario, the credible volume would now serve as a starting point to build a structural model of the protein in question. This model can then be used as prior information in a new Fast-NPS analysis with the structure of the protein now being excluded together with the crystal structure. The resulting credible volume then reflects the position of the dye in the presence of the modeled protein. This not only enables an incremental refinement of the built model but also increases the reliability for the resulting structure. In our benchmark study, we repeated the Fast-NPS analysis with the PDB file now containing the structure of TBP, leading to reduced credible volumes in the region where TBP is located but an overall unchanged position (Fig. S6 of the supplementary material).

IV. DISCUSSION

The smFRET and Fast-NPS data presented here show the position precision and accuracy achievable by the Fast-NPS analysis software by localizing a supposedly unknown position on a dsDNA and one position on a protein in a DNA-protein complex and comparing the NPS derived information about the position to the position expected from structural knowledge. Additionally, we tested whether individual assignment of different dye models depending on the environmental and conformational behavior of the dye and the linker used for its attachment could improve precision and accuracy.

In the dsDNA study, we localized the position of a dye molecule attached to $t(-9)$ on a dsDNA with a precision well below a single base pair when looking along the direction of the DNA long axis. Also, the accuracy along this direction was below the dimensions of a single base pair, assuming that the meanpos-iso model describes the dye molecule behavior well. However, Fast-NPS analysis of this benchmark DNA also points to some of the problems when using smFRET data in a quantitative manner for structural biology tools. First, the precision is highly dependent on the geometry of the network. For the case of the DNA, the precision along the long axis of the DNA is an order of magnitude better than the precision in 3D. Thus, the achievable precision in any experiment will be highly dependent on the relative positions of dye molecules. Second, the choice of different models for the dye molecules can improve the precision, but very restrictive models might no longer be accurate. Also dye interactions with the DNA or protein surface are likely leading to deviations since such interactions could also lead to restriction in the dye rotation, and hence the $\chi^2 \approx 2/3$ assumption would no longer be valid. Thus one should be very careful when using the results of the analysis for quantitative discussions.

An important feature of the Fast-NPS system is a consistency check to test whether the posterior calculated for a particular model combination is consistent with the data. For the DNA benchmark study, all tested models were consistent with the data. However, it should be noted that in case a particular model is consistent with the data, it could still be that some of the assumptions are invalid. For instance, assuming that a meanpos-iso model in the DNA benchmark study is consistent with the data, however, the computed meanpos-iso position is far away from the center of the probability distribution and not inside of the 68% credible volume.

In the past years, data from smFRET measurements have been used in a variety of hybrid structural biology approaches for determining structural models of biological complexes (Brunger et al., 2011; Kalinin et al., 2012; Nagy et al., 2015; Dimura et al., 2016; and Hellenkamp et al., 2017). The advantage of the Fast-NPS compared to these other approaches is that dye models can be adopted for each individual labelling site, thus optimizing both precision and accuracy in structural biology using smFRET data. The benchmark study for the protein-DNA complex of TBP, TFB, and promoter DNA is a good example to show how useful it can be to use hybrid models where dyes attached to different positions are treated individually in order to reflect variations in the local environment of the dye molecule. We found that a dye molecule attached to position TBP-S71C can be localized using the classic or var-meanpos models with a slightly improved precision. By contrast, when localized with the meanpos-iso model, the precision is dramatically improved; however, the analysis is no longer consistent with the data since the assumptions made by the model are not fulfilled. However, when we used a hybrid model where the dye molecule attached to nt(-24) and the antenna dye are described by the var-meanpos-iso model and all other dyes are described by the meanpos-iso model, we were able to obtain a high precision of around 1.0 Å and 3.5 Å in the two directions perpendicular to the long extension of the posterior and a precision of 12 Å in 3D. Moreover, this hybrid model is consistent with the data and the position of the posterior overlaps with the meanpos-iso model, both accurate and precise, where the precision is mostly limited to the arrangement of dye molecules. For ideal arrangement of satellite dyes, a precision of 1-3 Å can be expected in all directions.

V. CONCLUSION

In conclusion, we could show that the Fast-NPS is a powerful analysis software to localize unknown dye positions in common dsDNA and DNA-protein complex scenarios and is well suited for a broad application in various structural FRET approaches. With only a handful of FRET pairs, it was possible to get a precision of around 1 bp in 1D and a 1D accuracy of better than 1 Å in the dsDNA study, when taking the mean of all possible positions within the AV as a ground truth. Moreover, we want to stress that while it is important to use informative models for analyzing the data, it is important to always compare the resulting credible volumes to those from the classic model. In case there is no overlap with the credible volumes of the classic model, the results must be handled with caution, even though the calculation was consistent, since the classic model makes very conservative assumptions. The larger the smFRET network and the better the satellites are distributed in the three dimensions, the more precise and accurate will the Fast-NPS results be.
SUPPLEMENTARY MATERIAL

See supplementary material for additional information (figures and tables).

ACKNOWLEDGMENTS

We would like to thank D. Grohmann and S. Schulz for the preparation of the labeled and unlabeled archaeal transcription factors. J.M. was supported by the DFG grant (No. MI 749/4-1) and by the European Union through the ERC starting grant remodeling (No. 243256). The authors thank the Center for Translational Imaging MoMAN for its support.


Title: Complete Kinetic Theory of FRET

Authors: Tobias Eilert, Eleni Kallis, Carlheinz Röcker and Jens Michaelis

Journal: Journal of Physical Chemistry B

Copyright: American Chemical Society

Reprinted with permission from J. Phys. Chem. B 2018, 122, 49, 11677-11694 Copyright 2018 American Chemical Society. The use of the article in a PhD thesis is kindly allowed by the American Chemical Society by the Journal Publishing Agreement Form A according to the “Ethical Guidelines to Publication of Chemical Research”. 
**Author’s Contribution**

The author of the thesis at hand developed the ideas leading to this publication predominantly by himself, but also sought help by the co-authors and in the literature. He structured and wrote the publication on his own except for corrections and comments by the co-authors. The collection, combination and expansion of the theoretical information about FRET were his work. He implemented and conducted all of the simulations.
Complete Kinetic Theory of FRET
Tobias Eilert,† Eleni Kallis,† Julia Nagy,‡ Carlheinz Röcker,† and Jens Michaelis∗†
1Institute for Biophysics, Ulm University, Albert-Einstein-Allee 11, Ulm 89081, Germany
2Center for Translational Imaging (MoMAn), Ulm University, Albert-Einstein-Allee 11, Ulm 89091, Germany
† Supporting Information
ABSTRACT: Förster resonance energy transfer (FRET) can be used to measure distances and infer structures at the molecular level. However, the flexible linkers with which the fluorophores are attached to a macromolecule introduce a lack of knowledge. Both the dye’s geometry and kinetics give rise to uncertainties. Whereas the impact of the geometry is already well understood, the real extent of the kinetics has not been investigated thoroughly. Here, we present a single-molecule (sm)FRET theory that defines the kinetics of dye movements in a complete form. We introduce a formal nomenclature and provide a recipe for the calculation of the corresponding FRET efficiency. We further analyze experimental data in order to obtain parameters characterizing the geometry and kinetics of the FRET dyes and use them to resimulate the FRET efficiencies by diffusion of fluorophore and linker movement. We show in a real case scenario of dye molecules attached to dsDNA that when making geometrical and kinetic assumptions commonly used in the FRET community one obtains results differing from the experimental data. In contrast, our stochastic simulations taking kinetic parameters from experiments into account reproduce the correct FRET efficiencies. Furthermore, we present a method enabling us to classify the kinetics of the dyes by investigating single realizations of the simulated transfer process. The results support our notion that the common kinetic assumptions are not appropriate over the whole range of distances inferred by FRET even for the simple situation of dyes attached to DNA where few interactions occur.

INTRODUCTION
Structural biology as a key field of life sciences aims to obtain quantitative 3D models of molecular machines at the atomic level. However, in order to understand nature, it is not only important to resolve the static structure of macromolecules but also to elucidate structural dynamics or even transiently forming complexes in solution. Here, Förster resonance energy transfer (FRET)† is a powerful tool. FRET has been termed a molecular ruler because the distance range that can be measured is on the length scale of, e.g., that of proteins.† Furthermore, single-molecule FRET (smFRET)‡ has become a widely used technique for studying the structure and dynamics of even transient macromolecular complexes.3–6 In order to accomplish this, smFRET measurements are utilized to infer distances within macromolecules and their complexes.7–11

However, a major challenge in the inference of distances/structures from smFRET measurements is that the fluorophores are attached to the macromolecule of interest via flexible linkers.12 The linker and the molecular environment introduce a lack of knowledge regarding the freedom of orientation and position but also regarding the time scale of rotation and translation of the dye molecule relative to the FRET process. These factors can have a tremendous effect on the smFRET efficiency distribution as well as on its mean;13 however, the extent of the kinetics has not been investigated on real data so far.

There are several approaches to account for the relative geometry and kinetics of the dyes.12–22 However, in the current state of theory, there is none which rigorously addresses the degree of geometrical freedom and relative time scale of rotation as well as translation. Until now, it has commonly been assumed that the donor and acceptor rotate freely much faster and translate much slower than the donor is de-excited. Because especially the FRET rate can change drastically from sample to sample, this might not be true. In case these assumptions are not valid, the inference of distances and structures will be biased or even incorrect.

In light of these difficulties, we present a theory that directly addresses the kinetics occurring during an experiment, conducted to determine a distance (or a structure, if a network of distances is analyzed) from measured average smFRET efficiencies.6 In particular, we will use all information obtained in the fluorescence experiments and use as few assumptions as possible.6 The presented theory allows prediction of the average smFRET efficiency (or even the distribution of smFRET efficiency) in an efficient way and even categorization of the motions of the FRET dyes. As a result, one can infer the correct distances from the experimental data. Interestingly, all of the information that we need was discussed some 40 years ago but has been ignored persistently.19,20–30

Special Issue: William A. Eaton Festschrift
Received: August 8, 2018
Revised: October 11, 2018
Published: October 15, 2018
DOI: 10.1021/acs.jpcb.8b07719
GEOMETRY AND KINETICS OF FRET

Förster resonance energy transfer is a quantum mechanical phenomenon between two molecules. During FRET, energy is transferred from one fluorophore, called the donor, to another, called the acceptor. The donor is excited and with a certain probability transfers energy to the acceptor through nonradiative dipole–dipole coupling. Although FRET is a quantum electrodynamical process, Förster derived a purely classical mechanical description. Here the position and orientation of the fluorophore’s transition dipole moment itself are represented by a point in space and a unit directional vector, respectively (see Figure 1). For fixed positions and orientations of the donor and acceptor, further called configuration (we use this term for one, two, or even several dyes), Förster came up with the following simple formula for the rate of transfer \( k_T \)

\[
k_T = k_{\text{DD}} \frac{3}{2} \chi^2 \left( \frac{n_{\text{iso}}}{d} \right)^6
\]

where \( \chi^2 \) is the so-called orientation factor depending on the relative orientation of the dipoles and \( d \) is the distance between the dipoles. The rate of depopulation of the excited state of the donor in the absence of the acceptor, \( k_{\text{DD}} \), is given by the sum of the rate of fluorescence, \( k_{\text{FD}} \), and the rate of nonradiative processes, \( k_{\text{NR,DD}} \), i.e., \( k_{\text{DD}} = k_{\text{FD}} + k_{\text{NR,DD}} \). The so-called isotropic Förster radius, \( R_{\text{iso}} \), is a FRET pair characteristic, i.e., the distance at which the transfer rate equals the rate of depopulation of the excited state of the donor in the absence of the acceptor, if in addition donor and acceptor orientations are averaged dynamically. For a fixed configuration, the FRET efficiency, \( E \), is then given by

\[
E = \frac{k_T}{k_{\text{DD}}} = \frac{k_T}{k_{\text{DD}}} + k_T
\]

In a real-case scenario, we will not have only one configuration. The linker and environment of an attached dye determines its allowed orientations and positions as well as how often the dye resides at a specific configuration. This is summarized in its equilibrium distribution (see Figure 1). Because we assume that the motions of the donor and acceptor are independent, their combined configuration follows the product of their equilibrium distributions, the so-called joint equilibrium distribution.

Although the effect of the relative geometry of the dyes on FRET is well understood, relative kinetic effects have not been described and investigated in a rigorous and quantitative fashion. It was recently shown by Wallace and Atzberger that relative kinetics, i.e., the diffusion of dye orientation and position relative to the FRET process (see Figure 1), have a great impact on the measured smFRET efficiency, leading to changes in \( E \) up to 20 and 50%, respectively.

To understand the effect of dye diffusion on FRET, it is essential to understand the kinetics of the transfer process as well. After absorption of a photon by the donor, we distinguish between three possible paths of de-excitation. The donor may be depopulated by nonradiative processes, it may emit a photon, or the energy may be transferred to the acceptor. The distribution into these paths is given by the relative contribution of the corresponding rates, \( k_{\text{NR,DD}} \), \( k_{\text{FD}} \), and \( k_{\text{FC}} \). However, it is important to understand that the time until the donor is depopulated by nonradiative processes, the time when a photon is emitted, and the time when the energy is transferred to the acceptor follow the same exponential distribution with rate \( k_{\text{DD}} = r_{\text{DA}}^1 = k_{\text{NR,DD}} + k_{\text{FD}} + k_{\text{FC}} \). Here, \( r_{\text{DA}} \) is the depopulation rate of the donor in the presence of the acceptor (Figure 2). We want to emphasize that the average time until transfer or, for short, the average transfer time, \( \tau_T \), or a single realization, \( \tau_{\text{DA}} \), is defined by the time from the moment of donor excitation until the moment the donor is de-excited by transfer. Therefore, the reciprocal average transfer time, \( \tau_T^{-1} = k_{\text{DD}} \), is not equal to the transfer rate, \( k_T \).

The next section gathers and combines the current understanding regarding the geometric and kinetic effects on FRET and presents a single condensed theory.

PHYSICAL DESCRIPTION OF TRANSFER REGIMES

In the current kinetic FRET theory, three rates are important to classify the kinetic behavior of a dye relative to the FRET process into translational and orientational FRET regimes, also noted as transfer regimes, averaging regimes, or just regimes: the rate of translation, \( k_{\text{T,Transl}} \), the rate of rotation, \( k_{\text{T,Rot}} \), and the average rate of donor depopulation in the presence of the acceptor, \( k_{\text{DA}} \) (see the Geometry and Kinetics of FRET section for an explanation). With the help of these rates, each motion of the donor and acceptor is independently classified as dynamic (\( k_{\text{T,Transl}} k_{\text{T,Rot}} \gg k_{\text{DD}} \)), intermediate (\( k_{\text{T,Transl}} k_{\text{T,Rot}} \approx k_{\text{DD}} \)), or static (\( k_{\text{T,Transl}} k_{\text{T,Rot}} \ll k_{\text{DD}} \)).
relative to the FRET process. Therefore, the donor and acceptor
are either dynamically averaged, correlatedly (intermediately)
averaged, or statically averaged. Averaging occurs over positions
and over orientations (see Figure 3 for a schematic illustrating
the translational transfer regimes).

More generally, the FRET efficiency depends on the effective
time scale of the dyes’ motions relative to the apparent times
the donor is de-excited. Hereby, the configuration space, which
the dyes span in given molecular environments, defines the FRET
rates that are possible to encounter. How often a configuration
and thus transfer rate is probed depends on the joint equilibrium
distribution, i.e., the distribution over positions and orientations
of both FRET dyes. However, how often which configurations
or transfer rates lead to transfer is determined by the diffusional
velocity of the motions relative to the effective width of the
equilibrium distributions and relative to the apparent times
that the donor is depopulated. In order to account for these effects,
we define the transfer regimes at first in a physical sense and
afterward write down an appropriate mathematical description.
This distinct nomenclature then yields also a recipe for the
computation of the corresponding average of the smFRET
efficiency or its distribution.

In the following sections, we will define the conditions of the
different regimes. We treat rotation and translation simulta-
neously because the general conditions for the corresponding
regimes are the same. Because the FRET rate depends on
the configuration of the donor and also the acceptor, both of
their regimes have an effect on the smFRET efficiency. Thus,
in the following, the term dye may stand for donor and/or
acceptor.

**Static Transfer Regime.** In the static regime, the dye
rotates/translations so fast relative to the FRET process and to
the effective width of its orientation/position equilibrium distri-
bution that the orientation/position at the time the donor has
been excited does not change until donor de-excitation.

It donates/accepts from this configuration. Thus, the dye’s
influence on the FRET process is FRET-efficiency-weighted.

**Dynamic Transfer Regime.** In the dynamic regime, the dye
rotates/translations so fast relative to the FRET process and to
the effective width of its orientation/position equilibrium distri-
bution that its orientation/position at the time the donor has been
excited is irrelevant. It donates/accepts from an orientation/position
with a probability proportional to the transfer rate at this
configuration (weighted by its joint equilibrium distribution). Thus, the dye’s influence on the FRET process is trans-
fer-rate-weighted.

Translating this notion to a stochastic view means that the
orientation/position at the moment of donor de-population is
independent of the orientation/position at the moment of donor
excitation. Thus, in a simulation, we can draw independently
from the equilibrium distribution of orientation/position. In
other words, this means that a dye started at one orientation/position
(represented by a Dirac delta) had already enough time
to converge to its equilibrium distribution of orientation/position
(for a justification, see the Appendix section Conditions
of the Dynamic Regime). Roughly, we say that the dye had
sufficient time to be able to sample the entire position/orientation
space weighted by its equilibrium distribution.

**Intermediate Transfer Regime.** In the intermediate regime,
the orientation/position in the moment of donor de-excitation
depends on the orientation/position at the time the donor has
been excited. That means that the start and end orientations/
positions are correlated but not identical. Thus, stochastic dif-
ferential equations need to be solved by numeric simulations.

### MATHEMATICAL DESCRIPTION OF TRANSFER
REGIMES

In the preceding section, we defined the transfer regimes. From
the definitions, it is clear that the dynamic and static regimes
do not depend on time: the motions are completely uncorrelated
with the FRET process. This characteristic leads to a similar
mathematical description; therefore, we combine them in one
category and call them limiting regimes. If at least one motion of
a dye in the FRET pair is correlated with donor depopulation,
i.e., falls in the intermediate regime, the solution will be time-
derpendent. In this section, we present a notation that constitutes
a bijective mapping to every regime combination that there is,
which in turn gives a protocol for its computation. For the sake
of readability and clarity of this notation, we first rewrite the
transfer rate in terms of factors separating the influence of the dyes’
positions and orientations on the FRET efficiency in the best
possible way.

In order to separate the dependence of the transfer rate on the
orientations and positions of the dyes, we rewrite the Förster
Formula in terms of the orientation factor of transfer rate,
\[ \Gamma^2 = 2 \kappa^2 \in [0, 6] \] and the distance factor of transfer rate,
\[ D^6 = \left( \frac{R_{0}^{w}}{d} \right)^6 \in (0, \infty) \]:
\[ k_T = k_{0} \cdot \frac{3}{2} \left( \frac{R_0^{w}}{d} \right)^6 = k_{0} \cdot \Gamma^2 D^6 \] (3)
Both factors are dimensionless: The FRET efficiency can then be
expressed by
\[ E = \frac{\Gamma^2 D^6}{1 + \Gamma^2 D^6} \] (4)
This equation simplifies the mathematical description of the transfer regimes presented in the following section.

Notation for the Combination of Transfer Regimes. It is clear from the Physical Description of Transfer Regimes section that the average smFRET efficiency, regarding a combination of dynamic and static regimes, is just a (complicated) expected value. Thus, we can define a notation that gives us a recipe of how to compute the corresponding average smFRET efficiency.

The configuration of a dye follows an equilibrium distribution, which means that the possible positions and orientations are weighted. Thus, the average smFRET efficiency is a weighted expected value. In general, an expected value of a function \( f \) weighted by a function \( g \) is given by the integral of the inner product of \( f \) and \( g \) denoted by \( \langle f \rangle \Omega \equiv \int \Omega g (\omega) f(\omega) \, d\omega \), where \( \Omega \) is the space of integration and \( \omega \in \Omega \).

In the FRET context, e.g., a dynamically averaged donor and a statically averaged acceptor lead to an average smFRET efficiency given by the following weighted expected value

\[
\langle E \rangle = \frac{\langle g_D \Gamma^D \rangle_{\Omega_D} \langle g_A \Gamma^A \rangle_{\Omega_A}}{1 + \langle g_D \Gamma^D \rangle_{\Omega_D} \langle g_A \Gamma^A \rangle_{\Omega_A}} \Omega_{\Delta}.
\]

Here, \( g_{D/A} \) denote the equilibrium distributions over the configuration space of the donor or acceptor, \( \Omega_{D/A} \). As described in the Physical Description of Transfer Regimes section, the dynamic motion of the donor leads to transfer rate weighting, i.e., \( g_D \) acts on the factors of the transfer rate, and the static behavior of the acceptor means FRET efficiency weighting, i.e., \( g_A \) acts on the FRET efficiency. In a real-case scenario, the equation of the expected value gets deeply nested because not only may the two FRET dyes behave differently but also their rotational or translational motions may fall into different averaging regimes.

There exists an analytic solution for the dynamic regime regarding rotation, but it neglects many details of the FRET process (see the Summary and Discussion). However, if we consider averaging over positions, because of the nonanalytical shape of the accessible volumes (AVs), we are forced to use numerical approaches. The nature of the FRET geometry, which is obviously strongly determined by distributions, suggests integration by Monte Carlo methods.

If at least one motion of a dye in the FRET pair is correlated with FRET, the solution will be time-dependent. This leads to an expected value depending on the time evolution of the processes (see the Simulation of smFRET Efficiency section). However, our notation remains valid, and we will indicate this dependence by a \( t \) for time in the superscript of the concerned weight function, e.g., \( g_{D,t} \). As a convention in our notation, we position the weight function of the time-dependent dye according to the dynamic regime, i.e., acting on the transfer rate. Consequently, this notation represents a bijective mapping from any regime combination to a notation of the expected value. In conclusion, there are \( 3^4 = 81 \) distinct regime combinations (3 regimes, 2 dyes, and 2 modes of motion, translation and rotation) for a FRET pair.

Strictly speaking, the limits, i.e., dynamic and static, will never be completely fulfilled. The stochastic simulation (see Simulation of smFRET Efficiency) of the combined processes of rotation and translation of both dyes, fluorescence and transfer, yields correct results even for the approximate limits but might be highly inefficient. This is, because, e.g., in the dynamic transfer regime regarding rotation, we need to simulate the rotation for a long time until the excited state of the donor is depopulated. Thus, evaluation of the corresponding expected value is much more efficient.

The next section summarizes the molecular system that we use to test our kinetic theory.

## TEST SYSTEM

As a test system, we use dyes attached to dsDNA. We use some of the measurements and experimental data of the smFRET network described in Nagy et al.\(^3\) The two strands of the DNA are termed template (t) and nontemplate (nt) strands. The dyes are named after the position in the nucleotide sequence, e.g., the dye attached to the nontemplate strand at the thymine at position +7 is abbreviated as nt-7 (see Figure 4). The donor dye (Cy3B) was attached to either +7, −18, or −30 on the nontemplate strand. The acceptor dye (Alexa647) was attached to the template strand at position −9 (t−9), acting as the acceptor in all FRET measurements.\(^4\) The analyzed parameters of data from the fluorescence lifetime and time-resolved anisotropy experiments of these dyes (see the Supporting Information) are found in Table 1. We omitted the lifetime of the acceptor dye because it is not influencing the FRET efficiency and is, thus, not an essential parameter for the simulations. The AVs of the dyes for the software tool FastNPS.\(^5\) The dye and linker parameters for the AV computation can be found in the Supporting Information. The isotropic Förster distance is a parameter that depends on the chosen donor and acceptor dye pair. It is listed with the donor because for the described measurements only the attachment point of the donor changes.

The next sections, we review how we can analyze typical fluorescence experiments in order to obtain information about the geometry and kinetics of the donor and acceptor, so that we can predict the correct smFRET efficiency by simulating the intermediate regime.

Table 1. Parameters Gained from Fluorescence Lifetime and Time-Resolved Anisotropy Measurements

<table>
<thead>
<tr>
<th>dye</th>
<th>( r_{\pi}/(\text{rad/ns}) )</th>
<th>( \tau_{\text{AVS/NS}} )</th>
<th>( r_{\omega} )</th>
<th>( R_0' )/nm</th>
</tr>
</thead>
<tbody>
<tr>
<td>t−9</td>
<td>2.65</td>
<td>1.13</td>
<td>0.9</td>
<td>0.057</td>
</tr>
<tr>
<td>nt+7</td>
<td>2.63</td>
<td>1.13</td>
<td>0.122</td>
<td>7.19</td>
</tr>
<tr>
<td>nt−18</td>
<td>2.82</td>
<td>1.29</td>
<td>0.198</td>
<td>7.26</td>
</tr>
<tr>
<td>nt−30</td>
<td>2.58</td>
<td>1.10</td>
<td>0.083</td>
<td>7.17</td>
</tr>
</tbody>
</table>
FLUORESCENCE AND DEPOLARIZATION

Analysis of Fluorescence Lifetime Experiments. Fluorescence lifetime experiments enable us to determine the rate of depopulation of the excited state of a fluorophore, \( k = r^{-1} \), in a specific environment. It is the sum of the rate of fluorescence, \( k_0 \), and the rates of nonradiative processes, \( k_{NR} \), i.e., \( k = k_0 + k_{NR} \). In case of a single-exponential decay of the fluorescence, the fit formula is

\[
I(t) = I_0 \exp(-kt)
\]

where \( I(t) \) is the fluorescence intensity at time \( t \) and \( I_0 = I(0) \).

Rotational Equilibrium Distribution. In order to analyze time-resolved anisotropy measurements for the correct simulation of the rotational motion, we need to assume that the dye follows a certain rotational probability distribution function at equilibrium. Here, we propose that the transition dipole moment of the dye, denoted by the unit vector \( \alpha \), diffuses uniformly in a spherical cone of semiangle \( \Psi \) directed along the unit mean axis \( \mathbf{a} \) (see Figure 5 for a schematic). We denote the cap surface of the directed spherical cone by \( C_\Psi \equiv \{ \alpha : \Psi \leq \Psi \} \), where \( \Psi \) is the angle between \( \alpha \) and \( \mathbf{a} \). Therefore, the equilibrium distribution is

\[
P_{C_\Psi}(\alpha) \propto \begin{cases} 1 & \text{if } \alpha \in C_\Psi \\ 0 & \text{else} \end{cases}
\]

Analysis of Time-Resolved Anisotropy Experiments. Time-resolved anisotropy measurements yielding \( r(t) \), i.e., the anisotropy decay over time \( t \), serve to provide information about the diffusional velocity (rotational correlation time, \( \tau_{rot} \)) and the confinement of rotation (limiting anisotropy, \( r_{\infty} \)) of a dye. The approximate fit formula for the anisotropy decay over time under a strict cone model is given by

\[
r(t) = (r_0 - r_{\infty}) \exp(-t/\tau_{rot}) + r_{\infty}
\]

if we assume that the macromolecule itself rotates much slower. Because we are not interested in the fundamental rotational \( r_0 \equiv \lim_{t \to 0} r(t) \), we like to think in depolarization, \( A(t) \), that is

\[
A(t) = \frac{r(t)}{r_0} = (1 - A_{\infty}) \exp(-t/\tau_{rot}) + A_{\infty}
\]

where \( A_{\infty} = \lim_{t \to \infty} A(t) \) is the limiting depolarization. Assuming the strict cone model, we can relate \( A_{\infty} \) to \( \Psi \), i.e., the semiangle of the cone, by

\[
A_{\infty}(\Psi) = (0.5 \cos(\Psi)(1 + \cos(\Psi)))^2
\]

The limiting depolarization due to a semiangle \( \Psi \) in the cone model is plotted in Figure 5.

Besides the limiting depolarization, the rotational diffusion coefficient, \( D_{rot} \), is the second and last parameter needed for simulation of the dye rotation in a strict cone model. It is reciprocal to the rotational correlation time

\[
D_{rot} = \frac{\langle \sigma \rangle}{\tau_{rot}}
\]

where \( \langle \sigma \rangle \) is a factor depending solely on the restriction by the cone with unit rad\(^2\), i.e.,

\[
\langle \sigma \rangle = \frac{D(\Psi(A_{\infty}))}{1 - A_{\infty}}
\]

The factor \( D(\Psi(A_{\infty})) \) was derived by Lipari and Szabo given the strict cone model. Alternatively, the rotational diffusion coefficient can be obtained by

\[
D_{rot} = \frac{D(\Psi(A_{\infty}))}{E(A_{\infty}) \tau_{rot}}
\]

where \( E(A_{\infty}) \) is the area between the anisotropy decay and the constant line of the limiting anisotropy, (the dashed area in Figure 6B) divided by the fundamental anisotropy.

The next section presents the stochastic procedure for how we simulate the dye rotation in a spherical cone.

Simulation of Fluorescence Lifetime and Time-Resolved Anisotropy Experiments. A major step toward predicting smFRET efficiencies is to simulate the rotational motion of a dye based on parameters obtained from time-resolved anisotropy experiments. The rotational diffusion can be described by the following Itô drift–diffusion process

\[
dr = \sigma^2 \frac{1}{2 \tan(\Psi)} dt + \sigma dw_t
\]

where \( \sigma \) is the drift coefficient and \( w_t \) is the Wiener process.

The integration time is set to \( \Delta t = \min \left\{ \frac{\Psi}{2D_{rot}}, \frac{\tau}{\sigma^2} \right\} / 100 \), such that we account for the fastest process in the simulation. The fluorescence of the dye constitutes a homogeneous Poisson point process and is independent of rotation. Thus, we simulate the rotation with time steps \( \Delta t \) and check after every step if the dye has emitted with probability \( k \Delta t \), where the rate of the Poisson process is the depopulation rate of the excited state of the dye, i.e., \( k = r^{-1} \). Because the rate of excitation in a

Figure 5. Limiting depolarization due to a semiangle \( \Psi \) in a strict cone model. The inlet shows a schematic of a cone highlighting its mean axis \( \mathbf{a} \) (see Figure 5 for a schematic).
fluorescence microscope is low \( (k_{\text{Exc}} \approx \times 10^5−10^7 \text{ s}^{-1}) \) compared to the molecular processes, we can draw for every excitation independently from the cone and simulate the trajectory until the excited state of the dye is depopulated.

In order to test whether our simulation can accurately recover input values, we analyzed the fluorescence and anisotropy decay from confocal data, as described in the Supporting Information. The characteristic fit parameters for the example of dye nt+7 were fed into the simulation, and we then analyzed the synthetic data as discussed in the previous section. The synthetic data are shown in Figure 6, and the corresponding fit results are summarized in Table 2.

Table 2. Fit Parameters of Experimental and Simulated Data

<table>
<thead>
<tr>
<th>parameter</th>
<th>experiment</th>
<th>simulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \tau_\text{ns} )</td>
<td>2.65</td>
<td>2.64 ± 0.00</td>
</tr>
<tr>
<td>( \tau_\text{ex}/\text{ns} )</td>
<td>1.13</td>
<td>1.08 ± 0.06</td>
</tr>
<tr>
<td>( A_\text{oo} )</td>
<td>0.31</td>
<td>0.33 ± 0.01</td>
</tr>
<tr>
<td>( \Psi/\text{deg} )</td>
<td>48.41</td>
<td>46.84 ± 0.37</td>
</tr>
<tr>
<td>( D_{\text{rot}}/(\text{rad}^2/\text{ns}) )</td>
<td>0.14</td>
<td>0.14 ± 0.01</td>
</tr>
</tbody>
</table>

The simulation error is estimated from the analysis of 10 simulations each with \( 10^6 \) photons.

Because we are now able to simulate the rotation of a dye consistently, we can turn to translational diffusion and finally to simulation of the complete evolution over time of a FRET pair, such that we can compute the time-dependent expected value (see the Mathematical Description of Transfer Regimes section).

## SIMULATION OF SMFRET EFFICIENCY

### Simulation of a Trajectory

In the intermediate regime, the orientation/position at the moment of donor de-excitation depends on the orientation/position at the time the donor was excited. Thus, progression of position and orientation over time is necessary in order to predict the correct average smFRET efficiency (see the Appendix section Statistics of FRET).

Simulation of the rotational diffusion is described in the Simulation of Fluorescence Lifetime and Time-Resolved Anisotropy Experiments section. We now address the simulation of the translational diffusion, which is again described by a Itô drift–diffusion process given by

\[
\text{d}X = \sigma \text{d}W
\]

where \( W \) is a random vector of independent standard Wiener processes. This introduces the diffusion, such that \( \text{d}X \) is a three-dimensional drift–diffusion process. The variance, \( \sigma^2 \), of the translational diffusion process is given by \( 6D_{\text{Trans}} \text{dt} \) for a time duration \( \text{dt} \), where \( D_{\text{Trans}} \) is the translational diffusion coefficient.

There is little information about \( D_{\text{Trans}} \) of dyes relative to a macromolecule in the literature. Peulen et al.\(^ \text{24} \) determined \( D_{\text{Trans}} \approx 10 \text{ Å}^2/\text{ns} \) for Alexa488 in an MD and Brownian Dynamics simulation approach. For simplicity we assume this diffusion coefficient for all dyes attached to DNA.

We simulate rotational and translational diffusion simultaneously but independently (see the Appendix section Assumptions and Deductions). The integration time of the simulation is set to

\[
\Delta t = \min \left\{ \frac{\Psi_\text{D}^2}{2D_{\text{Rot}}}, \frac{\Psi_\text{A}^2}{2D_{\text{Rot}}}, \frac{\rho_{\text{D}}^2}{6D_{\text{Trans}}}, \frac{\rho_{\text{A}}^2}{6D_{\text{Trans}}}, \frac{(k_{\text{exc}} + k_{\text{em}})^{-1}}{100} \right\}
\]

in order to approximately account for the fastest process, where \( \langle k_t \rangle = k_{\text{exc}}/(E_{\text{exp}} + k_{\text{em}} - 1) \) is the average transfer rate obtained from the experimental average smFRET efficiency, \( E_{\text{exp}} \) and \( R_{D/A} \) are the radii of spheres with the same volume as the donor and acceptor AVs. Because the FRET rate is time-dependent, the whole simulation is a time-inhomogeneous Poisson process (see Appendix section Statistics of FRET).

After every simulated time step \( \Delta t \), we decide at first what de-excitation occurs with probability \( p_1(t) = \Delta t(k_{\text{exc}} + k_{\text{em}}) \) at time \( t \) and then what, transfer with probability \( p_2(t) = k_{\text{exc}}(k_{\text{exc}} + k_{\text{em}})^{-1} \) or emission and nonradiative depopulation with probability \( 1 - p_1(t) \). Figure 7 shows the transfer rate and its factors over time for a simulated trajectory of nt+7 and \( t=9 \).

This procedure we can now apply to simulate data obtained on immobilized complexes measured in, e.g., a TIRF microscope (or also for molecules freely diffusing through the confocal volume of a confocal microscope).

### Simulation of TIRF Experiments

Because the rate of excitation in a TIRF microscope is low \( (k_{\text{Exc}} \approx 10^7−10^9 \text{ ns}^{-1}) \) compared to the molecular processes, we can draw independently

![Figure 6](image-url)
from the cones and AVs for each excitation event and simulate the trajectories of the donor and acceptor until transfer, emission, or relaxation by nonradiative processes occurs. Because the expected photon detection rate is $\approx 10^3 - 10^4$ photons/s, we use a rate of $5 \times 10^3$ photons/s. We assume in a good approximation that the absorption of photons by the donor is Poissonian. Hence, we can simulate the arrival times of the photons and predict signals that would be detected in a TIRF movie.

As a real-case scenario, we simulated the TIRF measurement between $nt-30$ and $t-9$ for 60 s. Figure 8A shows the intensity traces of the donor ($nt-30$) and acceptor ($t-9$) shown in green and red, respectively. The corresponding smFRET trace is shown in blue. The integration time is 30 ms. Figure 8B shows the resulting FRET histogram with a binning of 0.025.

The Journal of Physical Chemistry B

Figure 7. Transfer rate $k_T$, orientation factor of transfer rate $\Gamma^2$, and distance factor of transfer rate $D^6$ for a single exemplary simulated trajectory of $nt+7$ and $t-9$. Because dipole positions and orientations of the donor and acceptor are stochastic, also $D^6$ and $\Gamma^2$ are stochastic. This means that transfer is a time-inhomogeneous Poisson process with a stochastic rate.

Figure 8. Analysis of a simulated TIRF measurement between $nt-30$ and $t-9$. (A) Intensity traces of the donor ($nt-30$) and acceptor ($t-9$) shown in green and red, respectively. The corresponding smFRET trace is shown in blue. The integration time is 30 ms. (B) SmFRET histogram obtained by a binning of 2.5%.

Comparison of the Experiment with Simulations and Limiting Models. The main aim of this publication is to compare the various models applied in order to infer distances/structures from FRET with the experimental data. We contrast the experimental smFRET efficiencies with six models consisting of the limiting regimes and the time-dependent simulations. In the limiting models, we assume, as is commonly done, that the donor and acceptor follow the same regimes. The four limiting models are dynRotStatTrans, i.e., dynamic, uniformly unrestricted rotation and static translation, statRotTrans, i.e., static, uniformly unrestricted rotation and static translation, dynRotTrans, i.e., dynamic, uniformly unrestricted rotation and dynamic translation, and statRotDynTrans, i.e., static, uniformly unrestricted rotation and dynamic translation. In the abbreviations, we omit a note that the dyes...
rotate uniformly over the whole (hemi)sphere (unrestricted) because here it is assumed for all limiting regimes.

A special difficulty of the time-dependent simulation is that, although we know the extent of the spherical cone due to time-resolved anisotropy measurements, we do not know the orientation of the cone mean axis. Thus, we have two cases for the time-dependent simulation: the most simple model is called randAxSim, i.e., the mean axes of the donor and acceptor are randomly distributed for every trial (see below), and a more specific model called geomAxSim, i.e., using geometric assumptions to infer mean cone axes of the donor and acceptor. For the randAxSim model, we do not expect to get the correct smFRET efficiency because varying mean axes violate the anisotropy data put in. This simulation is just a means to assess the systematic error due to the unknown cone axes. For the geomAxSim model, we infer an average orientation axis from the local environment of the dye at the macromolecule. Because it was shown that the hydrophobic dyes do not interact with DNA, the only reason for restriction in orientation is steric hindrance. Thus, we propose a model in which the normalized eigenvector with the greatest eigenvalue of a principal component analysis (PCA) of the AV is a good approximation for the average axis of the spherical cone. Here, we assume a correlation between position and orientation, which might not be given.

We have three smFRET measurements for our comparison (see the Test System section). In Table 3, we summarized the experimental, \( \langle E \rangle^{\text{exp}} \), and average smFRET efficiencies predicted by the simulations, \( \langle E \rangle^{\text{pred}} \). The predicted smFRET efficiencies are an average out of 1000 trials, each with 1000 random starting configurations. Since Nagy et al.\textsuperscript{12} derived a relative normal error of 4\% for the isotropic Förster radius, every trial is performed with a varying radius drawn from \( N(R_{\text{iso}}^{0.04}, 0.04R_{\text{iso}}^{0.04}) \).\textsuperscript{12} Thus, the sample standard deviations of the simulation results, \( \sigma_{\text{exp}}^{\text{pred}} \), do not only estimate the variation caused by Poisson statistics but also the error of the isotropic Förster radius. The sample standard deviations of the randAxSim model additionally contain the uncertainty due to unknown mean cone axes. The binomial nature of the FRET process (see the Appendix section Statistics of FRET) requires that the statistical error decreases to higher and smaller FRET efficiencies. This explains the outstanding small standard deviations of the measurement between \( n_{t−18} \) and \( t−9 \).

In order to summarize the level of agreement between the experimental and simulated average smFRET efficiencies, we compute two types of effect size. The mean of the absolute differences in predicted and experimental average smFRET efficiencies, \( |\Delta| = (1/3) \sum_{i=1}^{3} |\langle E \rangle_{i}^{\text{pred}} − \langle E \rangle_{i}^{\text{exp}}| \), gives a notion of how strong the average shift in smFRET efficiency is, independent of their errors. However, its standardized version, i.e., \( \mu \) and \( \sigma \) are the mean and standard deviation of the experimental data. This shift is then a measure for the discrepancy in distance of the models compared to the data for the situation in the dsDNA.

Table 3. Comparison of Experimental Data and Simulation Results (in %) with Standard Measurement Errors and Sample Standard Deviations, Respectively

| model         | donor       | nt+7 | nt−18 | nt−30 | \( |\Delta| \) | \( |\Delta| \) |
|---------------|-------------|------|-------|-------|-------------|-------------|
| experimental FRET efficiency | 66 ± 2      | 89 ± 2 | 46 ± 2 |
| dynRotStatTrans | 74.5 ± 4.3 | 96.4 ± 1.0 | 53.5 ± 5.4 | 7.8 ± 0.6 | 1.4 ± 0.1 |
| statRotTrans | 57.7 ± 3.8 | 85.1 ± 2.0 | 40.8 ± 4.0 | 5.8 ± 2.3 | 1.1 ± 0.3 |
| dynRotTrans | 84.4 ± 5.5 | 99.4 ± 0.3 | 61.0 ± 5.7 | 14.6 ± 4.0 | 3.3 ± 1.3 |
| statRotDynTrans | 65.3 ± 5.7 | 92.9 ± 1.5 | 45.8 ± 4.5 | 1.6 ± 2.1 | 0.4 ± 0.1 |
| randAxSim | 69.2 ± 4.5 | 93.5 ± 1.5 | 49.3 ± 5.4 | 3.6 ± 0.7 | 0.7 ± 0.5 |
| geomAxSim | 62.6 ± 4.1 | 91.7 ± 1.6 | 45.8 ± 5.0 | 2.1 ± 1.7 | 0.5 ± 0.4 |

The last two columns show the mean of absolute differences (in %) and its standardized version as two alternative effect sizes along with their standard deviations.

As expected from the previous findings, dynRotStatTrans, statRotTrans, and dyndRotTrans show the greatest shifts on average (\( |\mu| > 5 \) Å). Further, their values for \( \mu \) and \( |\mu| \) are equal (except for sign), which means that the predicted distances are always too short (statRotTrans) or too long (dyndRotStatTrans and dynRotTrans). This suggests that the results are biased. The systematic shift might be due to the unknown mean axes because the true mean
axes are just one possibility of the randomly chosen ones. However, with an average shift of $\sim 4 \text{ Å}$, the model still performs well. The limiting model *statRotDynTrans* and the *geomAxSim* model yield the best results, indicated by absolute average deviations of $\sim 2 \text{ Å}$. However, the mean of the distance shifts, $\mu$, shows that the *statRotDynTrans* is more biased to greater distances.

We have seen that some limiting models deviate strongly from the experimental results. Because the AVs are the same for all models, the reason for the deviations can only be due to misconceived kinetics or restriction of rotation. Further, it came as a particular surprise that the results of the *statRotDynTrans* model agree with the experimental data well. Because the usually applied model *dynRotStatTrans* assumes exactly the opposite fusion coe$cients$. Instead, rotational correlation times, and geometrical constraints implied by the limiting anisotropies. Errors obtained for one particular situation could be significantly different from those for another one.

Therefore, we proceed, after having established a solid time-dependent simulation that accounts for geometrical and kinetic effects by investigating to what extent the common kinetic assumptions used for distance and structure determination by FRET are valid. For this, we derive in the following sections a method in order to determine which fraction of FRET events follows which regime.

### CLASSIFICATION INTO TRANSFER REGIMES

As described in the Physical Description of Transfer Regimes section, the current kinetic theory of FRET relates the average rates of the processes (rotational and translational diffusion and donor depopulation) to classify a motion of a dye completely as, e.g., dynamic. The most common model found in the literature for the dyes is based on the following assumptions:

$$k_{\text{DynTrans}} \gg k_{\text{DA}} \gg k_{\text{Stat}}$$

Given that these inequalities are true, it is deduced that the donor and acceptor (1) follow the dynamic regime regarding rotation and (2) are static regarding translation. However, there are some misconceptions regarding this procedure.

A major problem of the inequality given above is that it utilizes averages. However, one should look at single events rather than ensemble averages. The rate of donor depopulation, $k_{\text{DA}}$, for example, stands for many FRET events for which the times are (at least approximately) exponentially distributed. Because rotation and translation are stochastic processes as well, the times when we enter or leave a regime are random variables, too. Because the distribution of de-excitation times may overlap with these regime distributions, it is probable that one dye may behave static for one FRET event but dynamic for another. Thus, we expect to have a mixture of static, intermediate, and dynamic FRET occurrences of one dye.

Further, FRET is a process that depends on two different dyes in different environments. Thus, it is likely that we meet situations in our experiment where the kinetics of a donor's motion vary significantly from the acceptor's. Therefore, in contrast to the common procedure mentioned above, the donor's and acceptor's kinetics should be treated separately. Another point of criticism is that there exists no specific rate of rotation or translation as they are utilized in the inequality shown above. Instead, diffusion is characterized by the corresponding diffusion coefficients.

In light of these obvious shortcomings, we developed a method that—with help of our stochastic simulation (see the Simulation of smFRET Efficiency section) and applied statistics—decides how many FRET events can be classified as dynamic, intermediate, and static of the donor and acceptor separately. It is important to clarify that a FRET event—or we also term it a de-excitation or depopulation event—is a single realization of a donor de-excitation process, i.e., it might end in nonradiative relaxation, emission, or transfer. However, all of these realizations are important for the averaging process, not only those that end with transfer.

The next sections show how we translate the notions of each regime into so-called regime times telling us when we enter or leave a regime and how to compare these to the times until de-excitation of the donor in order to get the average fractions in the regimes.

#### Fractions of FRET Events within a Regime

The general procedure is the same for translation and rotation; therefore, we...
concentrate on rotation only. Given the rotational diffusion coefficient and the cone of a dye, the regime times telling when an occurrence is not static any more, i.e., entering the intermediate regime, or starts to behave dynamically depend only on the orientation in the cone at the time of donor excitation. Let us assume we have appropriate static regime times, $\tau_{\text{stat}}(\alpha)$, i.e., the times until the dye acts as static regarding rotation for every starting orientation $\alpha$ in the dye’s cone. Then, the expected static fraction, $x_{\text{stat}}$, is given by

$$ x_{\text{stat}} = \int \mathbb{H}(\tau_{\text{DA}}(\alpha) < \tau_{\text{stat}}(\alpha)) \, \mathrm{d}A / |C| $$

where $\mathbb{H}(A)$ is the indicator function, i.e., it returns 1, if the condition $A$ is true; otherwise, it is 0. The cap surface area of the cone is given by $|C| = \int \mathrm{d}A$, and the average time until donor de-excitation, or for short the average de-excitation time, from the starting orientation $\alpha$ is denoted by $\tau_{\text{DA}}(\alpha)$. It is noteworthy that the latter is not only an average over all possible paths from $\alpha$ but also an average over all paths from all allowed starting orientations of the FRET partner.

The expected dynamic fraction is obtained in a similar way. Let us again assume that we have appropriate dynamic regime times, $\tau_{\text{Dyn}}(\alpha)$, i.e., the times after which the dye acts dynamic regarding rotation for every starting orientation $\alpha \in C$. Then, the expected dynamic fraction, $x_{\text{dyn}}$, is given by

$$ x_{\text{dyn}} = \int \mathbb{H}(\tau_{\text{Dyn}}(\alpha) < \tau_{\text{Dyn}}(\alpha)) \, \mathrm{d}A / |C| $$

The fraction following the intermediate regime is then simply given by

$$ x_{\text{inter}} = 1 - (x_{\text{stat}} + x_{\text{dyn}}) $$

because a dye started at orientation $\alpha$ follows the intermediate regime, if it is not dynamic nor static.

Given appropriate regime times regarding translation, we obtain with a similar procedure the fractions in the respective translational regimes.

**Regime Times.** Determining the fractions in the regimes is simple. The hard part is defining regime times telling when we enter or leave a regime. Here, we need to translate the notions of the regimes to math. For the sake of readability, we focus again on rotation only. However, the concept is the same for translation. The detailed definition of the regime times regarding translation can be found in the Supporting Information.

**Static Rotation.** The definition of the static regime regarding rotation says that the dye did not rotate away from its initial orientation before the donor was de-excited. As a conservative approximation, we say that a dipole started at orientation $\alpha \in C \Psi$ is not static any more if it has left a cone with a cap surface, $C \Psi$, centered at $\alpha$ of semiangle $\Psi := \Psi / 10$. Therefore, an occurrence is not static any more if it has left a cone with a semiangle of 10% of the whole cone. The surfaces, $C \Psi$, to leave are of course intersections of the small cone with the whole cone, such that we have $C \Psi = C \Psi \cap C \Psi'$. Hence, we may have a reflecting boundary due to the cone border.

Therefore, we define regime times showing when the major probability mass of a dye started at $\alpha$ (we start off here with a Dirac delta) is outside of $C \Psi$, given formally by

$$ \tau_{\text{stat}}(\alpha) = \min \{ t \mid \int_{C \Psi} p^\Psi_{\Psi}(\psi, \varphi, t) \, \mathrm{d}A < \tau_{\text{Dyn}}(\alpha) \} $$

where $\min \{ \} \}$ denotes the minimum,“$A \mid B$ means $A$ such that $B$, and $\tau_{\text{Dyn}}(\alpha)$ is a stopping constant of choice. The probability distribution function $p^\Psi_{\Psi}(\psi, \varphi, t)$ over time $t$, the zenith angle $\psi$, and azimuth angle $\varphi$ of the corresponding spherical coordinate system aligned with the mean axis $\alpha$ is the solution to Fick’s second law with appropriate constraints, i.e., starting orientation and spherical cone border. The restricted rotational diffusion in a spherical cone started from any orientation $\alpha \in C \Psi$ does not yield a closed-form solution. Thus, we utilize stochastic simulations to estimate the static regime times for a set of starting positions.

In order to estimate $\tau_{\text{stat}}(\alpha)$ by $\tau_{\text{stat}}(\alpha)$, we start with $N = 1000$ dipoles at $\alpha$. The dipoles, $\alpha_i$, will spread, and we stop the simulation when less than $\tau_{\text{Dyn}} = 10\%$ of the dipoles are left within $C \Psi$. Therefore, we formally have

$$ \tau_{\text{dyn}}(\alpha) = \min \{ t \mid \sum_{i=1}^{N} \mathbb{H}(\alpha_i \in C \Psi) / N < \tau_{\text{Dyn}}(\alpha) \} $$

The procedure to determine the static regime time starting from a specific orientation is shown in Figure 10 for $n = \tau_{\text{Dyn}}$. We do this for $M = 1000$ starting orientations uniformly distributed in the cone. The result is a map of static regime times shown in Figure 12.

**Dynamic Rotation.** According to the definition of the dynamic regime regarding rotation, the dye needs to have sufficient time to be able to sample the whole cone of semiangle $\Psi$ until the donor is de-excited. Therefore, we define regime times that tell us when the probability distribution of a dye started from a specific orientation, $\alpha$, in the cone is approximately uniform over the whole cone, given formally by

$$ \tau_{\text{dyn}}(\alpha) = \min \{ t \mid p^\Psi_{\Psi}(\psi, \varphi, t) < \tau_{\text{Dyn}}(\alpha) \mid \forall \alpha \in C \Psi \} $$

where $|C \Psi| = (2\pi(1 - \cos(\Psi)))$ is the cap surface area of the cone and $\tau_{\text{Dyn}}(\alpha)$ is a stopping constant of choice. As mentioned above, the restricted rotational diffusion in a spherical cone started from a specific orientation $\alpha$ does not yield a closed-form solution. Thus, we estimate the dynamic regime times for a set of starting positions by means of stochastic simulations.

In order to estimate $\tau_{\text{Dyn}}(\alpha)$ by $\tau_{\text{Dyn}}(\alpha)$, we start with $N = 1000$ dipoles at $\alpha$. The dipoles will spread until they are uniformly distributed in the cone. As a measure to monitor the convergence to uniformity, we utilize the sum over all pairwise angles, i.e., $a(t) := \sum_{i=1}^{N} \sum_{j=1}^{N} a_{ij}(t)$, where $a_{ij}(t)$ is the angle between the dipoles of dyes $i$ and $j$ at time $t$. We say, uniformity from orientation $\alpha$ has been reached if we have $a(t) \geq 0.9(a)$, where $\langle a \rangle$ is the average of the sum over all pairwise angles of 1000 dipoles distributed uniformly on $C \Psi$ out of 100 trials. With this, we account for stochastic fluctuations. Therefore, we formally have

$$ \tau_{\text{dyn}}(\alpha) = \min \{ t \mid a(t) \geq 0.9(a) \} $$

The procedure to determine the dynamic regime time starting from a specific orientation is shown in Figure 11. We do this for $M = 1000$ starting orientations uniformly distributed throughout the cone. The result is a map of dynamic regime times shown in Figure 12.

**Estimating the Fractions of FRET Events within a Regime.** The general procedure is again the same for translation and rotation; therefore, we again concentrate on rotation only. Let's say, we have a set of $M$ starting configurations $\{q^M_i, \varphi^M_i, \alpha^M_i\}$ in $A \Psi \times C \Psi \times A \Psi \times C \Psi$. The donor (D) and
acceptor (A). Then, with the procedure defined above, we get a set of estimates, \( \omega(t) \), of static regime times regarding rotation. Hence, we can compute the static fraction of the donor by its Monte Carlo estimate.

Figure 10. Procedure to determine the static regime time regarding rotation. The starting orientation, \( \omega \), is shown as a red asterisk in the left plot. The border of the cone for all allowed angles is indicated by the blue ring. All simulated dipole orientations start from the red asterisk and spread over the spherical cone. After every time step, we check which dipoles are in \( \mathcal{C}_{\Psi} \), shown as green dots, and which are outside, shown as blue dots. The fraction outside of \( \mathcal{C}_{\Psi} \) is shown over time in the right plot. We say the static regime has ended for this specific starting orientation if more than 90% of the simulated dipole orientations have left \( \mathcal{C}_{\Psi} \), indicated by a red line. The distribution of dipole orientations found by this criterion is shown in the left plot.

Figure 11. Procedure to determine the dynamic regime time regarding rotation. The starting orientation is shown as a red asterisk in the left plot. The cone border is indicated by the blue ring. All dipoles start from the red asterisk and spread across the spherical cone during the simulation. After every time step, we calculate the normalized sum over all pairwise angles, \( \frac{1}{n} \sum a(t) / \langle a \rangle \), shown in blue in the right plot. The red line in the right plot shows the stopping criterion for the dynamic regime, namely, 0.9. The distribution of dipoles as an approximation for uniformity over the cone surface found by this criterion is shown by the blue dots in the left plot.

Figure 12. Maps of regime times regarding rotation. As an example, the simulation results of nt+7 are shown. For each starting orientation of a simulation (circle), the color gives the corresponding regime time in ns.
donor and acceptor mean regime times for translation and rotation (see the Supporting Information) are equal to or less than 100 ps. However, the average de-excitation time (see the Supporting Information) varies from 200 ps to 1.3 ns. Thus, in our test system, the time of donor de-excitation seems to be the major factor defining the fractions in the regimes.

This finding motivates us to investigate how the fractions change if the distance between the dyes and, thus, the FRET efficiency or average transfer rate change. Therefore, we artificially shifted the AV of the acceptor relative to the donor parallel to the DNA axis for the dye pair nt−30 and t−9 and computed for every distance the rotational and translational regime fractions of nt−30. Because the regime time maps do not change, if the AV is translated, we only need to perform the FRET simulation for every distance anew (see Figure 13).

At first, we notice that in the whole distance range, in which we usually use FRET to infer distances, we have a mixture of regimes. Furthermore, especially for rotation, the intermediate regime is dominating. Another striking finding is that for translation and rotation the dynamic fraction converges for greater distances, but not to 100% as was naively expected. This effect is due to the convergence of the average time until de-excitation, \(\tau_{DA} = (k_{D0} + k_\gamma)^{-1}\), to the characteristic time of donor depopulation in the absence of the acceptor, \(\tau_{D0}\), although the average transfer rate, \(k_\gamma\), still decreases. This phenomenon is shown in Figure 14 and will be further discussed in the next section.

### SUMMARY AND DISCUSSION

Distance or structure determination by means of smFRET measurements is a useful tool to complement other structural methods. However, the main source of uncertainty is that the fluorophores are attached to the macromolecule of interest via flexible linkers. A common approach is to choose the linker to be long enough such that the dye can rotate freely and fast, which means one can apply uniformly unrestricted and dynamic averaging and use \(\langle \kappa' \rangle = 2/3\). Additionally, the linker is chosen to be short enough that the uncertainty in position is minimized, which is accounted for by AV computation algorithms. Further, translation of the dye is oftentimes assumed to be statically averaged, which leads to one of the limiting models discussed in the Classification into Transfer Regimes section.

Although this model is predominantly used in the FRET community, one needs to test whether these assumptions are true. First, there is oftentimes experimental evidence that rotation is not unrestricted, i.e., if the time-resolved anisotropy measurement shows an anisotropy that is not decaying to zero. This case could be also accounted for by another limiting model discussed in the Notation for the Combination of Transfer Regimes section or as shown by Eilert et al.\textsuperscript{23}

However, if at least one motion of the donor or acceptor is in the range of the average depopulation time, no limiting model is applicable any more. The configuration of the FRET pair at donor de-excitation is then dependent on the configuration at donor excitation. Therefore, in order to predict the correct smFRET efficiency, there is a necessity to find a way to treat diffusion on the time scale of FRET.

The integration time during the smFRET measurement is in a TIRF setup at around 30−100 ms and in a confocal microscope at ~1 ms. This is long compared to the important processes of the dyes (emission, transfer, rotation, and translation), which are on the nanosecond time scale. Thus, we measure an average smFRET efficiency. However, even if we would use the exact arrival time of every acceptor photon, the relative dye kinetics

\[
\sum_{i=1}^{M} H_{i}^{DA}(p, o, p', o') < \varepsilon_{i}^{DA}(o_D)
\]  

(24)

where \(\{f_{DA}(p, o, p, o')\}\) is the set of realizations of the corresponding times until donor de-excitation. Progressing similarly for the dynamic regime, we then also obtain the estimated expected fraction of the intermediate regime. In the same manner, we obtain the fractions for the acceptor.

In the next section, we will apply this method to a realistic example of FRET measurements between dyes attached to various positions along dsDNA in order to investigate whether or when which kinetic assumptions may be appropriate.

### Exemplary Real-Case Study

For the investigation of a real-case scenario, we use again the three measurements of our test system. At first we simulate 1000 FRET trajectories in the geomAxSim model and memorize their starting configurations and de-excitation times. From these configurations, we start the simulations as described above to obtain estimates of the corresponding regime times for each motion. Subsequently, we estimate the fractions in the different regimes, as explained in the previous section. Table 5 summarizes the results for the FRET pairs. We also compute the means of the times of interest, i.e., the de-excitation time and dynamic and static regime time for translation as well as for rotation, which can be found in the Supporting Information.

The first thing that we notice is that our initial argument was right: in most of the cases, translation and rotation show a mixture of regimes. However, it might come as a surprise that for the majority of the dyes the intermediate regime dominates. This suggests that for these measurements no limiting model is applicable. In particular, this means that the good agreement of the results of the staticRotDynTrans model with the experimental data came by chance, e.g., due to the assumption of unrestricted rotation. Another striking observation is that the regime fractions change drastically for the measurement to nt−18, although the diffusion coefficients and the cap surface areas of the cones do not change considerably and the volumes and the shape of the AVs do not change at all, compared to the remaining measurements. The only major difference here is the increasing (average) smFRET efficiency from around 40 to around 90%. As a result, the average time until de-excitation is decreased drastically, which gives the dyes less time to diffuse. Thus, the fractions of the static regimes are increased, and the other regimes are depopulated as expected. The standard deviations of

### Table 5. Percent of FRET Events in the Different Regimes\textsuperscript{6}

<table>
<thead>
<tr>
<th>motion</th>
<th>regime</th>
<th>nt=30</th>
<th>t=9</th>
<th>nt+7</th>
<th>t=9</th>
<th>nt=18</th>
<th>t=9</th>
</tr>
</thead>
<tbody>
<tr>
<td>translation</td>
<td>static</td>
<td>4 ± 1</td>
<td>7</td>
<td>12</td>
<td>38</td>
<td>49</td>
<td></td>
</tr>
<tr>
<td></td>
<td>intermediate</td>
<td>84 ± 8</td>
<td>87</td>
<td>87</td>
<td>62</td>
<td>51</td>
<td></td>
</tr>
<tr>
<td></td>
<td>dynamic</td>
<td>12 ± 5</td>
<td>6</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>rotation</td>
<td>static</td>
<td>12 ± 11</td>
<td>19</td>
<td>61</td>
<td>60</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>intermediate</td>
<td>52 ± 48</td>
<td>57</td>
<td>57</td>
<td>37</td>
<td>37</td>
<td></td>
</tr>
<tr>
<td></td>
<td>dynamic</td>
<td>36 ± 24</td>
<td>27</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>

\textsuperscript{6}We ordered the dyes (at first donor, then acceptor) of the measurements from smallest to highest average smFRET efficiency. However, even if we would use the exact time of every acceptor photon, the relative dye kinetics.

The Journal of Physical Chemistry B

Article

DOI: 10.1021/acs.jpcb.8b07719
J. Phys. Chem. B XXXX, XXX, XXX−XXX

L

80
would still need to be taken into account in order to predict the correct smFRET efficiency (see the Appendix section Statistics of FRET). Thus, even time-resolved FRET measurements do not bypass the effects of the kinetics on FRET.

There are some publications that partially recognized the difficulty arising due to dye diffusion and tried to address it by different approaches. However, to date, a complete kinetic theory has not been reported. The seminal work of Dale and Eisinger exhaustively treated the dynamic regime even with geometrical aspects from time-resolved anisotropy measurements. However, these approaches are limited to the case that both dyes are in the same limiting regime and only rotation is considered.

Wieb van der Meer was one of the first who summarized and addressed the limiting regimes in a review but only with respect to rotation and for both dyes in the same regime. Before his review, interestingly, he with colleagues developed a strategy to deal with donor–acceptor systems in which transfer and diffusion take place simultaneously. Van der Meer et al. reduced the configuration space of the donor and acceptor to a discrete number of states and designed rate matrix models for the jumps between these states. They presented three examples, which all consisted of somewhat artificial geometries. Although these situations are strongly simplified, i.e., they treat orientation (just 3 orientational states for each dye or isotropic averaging) and translation again separately, the corresponding matrix models explode in complexity quickly.

Walczewska-Szewc et al. apply MD simulations and an AV approach in order to account for rotational and translational diffusion. MD simulations are not feasible for everyone and bring no better theoretical understanding, and the force fields for dye molecules are not developed sufficiently yet. The AV approach is partially similar to the one presented here, but in that work, both dyes are in the same regime, translation is addressed as dynamic or static, but rotation is only dynamic and uniformly unrestricted. Therefore, they discuss only two of the 81 possible regime combinations (see the Mathematical Description of Transfer Regimes section).

Badali et al. became aware that it is necessary to treat rotational diffusion simultaneously with translational diffusion. They performed Monte Carlo integration to solve 3 out of the 81 possible regime combinations (see the Mathematical Description of Transfer Regimes section). Moreover, they used somewhat artificial geometrical models where the linker was a rigid line, such that the position space of the dye was a hemisphere and the mean axes of the cones of the transition dipole moments were for every position aligned with the linker.
direction, resulting in atypical shapes for the simulated smFRET efficiency histograms.

Our work is partially inspired by Wallace and Atzberger. They understood that rotation and translation are Ito drift—diffusion processes and provided insights into dye kinetic effects on smFRET. However, again rotation and translation were investigated separately. They simulated rotation of the fluorophores without restriction and defined an artificial rotational diffusion coefficient such that it was not clear if the found effects were apparent in real experiments.

In order to overcome the described limitations, we established a theory that addresses every facet of FRET. Our theory includes both the relative geometry and kinetics of the dyes because these two are inherently linked. We present the notion of the transfer regimes in a clear way and, by rewriting the Förster formula, obtain a unique nomenclature that translates every regime combination into an expected value. If we require that we only have limiting regimes, the expected value is time-independent and the average smFRET efficiency can be computed by Monte Carlo integration.

In case where we have at least one motion in the time scale of transfer, we are forced to apply the intermediate regime. Hence, we established a method to simulate a complete trajectory of both FRET dyes and validated it on real data.

For simulation of the translational diffusion, we assume for all dyes a uniform distribution over the computed AVs. While this is a good assumption for dyes attached to nucleic acids, this might not be the case for other macromolecular environments, such as, e.g., proteins. In order to adapt our time-dependent simulation to these situations, one could apply one of the previously discussed dye distributions on AVs.

We assume in our simulation that rotation and translation are independent processes, i.e., the dye can be anywhere in its directed spherical cone at every position in its AV and vice versa. Because the dye’s freedom of orientation will mainly be limited by its local steric environment, this assumption is not strictly correct. In comparison, the limiting models used so far in the FRET community assume that the dye is rotating over the whole (hemisphere) at every position, which is a much stronger assumption than what is applied here. A more rigorous approach would be the ab initio simulation of anisotropies from macromolecular structures, which would need to be validated by experiments.

**Test Simulations.** Nagy et al. presented a benchmark study in which the smFRET efficiencies between dyes attached to bases at known positions of a linear double-stranded DNA were measured in a TIRF microscope. We used three of these measurements to evaluate the accuracy of a set of limiting models and the novel simulations. As a reminder, the limiting models assume uniformly unrestricted rotation and that both dyes are in the same regime regarding rotation or translation.

It is remarkable that the dynRotStatTrans model, which is commonly used for hydrophobic dyes, differs from the experimental outcome for the three investigated distances. The difference is also hinted because the angular deviations are longer than expected. Thus, using this model in structural modeling could result in a distorted macromolecular structure. The same is true also for the two other limiting models, dynRotTrans and statRotTrans. The statRotDynTrans model, however, represents the three test distances well. One interpretation is that this model correctly represents the kinetic conditions. However, this is inconsistent with our findings in the Classification into Transfer Regimes section, which tell us that no limiting model is applicable. Thus, the more likely interpretation is that the incorrectly assumed unrestricted rotation (it is incorrect because the experimental limiting anisotropies are greater than zero) shifts the predicted smFRET efficiencies because orientations are sampled that would not be allowed by a restricted rotation. Here, care needs to be taken because for another sample, which does not show a collinear geometry of dyes, the outcome could be very different, as discussed above.

A special difficulty of the time-dependent simulations is that we have no experimental evidence about the average axis of the spherical cone. Thus, in order to incorporate this as an error source, we performed 1000 simulations (each with 1000 trajectories), each with random average axes for donor and acceptor. The deviations of the predicted smFRET efficiencies provide a measure for the error due to unknown mean axes. This can be used to assess the error in distance determination by FRET due to the unknown orientation factor directly.

In our investigation, we used a straight and rigid dsDNA for the computation of the AVs. Gross et al. showed that the persistence length of dsDNA is ~39 nm. Our dsDNA is, with a total length of 66 base pairs, ~1.7× that large. This allows for some bending effects. However, not the total length of the dsDNA but the distance from donor to acceptor, which is here maximal 22 base pairs, is important for reduced distances by bending. These distances are way below the persistence length, such that the assumption of a linear, rigid geometry is a good approximation. Furthermore, Hellemans et al. showed for these distances that the bending effect on FRET is small, i.e., the difference in FRET efficiencies between the linear, rigid geometry and the flexible dsDNA is maximally 3% but usually less. However, it is striking that also in these studies the difference between the FRET efficiency obtained from a flexible dsDNA, calculated with the dynRotStatTrans model, from the experimental outcome is up to 6%. This suggests that an additional effect leads to a shift of the experimental FRET efficiency compared to the dynRotStatTrans model, which, as we showed, can be explained by dye kinetics.

A major advantage of our time-dependent simulations is that they are adaptive. This means that in another experimental system, e.g., the dyes are in a protein environment, another one (or even worse none) of the limiting models might be representing the molecular situation well. Our simulations, however, take experimental parameters into account to describe the kinetic situation from sample to sample individually.

**Classification into Transfer Regimes.** One of the main aims of this publication was to derive a method that allows us to elucidate by experimental data and stochastic simulations in which regimes the dyes are for a given FRET geometry and kinetics. The major conceptual step here is to switch from averages to single occurrences: we look at one realization of a transfer process, i.e., a specific configuration at excitation and its specific trajectory taken, and compare the resulting time until donor de-excitation with the corresponding dynamic and static regime times regarding rotation and translation. The latter were rigorously defined and accessed by stochastic simulations.

Performing this procedure for many starting configurations enabled us to provide regime time maps and subsequently estimate the expected fractions of how often a dye follows which regime. As expected, we observe a mixture of translational and rotational regimes, and most frequently, the intermediate regime is dominating. Thus, for these measurements, no combination of limiting regimes is applicable

A striking finding, however, is that we see a correlation between the FRET efficiency and the fractions in the regimes.
Therefore, for further investigation, we varied the distance between the AVs of nt−30 and t−9 in order to alter the FRET efficiency and computed the corresponding regime fractions. Toward smaller distances, the static fraction increases, but the intermediate and the dynamic fractions are reduced for translation as well as rotation (see Figure 13). For increasing distances, the static regime decreases, but the intermediate and the dynamic fractions increase. The fractions in the intermediate regime obtain at some distance their maxima and decrease then again. Here, it is important to realize that for the range of commonly measured FRET efficiencies (5−95%) no limiting regime dominates, neither for rotation nor translation. Therefore, for the dyes attached to DNA over the whole range of practically useful FRET efficiencies, there is no limiting model applicable. For other experimental samples investigated by FRET, we propose that the experimentally determined average time of the depopulation of the donor in the presence of the acceptor should be compared to the average regime times that are provided in the Supporting Information. With these, one can approximately decide whether a specific model of limiting regimes is applicable or our sophisticated simulations are necessary to predict the correct FRET efficiency.

One important caveat that becomes evident by our simulation is that it is not possible to completely reach the dynamic regime. In Figure 13, one can see that for increasing distances the fractions of the regimes converge but not to 100% of the dynamic regime, as one might naively have expected. This can be explained by photophysics (see Figure 2) by distinguishing between three possible paths of de-excitation: The donor may be depopulated by nonradiative processes, it may emit a photon, or the energy may be transferred to the acceptor. The distribution into these paths is given by the relative contribution of the corresponding rates. However, the distribution of the times of these three processes is the same exponential distribution with rate $k_{DA} = k_{NR,D} + k_{EL,D} + k_{T,D}$. Thus, if the average distance $d$ between the donor and acceptor increases and hence the FRET rate $k_T$ decreases, we have $k_{DA} \rightarrow \infty \rightarrow k_{NR,D} + k_{EL,D} = k_{DO}$. Therefore, the average of the times until transfer has an upper bound defined by $\tau_{DO}$, although the rate of transfer further decreases (see Figure 14). The durations for which the dye may diffuse do not increase any more; hence, the fractions in the regimes stagnate, too. As already discussed in the Geometry and Kinetics of FRET section, we need to distinguish between the inverse of the average time until transfer, $\tau^{-1} = k_{DA} \left( k_{DO} \infty \right)$ and the transfer rate $k_{T} \in \left[ 0, \infty \right)$. Hence, whether we can reach the dynamic regime completely depends therefore on the depopulation rate of the donor in the absence of the acceptor, $k_{DO}$, in comparison to the diffusion coefficient and the extent of the configuration space. Although this finding is basic and follows from the fundamental photophysics, most publications concerning FRET or specifically the transfer regimes define the latter by relating the diffusion rates with the transfer rate $k_T$ and not $k_{DA}$.

## CONCLUSIONS AND OUTLOOK

We provide a complete definition of the transfer regimes, the inherently connected combination of the relative geometry and kinetics of the donor and acceptor. Moreover, we find an appropriate mathematical description, i.e., a possibly time-dependent expected value weighted by the joint equilibrium distributions of rotation and translation. This complete theory will in the future simplify discussions between FRET scientists and bring better understanding to probable error sources in the analysis of FRET experiments.

There are several approaches to support or analyze smFRET measurements by simulations.\textsuperscript{2−4} These are throughout ab initio. We, however, presented a simulation procedure that is mainly guided by parameters gained from common fluorescence lifetime and time-resolved anisotropy experiments. This makes our approach less sensitive to simulation-related artifacts, e.g., wrong force fields in MD simulations for hydrophobic dyes that show an extended delocalized electron system.

In contrast, our approach is mainly limited by assumptions regarding the unknown equilibrium distributions of orientation and position. For example, rotation does not need to be uniform in a cone. However, we draw information out of experiments in order to adapt the size of the cone and the diffusion coefficient therein.

We showed that the most commonly applied kinetic dye model deviates strongly from the experimental data, resulting in distance discrepancies of up to 13 Å. This might lead to distortions of the structure analyzed by FRET. However, our simulations showed to be highly consistent with the data, with small discrepancies for unknown and inferred mean axes of rotation. Thus, our simulation method provides the key tool for unbiased determination of macromolecular structures by FRET.

Because the simulation enables us to look in detail at the processes that are hidden in the experiment, there will be many applications in which we get a better physical understanding of the molecular processes, e.g., conformational changes. Furthermore, our simulation method is not limited to FRET but could also be applied to electron spin localization using double electron−electron resonance (DEER) or electron pair resonance (EPR).

The rigorous translation of the notion of the transfer regimes to regime times enabled us to show that for common dyes attached via flexible linkers to dsDNA no limiting model applies over the usual range of FRET efficiencies. Further, we found in simulation and theory that the dynamic regime is not only determined by the dye’s diffusional velocity and the effective extent of the weighted configuration space but is actually limited by the inherent and environmental fluorescence characteristics of the donor. This prohibits application of the uniformly unrestricted, dynamic average orientation factor ($\chi^2$) of 2/3 for FRET measurements with donors showing a characteristic lifetime of at least less than 3 ns.

While assuming a comparatively large error in the Förster radius due to the uncertainties of the orientation factor\textsuperscript{27−30} can provide a practicable means to overcome these complications, the described theory and simulations can help by providing both a better theoretical understanding and a more precise definition of experimental uncertainties as well as biases.

### APPENDIX

#### Assumptions and Deductions

Assumptions regarding AV computation:

- the fluorophore is modeled as a sphere
- the center of the fluorophore is the position of the transition dipole moment
- the transition dipole moment is represented by a unit vector
- the linker is modeled as a flexible rod
- the macromolecule of interest is rigid

Assumptions regarding dye motions:
rotation and translation are assumed to be independent, i.e., the position of the dipole does not influence its orientation and vice versa.

- the equilibrium distribution of a dye’s motion can be expressed as a product of the equilibrium distributions of rotation and translation
- rotation and translation can be simulated independently
- the equilibrium distributions of rotation are uniform over spherical cones
- the equilibrium distributions of translation are uniform over the corresponding AVs

Assumptions regarding photophysics:
- the rate of excitation is low compared to the molecular processes, i.e., rotation, translation, and depopulation
- for every excitation, the donor and acceptor configuration can be drawn independently from the joint equilibrium distribution
- photon absorption is a Poissonian process

Statistics of FRET

If the donor and acceptor have a fixed configuration, we have a constant rate of depopulation of the donor excited state, \( k_{\text{DA}} = k_{\text{D}0} + k_1 \), and the depopulation process is Poissonian, i.e., \( X = X_{\text{D}0} + X_1 + X_{\text{R}} \sim \text{Pois}(k_{\text{DA}} t) \). For a short time duration \( \Delta t \), the probability for depopulation of the donor is then \( k_{\text{DA}} \Delta t + o(\Delta t) \), where \( o(\Delta t)/\Delta t \rightarrow 0 \) as \( \Delta t \rightarrow 0 \). The depopulation times of the donor are exponentially distributed with rate \( k_{\text{DA}} \). Because nonradiative relaxation, emission, and transfer are competing processes, the success probability to get transfer is \( p = \frac{k_1}{k_{\text{DA}}} \). If we have \( N \) photons absorbed, we have \( X_{\text{NIR}} + X_1 + X_{\text{E}} = N \). Conditional on that we have \( X_{\text{E}} \sim \text{Binom}(N, p) \). So the FRET efficiency, \( E = \frac{N_p}{N} \), follows a transformed binomial distribution with probability mass function \( p(E|N) = \binom{N}{N_p} p^{N_p} (1 - p)^{N - N_p} \). The sample space of the FRET efficiency is a subset of the rational numbers \( \mathbb{Q} = \left\{ \frac{z}{1} : z \in \{1, \ldots, N\} \right\} \subseteq \mathbb{Q} \). Its expected value is \( E = (E) \), i.e., the average FRET efficiency, and the variance is given by \( E(1 - E) \).

In a real-case scenario, but with a combination of limiting regimes, we will have no fixed configuration but a joint equilibrium distribution of orientations and positions of donor and acceptor dipoles. For the sake of simplicity, we discuss static averaging of translation and rotation of the donor and acceptor as an example only. Here, we do not have a single but a distribution of transfer rates derived from the joint equilibrium distribution. Given a configuration, \( \omega \), drawn from the joint equilibrium distribution, the success probability to get transfer is \( p(k_1(\omega)) = k_1(\omega)/k_{\text{DA}}(\omega) + k_2(\omega) \). If we have \( N \) photons absorbed, we have drawn \( N \) configurations, \( \{\omega_i\} \), independently from the joint equilibrium distribution, such that we get \( N \) independent transfer rates \( \{k_1(\omega_i)\} \). Thus, the FRET efficiency follows a transformed Poisson binomial distribution (which is the discrete probability distribution of a sum of independent but not identically distributed Bernoulli trials) with success probabilities \( \{p(k_1(\omega_i))\} \). Conditional on that the probability mass function of the FRET efficiency is given by

\[
p(E|k_1(\omega), i \in \mathbb{N}) = \sum_{A \in F_E} \prod_{i \in A} p(k_1(\omega_i)) \prod_{j \in A^C} (1 - p(k_1(\omega_j)))
\]

(25)

where \( A^C = \{1, \ldots, N\}/A \). The family of sets \( F_E \) is the set of all subsets of \( N \)-E integers that can be selected from \( \{1, \ldots, N\} \).

In an intermediate scenario (meaning nonlimiting regimes), the FRET dyes translate and rotate on the time scale of transfer. As they are described by Ito drift–diffusion processes, the transfer rate itself is a stochastic process (see Figure 7). Let us define a trajectory, \( T(t) \) with \( t \in [0, \infty) \), as a time-parametrized path of translation and rotation of both dyes. The probability to see a photon at time \( T \) after excitation (\( t = 0 \)) in a short time period \( \Delta t \) along the trajectory is \( (k_1(T) + k_{\text{D}0}) \Delta t + o(\Delta t) \). Therefore, the depopulation of the donor excited state is a time-inhomogeneous Poisson process. It is noteworthy that the depopulation times of the donor are not necessarily exponentially distributed any more.\(^4\) The success probability is then given by \( p(k_1(T)) = k_1(T) / (k_1(T) + k_{\text{D}0}) \). Thus, the trajectory determines when and therefore also what happens, i.e., transfer, nonradiative relaxation, or emission. If we have \( N \) photons absorbed, we have drawn \( N \) start configurations independently from the joint equilibrium distribution. From these, we follow \( N \) trajectories, \( \{T_i\} \), until the excited state of the donor is depopulated at times \( \{T_i\} \), such that we get \( N \) independent transfer rates \( \{k_1(T_i)\} \). Now it is clear that also the FRET distribution depends on the trajectories taken

\[
p(E|k_1(T_i)), i \in \mathbb{N}) = \sum_{A \in F_E} \prod_{i \in A} p(k_1(T_i)) \prod_{j \in A^C} (1 - p(k_1(T_j)))
\]

(26)

Conditions of the Dynamic Regime

We apply a less strict definition of the dynamic regime compared to Dale and Eisinger.\(^5\) They demand that the transition dipoles sample all orientations in a short time compared with the depopulation time (or, expanding their theory, also all positions). That means that the dye needs to have effectively visited all allowed configurations before depopulation occurred, such that it can be averaged dynamically. We think that this requirement is too strong because the ergodic hypothesis states if in a given time duration, say \( T \), a dye may have sampled the whole position/orientation space, the ensemble has sampled the whole position/orientation space. This means that after this time duration \( T \) we can already draw independently from the equilibrium distribution stating that the dynamic averaging is yet applicable. Therefore, in the experiment (given the data of many molecules and/or many excitations of the same molecule) and, thus, also in the simulation, the less strict condition already yields the same results as the more strict one.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jpcb.8b07719.

Definition of regime times regarding translation, lifetime and time-resolved anisotropy measurements, mean times and their standard deviations from the simulation, and computation of AVs (PDF)

DOI: 10.1021/acs.jpcb.8b07719
The Journal of Physical Chemistry B

AUTHOR INFORMATION

Corresponding Author
*E-mail: jens.michaelis@uni-ulm.de. Phone: +49 (0) 731 5023050.

ORCID
Jens Michaelis: 0000-0002-2739-4172

Notes
The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was funded by the DFG through Grant 286540087. The authors thank the Ulm University Center for Translational Imaging MoMAN and the doctoral fellowship of the Carl-Zeiss-Foundation for their support.

ADDITIONAL NOTES

(1) The theory defined here is applicable to single-molecule as well as bulk experiments.
(2) In order to provide clarity, we give in the Appendix section Assumptions and Deductions, a list of all of the assumptions applied and what we deduce from them.
(3) Without loss of generality, we assume here that the transfer rate is constant over time.
(4) It is noteworthy that in the case of a time-dependent transfer rate the paths with greater transfer rates, i.e., also with higher depopulation rates, will more likely end with transfer than with the other processes. Therefore, the average time until transfer will be shorter than, e.g., the average time until emission. Thus, even the stated equality does not necessarily hold any more, i.e., $k_{DA} \neq r^2$. However, we investigated the extent of this effect (see below) and found that it is considerably small, such that we have $k_{DA} \approx r^2$.
(5) It is noteworthy that in a strict sense there are no rates of rotation or translation but diffusion coefficients. We could obtain a rate if we know an average angle/distance to diffuse. For more advanced treatment of this problem, see the Classification into Transfer Regimes section.
(6) For convenience, we will not mention the weighting of the configuration space by the corresponding equilibrium distribution again but assume that the reader keeps this in mind.
(7) The numbering of bases was chosen following Nagy et al.32
(8) Here, we treat every motion as if it is in the intermediate regime. However, we could also simulate only some motions and treat the remaining as described in the Notation for the Combination of Transfer Regimes section.
(9) In a complete description of a drift-diffusion process, we have a drift term, $\mu$. The drift term is the force implied by the potential in position space. Because we have knowledge about only the allowed positions that the dye may occupy relative to the macromolecule to which it is attached but we do not know anything about its preferred locations, we assume the equilibrium distribution over the AV to be uniform. Hence, the drift term is zero within the AV. If the dye tries to leave the AV, the proposed translation is prohibited and a new proposal is drawn.
(10) For calculation of the limiting models, we use Monte Carlo simulations guided by the nomenclature discussed in the Notation for the Combination of Transfer Regimes section. However, we treat them here as Poisson counting processes, such that they have the same error source as the time-dependent simulations.

REFERENCES


5 Summary and Discussion

The cumulative thesis at hand comprises roughly three topics distributed over four publications: the development and promotion of the Fast-NPS software and its underlying algorithms and methods, the development of progressively realistic dye models and validation of the results of the structural inference by Fast-NPS.

Development and Promotion of Fast-NPS

The theoretical basis of the Fast-Nano-Positioning System (Fast-NPS) is a probability distribution, the so-called posterior, that tells us how probable a configuration of dyes, i.e. their positions and orientations, is, given a set of experimentally measured average smFRET efficiencies. The posterior contains the structural information in the form of the vector of expected values, marginal distributions and their credible volumes of dye positions. For various reasons these cannot be obtained as mathematical expressions and calculated analytically. Thus, we need to estimate the characteristics of the posterior distribution. This is possible by collecting samples, i.e. realizations, from the corresponding probability density function (see Section 3.2.3). These samples allow then the visualisation of the marginal distributions, approximate the credible volumes and compute the vector of expected values. In more detail, the fundamentals of NPS can be found in Sections 1 and 2 of publication 4.1. There are several approaches to drawing samples from a probability distribution. We chose the algorithm class of Markov Chain Monte Carlo (MCMC) methods. The loose framework of MCMC algorithms enabled us to selectively cope with problems that arose mainly because of the complicated prior distribution induced by the structure of the macromolecules of interest (see Section 3.1 in publication 4.1). We analyse the convergence of the cross entropy statistically in order to determine when the burn-in phase is over. We adapt the proposal kernel of the sampler for every network to the posterior distribution, such that we efficiently collect samples from the posterior distribution locally. In order to assure the ergodicity on the complicated state space of the diverse networks globally, we equipped the sampler with a parallel tempering scheme. It is fully adaptive with respect to optimizing the swap probabilities between the neighbouring temperature chains and the number of chains resulting in a complete determination of the temperature ladder. Here, we exploit the compactness of the support of the posterior. All adaptation phases are iterative or partially nested, i.e. they are stopped by specialized criteria. On a common desktop PC the runtime varies from 5 min to up to 5 h.

The prior, i.e. the distribution of allowed dye configurations and dye model assumptions, is knowledge that we apply beforehand. It is based on experimental evidence, e.g. crystal structures or data from additional fluorescence experiments, or may depend on rough assumptions. Because of the latter, there is a chance that the posterior may not represent the smFRET data we have put in. If this case is met, we verify with the consistency check in which we exploit the asymptotic multivariate normality of the average smFRET efficiencies (see
Section 3.3 in publication 4.1).
All these features make Fast-NPS a fast, consistent and easy-to-use tool for the inference of macromolecular structures by smFRET measurements. However, it is important to promote the software such that the whole target community is aware of the possibilities which Fast-NPS provides, and that we can encourage other scientists to conduct smFRET experiments and use Fast-NPS for structure determination. Hence, we developed a manuscript (see publication 4.2) that presents the setup of a TIRF microscope, the experimental procedure of smFRET measurements and the data analysis with Fast-NPS. The main and critical parts of the manuscript were visualised in a video. This reduces the contact barrier for scientists which are new to Fast-NPS or even to smFRET experiments.

**Primitive Dye Models**

One of the major difficulties in the structural analysis by smFRET data is the correct transformation of the average smFRET efficiencies to distances and vice versa. The fluorophores between which energy transfer happens are covalently bound to the macromolecule of interest via flexible linkers. The latter cause uncertainties in the position and orientation of the dyes’ transition dipole moments, even though for example the macromolecule is in a fixed configuration. This is the reason for a lack of knowledge about the exact transformation. Thus, we do not only have to take the statistics of the FRET process itself into account, but also the geometry and kinetics of the relative dye motions. For structure determination by smFRET it is widely assumed that the dyes orient freely at every point in the allowed position space, which is determined by AV algorithms. The equilibrium distribution on this configuration space is usually assumed to be uniform. For the translational diffusion it is assumed that the dye does not move on the timescale of the FRET process, i.e. translation is static. The rotational diffusion, however, is presumed to lead to a dynamic averaging, i.e. the rotation is much faster than the FRET process. Given that these assumptions reflect the truth, this gives us a bijective mapping from the average smFRET efficiency to the average distance. Since some of these points cannot be shown so far and others might be even proven wrong by experiments, we developed a set of primitive dye models varying and/or mitigating the assumptions mentioned above. These dye models were at first applicable network-wise, i.e. every dye in a network behaves in the same way. Here, we could show in an example network that only the two least restrictive models are consistent with the experimental smFRET efficiencies. This states that the commonly applied model described above is incorrect for at least one dye in this network (see sections Representative Results and Discussion in publication 4.2).

FRET dyes are attached to sites of the macromolecular complex which constitute different local environments, e.g. proteins, DNA, RNA or combinations,

---

\[c\] We call them primitive, because we have no profound evidence that can justify the application of one model in comparison to a another one.
that may restrict the rotation and/or effect the translation of the dyes individually. Hence, we developed a computationally fast Monte Carlo integration method that can provide for any dye model combination an individual conversion polynomial for the transformation from the FRET efficiency to the interdye distance. With that each dye in a network can be described by the best suited model.

Since the choice, which dye is assigned which model, is still based on rough biochemical reasoning, we expanded the consistency check to the marginal likelihood (see Section 3.3 in publication 4.1). Marginalization projects the likelihood to the distribution of only one measurement. Thus, we are able to find out which measurements are violate by the dye model assumptions. If we find that several violated measurements contain the same dye, it is likely that the model assumptions of this dye are causing the inconsistency. Hence, we can try a more suited model for this dye. We applied this method to an example FRET network. Here, we could achieve a 4-fold increase in precision on average compared to the most conservative model while staying consistent with the data. The highest observed improvement in precision was 18-fold.

**Validation**

In order to validate our structure determination tool equipped with the above mentioned primitive dye models, we conducted a benchmark study (see publication 4.3). Therefore, we performed smFRET measurements at a DNA only and DNA-protein construct in order to localize a supposedly unknown position. We could show that the individual assignment of the primitive models can improve the precision and accuracy of the dye localization strongly. With dye models individually applied according to the biochemical environment, Fast-NPS inferred the correct position.

Although these models and their dye-wise application provide a more selective approach to model the behaviour of the dyes, they are still problematic. The assignment of the models to dyes is not guided by experimental evidence, i.e. they are still parameters which are set subjectively. Thus, it may occur that a consistent model combination is yet not the true one. In the benchmark study we could show that this situation is met in the localization of a single dye at linear dsDNA (see Section III A1 in publication 4.3). In the analysis of a network of dyes this may lead to consistent, but distorted, thus incorrect structures. Further, the primitive models only provide variation in geometrical effects for rotation and translation, but do not account for the individual kinetics of the dye’s motions.

**Advanced Dye Models**

In our search for more realistic dye models, we understood that the FRET geometry and kinetics completely determine the smFRET distribution and the average smFRET efficiency. We elucidated at first that the FRET geometry is
given by the joint equilibrium distribution of donor and acceptor, i.e. the product of their independent equilibrium distributions of orientation and position. The FRET kinetics, however, are defined by the diffusion coefficients regarding rotation and translation (see sections Geometry and Kinetics of FRET and Physical Description of Transfer Regimes in publication [4.4]). We showed in the case of a known FRET geometry and limiting regimes, i.e. the diffusion processes are much faster and/or slower than the donor de-excitation process, that the average smFRET efficiency is obtained by a time-independent expected value. Here, we established a nomenclature that translates any limiting regime combination to the corresponding expected value and gives implicitly a manual for its numeric computation (see section Mathematical Description of Transfer Regimes in publication [4.4]).

In the case of non-limiting regimes, i.e. the diffusion processes are in the time scale of the depopulation process, we have a time-dependent expected value. Here, we rigorously defined the translational and rotational diffusion as Itô drift-diffusion processes and implemented a simulation method that takes auxiliary experimental data into account (see sections Fluorescence and Depolarisation and Simulation of smFRET Efficiency in publication [4.4]). This has never been done so far. All other approaches are either based on assumptions, e.g. the primitive or limiting models with other geometries, or are ab-initio. We tested our stochastic simulation on a network of dyes attached to dsDNA where the FRET geometry is approximately known. The auxiliary data supplying us the information about the geometric constraints and kinetics of rotation were obtained from fluorescence lifetime and time-resolved anisotropy measurements. For the simulation of translation, we used values from the literature. Subsequently, we compared the experimental results obtained at a TIRF microscope with our simulation and a set of limiting models. We find that three of the limiting models – including the most commonly applied model described above – deviate strongly from the experiments. Our simulations, however, are in close agreement with the experimental outcome.

A special feature of our simulations is that we assume that the dyes’ orientations are confined to spherical cones. As indicated above, we have experimental evidence about the extent of the cone, but do not know how it is directed in the reference frame of the macromolecule. If we apply random average cone axes of donor and acceptor for every FRET trajectory, i.e. the time-parametrised path of configuration of both dyes, we get close to the experimental data. Further, this simulation procedure explicitly accounts for the error in FRET efficiency due to the unknown distribution of the orientation factor $\kappa^2$. Another option is to infer average cone axes from the molecular environment. Since it is presumed that the usual FRET dyes do not interact with dsDNA rather than steric repulsion, it is natural to assume a correlation between dye position and orientation. Thus, we perform a principal component analysis (PCA) of the accessible volume of the respective dye and propose the normalized eigenvector with the greatest eigenvalue as the cone axis. The simulations with these axes showed to increase the agreement with the experiments drastically (see section Simulation of smFRET Efficiency in publication [4.4]). Although, this
procedure seems to work well for dyes attached to dsDNA, care need to be taken, if we analyse FRET measurements between dyes that are in a protein environment. Here, the dyes might stick to hydrophobic patches violating the assumption of steric repulsion only. Hence, the PCA procedure might lead to strongly biased distances and in the end to incorrect structures.

The exact trajectories of donor and acceptor configurations and thus the regime times, i.e. the times when a dye leaves or enters a regime, are hidden within the experiment. Our simulation – of course under the assumption that it reflects the truth – makes them accessible for analysis and interpretation. Thus, we are not restricted to determining which regime a dye’s motion follows on average, but we can classify any single occurrence. This enables us to give fractions in the different averaging regimes, what on the other hand makes it possible to study if the usually applied kinetic models are reasonable for a given dye combination and environment. This analysis, applied to the test network, revealed that we have a mixture of regimes in all measurements and that the intermediate regime is mostly dominating (see section Classification into Transfer Regimes in publication [4]). This finding stresses that no limiting dye model is applicable for the investigated dyes attached to dsDNA.

The regime fractions depend on three dependent characteristics: the mobility of the dyes, their equilibrium distributions and the donor depopulation process. We studied the effect of varying the FRET efficiencies, and thus of the depopulation times, on the regime fractions. That means, we changed the average distance between the dyes’ AVs, and kept the shape and relative orientation of the configuration space, and the diffusional mobility of the dyes unchanged. The result shows that we find mixtures of regime fractions over the complete distance range commonly used for measuring the average smFRET efficiencies, i.e. no limiting regime is reasonably applicable. We further saw that it is actually possible that a dye might not reach the dynamic regime completely, because the time the dye is able to diffuse is upper bounded by the characteristic lifetime of the donor in the absence of the acceptor (see Figure 13 in publication [4]).
6 Conclusion and Outlook

Single-molecule FRET data can be used to complement and even extend other methods for structure determination of dynamic macromolecules in biophysics. The cumulative thesis at hand contributes to key points of the analysis of smFRET experiments.

Development of Fast-NPS

On the one hand, the software Fast-NPS allows the user to analyse structures by smFRET data in a time duration of minutes to hours due to its novel, fast sampling engine. On the other hand, the adaptive MCMC algorithm of Fast-NPS enables the user to use Fast-NPS as a black box, i.e. it is not necessary to set any algorithmic parameters. In sum, this makes Fast-NPS an attractive and easy-to-use structure analysis tool. With this we contribute to the advance of smFRET measurements as a method for structure determination of biomolecules.

In the following paragraphs we will present selected improvements of the sampling engine that have the potential to decrease the correlation between subsequent samples and thus increase the speed of structural inference. Then we will address conceptual advances in the structure analysis by Fast-NPS.

The sampling efficiency locally is determined by the proposal kernel. The better it is adapted to the posterior, the less correlated the successive samples are. In the current version of the algorithm, we only adapt the diagonal entries of the covariance matrix of the multivariate normal proposal, i.e. the principal components coincide with the axes of the macromolecular coordinate system. This might not represent the shape of the posterior well. A simple future improvement would be to extend the estimation to the full Hessian of the posterior distribution. This would allow efficient proposals directed along the principal directions of the posterior in any case. A disadvantage, however, is that a multivariate normally distributed proposal would then need the Cholesky decomposition of the correlated covariance matrix which needs to be computed by e.g. a singular value decomposition. Since this is necessary only once before the actual inference though, it should not increase the run time remarkably. The proposals drawn from this improved distribution might decrease the correlation between the subsequent samples drastically and in turn increase the effective sample such that the overall runtime of the inference is reduced.

On a global scale, the efficiency of the sampler is enhanced by a parallel tempering scheme. After some local updates, we try to swap neighbouring replicas starting at the lowest temperature chain, progressing until we reached the highest temperature chain. We do this because it is more likely that neighbouring chains have a similar energy, i.e. are easier to swap. However, another swapping scheme was proposed by Lacki et. al. and suggests that prior to a swap we order the chains by their current energy. This ordering increases the mixing of chains which are further apart in temperature and thus enhances the exploration of the state space. As simple as this feature is, we expect it to
have a great impact especially on networks with a complicated support.

The state of Fast-NPS according to the publications of this thesis is that dyes of unknown position are structurally independent of each other. This enables us to localize flexible domains or determine the structure of a second flexible macromolecule, e.g. the path of a DNA/RNA, relative to the primary biomolecule.

However, sometimes the relative configuration of two or more almost rigid macromolecules to each other is of interest. This is called rigid-body docking. Here, each molecule defines a coordinated system, usually called reference frame. One of them is chosen to be the root frame containing the root molecule relative to which we rotate and translate the other molecules, usually called ligands. Each molecule should be labelled at least with four dyes such that the relative configuration can be determined unambiguously. It is important to mention that the dyes within one reference frame are now structurally dependent. Through their structural dependence, the posterior of the dye configurations induces a distribution defined over the positions and orientations of the reference frames relative to the root frame which we call reference frame posterior. The position of a molecule is taken to be its geometric centre which usually coincides with the origin of its reference frame by construction. The relative orientation of a ligand to the root molecule can be described by a set of three Euler angles, in our case yaw, pitch and roll. In sampling the posterior of dye configurations, we implicitly collect samples from the reference frame posterior. From these samples we can construct credible volumes of the ligand positions or present a selection of the sampled configurations in the root frame. The latter can be shown as a collection of configurations or iteratively in a video. We implemented two versions of increasing complexity: the first, called independent docking, models the ligands as spheres, the second, called dependent docking, takes their detailed molecular structure into account. The runtime of the independent docking does not vary significantly from the non-docking inference. The runtime of the dependent docking, however, usually more than triples. Thus, for early investigation of macromolecular complexes we recommend the independent version, but for precise inference of molecule configurations the latter should be preferred.

Oftentimes the proper docking of macromolecules requires that the local or even the global structure changes. One way to handle this problem is to produce for all docked molecules an ensemble of likely conformers and dock every possible combination. We understand a distinct combination of ensemble elements as a model, such that the ensemble docking can be comprehended as a model selection problem. This of course constitutes an extensive sampling problem. For a vast space of models we would apply the reversible-jump MCMC.\textsuperscript{31} It proposes jumps between the elements of an ensemble and accepts them with probability proportional to the ratio of their posterior values at the current docked configuration. On the long run the fraction of effective samples within a distinct combination of ensemble elements gives us an estimate for the evidence of this particular model relative to all other combinations.
Another way to account for structural flexibility is the continuous change of the conformation of each docked macromolecule. Here, we would need to define an appropriate force field as an advanced prior distribution, which then need to be multiplied with the already defined FRET likelihood, such that we obtain the posterior distribution. However, the drastically enlarged configuration space requires a more advanced sampling technique such as e.g. Hamiltonian Monte Carlo, again equipped with a replica exchange algorithm. A convenient solution would be the combination of already established flexible docking software tools with Fast-NPS.

Development and Application of Dye Models

The development of progressively realistic dye models is a great necessity in the field of structural modelling by smFRET data. Thus, the publications dealing with dye models contribute to the progress in a current topic of major interest. In the following paragraphs we show how we will further use the developed dye models, present their first applications and give ideas for their future development.

We started with a single conservative model and went to a selection of predefined models which describe the most probable molecular situations for dyes attached to macromolecules. Since their application is based on rough biochemical reasoning and testing statistical consistency, but follows no experimental evidence, we progressed to the development of a simulation method that incorporates parameters obtained from fluorescence experiments. The latter give information about the individual kinetic and geometric situation met by a dye in a given macromolecular environment. We showed in an example network that the stochastic simulation agrees best with the experimental data when compared to the limiting models. Thus, the next step is to implement the simulation in the algorithm of Fast-NPS and apply it to infer macromolecular structures.

The stochastic simulation has been applied first in structure determination for the investigation of PARP-1, an enzyme which is involved in the repair of damaged dsDNA. It comprises two zinc fingers, F1 and F2, binding to the damaged DNA. It has been shown that PARP-1 kinks DNA with a single-strand break to a certain degree. Kallis et. al. investigated the special case of single-strand breaks leading to a nick (see publication [v] in Section [I]). In order to illuminate the effect of the different domains on the kinking angle, we attached a donor on one site and an acceptor on the other site of the nick in the dsDNA. In a confocal microscope, we obtained the average smFRET efficiencies from several constructs: the nicked DNA only, DNA with F2 and DNA with F1 and F2. Simultaneously, we determined the auxiliary parameters, which are a prerequisite for the stochastic simulation, from fluorescence and anisotropy decays of both dyes from all three constructs. For the analysis, we produced an ensemble of sterically possible kinked dsDNA molecules only, with F2 and with F1 and F2 bound. The latter is a structurally narrow ensemble inferred
from NMR experiments. We computed for every element of the ensembles the accessible volumes for donor and acceptor and simulated the average smFRET efficiency. For each of the three experimental average smFRET efficiencies we then selected these elements of the ensembles which show a simulated efficiency within some error. The analysis with help of the stochastic simulation revealed three distinct angular states of kinked DNA which are consistent with the ensembles of DNA only and of DNA with F2. The ensemble from the NMR experiments is only consistent with the experimental smFRET efficiency of DNA with F1 and F2 as expected. These results led to a better understanding of the mechanism of the damage recognition by PARP-1.

Although the stochastic simulation seemed to describe the experimental data best in the publication [4.4], we applied several assumptions that might not be true. The underlying equilibrium distribution of dye orientation is assumed to be uniform over a spherical cone. Although we substantiated that this is a good approximation given the experimental error, there is evidence obtained by MD simulations that the more likely distribution is a bivariate normal wrapped around the hemisphere. A similar to the cone, the width or variance of the normal distribution can be inferred by the limiting anisotropy obtained from time-resolved anisotropy measurements. However, the relation to the corresponding rotational diffusion coefficient, which depends on the rotational correlation time as well as on the limiting anisotropy, has no analytic solution. In a forward approach, we could simulate scenarios of varying width and diffusion coefficient and find a functional relation to the parameters of the anisotropy decay.

A major uncertainty in the stochastic simulation is an experimental shortcoming: so far there is no experimental procedure that allows us to analyse the translational diffusion coefficient of a dye relative to the macromolecule it is attached to. Thus, we are forced to apply a diffusion coefficient obtained from MD simulations to any dye in any environment. This limits the individual treatment of FRET dyes. Here, a more sophisticated experimental method need to be designed.

With these further improvements, the major results of this thesis might contribute even more to the research field of structure determination of biomolecular complexes by FRET measurements!
7 Donor Intensity Trace is Asymptotically a White Noise Process with Offset

The donor intensity trace is a discrete-time stochastic process of random intensities, \((I_t)_{t \in \mathbb{T}}\), each taking values in \(\mathbb{N}_0\). The set of times is a countable subset of the positive real line, i.e. \(\mathbb{T} = \{i\Delta t : i \in \mathbb{N}\}\), where \(\Delta t\) is the integration time of the TIRF microscope. We analyse at first the inherent dependencies of the renewal process and show then that the random donor intensities are asymptotically independently and identically distributed with finite mean and variance, i.e. the donor intensity trace is asymptotically a white noise process with offset. If not stated differently, all of the applied theory is taken from [39].

The arrival pattern of photons within a laser beam in a TIRF microscope is a Poisson counting process. This means that the random times for donor excitations, \((T_{\text{exc}}^i)_{i \in \mathbb{N}}\), are independently and exponentially distributed with rate \(k_{\text{exc}}\). If we assume that the de-excitation of the donor has a constant rate, \(k_{\text{DA}}\), then it is an exponential decay process and the corresponding random times, \((T_{\text{DA}}^i)_{i \in \mathbb{N}}\), are independent and exponential as well with rate \(k_{\text{DA}}\). Here, we usually have \(k_{\text{exc}} < k_{\text{DA}}\). The interarrival times between two excitation–de-excitation events, \(T_i = T_{\text{exc}}^i + T_{\text{DA}}^i\), are then iid, too. Since \(T\) is a sum of independent random variables, its probability density function, also called the interarrival time density, is given by the convolution - denoted by \(*\) - of the two exponential densities:

\[
    f_T(t) = (f_{T_{\text{exc}}} * f_{T_{\text{DA}}})(t) = \frac{k_{\text{DA}}k_{\text{exc}}}{k_{\text{DA}} - k_{\text{exc}}} \left(e^{-k_{\text{exc}}t} - e^{-k_{\text{DA}}t}\right). \tag{24}
\]

This implies that the corresponding counting process does not have stationary nor independent increments any more. In order to show that, we compute the hazard function, \(h\):

\[
    h_T(t) = \frac{f_T(t)}{R(t)} = k_{\text{DA}}k_{\text{exc}} \frac{e^{-k_{\text{exc}}t} - e^{-k_{\text{DA}}t}}{k_{\text{DA}}e^{-k_{\text{exc}}t} - k_{\text{exc}}e^{-k_{\text{DA}}t}}, \tag{25}
\]

where \(R(t) = 1 - \int_0^t f_T(x)dx\) is the survival function, i.e. the probability that no event occurred until time \(t\). The hazard function is thus the momentary rate for a excitation–de-excitation event at time \(t\) after the previous event. An example of the hazard function is plotted in Fig. 5. For independent and exponential random variables with rate \(k\), for example, we have \(h(t) = k\), that means the rate for an event is independent of the time when the last event occurred. This is also known as the memoryless characteristic of the exponential distribution. In our case, the hazard rate is time-dependent, i.e. the rate for an event depends on the time when the last event happened.

\(^d\) This assumption is generally not true as shown in the publication [44] but it is adequate in order to study the characteristics of the donor intensity trace.

\(^e\) The index \(DA\) stands for the de-excitation of the donor (D) in the presence of the acceptor (A).
Hence, the process does not have stationary nor independent increments any more. However, after some time the process approaches to the memoryless characteristic and behaves like an ordinary time-homogenenous Poisson process again. This is easily seen by the fact that for \( t \to \infty \) we have \( h_T(t) \to k_{\text{exc}} \).

After we have revealed the type of dependence of successive events, we want to derive the distribution of number of photons in a specified time interval, the so-called count distribution, at first exactly and then asymptotically. Here we assume that the process starts at \( t = 0 \). In this context, we should note that the probability for less than \( n \) excitation–de-excitation events in a time interval \( t \) is the same as the probability that the time of the \( n \)-th arrival, \( S_n = \sum_{i=1}^{n} T_i \), is equal or greater than \( t \), i.e. \( P(N_t < n) = P(S_n \geq t) \). So the count distribution is derived by

\[
P(N_t = n) = P(N_t < n + 1) - P(N_t < n) = P(S_{n+1} \geq t) - P(S_n \geq t) = 1 - F_{n+1}(t) - (1 - F_n(t)) = F_n(t) - F_{n+1}(t),
\]

where \( F_n(t) = \int_0^t f_{S_n}(x)dx \) is the cumulative distribution function of \( S_n \). The density of the time for the \( n \)-th arrival is given by the iterated convolution:

\[
f_{S_n} = f_{T_1} \ast \cdots \ast f_{T_n}
= \left( f_{T_{\text{exc}}^1} \ast f_{T_{\text{DA}}^1} \ast \cdots \ast f_{T_{\text{exc}}^n} \ast f_{T_{\text{DA}}^n} \right)
= \left( f_{T_{\text{exc}}^1} \ast \cdots \ast f_{T_{\text{exc}}^n} \right) \ast \left( f_{T_{\text{DA}}^1} \ast \cdots \ast f_{T_{\text{DA}}^n} \right)
= f_{S_{\text{exc}}} \ast f_{S_{\text{DA}}}.
\]  

(27)

Here, we exploited that \((\mathcal{L}^1, +, \ast)\) is a commutative ring, where \(\mathcal{L}^1\) is the space of Lebesgue integrable functions. The last two densities are each the convolution of \(n\) independent exponential densities, such that the random variables, \(S_{\text{exc}}^n\) and \(S_{\text{DA}}^n\), follow an Erlangian distribution with parameters \(n\) and \(k_{\text{exc}}\) or \(k_{\text{DA}}\), respectively. The Erlangian distribution is a special Gamma distribution with the condition that \(n \in \mathbb{N}_0\). The Erlangian density of e.g. \(S_{\text{exc}}^n\) is given by

\[
f_{S_{\text{exc}}^n}(t) = \frac{k_{\text{exc}}^n \exp(-k_{\text{exc}}t) t^{n-1}}{(n-1)!},
\]

(28)

such that the density of the interarrival time between \(n\) events can be derived with (27) as

\[
f_{S_n}(t) = \left( \frac{k_{\text{DA}} k_{\text{exc}}}{k_{\text{DA}} - k_{\text{exc}}} \right)^n \sum_{i=0}^{n-1} \binom{n-1}{i} \frac{t^{n-i-1} \exp(-k_{\text{exc}}t)}{(-k_{\text{DA}} - k_{\text{exc}})^i} \gamma(n+i, (k_{\text{DA}} - k_{\text{exc}})t),
\]

(29)

where \(\gamma(s, x) = \int_0^x \exp(-t)t^{s-1}dt\) is the lower incomplete Gamma function. For \(s \in \mathbb{N}\) it has the form \(\gamma(s, x) = (s-1)! \exp(-x) \sum_{i=0}^{s-1} \frac{x^i}{i!}\).
Since we cannot derive a handy expression for the count distribution (26), we look at asymptotic results in order to make statements about its characteristics. Sir David R. Cox proved that the distribution of \( N_t \) is asymptotically normally distributed with mean \( t/\mu \) and variance \( \sigma^2 t/\mu^3 \), where \( \mu \) and \( \sigma \) are the mean and standard deviation of the interarrival time distribution.\(^{39}\) For the density in (24), we have \( \mu = (k_{exc} + k_{DA})/(k_{exc}k_{DA}) \), \( \sigma = \sqrt{(k_{exc}^2 + k_{DA}^2)/(k_{exc}k_{DA})} \) and both are finite. We conclude that the count distribution (26) has asymptotically finite mean and variance, if the mean and variance of the interarrival time distribution are finite.

Further, it is straightforward to see that the squared ratio of variance to mean, say \( \varphi \), of the count distribution tends asymptotically to the coefficient of variation of the interarrival time distribution, \( \nu = \sigma/\mu \), as \( t \to \infty \). In our case we have \( \nu = \sqrt{(k_{exc}^2 + k_{DA}^2)/(k_{exc} + k_{DA})} < 1 \forall 0 < k_{exc} < k_{DA} \). Thus, the count distribution in (26) is always underdispersed compared to the Poisson distribution having \( \varphi = 1 \).\(^{40}\)

\[
\begin{align*}
\text{Figure 5: Hazard function} & \quad h_T \text{ of the excitation–de-excitation process normalised by its limiting value, here } k_{exc}. \text{ As a realistic example we chose } k_{DA} = 1/\text{ns} \text{ and } k_{exc} = 1/\mu \text{s. It rapidly converges to its limiting value stating that the process quickly regains its memoryless property after every event.} \\
\text{In a TIRF setup the rate of donor excitation is usually much smaller than the rate of de-excitation. For the realistic example of } k_{DA} = 1/\text{ns} \text{ and } k_{exc} = 1/\mu \text{s, the hazard rate has reached } 99\% \text{ of its limit } \lim_{t \to \infty} h(t) = k_{exc}, \text{ already after } < 5 \text{ ns compared to an expected time between two events of } \mu = 1001 \text{ ns. That means the process rapidly recovers its memoryless property. So we can assume that the random times between two excitation–de-excitation events are}
\end{align*}
\]
exponentially distributed with rate $k_{exc}$. Since the integration time is usually $30 - 100 \text{ ms} \gg \mu$, we can appropriately use the asymptotic results. For these example parameters we have $\mu \approx \sigma \approx 1/k_{exc}$, thus the expected value and the variance of the count distribution are asymptotically equal to $\Delta t k_{exc}$. Hence, we do not see a strong underdispersion of the count distribution, either, i.e. $\nu > 0.999$. We conclude that the counting process can be adequately assumed to be Poissonian with rate $k_{exc}$.

The actual rate of the detection of a donor photon is given by $k_{Det} = \Phi_{DA} \eta k_{exc}$, where $\Phi_{DA}$ is the quantum yield of the donor in the presence of the acceptor and $\eta$ the detection efficiency of the camera. This means that only with probability $\Phi_{DA} \eta$, a donor de-excitation leads to the detection of a donor photon. However, a thinned Poisson process is again a Poisson process.

A Poisson process is a Levý process, i.e. it has stationary and independent increments. The latter condition states that the random number of donor photons within disjoint intervals of finite length are independent. Thus, the random donor intensities in disjoint time intervals of length $\Delta t$ given by $I_{i\Delta t} = N_{i\Delta t} - N_{(i-1)\Delta t}$ are independent for all $i \in \mathbb{N}$. Further, stationarity of increments tells us that $N_{i\Delta t} - N_{(i-1)\Delta t}$ is equal in distribution to $N_{\Delta t}$. Thus, the random donor intensities are asymptotically equally distributed by $I_{i\Delta t} \sim Pois(k_{Det}\Delta t)$ for all $i \in \mathbb{N}$ with $k_{Det}\Delta t$ as mean and variance. We conclude that the donor intensities in a TIRF microscope are asymptotically independently and identically distributed with finite mean and variance, i.e. their trace is represented well by a white noise process with offset.
## List of Figures

1. Jablonski diagram for FRET ............................................. 8
2. Relative geometry of the fluorophore transition dipole moments 10
3. FRET efficiency as a function of the distance .................... 11
4. Inverse transform method ............................................... 17
5. Hazard function .......................................................... 99
References


List of Publications


## List of Conferences

<table>
<thead>
<tr>
<th>Event</th>
<th>Location</th>
<th>Date</th>
<th>Presentation Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biennnial Meeting</td>
<td>German Biophysical Society</td>
<td>2018, August, Düsseldorf, Germany</td>
<td>Complete Kinetic Theory of FRET</td>
</tr>
<tr>
<td>Oral Presentation:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Poster Presentation:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Förster Resonance Energy Transfer in Life Science II</td>
<td>Max-Planck Institute for Biophysical Chemistry</td>
<td>2016, April, Göttingen, Germany</td>
<td>A specialized Sampling Engine for the Bayesian Inference Software Fast-NPS</td>
</tr>
<tr>
<td>Poster Presentation:</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Curriculum Vitae

2016, March – Present: Study of Mathematics
University of Ulm, Germany

2015, February – Present: PhD Student
Institute for Experimental Biophysics
University of Ulm

2012, October – 2014, December: Master of Science
Biochemistry
University of Ulm, Germany

2009, October – 2012, July: Bachelor of Science
Biochemistry
University of Ulm, Germany

Heinrich-Heine Gymnasium
Bottrop, Germany
Danksagung


Mein Dank gebührt ebenfalls meinem Zweitgutachter und Prüfungsvorsitzenden Prof. Dr. Christof Gebhardt. Als Professor im selben Institut betreuten wir zusammen ein biophysikalisches Modul bestehend aus Vorlesung und Seminar. Die Zusammenarbeit war kreativ und produktiv, sodass das Modul um einen neuen Programmier-Teil erweitert wurde.

Ich möchte auch Prof. Dr. Peter Reineker und Prof. Dr. Kay Gottschalk dafür danken, dass sie sich bereit erklärt haben als Beisitzer zu dienen.


Eine Person, die alle meine Hoch- und Tiefpunkte hautnah mitbekommen hat,
ist Tuğçe Demiral. Ohne ihre Unterstützung hätte ich manche schwierigen Momente nicht bewältigt. Danke!
Neben Freunden ist der wichtigste Rückhalt im Leben die Familie. Ich danke meiner Schwester Dr. Janine Eilert dafür, dass sie mir immer mit Rat und Tat zur Seite stand und auch nicht von mir gewichen ist, wenn ich mal schwierige Phasen hatte. Es mag wie eine Floskel klingen, jedoch ist mir erst spät in meinem Leben klar geworden, dass ich all jenes was ich bisher erreicht habe einzig und allein meinen Eltern, Walter und Ortrud Eilert, zu verdanken habe.