Supplementary Information for: Complex RNA folding kinetics revealed by single molecule FRET and hidden Markov models

Bettina G. Keller[1*], Andrei Kobitski[2], Andres Jäschke[3], G. Ulrich Nienhaus[2,4], Frank Noé[5*]

[1] Freie Universität Berlin, Institute of Chemistry and Biochemisty, Takustr. 3, 14195 Berlin, Germany
[2] Institute of Applied Physics, Center for Functional Nanostructures, and Institute of Toxicology and Genetics, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany
[3] Institute of Pharmacy and Molecular Biotechnology, Heidelberg University, Heidelberg, Germany
[4] Department of Physics, University of Illinois at Urbana-Champaign, Urbana, IL 61801, United States
[5] Freie Universität Berlin, Institute of Mathematics, Arnimallee 6, 14195 Berlin, Germany
[*] Correspondence to bettina.keller@fu-berlin.de or frank.noe@fu-berlin.de

1 Likelihood and Hidden Markov model

1.1 Model of the emission process

The framework of Markov-modulated Poisson processes operates on single-photon arrival times and is particularly suited to model fluorescence-based experiments.[1–3] In this framework, conformational dynamics are modeled as a Markov jump process between distinct conformational states which can only be observed via an experimental signal (hidden states). The states are not associated with a single FRET value of the signal but to a value range, both due to shot noise and conformational fluctuations. The value ranges of different states can overlap considerably. As a consequence, most values of the observed signal are associated with more than one hidden state, and the sequence of states cannot be directly inferred from the sequence of observed signal values. Instead one aims at finding the most likely sequence of states given the observed time series (Bayesian logic). For this, the emission of the signal is modeled by an observation likelihood, in which parameters adjust the range of possible values in a given state and their relative probabilities. Finding the optimal observation parameters for each state is part of the HMM optimization. In short: the functional form of the observation likelihood is pre-defined by the user, the optimal parameters are found by the HMM algorithm. The convergence of the HMM algorithm and the validity of the results crucially depends on how closely the functional form of the observation likelihood models the actual physical process of signal emission.

Poisson-shaped observation likelihoods model the photon emission process, which gives rise to the observed photon-by-photon trace in a single-molecule FRET experiment.

$$\mathbb{P}(n_a, n_d) \quad = \quad \mathrm{Pois}(k_a; n_a) \cdot \mathrm{Pois}(k_d; n_d)\,, \qquad (1)$$

n_a is the number of observed acceptor photons, n_d is the number of observed donor photons, k_a and k_d represent the emission rates of acceptor and donor photons, respectively. k_a and k_d are the parameters, which are adjusted to represent the photon emission in a particular hidden state. Alternatively, the donor and acceptor photon trace can be combined to yield a time series of FRET efficiencies. The "emission" of FRET efficiencies is often modeled by Gaussian functions which are centered at certain FRET values E and cover a symmetrical range of FRET values according to their widths σ_E.

$$\mathbb{P}(n_a, n_d) = \mathbb{P}\left(E = \frac{n_a}{n_a + n_d}\right) \quad = \quad \mathcal{N}(E, \sigma_E)\,. (2)$$

If the experimental setup only measures FRET efficiencies, one indeed has no choice but to model the FRET efficiency distributions (emission probabilities) in the hidden states by Gaussian probability distributions (or similar distributions which are functions of E). However, in the context of single-molecule FRET one virtually always measures the arrival time of individual photons, either photon by photon or binned into time slots. In this situation, one can formulate the emission probability either on the photon-level using Poisson distributions (eq. 1) or on the level of FRET efficiencies using Gaussian distributions (eq. 2). However, these two formulations are by no means equivalent. Eq. 2 is a rather severe approximation to Eq. 1 because the information how many photons have been recorded within a given time window is discarded at the level of FRET efficiencies.

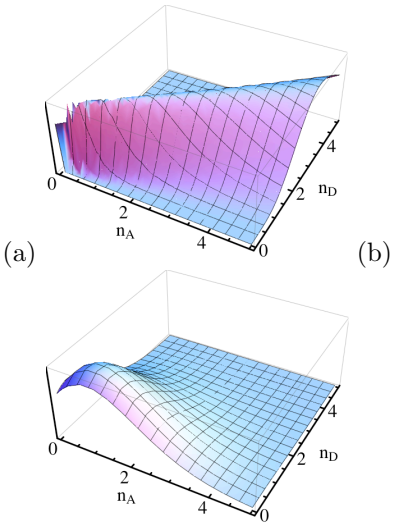

Figure S1: Comparison of emission probabilities based on (a) Poisson distributions (Eq. 1) and (b) Gaussian distributions (Eq. 2)

This is illustrated in Fig. S1, which shows the probability density of Eq. 1 for $k_a = 1.4\,s^{-1}$ and $k_D = 1.0\,s^{-1}$ (Fig. S1.a), and shows the probability density of Eq. 2 with $E = k_a/(k_a + k_d) = 0.58$ and $\sigma = 0.1$ (Fig. S1.b). Eq. 2 assigns high probabilities to certain observations with four and more photons per second, although these observations are highly unlikely given a total emission rate of $k_{tot} = k_a + k_d = 2.4\,s^{-1}$. Moreover, the FRET efficiency distribution cannot be symmetrical if the expectation value of the FRET efficiency either close to 0 or 1. Consequently, the Gaussian distribution does not approximate the FRET distribution well in these value ranges.

In summary, the choice as to which type of distribution is used to model the emission process is critical to the success of the HMM optimization. Gaussian distributions are approximations to Poisson distributions and should be used with caution. We therefore model the observation likelihood by a Poisson process. Fig. S4 shows that the observed photon counts in the present data sets are very well approximated by a Poisson process.

1.2 Photon likelihood including background photons

We derive a likelihood for observing (n_a, n_d) photons during a time span Δt in the acceptor and donor channel, respectively, given that a molecule labeled with a FRET pair is observed in the presence of background noise. In this situation, we have multiple photon-emitting processes, namely:

1. The molecule with current FRET efficiency E is irradiated with a certain laser intensity. This generates "molecule" photons (either emitted from donor or acceptor) with a rate $\frac{k_{mol}}{\Delta t}$.
2. The probability of such a photon to be an acceptor photon is the current apparent FRET efficiency E. Note that this efficiency E contains no background noise, but it may still contain effects like crosstalk and gamma. Important here is that we assume these effects to be a function of molecular state, but not a function of the trajectory (whereas the noise is).
3. For each trajectory being recorded we know that background photons are recorded in the acceptor and donor channels with rates $\frac{k_{a,bg}}{\Delta t}$ and $\frac{k_{d,bg}}{\Delta t}$, estimated from the bleached phase of each trajectory.

For each trajectory we estimate k_{mol}, in units of photons per Δt as follows:

$$\begin{aligned} k_{tot} &= \frac{\sum_{t=0}^{t_{tot}} n_{a,i} + n_{d,i}}{t_{tot}} \\ k_{mol} &= k_{tot} - k_{a,bg} - k_{d,bg}. \end{aligned} \tag{3}$$

The rate processes here generate photons according to Poisson processes. We now want to calculate the probability of observing a certain number of acceptor and donor photons n_A, n_D at time step i with given efficiency E and given rates k_{mol}, $k_{a,bg}$, $k_{d,bg}$ (for the estimation step of an HMM).

Consider a bin with n_a, n_d photons in the acceptor and donor channel, respectively. In a scenario where $n_{a,bg} \leq n_a$ and $n_{d,bg} \leq n_d$ background photons are amongst the observed photons, the molecule has emitted the remaining $n_{a,mol} = n_a - n_{a,bg}$ and $n_{d,mol} = n_d - n_{d,bg}$ photons, i.e., it has emitted a total of $n_{mol} = n_{a,mol} + n_{d,mol}$. We can compute the likelihood to observe n_a, n_d given that the molecule has a structure with efficiency E and the photon rates are $k_{a,bg}, k_{d,bg}, k_{mol}$ by summing up all possible scenarios of photons from background or molecule:

$$\begin{aligned} & p_{AD}(n_a, n_d) \\ &= \mathbb{P}(n_a, n_d \mid E, k_{mol}, k_{a,bg}, k_{d,bg}) \\ &= \sum_{n_{a,bg}=0}^{n_a} \sum_{n_{d,bg}=0}^{n_d} \mathbb{P}(n_{a,mol}, n_{d,mol} \mid n_{mol}) \\ &\qquad \times \mathbb{P}(n_{mol})\, \mathbb{P}(n_{a,bg})\, \mathbb{P}(n_{d,bg}) \\ &= \sum_{n_{a,bg}=0}^{n_a} \sum_{n_{d,bg}=0}^{n_d} \mathrm{B}_{n_{a,mol}}(n_{mol}, E) \\ &\qquad \times \mathrm{Pois}_{n_{mol}}(k_{mol}) \\ &\qquad \times \mathrm{Pois}_{n_{a,bg}}(k_{a,bg}) \mathrm{Pois}_{n_{d,bg}}(k_{d,bg}) \\ &= \sum_{n_{a,bg}=0}^{n_a} \sum_{n_{d,bg}=0}^{n_d} \begin{pmatrix} n_{mol} \\ n_{a,mol} \end{pmatrix} (E)^{n_{a,mol}} (1-E)^{n_{d,mol}} \\ &\qquad \times \frac{\mathrm{e}^{-k_{mol}} (k_{mol})^{n_{mol}}}{(n_{mol})!} \\ &\qquad \times \frac{\mathrm{e}^{-k_{a,bg}} (k_{A,BG})^{n_{a,bg}}}{(n_{a,bg})!} \\ &\qquad \times \frac{\mathrm{e}^{-k_{d,bg}} (k_{d,bg})^{n_{d,bg}}}{n_{d,bg}!} \end{aligned} \tag{4}$$

where $\mathrm{B}_x(n, p)$ is the binomial distribution in x with n number of trials and success probability p, and $\mathrm{Pois}_x(k)$ is the Poisson distribution in x with rate k. Using some algebraic rearrangements, it can be shown that the above expression is equivalent to

$$\begin{aligned} p_{AD}(n_a, n_d) &= \mathrm{e}^{-k_a} \frac{k_a^{n_a}}{n_a!} \mathrm{e}^{-k_d} \frac{k_d^{n_d}}{n_d!} \\ &= \mathrm{Pois}_{n_a}(k_a) \mathrm{Pois}_{n_d}(k_d) \end{aligned}$$

with rates

$$\begin{aligned} k_a &= E k_{mol} + k_{a,bg} \\ k_d &= (1-E) k_{mol} + k_{d,bg} \end{aligned} \tag{5}$$

Eq. (5) can be interpreted as follows: in the red (acceptor) channel, n_A photons are recorded. This is the result of a sum of two Poisson processes which output with a total rate of $E k_{mol} + k_{a,bg}$. Independently, n_D photons are recorded in the green (donor) channel from a Poisson process with rate $(1-E)k_{mol} + k_{d,bg}$.

For the special case of background absence ($k_{a,bg} = k_{d,bg} = 0$) we have $k_a = E k_{mol}$ and $k_d = (1-E) k_{mol}$

and thus p_{AD} can be simplified to:

$$\begin{aligned} p_{AD}(n_a, n_d) &= \mathrm{e}^{-(Ek_{mol})} \frac{E^{n_a} k_{mol}^{n_a}}{n_a!} \\ &\quad \times \mathrm{e}^{-(1-E)k_{mol}} \frac{(1-E)^{n_d} k_{mol}^{n_d}}{n_d!} \\ &= \mathrm{Pois}_n(k_{mol})\, \mathrm{B}_{n_a}(n, E) \\ &\propto E^{n_a}(1-E)^{n_d} \end{aligned} \quad (6)$$

with $n = n_a + n_d = n_{mol}$.

1.3 Background-corrected FRET histograms

The likelihood described in Sec. "Photon likelihood including background photons" can be used to construct FRET efficiency histograms, or distributions. Let $\{n_a^{(\Delta t)}(j,i), n_d^{(\Delta t)}(j,i)\}$ be a set of bins indexed by $i \in \{1, ..., N_j\}$, each containing the number of photons recorded in the acceptor and donor channels during a time Δt in trajectory $j \in \{1, ..., M\}$. Using Eq. (5), the FRET efficiency distribution for time bin Δt is estimated as:

$$\begin{aligned} &p^{(\Delta t)}(E) \\ &= \frac{1}{M} \sum_{j=1}^{M} \frac{1}{N_j} \sum_{i=1}^{N_j} \mathrm{Pois}_{n_a^{(\Delta t)}(k,i)}(Ek_{mol,j} + k_{a,bg,j}) \\ &\quad \times \mathrm{Pois}_{n_d^{(\Delta t)}(k,i)}((1-E)k_{mol,j} + k_{d,bg,j}), \end{aligned} \quad (7)$$

where for each bin one count is distributed over a histogram of E-values, and this histogram is then normalized by the number of counts to yield a distribution in E. The E-distributions shown in the main manuscript use FRET efficiency binning widths of 0.02.

For calculating the statistical error of FRET histograms, we use a bootstrapping procedure. 500 samples of trajectories are drawn with replacement from the full set of data trajectories. Each sample is generated such that it amounts to approximately the same sampling time as the full data set, i.e., it has the same information content. For each sample, $p^{(\Delta t)}(E)$ is estimated, and the sample mean and standard deviation is reported in Fig. 1a of the main text.

1.4 Hidden Markov models

For a comprehensive review of Hidden Markov models and algorithms, please refer to ref.[6] Here and in the subsequent sections, we will only describe the aspects of HMMs and HMM optimization algorithms that are specific to the present model.

We use an HMM for single-molecule FRET analysis for which we seek for a set of parameters $\lambda = \{\mathbf{T}, \boldsymbol{\pi}, \mathbf{e}\}$, that maximize the value of a likelihood function $\mathbb{P}(O|\lambda)$, which depends on these parameters. O is the observation, in this case the production phase of the measured smFRET traces. $\mathbf{T}$ is the matrix of transition probabilities between the conformational states within the time resolution Δt of the observed traces O, $\boldsymbol{\pi}$ is the vector of equilibrium populations for each state, and $\mathbf{e}$ is the vector of the FRET efficiencies of each state. Note that instead of rate matrices for modeling the hidden Markov process one can also use rate matrices which operate on continuous time.[3] However, rate matrices can invoke numerical difficulties when used in conjunction with many hidden states.[4,5]

The likelihood function has the form

$$\mathbb{P}(O|\lambda) = \sum_{\text{all possible } S} \mathbb{P}(O|\lambda, S), \quad (8)$$

where

$$\begin{aligned} \mathbb{P}(O|\lambda, S) &= \pi_{s_i} \mathbb{P}(o_1|e_{s_1}) \cdot \\ &\quad \prod_{i=2}^{T} T_{s_{i-1}\, s_i} \cdot \mathbb{P}(o_i|e_{s_i}). \end{aligned} \quad (9)$$

S is a particular trace over the conformational (hidden) states, and the summation over all S in the first equation is done using the forward algorithm.[6] Then, s_i is the particular conformational state which is visited at time $t = i\Delta t$. π_{s_i} is the equilibrium population and e_{s_i} is the FRET efficiency of this state. $T_{s_{i-1}s_i}$ is equal to the element of the transition matrix $\mathbf{T}$ which denotes the transition probability between the state at time $t = (i-1)\Delta t$, s_{i-1}, and the current state s_i. The product in the second equation runs over all T time steps in the observed trace. $\mathbb{P}(o_i|e_{s_i})$ is the emission probability and denotes the probability of observing a particular combination of $n_{D,i}$ donor and $n_{A,i}$ acceptor photons ($o_i = (n_{D,i}, n_{A,i})$), given that the molecule is in a state with FRET efficiency e_{s_i}. How well this function models the actual physical processes in the experiment, determines to a large extent the quality of the HMM results.

A crucial component in $\mathbb{P}(o_i \mid e_{s_i})$ is the background noise, which is not (entirely) determined by the experimental setup and does differ from trace to trace. In order to precisely distinguish states with similar FRET efficiencies, and recognize identical states visited in different trajectories as identical, it is crucial to account for the different background intensities explicitly in the stochastic model. This is even the case when the differences in background noise are rather small. We therefore use the likelihood function including background photons given by Eq. (5) for $\mathbb{P}(o_i \mid e_{s_i})$.

Note that additional measurement errors such as spectral cross-talk and differences in the detection efficiencies of the detectors are determined by the experimental setup and therefore do not differ from trace to trace, or change during a given trace. Differences in the quantum yields of the chromophores might in principle change during a given trace due to interaction with the biomolecule. However, if flexible linkers for the chromophores are used (as in the present case), the changes are averaged out on the time scale of the

sampling rate of the trace (10 μs). Therefore, also the differences in chromophore quantum yield can be regarded as constant over the set of traces and within a given trace. Measurement errors which are constant result in a constant shift of the FRET efficiency of each state and do not impair its usefulness as an order parameter to classify the states. Since distinguishing different states is the principal aim of this study (rather than measuring their exact chromophore distance), we do not incorporate these effects into the stochastic model underlying the HMM analysis. Rather, crosstalk is corrected on the level of conformational states after the HMM analysis.

Algorithm 1: Expectation-Maximization algorithm

Input: Observed trajectories $O = O^{(1)}, ..., O^{(M)}$,
initial parameter set $\lambda^0 = (\mathbf{E}^0, \mathbf{T}^0, \boldsymbol{\pi}^0)$,
Output: Parameters λ and likelihood $L(\lambda)$

1. For $i = 1, ..., N_{\text{step}}$:

1.2. **E-step** – find conditionally optimal hidden trace S^i

$$S^i = \arg\max_S \mathbb{P}(O \mid S, \lambda^{i-1})$$

1.2. Compute likelihood:

$$L^{i-1}(\lambda) = \mathbb{P}(O \mid S, \lambda^{i-1})$$

1.3. If $i < N_{step}$ and likelihood not yet converged: **M-step** – find conditionally optimal parameter set λ:

$$\lambda^i = \arg\max_\lambda \mathbb{P}(O \mid S^i, \lambda)$$

2. Return $\lambda^{N_{step}}$

For maximizing $\mathbb{P}(O|\lambda)$, we use the Expectation-Maximization (EM) approach, which iterates an expectation step (E-step) in which the hidden path S is optimized given the current set of parameters λ, and the maximization step (M-step) in which the parameters λ is optimized given the current hidden path S.[6,7] Iterating the EM algorithm to convergence leads to a local maximization of $\mathbb{P}(O|\lambda)$. In order to ensure that the local maximum is also a global maximum that can be found reproducibly, it is advisable to run the EM algorithm multiple times from different starting conditions (see Sec. "HMM initialization and optimization protocol"). The components of the EM algorithm (see algorithm 1) are described in the subsequent sections.

1.5 HMM path estimation (E-step)

Here we will first first discuss the E-step, which is done using a slightly modified forward-backward algorithm.

Forward algorithm. Now we consider all time points t (including those where no photon was observed), and treat t as a discrete count from 1 to t_{max}. For all possible pathways the forward variables α evaluate the following probability:

$$\alpha_t(i) := \mathbb{P}(o_1, ..., o_t, s_t = i \mid \lambda)$$

and are calculated by the Forward algorithm:[6]

$$\begin{aligned}\alpha_1(j) =& \pi_j \mathbb{P}(o_1 \mid s_1 = j)\\ \alpha_t(j) =& \mathbb{P}(o_t \mid s_t = j) \sum_{i=1}^{m} \alpha_{t-1}(i)\, T_{ij}.\end{aligned}$$

We define the vectors $\boldsymbol{\alpha}_t^T = [\alpha_t(1), ..., \alpha_t(n)]$ and the output matrix $\mathbf{P}_t = \text{diag}(\mathbb{P}(o_t \mid s_t = 1), ..., \mathbb{P}(o_t \mid s_t = n))$, and can rewrite this in:

$$\begin{aligned}\boldsymbol{\alpha}_1^T &= \boldsymbol{\pi}^T \mathbf{P}_1\\ \boldsymbol{\alpha}_t^T &= \boldsymbol{\alpha}_{t-1}^T \mathbf{T}\, \mathbf{P}_t\end{aligned}$$

Consider the time interval from time step $t_1 < t_2$ to t_2:

$$\boldsymbol{\alpha}_{t_2}^T = \boldsymbol{\alpha}_{t_1}^T \mathbf{T}\, \mathbf{P}_{t_1+1} \mathbf{T}\, \mathbf{P}_{t_1+2} \cdots \mathbf{T}\, \mathbf{P}_{t_2-1} \mathbf{T}\, \mathbf{P}_{t_2}$$

and assume that we observe photons on t_1 and t_2 but no photons in between. Then we have the output matrices $\mathbf{P}_t \propto \mathbf{Id}$ for no-photon times $t \in (t_1, t_2)$ and can simplify to $\boldsymbol{\alpha}_{t_2}^T \propto \boldsymbol{\alpha}_{t_1}^T \mathbf{T}^{(t_2-t_1)} \mathbf{P}_{t_2}$. As a result, the forward algorithm can be formulated only on the observation time points k, each of which has at least one photon:

$$\boldsymbol{\alpha}_{k+1}^T \propto \boldsymbol{\alpha}_k^T \mathbf{T}^{d_k} \mathbf{P}_{k+1}.$$

where d_k is the time interval between observation k and $k+1$.

Likelihood evaluation. The total likelihood of the trajectory can be evaluated as:

$$\mathbb{P}(O \mid \lambda) = \sum_{i=1}^{m} \alpha_N(i) = \sum_{i=1}^{m} \pi_i \beta_1(i)$$

Backward algorithm. In the same manner we define the **Backward Variables**:

$$\beta_t(i) = \mathbb{P}(o_{t+1}, ..., o_T \mid s_t = i, \lambda)$$

and calculated by

$$\begin{aligned}\beta_{t_{max}}(i) :=& 1\\ \beta_t(i) :=& \sum_{j=1}^{m} \beta_{t+1}(j)\, T_{ij}\, \mathbb{P}(o_{t+1} \mid s_{t+1} = j).\end{aligned}$$

In analogy to the forward variables, the backward variables can be calculated on the observation time points k which have at least one photon:

$$\boldsymbol{\beta}_k \propto \mathbf{T}^{d_k}\, \mathbf{P}_{k+1}\, \boldsymbol{\beta}_{k+1}.$$

Most probable hidden pathway. Based on the forward-backward variables we can now calculate the probability in each timestep to be in a given state:

$$\begin{aligned}\gamma_t(i) :&= \mathbb{P}(s_t = i \mid o_0, ..., o_T, \lambda) \\ &= \frac{\alpha_t(i)\beta_t(i)}{\mathbb{P}(o_0, ..., o_T \mid \lambda)} = \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{m} \alpha_t(j)\beta_t(j)}\,. \quad (10)\end{aligned}$$

The most likely path is calculated here using the direct approach:

$$\hat{s} = [\arg\max\{\gamma_1(i)\}, ..., \arg\max\{\gamma_T(i)\}]. \quad (11)$$

An alternative approach to calculate the most likely paths are the Viterbi algorithm, which avoids some spurious re-crossings and is therefore often preferred to the direct maximium likelihood path $\hat{s}$. However, theses spurious re-crossings are one (but not the only) aspect of non-Markovian behavior at short timescales. This effect is here corrected for a posteriori, by using a sufficiently long lagtime τ in the transition matrix estimated in the post-processing (see below).

1.6 HMM parameter estimation (M-step)

While the E-step is general for all Hidden Markov models (here the only modification was to derive a simplified version for the possibly irregular photon arrival times), the M-step differs for different HMM implementation, as it estimates the parameters for the specific HMM. Here, we estimate a statistically reversible transition matrix between hidden states, estimate the initial distribution of hidden states as the stationary distribution of that transition matrix, and estimate the background-corrected FRET efficiencies.

Count matrix estimate with Baum-Welch. For updating the estimate of the transition matrix, the number of transitions between hidden states are counted. The Baum-Welch estimate[7] is used, giving rise to the counts between states i and j at time point t:

$$c_t(i,j) = \frac{\alpha_t(i)\, T_{ij}\, \beta_{t+1}(j)\, \mathbb{P}(o_{t+1} \mid s_{t+1} = j)}{\sum_{m,n} \alpha_t(m)\, T_{mn}\, \beta_{t+1}(n)\, \mathbb{P}(o_{t+1} \mid s_{t+1} = n)}\,. \quad (12)$$

In matrix-vector form this can be written as:

$$\mathbf{C}_t \;=\; N\{\boldsymbol{\alpha}_t \mathbf{T} \boldsymbol{\beta}_{t+1} \mathbf{P}_{t+1}\}\,, \quad (13)$$

where $N\{\}$ means element-wise normalization as given by Eq. (12). This approach is immediately applicable to estimate C when a Markov model is formulated on all time points t. In single-photon data, we prefer to estimate and store only the hidden variables on the time points where photons arrive. Consider that we have photons at time points t_1 and t_2, and no photons during the $d_k - 1 = t_2 - t_1 - 1$ time steps in between. We then consider the state probabilities to be fixed to $\boldsymbol{\alpha}_{t_1}$ during the first half-time, then undergo a transition to $\boldsymbol{\beta}_{t_2}$ which are fixed during the second half-time. This yields the count matrix:

$$\begin{aligned}\mathbf{C}_t \;=\; &\frac{d_k - 1}{2} N\{\mathrm{diag}(\boldsymbol{\alpha}_{t_1})\} + N\{\boldsymbol{\alpha}_{t_1} \mathbf{T} \boldsymbol{\beta}_{t_2} \mathbf{P}_{t_2}\} \\ &+ \frac{d_k - 1}{2} N\{\mathrm{diag}(\boldsymbol{\beta}_{t_2})\} \quad (14)\end{aligned}$$

The total count matrix is obtained by summing over all counts of all M trajectories of length N_j each:

$$\mathbf{C} = \sum_{j=1}^{M} \sum_{t=1}^{N_j} \mathbf{C}_t^{(j)}$$

Reversible maximum likelihood transition matrix. We assume the conformational dynamics of the molecule to be in thermal equilibrium. As argued in,[8] this implies that the transition matrix between conformational states should be statistically reversible, i.e., fulfill the detailed balance equations. Thus, given the count matrix $\mathbf{C}$, we estimate the transition matrix by maximizing its likelihood subject to the detailed balance constraints:

$$\hat{\mathbf{T}} = \arg\max \left\{ \prod_{i,j=1}^{n} T_{ij}^{c_{ij}} \mid \pi_i T_{ij} = \pi_j T_{ji} \;\; \forall i,j \right\}, \quad (15)$$

where $\boldsymbol{\pi} \in \mathbb{R}^n$ is the stationary distribution of $\mathbf{T} \in \mathbb{R}^{n \times n}$, i.e., $\boldsymbol{\pi}^T = \boldsymbol{\pi}^T \mathbf{T}$. Problem (15) is solved using the iterative optimization algorithm described in,[8] using the implementation in the EMMA library.[9]

Initial / stationary distribution. We assume the molecule studied to be governed by a stationary dynamical process. Stationarity here means that the rules of the dynamics don't change during measurement time, i.e., they are determined by the same transition matrix at all times. Consequently, the stationary distribution of the transition matrix estimate $\hat{\mathbf{T}}$ is used as an initial distribution for all trajectories in the subsequent E-step:

$$\hat{\boldsymbol{\pi}}^T = \hat{\boldsymbol{\pi}}^T \hat{\mathbf{T}}. \quad (16)$$

The solution $\hat{\boldsymbol{\pi}}$ of the eigenvalue problem (16) is obtained as a side-product of the solution of the maximum likelihood problem (15).

FRET efficiencies. For each trajectory j, the apparent efficiency of state s, $e_s^{(j)}$, and its statistical weight $w_s^{(j)}$, are given by:

$$e_s^{(j)} = \frac{1}{w_s^{(j)}} \sum_{t=1}^{t_{max}} \gamma_t^{(j)}(s)\, n_a^{(j)}(t).$$
$$w_s^{(j)} = \sum_{k=1}^{k_{max}} \gamma_t^{(j)}(s)\, (n_d^{(j)}(t) + n_a^{(j)}(t)).$$

These sums are taken only over actual photon events.

When there is background noise, $e_s^{(j)}$ needs to be corrected, before combining the estimates of the different trajectories. As above, let k_{mol} be the estimated total photon emission rate of the molecule, and $k_{a,bg}$ and $k_{d,bg}$ be the background photon emission rates. The apparent FRET efficiency is then given as a function of the background-corrected FRET efficiency E_s by

$$e_s = \frac{k_{mol}E_s + k_{a,bg}}{k_{mol}E_s + k_{a,bg} + k_{mol}(1-E_s) + k_{d,bg}}\,.$$

Rearranging leads to the estimate:

$$E_s^{(j)} = e_s^{(j)}(1 + \eta_a + \eta_d) - \eta_a\,,$$

where we have introduced the noise-signal ratios as $\eta_a = k_{a,bg}/k_{mol}$ and $\eta_d = k_{d,bg}/k_{mol}$.

Subsequently, the estimates of all trajectories are combined as following:

$$\hat{E}_s = \frac{\sum_{j=1}^{M} w_s^{(j)} E_s^{(j)}}{\sum_{j=1}^{M} w_s^{(j)}}.$$

1.7 HMM initialization and optimization protocol

As the EM-algorithm finds only local maxima, a multistart protocol is used, in which multiple EM-runs are started from different initial parameter sets. There are only two important input variables that guide the search for the HMM: the number n of hidden states, and an estimate of the characteristic timescale t of the relaxations in the data. Given these inputs, a random initial paramter set is generated by Algorithm 2

Step 2 in Algorithm 2 generates a reversible transition matrix with approximately equally distributed stationary probabilities and implied timescales on the order of t. For illustration, consider a 2×2 matrix whose second eigenvalue is given by $\lambda_2 = 1 - T_{12} - T_{21}$. Using the initialization algorithm above, the average eigenvalue of the matrix $\mathbf{T}^0$ evaluates to:

$$\langle \lambda_2 \rangle = 1 - 2\langle T_{ij}^0 \rangle = 1 - 2\frac{0.5}{0.5+t} \approx 1 - \frac{1}{t}$$

Algorithm 2: Generation of initial parameters λ_0

Input: Number of states n, timescale t
Output: $\lambda_0 = (\mathbf{E}^0, \mathbf{T}^0, \boldsymbol{\pi}^0)$

1. Generate a vector of initial FRET efficiencies, $\mathbf{E}^0 \in \mathbb{R}^n$, with

$$E_s^0 \sim U(0,1)$$

2. Generate initial transition matrix $\mathbf{T}^0 \in \mathbb{R}^{n\times n}$ by:

2.1. Let $\mathbf{X} \in \mathbb{R}^{n\times n}$ be a symmetric matrix of pseudocounts. Initialize by:

$$x_{ij} = t\delta_{ij}$$

2.2. For all $i < j$:

$$x_{ij} \sim \tfrac{1}{n-1} U(0,1)$$
$$x_{ji} = x_{ij}$$

2.3. Normalize:

$$T_{ij}^0 = \frac{x_{ij}}{\sum_k x_{ik}}$$

3. The initial and stationary distribution is given as the solution of

$$\boldsymbol{\pi}^{0,T} = \boldsymbol{\pi}^{0,T}\mathbf{T}^0$$

4. Return $\lambda^0 = (\mathbf{E}^0, \mathbf{T}^0, \boldsymbol{\pi}^0)$.

The approximation is valid when t is much larger than 1. The associated implied timescale is then given by:

$$t_2 \approx \frac{-1}{\ln(1-t^{-1})} \approx t.$$

Given the initial parameters, the search for the optimal HMM can be started. Because many initial parameters will lead to poor local maxima, we split the optimization in two phases: An exploration phase, where many initial parameter sets are optimized for a few steps only and are ranked according to their likelihood; and an refinement phase in which the best parameter set is optimized to convergence. Algorithm 3 summarizes these steps.

Each HMM reported in the main manuscript is an output of Algorithm 3 using $N_{\text{explore}} = 100$, $S_{\text{explore}} = 3$, and $S_{\text{opt}} = 100$. For each Mg^{2+} concentration, the entire Algorithm 3 was repeated 10 times to test whether the optimum could be found reproducibly. To determine n, we repeated this procedure for $n = 2, 3, 4, ...$, and used the largest value of n for which the optimum value could be found at least 2 out of the 10 times.

Algorithm 3: HMM optimization

Input: Number of states n, timescale t. Parameters $N_{\text{explore}}, S_{\text{explore}}, S_{\text{opt}}$

Output: HMM parameters λ

1. For $i = 1, ..., N_{\text{explore}}$

2.1. Generate initial parameters λ_i^0 using Algorithm 2

2.2. Optimize λ_i^0 for S_{explore} steps to λ_i^e using Algorithm 1. 3. Let

$$\lambda_{\text{best}}^e = \arg\max_i \{L(\lambda_i^e)\}$$

be the best parameter set after the exploration phase. Optimize this parameter set to convergence or up to a maximum of S_{opt} steps using Algorithm 2:

$$\lambda = EM(\lambda_{\text{best}}^e, S_{\text{opt}})$$

4. Return optimized parameters λ_{best}^e

2 HMM validation and refinement methods

2.1 Lifetime distributions

Let $\hat{s}^{(j)}$ be the maximum likelihood path of trajectory j, according to Eq. (11). All paths are dissected in segments in which the path is in the same state, resulting in segments $s^{(k)}$, each having a length $l(s^{(k)})$. The last segment in each trajectory is discarded because its duration is determined by the end of the measurement rather than by the lifetime distribution of the state. We then select all segments which are in state s and compute the cumulative lifetime distribution as:

$$p(t) = \frac{\left\|\left\{s^{(k)} \mid l(s^{(k)}) \geq t\right\}\right\|}{\left\|\left\{s^{(k)}\right\}\right\|}.$$

where $\|S\|$ is the size of set S. The cumulative lifetime distribution is used rather than the direct lifetime distribution, because it is numerically more robust to estimate. In Markovian dynamics, the lifetime distribution is exponential, and hence the cumulative lifetime distribution is exponential, too. However, as stated above, using the direct maximum likelihood path $\hat{s}^{(j)}$ can create spurious transitions and thus induce some non-exponentiality in the lifetime distributions. Hence, the degree of exponentiality of $p(t)$ is only used as an rough indicator of the quality of the discretizations, whereas the subsequently described comparison of the time-binned FRET efficiency distribution is considered to be the test of the stationary and kinetic consistency of the HMM.

2.2 Implied timescale test

The eigenvectors of a detailed balanced and irreducible transition matrix $\mathbf{T}(\tau)$ form a complete basis of $\mathbb{R}^N$, where N is the dimension of the transition matrix, any probability on the associated conformational space $\mathbf{p}(t = n\tau)$ can be expressed as a linear combination of these eigenvectors:

$$\mathbf{p}(t = n\tau) = \sum_i c_i \lambda_i^n(\tau)\psi_i\,, \tag{17}$$

where $\lambda_i(\tau)$ is the eigenvalue associated to the ith eigenvector ψ_i.[10] The probability distribution $\mathbf{p}(t)$ can be interpreted as consisting of modes $\{\psi_i\}$ which have time-dependent amplitudes $c_i\lambda_i^n(\tau) = c_i\lambda_i^{t/\tau}(\tau)$. More precisely, $\lambda_i^n(\tau)$ encodes an exponential decay (in transition matrices, $\lambda_i(\tau) \leq 1 \;\; \forall\, i$)

$$\lambda_i^n(\tau) = \exp\left(\frac{t}{\tau}\ln\left(\lambda_i(\tau)\right)\right) = \exp\left(-\frac{t}{\mu_i(\tau)}\right) \tag{18}$$

where the implied timescale μ_i is given as

$$\mu_i(\tau) = -\frac{\tau}{\ln(\lambda_i(\tau))}. \tag{19}$$

Eq. 19 can be used to identify deviations from Markovian dynamics. If the dynamics are indeed Markovian, the eigenvector expansion (eq. 17) can be based on a transition matrix $\mathbf{T}(\tau)$ with arbitrary lag time τ, without altering the implied timescale of the decay process, i.e. $\mu_i(\tau) = \text{const}$. Plotting the implied timescales of transition matrices with various lag times $n\tau$ will yield a set of constant functions if the underlying dynamics is Markovian.[10] In practice, one finds that this is only (approximately) true for lag times greater than a critical a value: $\tau \geq \tau_{Markov}$.

2.3 Time-binned FRET efficiency distributions

In order to test whether the estimated HMM is consistent with the stationary and kinetic behavior of the data, we compute FRET efficiency distributions for a series of time-binning lengths separately from the data using the method described in Sec. "Background-corrected FRET histograms" and from the HMM. Algorithm 4 summarizes how the FRET efficiency distribution is calculated from the HMM.

Algorithm 4: FRET efficiency distributions from HMMs

Input: k_{mol}, τ
Output: $p(E)$

1. For $i = 1, ..., N$

1.1. $n_a = 0$, $n_d = 0$

1.2. Generate trajectory $s_{1,...,\tau}$ with
$$s_0 \sim \boldsymbol{\pi}$$
$$s_t \sim \mathbf{T}_{s_{t-1}}$$

1.3. For $t = 1, ..., \tau$

1.3.1. With probability $\Delta t k_{mol}$: generate a (virtual) photon.

With probability E_{s_t}: assign it to the acceptor channel and increment n_a,

else: assign it to the donor channel and increment n_d.

1.4. Add a histogram count at $\hat{E} = \frac{n_a}{n_a+n_d+\epsilon}$ where ϵ is a small number.

2. Normalize histogram by dividing by N, and return $p(E)$.

We consider an HMM with parameters λ to have passed the consistency test, when data-estimated and HMM-generated time-binned FRET efficiency distributions agree for all time windows used, within two standard deviations standard error of the data-estimated distributions.

2.4 State splitting probability

Like a direct Markov model,[8] the estimated HMM will provide a consistent model of the stationary and kinetic properties of the data if the Markov states are sufficiently fine, i.e., if there are sufficiently many states such that the conformational dynamics is appropriately described by a Markov chain. We conducted tests of the model consistency by comparing directly estimated and HMM-generated FRET efficiency histograms on various timescales as described in Sec. "Time-binned FRET efficiency distributions", and, when this test fails we consider to split states that are most likely to be the course for the failure. This procedure is continued until the HMM test succeeded. To select candidates for splitting, we considered the lifetime distributions described in Sec. "Lifetime distributions" and identified those states whose lifetimes can clearly not be described by a single exponential.

Since the lifetime distributions are computed from a finite number of realizations, the decision whether an estimated lifetime distribution is single-exponential or not, must be based on statistics. Here, we develop a Markov Chain Monte Carlo (MCMC) algorithm that, for each estimated cumulative lifetime distribution $\hat{p}(t)$, performs a model selection between a single exponential generating model:

$$p_1(t) = \mathrm{e}^{-tk}$$

and a bi-exponential generating model:

$$p_2(t) = a\mathrm{e}^{tk_1} + (1-a)\mathrm{e}^{tk_2}.$$

The probability for either of these models to generate a sample of segment lengths $(t_1, ..., t_n)$ is given by:

$$\mathbb{P}(\lambda_x \mid t_1, ..., t_n) = \mathbb{P}_x(\lambda) \prod_i p_x(t_i)\,,$$

where we use Jeffrey's prior:

$$\mathbb{P}_1(k) = \frac{1}{k}$$

and

$$\mathbb{P}_2(a, k_1, k_2) = \frac{1}{ak_1 + (a-1)k_2}$$

When correctly defined, this MCMC algorithm will sample from each of the two models according to their respective probabilities to have generated the observed set of exit times. Such an MCMC algorithm requires at least four Monte Carlo steps: (1) a step that can sample new parameters k within the single-exponential model, (2) a step that can sample new parameters a, k_1, k_2 in the bi-exponential model, (3) a step to split a single-exponential model into a bi-exponential model, (4) a step to merge a bi-exponential model into a single-exponential model.

In order to implement the split and merge steps, we need to propose a rule by which the single-exponential parameter k and the bi-exponential parameters a, k_1, k_2 are related, and then compute the appropriate MCMC acceptance probabilities from this rule. Consider the following relation:

$$\begin{aligned} k &= ak_1 + (1-a)k_2 \\ &= a(k_1 - k_2) + k_2 \end{aligned} \tag{20}$$

and consider further the parametrization

$$k_1 = bk$$

with $a, b \in [0, 1]$ and $k_1 \leq k_2$. We define the splitting move by generating a, b as uniform random numbers in $[0, 1]$. We obtain

$$k_2 = \frac{1-ab}{1-a}k$$

In order to calculate the proposal probability of the splitting move consider the random number distributions

$$\begin{aligned} p(a) &= 1 \quad a\in[0,1] \\ p(b) &= 1 \quad b\in[0,1] \end{aligned}$$

we transform the variables (a, b) into (k_1, k_2):

$$\begin{aligned} a &= \frac{k-k_2}{k_1-k_2} \\ b &= \frac{k_1}{k} \end{aligned}$$

This involves the Jacobian:

$$\begin{aligned} \mid J \mid &= \left| \det \begin{pmatrix} \frac{d}{dk_1}a & \frac{d}{dk_2}a \\ \frac{d}{dk_1}b & \frac{d}{dk_2}b \end{pmatrix} \right| \\ &= \frac{k-k_1}{(k_1-k_2)^2}\frac{1}{k} \end{aligned}$$

This yields the splitting proposal density

$$p(k \to a, k_1, k_2) \propto \frac{k-k_1}{(k_1-k_2)^2}\frac{1}{k} = \frac{1-b}{(k_1-k_2)^2} \quad k_1 \leq k$$

while the merging proposal density is given by:

$$p(a, k_1, k_2 \to k) = 1$$

yielding the splitting acceptance probability:

$$p_{acc}^{\text{split}} = \frac{\mathbb{P}_2(a, k_1, k_2 \mid t_1, ..., t_n)}{\mathbb{P}_1(k \mid t_1, ..., t_n)} \frac{(k_1-k_2)^2}{(1-b)}$$

and the merging acceptance probability:

$$p_{acc}^{\text{merge}} = \frac{\mathbb{P}_1(k \mid t_1, ..., t_n)}{\mathbb{P}_2(a, k_1, k_2 \mid t_1, ..., t_n)} \frac{(1-b)}{(k_1-k_2)^2}$$

For the MCMC steps that change the parameters within a given model, we consider the straightforward and the uniform move $a \sim U(0,1)$, and the rate scaling move:

$$k' = (c+0.5)k = kc + 0.5k$$

with random number $c \sim U(0,1)$. The proposal densities are

$$\begin{aligned} p(k \to k') &= k \\ p(k' \to k) &= k' \end{aligned}$$

giving rise to the ratio:

$$\frac{p(k' \to k)}{p(k \to k')} = \frac{k'}{k} = c + 0.5\,.$$

This results in the sampling algorithm 5. When Algorithm 5 returns $n > 1.5$, the corresponding hidden state should be split.

2.5 Performing a state splitting

Without restriction of generality we consider that the n'th state will be split, and we generate a new set of state parameters for the childs $(n, n+1)$ as described below. The new parameters will serve as an input to an EM algorithm, in which the new full parameter set is optimized to convergence.

We find a separation of the nth exit time distribution in terms of

$$p(\tau) = a\mathrm{e}^{-k_n\tau} + (1-a)\mathrm{e}^{-k_{n+1}\tau}\,.$$

Given a transition element Q_{ii} we have the relationship

$$k_i = -\ln Q_{i,i}$$

suggesting diagonal matrix elements

$$Q_{i,i} = \mathrm{e}^{-k_i}\,.$$

We start with matrix $\mathbf{T}$ and stationary distribution $\boldsymbol{\pi}$, for which the corresponding correlation matrix is defined as

$$\mathbf{C} = \boldsymbol{\Pi}\mathbf{T}$$

with $\boldsymbol{\Pi} = \mathrm{diag}(\boldsymbol{\pi})$. Furthermore, let us assume we want to split the last state, n, such that the new states have diagonal elements given by

$$\begin{aligned} Q_{n,n} &= \mathrm{e}^{-k_n} \\ Q_{n+1,n+1} &= \mathrm{e}^{-k_{n+1}}\,, \end{aligned}$$

and the relative probabilities are given by

$$\frac{\pi_n}{\pi_{n+1}} = \frac{a}{1-a}$$

An new correlation matrix $\mathbf{D}$ with a modified state n and and additional state $n+1$ is obtained from the original matrix $\mathbf{C}$ as

$$\mathbf{D} = \begin{pmatrix} C_{11} & \cdots & C_{1,n-1} & D_{1,n} & D_{1,n+1} \\ \vdots & & \vdots & \vdots & \\ C_{n-1,1} & \cdots & C_{n-1,n-1} & \vdots & \vdots \\ D_{n,1} & \cdots & \cdots & D_{n,n} & \\ D_{n+1,1} & & \cdots & & D_{n+1,n+1} \end{pmatrix}$$

with the constraints:

1. $Q_{n,n} = \frac{D_{n,n}}{\sum_i D_{n,i}} = \mathrm{e}^{-k_n}$
2. $Q_{n+1,n+1} = \frac{D_{n+1,n+1}}{\sum_i D_{n+1,i}} = \mathrm{e}^{-k_{n+1}}$
3. $\frac{\pi'_n}{\pi'_{n+1}} = \frac{\sum_i D_{n,i}}{\sum_i D_{n+1,i}} = \frac{a}{1-a}$
4. $\pi'_n + \pi'_{n+1} = \pi_n$
5. $D_{ij} = D_{ji} \quad \forall i$

We intend to fulfill all these contraints, and additionally staying "close" to $D_{n,i} + D_{n+1,i} = C_{n,i} \quad i \in \{1, n-1\}$. From 3 and 4 we get the row sums:

$$\begin{aligned} \pi'_{n+1} &= \pi_n - \pi'_n \\ \pi'_n &= \pi'_{n+1}\frac{a}{1-a} \\ &= (\pi_n - \pi'_n)\frac{a}{1-a} \\ \pi'_n &= \pi_n \frac{\frac{a}{1-a}}{(1+\frac{a}{1-a})} = \pi_n a \end{aligned}$$

From the row sums and constraints 1 and 2 we get the diagonals:

$$\begin{aligned} D_{n,n} &= \pi'_n \mathrm{e}^{-k_n} = d_1 \\ D_{n+1,n+1} &= \pi'_{n+1}\mathrm{e}^{-k_{n+1}} = d_2 \end{aligned}$$

Algorithm 5: Number_of_exponentials$(t_1, ..., t_n)$

Input: A set of lifetimes $t_1, ..., t_n$

Output: $(n,\ k, a, k_1, k_2)$ where $n \in [1, 2]$ is the estimated number of exponentials required to fit the data, k is the rate parameter of the single-exponential model, and a, k_1, k_2 are the parameters of the bi-exponential model.

1. $n_{exp} = 1$, $n_1 = 0$, $k_{1,sum} = 0$, $a_{sum} = 0$, $k_{21,sum} = 0$, $k_{22,sum} = 0$

2. For $i = 1, ..., N_{sample}$

2.1. $r_1 \sim U(0, 1)$

2.2. If $n_{exp} = 1$

If $r_1 < 0.5$

Propose rate change $k \to k' = (r_2 + 0.5)k$ with $r_2 \sim U(0, 1)$. Accept with

$$p^k_{acc} = (r_2 + 0.5)\frac{\mathbb{P}_1(k' \mid t_1, ..., t_n)}{\mathbb{P}_1(k \mid t_1, ..., t_n)}$$

If $r_1 \geq 0.5$

$a \sim U(0, 1)$, $b \sim U(0, 1)$. Propose split $k \to \left(a,\ k_1 = bk,\ k_2 = \frac{1-ab}{1-a}k\right)$. Accept with:

$$p^{\text{split}}_{acc} = \frac{\mathbb{P}_2(a, k_1, k_2 \mid t_1, ..., t_n)}{\mathbb{P}_1(k \mid t_1, ..., t_n)} \frac{(k_1 - k_2)^2}{(1 - b)}$$

Else if $n_{exp} = 2$

If $r_1 < 0.5$

$r_3 \sim U(0, 1)$

If $r_3 < \frac{1}{3}$: Propose new amplitude $a' \sim U(0, 1)$. Accept with

$$p_{acc} = \frac{\mathbb{P}_2(a', k_1, k_2 \mid t_1, ..., t_n)}{\mathbb{P}_2(a, k_1, k_2 \mid t_1, ..., t_n)}$$

Else: propose rate change $k_{1/2} \to k'_{1,2} = (r_2 + 0.5)k_{1/2}$ with $r_2 \sim U(0, 1)$. Accept with

$$p^k_{acc} = (r_2 + 0.5)\frac{\mathbb{P}_1(k'_{1,2} \mid t_1, ..., t_n)}{\mathbb{P}_1(k_{1,2} \mid t_1, ..., t_n)}$$

If $r_1 \geq 0.5$

Propose merge $(a, k_1, k_2 \to k = ak_1 + (1 - a)k_2)$. Accept with:

$$p^{\text{merge}}_{acc} = \frac{\mathbb{P}_1(k \mid t_1, ..., t_n)}{\mathbb{P}_2(a, k_1, k_2 \mid t_1, ..., t_n)} \frac{(1 - b)}{(k_1 - k_2)^2}$$

3. Return:

$$\begin{aligned} n &= 1 + n_1/N_{sample} \\ k &= k_{1,sum}/N_{sample} \\ a &= a_{1,sum}/N_{sample} \\ k_1 &= k_{21,sum}/N_{sample} \\ k_2 &= k_{22,sum}/N_{sample} \end{aligned}$$

Next we fill the lower right block. In the ideal case we can maintain counts of the split diagonal elements.

$$\begin{aligned} C_{nn} &= D_{n,n} + 2D_{n+1,n} + D_{n+1,n+1} \\ D_{n+1,n} &= \frac{C_{nn} - D_{n,n} - D_{n+1,n+1}}{2} \end{aligned}$$

However, if that would results in $D_{n+1,n} < 0$ or $D_{n+1,n} \geq \min\{\pi'_n - D_{n,n}, \pi'_{n+1} - D_{n+1,n+1}\}$ this solution can't be used, and instead we resort to some number

$$D_{n+1,n} = \epsilon < \min\{\pi'_n - D_{n,n}, \pi'_{n+1} - D_{n+1,n+1}\}$$

We now have remaining counts to be distributed:

$$\begin{aligned} r_n &= \pi'_n - D_{n,n} - D_{n+1,n} \\ r_{n+1} &= \pi'_{n+1} - D_{n+1,n+1} - D_{n+1,n} \end{aligned}$$

and set the remaining elements as:

$$\begin{aligned} D_{n,i} &= r_n \frac{C_{n,i}}{\pi_n} \\ D_{n+1,i} &= r_{n+1} \frac{C_{n,i}}{\pi_n} \end{aligned}$$

Finally, we normalize $\mathbf{D}$ row-wise to get a transition matrix.

The new state parameters are used as an input for an EM algorithm, in which the HMM parameters are again optimized to convergence.

2.6 Lumping together of rapidly interconverting states

For the sake of clarity, models with many states were coarse-grained by lumping together rapidly interconverting states with similar FRET efficiencies. This amounts to constructing a coarse-grained transition matrix $\mathbf{T}_c$ from the original unconditional transition probability matrix, also called correlation matrix, which is given by

$$\mathbf{C} = \mathbf{\Pi T}\,.$$

$\mathbf{\Pi} = \mathrm{diag}(\boldsymbol{\pi})$ is a square matrix which contains the stationary probabilities of states on the diagonal. A coarse-grained correlation matrix is then obtained by

$$\mathbf{C}_c = \boldsymbol{\chi}\mathbf{C}\boldsymbol{\chi}\,,$$

where $\boldsymbol{\chi} \in \mathbb{R}^{m\times n}$ is a coarse-graining operator, which has column-sums of 1, and each element i, j indicates to which fraction state j is lumped into coarse state i. As described in,[11] the correct way of computing $\boldsymbol{\chi}$ is via the PCCA+ method.[12] However, when the dynamics is metastable, rapidly-interconverting substates can be merged by a simple $\boldsymbol{\chi}$-matrix, whose rows simply contain a value which xequals 1 for all states to be lumped into the corresponding macrostate. The application of this simple $\boldsymbol{\chi}$-matrix is then identical to summing up the unconditional probabilities in $\mathbf{C}$ over the joined groups of states. The resulting coarse-grained correlation matrix is turned into a transition matrix by normalizing its rows:

$$\mathbf{T}_c = \mathbf{\Pi}_c^{-1}\mathbf{C}_c\,.$$

3 Post-processing and analysis methods

3.1 Correcting the spectral crosstalk χ

The states of a HMM are characterized by apparent FRET efficiencies

$$E^{app} = \frac{k_a^{app}}{k_{tot}} = \frac{k_a^{true} + \chi k_d^{true}}{k_{tot}}\,, \qquad (21)$$

where, due to spectral crosstalk, the apparent photon count rate in the acceptor channel k_a^{app} is higher than the true photon count rate, k_a^{true}, emitted by the acceptor chromophore. χ is the spectral crosstalk factor estimated according to eq. 27 in the main text, and k_d^{true} is the (true) photon count rate emitted by the donor chromophore, which is related to the true FRET efficiency E^{true} by

$$E^{true} = \frac{k_a^{true}}{k_a^{true} + k_d^{true}} \quad \Leftrightarrow \quad k_d^{true} = \frac{1 - E^{true}}{E^{true}} k_a^{true} \qquad (22)$$

Note that the total photon count rate is not affected by the spectral crosstalk and we have $k_{tot} = k_a^{true} + k_d^{true}$. Inserting eq. 22 in eq. 21 and solving for E^{true} yields the crosstalk corrected FRET efficiency

$$E^{true} = \frac{E^{app} - \chi}{1 - \chi}\,. \qquad (23)$$

The FRET efficiencies of the states in Fig. 2a and 3b in the main text, in Figs. S10-S13, have been corrected for spectral crosstalk using eq. 23.

3.2 Free energy interpretation of HMM states and transition states

Free energy differences of HMM states (with respect to an arbitrary reference state R) are computed as

$$\Delta G_{iR} = -k_B T \ln\left[\frac{\pi_i}{\pi_R}\right]\,,$$

where π_i is the stationary probability of state i, and π_R is the stationary probability of the reference state. Both are obtained from the first eigenvector of the HMM transition matrix $\mathbf{T}$. k_B is the Boltzmann constant, and T is the absolute temperature, here set to 298 K. Higher eigenvectors of $\mathbf{T}$ indicate transitions between regions of the conformational space, associated to rates $\kappa_i^{\mathrm{HMM}} = -\Delta t^{-1} \ln \lambda_i^{\mathrm{HMM}}$, where λ_i^{HMM} are the eigenvalues of the HMM transition matrix $\mathbf{T}$.

Up to a constant, transition state energies and transition rates are related by: $G_i^{\ddagger} = -k_B T \ln \kappa_i$. Alternatively, one can calculate the transition state energies between pairs of hidden states. Using

$$\pi_i \frac{T_{ij}}{\tau} \approx \pi_i k_{ij} = \pi_{ij} = Z^{-1} \exp\left(-\frac{1}{k_B T} U_{ij}\right) \quad (24)$$

one obtains

$$U_{ij} \quad = -k_B T \ln \pi_i \tfrac{T_{ij}(\tau)}{\tau} - k_B T \ln Z \,, \quad (25)$$

where $-k_B T \ln Z$ is an unknown but constant offset. Results are shown in Fig. 6a of the main text.

3.3 Eigenvalue decomposition

Let $\mathbf{T}(\tau)$ be the transition matrix estimated from the HMM, where τ is the time step used to parametrize it. In the present study, $\tau = \Delta t = 10\,\mu s$ is the elementary time step at which photons are being recorded. We can perform an eigenvalue decomposition of $\mathbf{T}(\tau)$, and obtain:

$$\mathbf{T}(\tau)\mathbf{\Psi} = \lambda(\tau)\mathbf{\Psi} \,,$$

where

$$\mathbf{\Psi} = [\boldsymbol{\psi}_1, ..., \boldsymbol{\psi}_n]$$

contains the right eigenvectors as columns and $\lambda_1 = 1 > \lambda_2 \geq \lambda_3 \geq ... \geq \lambda_n$ are the eigenvalues. Each eigenvalue is associated with a timescale

$$t_i = -\frac{\tau}{\ln \lambda_i},$$

also called implied timescale[10] or relaxation time. At timescale t_i, conformational transitions occur which are specified by eigenvector $\boldsymbol{\psi}_i$: States with negative sign exchange with states with positive sign. For a detailed interpretation and illustration of eigenvalues and eigenvectors please refer to ref.[8] Figs. S10-S13, show the implied timescales and eigenvectors of the HMMs.

3.4 Transition path analysis

The folding pathways from misfolded to native states are here calculated using Transition Path Theory (TPT). TPT has been originally introduced in[13] and has been formulated for time-continuous Markov jump processes in.[14] Here we use transition path theory for Markov chains as described in,[15] employing the implementation in the EMMA software.[9]

TPT for Markov chains requires the transition matrix $\mathbf{T}(\tau)$, and a definition of the educt states, A, and product states, B, as an input. Here, the long-lived low-FRET-efficiency states (colored green and yellow in Fig. 2a in the main text, which presumably represent misfolded secondary structures, were chosen as A, while the high-efficiency state (colored black in Fig. 2a in the main text) was chosen as B state. From $\mathbf{T}(\tau)$, the stationary distribution $\boldsymbol{\pi}$ is computed as well as the committor probabilities, $\mathbf{q}$, whose elements q_i denote the probability that the dynamics will, starting from state i, go next to the set B rather than to the set A. In other words, q_i is the probability of folding of state i. $\boldsymbol{q}$ can be calculated from $\mathbf{T}(\tau)$ for a given definition of A and B by solving a linear system of equations specified in the supplement of.[15] We order states by increasing value of the commitor, and calculate the folding flux from state i to state j (when $q_i \leq q_j$) by

$$f_{ij}^{+} = \pi_i (q_j - q_i) T_{ij}(\tau) \,. \quad (26)$$

The folding flux is a rate in units of τ, and specifies a directed network whose edges quantify the flux of folding trajectories from A to B. The folding fluxes for the present HMMs are shown in Fig. 3b in the main text.

4 Diels-Alderase measurement and analysis protocol

4.1 Single-molecule FRET experiments and data processing

Single-molecule fluorescence data were obtained as described in Ref.[16] Briefly, DAse ribozyme molecules with covalently attached two fluorophores (Cy3 as donor and Cy5 as acceptor) and biotin were immobilized on a biocompatible surface that minimally affects RNA folding. The sample chamber was filled with buffer containing 50 mM Tris, 300 mM NaCl, and $MgCl_2$ in varying amounts. Additionally, an enzymatic oxygen scavenging system and triplet quencher were employed to extend the observation time. Fluorescence emission was measured by using a home-built laser scanning confocal microscope with two-channel detection. To ensure the presence of a functional donor and acceptor pair on the measured RNAs, confocal fluorescence images of sparsely immobilized molecules were first taken under red excitation. Subsequently, selected spots were - one by one - moved into the focus, and the fluorescent signal was recorded under green excitation. Single-molecule trajectories of the number of recorded donor and acceptor photons were stored at a time resolution of 10 μs. With a mean photon rate of 2 - 3 ms^{-1}, this protocol ensures acquisition of single-photon data. Only those trajectories containing a single-step photobleaching of first the acceptor and afterwards the donor were chosen for further analysis. The traces were split into three regions: (i) production phase (both chromophores active), (ii) acceptor bleached phase, and (iii) donor+acceptor bleached phase. The combined data set of all production phases (of a given Mg^{2+} concentrations) was subjected to HMM analysis without further modification. The total production time amounted to more than 230 s, 400 s and 220 s for 0, 5, and 40 mM Mg^{2+}, respectively. The smFRET data sets are summarized

in Table S1, and histograms of the length of the photon traces are displayed in Fig. S2. Fig. S4 shows that the statistics of the photon arrival times in the data set follows a Poisson process.

The rates of the background noise $k_{a,bg}$ an $k_{d,bg}$ in each channel were estimated from the donor+acceptor bleached phase of this traces as $k_{a,bg} = n_{a,bg}/T_{bg}$ and $k_{d,bg} = n_{d,bg}/T_{bg}$. $n_{a,bg}$ ($n_{d,bg}$) is the number of accpetor (donor) photons detected during this phase and T_{bg} is the duration of this phase.

The amount of spectral crosstalk χ from the donor to the acceptor channel was estimated as

$$\chi = \frac{\langle k_{a,\chi}\rangle - \langle k_{a,bg}\rangle}{\langle k_{a,\chi}\rangle - \langle k_{a,bg}\rangle + \langle k_{d,\chi}\rangle - \langle k_{d,bg}\rangle}\,, \qquad (27)$$

where $\langle k_{a,\chi}\rangle$ ($\langle k_{d,\chi}\rangle$) is the average acceptor (donor) photon count rate in the acceptor bleached phases, $\langle k_{a,bg}\rangle$ ($\langle k_{d,bg}\rangle$) is the average acceptor (donor) photon count rate in the donor+acceptor bleached phases of a given data set.

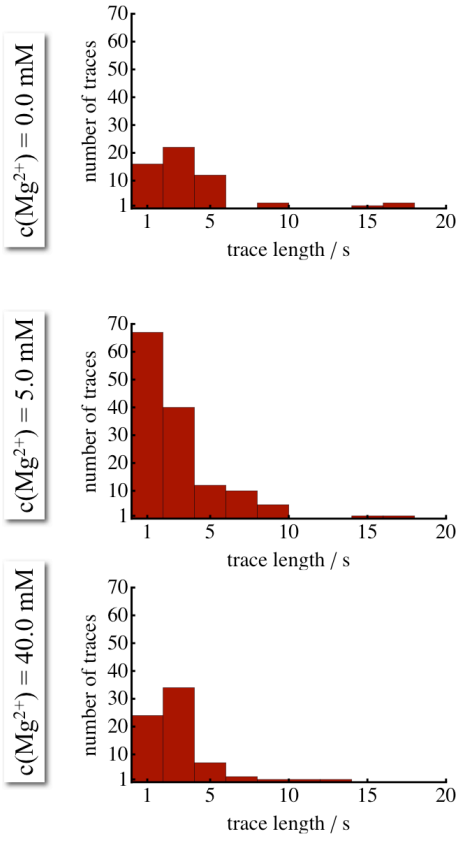

Figure S2: Histogram of the trace lengths in the three data sets from which the HMMs were constructed and which is further summarized in Tab. S1.

Table S1: Photon statistics of the production phase of the data sets in construct I

$c(Mg^{2+})$	$0.0\,mM$	$5.0\,mM$	$40.0\,mM$
General information			
time resolution	10 μs	10 μs	10 μs
total time	238.70 s	406.19 s	222.79 s
# traces	56	136	70
Number of photons			
don. channel	233'661	376'892	169'734
acc. channel	414'787	729'852	469'428
total	648'448	1'106'744	639'162
Average photon count rate / ms^{-1}			
don. channel	0.98	0.93	0.76
acc. channel	1.74	1.80	2.11
total	2.72	2.72	2.87

Table S2: Photon statistics of the production phase of the data sets in construct II

$c(Mg^{2+})$	$5.0\,mM$
General information	
time resolution	10 μs
total time	739 s
number of traces	53
Number of photons	
don. channel	592'890
acc. channel	2'567'705
total	3'160'595
Average photon count rate / ms^{-1}	
don. channel	0.80
acc. channel	3.47
total	4.27

4.2 Effects of surface Immobilization

To evaluate the effect of surface immobilization on the folding properties of the DAse ribozyme, we have performed ensemble fluorescence measurements of ribozyme molecules at different Mg^{2+} ion concentrations. The fluorescence spectra under 532 nm excitation were fitted with the known spectra of the Cy3 (donor) and Cy5 (acceptor) fluorophores, and the folded fraction was calculated as a function of Mg^{2+} concentration, based on their normalized spectral areas (shown in Fig. S3a). Fitting of the calculated data with the Hill equation (Eq. 5 in Ref.[16]) yields a transition midpoint concentration of 2.5 mM and a cooperativity parameter n = 2. From the surface-immobilized single-molecule data, we obtain very similar data, i.e., a midpoint concentration of 6.1 mM and n = 1.8 for the intermediate-to-folded state transition. Additionally, we have observed compaction of the intermediate state in the single-molecule data causing an increase of the average FRET efficiency, character-

ized by a transition midpoint at 3.8 mM and a cooperativity parameter of 4.2 (Fig. 8 in Ref.[16]). Thus, surface immobilization affects the folding properties of the studied DAse ribozyme only very weakly.

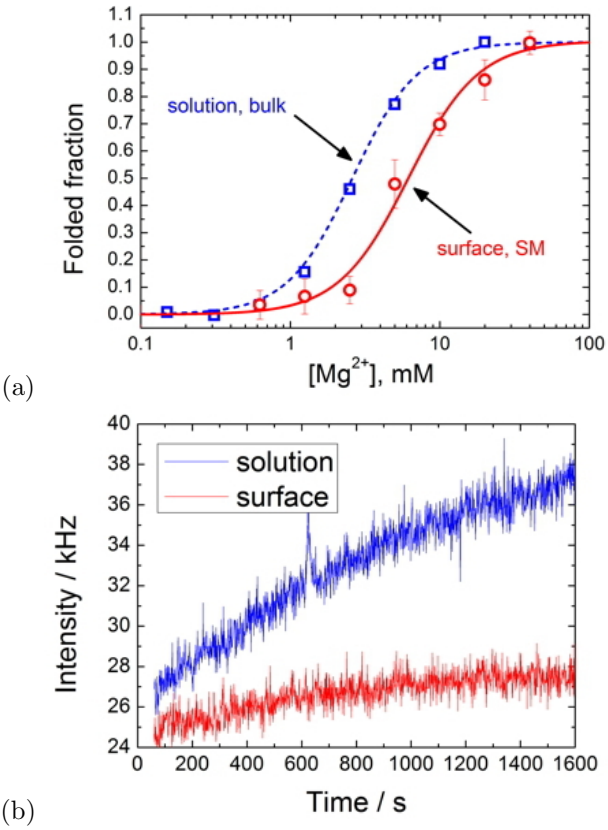

Figure S3: Effects of the surface immobilization on thermodynamics and kinetics of Diels-Alderase. (a) Folded fraction of Diels-Alderase as a function of Mg^{2+} ion concentration. *Blue squares:* ensemble fluorescence measurements. *Red circles:* surface-immobilized single-molecule measurements. *Lines:* Hill equation fitted to the respective data set. (b) Reaction progress curves monitored by a fluorescent sensor (anthracene-1-yl-BODIPY (1-AB)) for the biocatalytic Diels-Alder reaction. *Blue line:* Diels-Alderase in solution. *Red line:* Diels-Alderase molecules immobilized on BSA surface.

Furthermore, to estimate the influence of the DAse ribozyme immobilization on the reaction rate, we compared the reaction progress curves in solution and molecules immobilized on BSA surfaces under conditions as close as feasible to single molecule FRET measurements. To monitor the catalytic reaction, we have utilized the anthracene-1-yl-BODIPY (1-AB) substrate as a fluorescent sensor for the biocatalytic Diels-Alder reaction.[22] In these experiments, a variant of DAse ribozyme containing only a biotin linker has been used, but no fluorescent dyes. The results are reported in Fig. S3b. By mixing 200 nM of ribozyme with 100 nM of 1-AB and 10 μM of NPM, a clear increase in the fluorescence signal can be observed that indicates product formation. For surface immobilization, a solution containing 100 nM of DAse ribozyme was added to the sample chamber coated with biotinylated BSA and streptavidin, incubated for 10 min and then rinsed with buffer to remove all unbound molecules. Afterward, a solution containing 100 nM of 1-AB and 10 μM of NPM was added to the sample chamber, and the reaction progress was monitored. We note that the exact number of molecules attached to the surface is not known. The expected coverage is less than a monolayer but larger than what we use in single-molecule preparations, where we incubate with a concentration of 10-50 pM. Anyhow, a notable increase in fluorescence can be detected. A precise quantitative comparison of the experimental data presented below, that is, the extraction of rate coefficients for free molecules in solution and immobilized ones, is not feasible. However, the data qualitatively confirm the catalytic activity of immobilized molecules.

4.3 Statistics of the photon arrival times

The likelihood in eqs. 4 and 5 is based on the assumption that the arrival times of the photons at the detector follow a Poisson process. Whether this assumption is valid for a given trace can be tested by plotting the histogram of interphoton times, which is expected to show an exponential decay

$$p(t) \quad = \quad N \exp\left(-nt\right) . \tag{28}$$

n is the total photon count rate in this trace estimated as $N_{photons}/T$, where $N_{photons}$ is the total number of photons the trace, and T ist the length of the trace The histogram is normalized to the number of interphoton times.

$$N = \frac{N_{photons} - 1}{\int_{t=0}^{T} \exp^{-nt}\, dt} . \tag{29}$$

We mostly find that the photon arrival times are very well modeled by a Poisson process (Fig. S4a). In rare cases, the photon arrival times follow a Poisson process with a slightly different rate than the one that is expected from the total photon count rate (Fig. S4b).

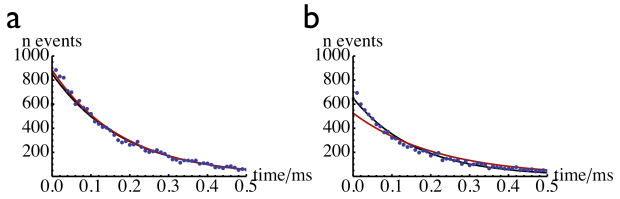

Figure S4: Histogram of interphoton times in example traces. *blue dots:* Estimated from the trace; *black line:* exponential fit; *red line:* exponential decay function (eq. 28) as expected from total photon count rate of the trace.

4.4 Information based criteria

It is common practice to determine the number of states of the hidden Markov model by using information based indicators, such as the Bayesian information criterion (BIC) or the Akaike information criterion (AIC). These indicators are designed to determine the optimal number of states in HMMs under certain assumptions. The most prominent assumption is that the hidden process is an an inherently discrete process with a fixed number of states. As discussed in the previous paragraph and in the main part of the manuscript, this is typically not the case for molecular dynamics. Rather, the hidden conformational changes are governed by a hierarchy of states and processes.

Second, information based criteria are valid only asymptotically, i.e. in the limit of infinite sampling. They are therefore of little use for finite data sets because the associated likelihood function does not have a single dominant maximum, but multiple local maxima. The Baum-Welch algorithm, however, only identifies the nearest local maximum, and as a consequence, the results of the HMM optimization for a finite data set vary depending on the initial parameters. If this variation for a given number of hidden states is larger than the change of the information criterion of choice over a certain range of numbers of hidden states, this criterion has hardly any significance.

This is illustrated in Fig. S5, which shows the BIC as a function of states used for the HMM optimization on the data set with c(Mg^{2+})=0.0 mM. For each number of states n, the HMM optimization has been repeated 20 times with varied initial parameters. This results in a range of models for a given n, each of which is represented by a dot in Fig. S5. The model with the highest log-likelihood at $n = 8$ indeed corresponds to the maximum in the BIC curve. However, the difference to the models with one state more or less is only on the order of 10 ($n = 7$: $-5.16388 \cdot 10^6$; $n = 8$: $-5.16383 \cdot 10^6$; $n = 7$: $-5.16386 \cdot 10^6$), whereas the differences between different minima of the HMM likelihood using the same number of states can be on the order of 10^3. These results also show that, due to the large variation in the HMM results for a given value of n, verifying the reproducibility of the model with largest log-likelihood is indispensable.

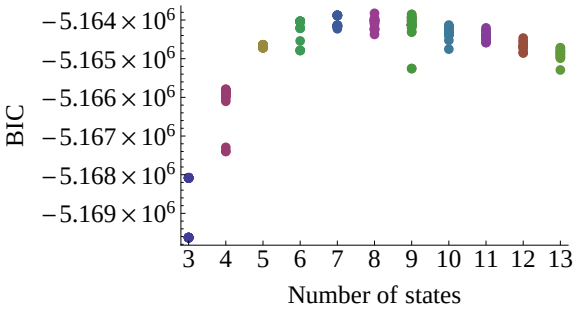

Figure S5: Bayesian information criterion as a function of hidden states. Each dot represents an independent HMM optimization of the data set at c(Mg^{2+})=0.0mM.

4.5 HMM generation, validation, and refinement

HMMs were constructed for the smFRET data sets summarized in Tables S1. and S2 Using the optimization protocol described in the Methods section, the largest number of states for which HMM's could be reproducibly obtained were 8, 8 and 7 at Mg^{2+} concentrations of 0, 5 and 40 mM, respectively.

Fig. 4 in the main text compares the predicted time-dependent FRET efficiency distributions of these HMM's with the data. It is apparent, that the HMMs for 0 and 5 mM Mg^{2+} predict time-dependent FRET efficiency distributions well, whereas the prediction of the 7-state HMM at 40 mM Mg^{2+} significantly disagrees with the data.

We assumed that the reason for the 7-state HMM at 40 mM Mg^{2+} failing the validation test was that it contained some states with non-Markovian behavior that should be splitted into substates. To confirm this, we computed the maximum likelihood hidden path, $\hat{s}$, for all Markov models and inspect the lifetime distributions of states, which is expected to be single-exponential if the dynamics is Markovian. Indeed, the maximum likelihood paths for 0 and 5 mM Mg^{2+} show a nearly perfect Markovian behavior (Fig. 4b in the main text), while the 7-state HMM at 40 mM Mg^{2+} contained three clearly non-exponential states (Fig. S8b). Thus, the non-exponential states of the 7-state model were split using the protocol described in Sections "State splitting probability" and "Performing a state splitting". With two iterations of splitting, a 13-state HMM for 40 mM Mg^{2+} was obtained which still exhibited nonexponential states (Fig. 4b in the main text), but now passed the validation test using the FRET efficiency distributions (Fig. 4a in the main text). To test whether the remaining non-exponentiality came from an actual non-Markovianity

of the discrete state dynamics or just from spurious transitions generated from the estimation of the maximum likelihood, we conducted the Markov tests as described in Sec. "Implied timescale test". The results in Fig. 4c in the main text show that the implied timescales of the maximum likelihood hidden paths, $\hat{s}(t)$ depend on τ for short lagtimes, indicating that these paths are non-Markovian at short timescales. However, at 10-30 ms the implied timescales become approximately constant in τ, and approximately agree with the timescales estimated from the HMM transition matrix. This indicates that the kinetics of all three HMMs, consisting of 7 states (0 and 5 mM Mg^{2+}) and 13 states (40 mM Mg^{2+}) are consistent with the data.

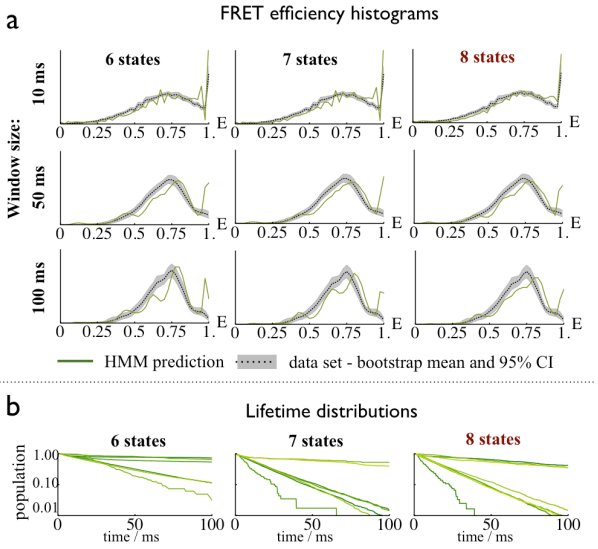

Figure S7: Validation of the hidden Markov models for c(Mg^{2+}) = 5.0 mM. (**a**) Time-window-dependent FRET efficiency histograms at window lengths of 10 ms, 50 ms, and 100 ms. *grey area:* 95% confidence interval of the data set as calculated from a bootstrapping sample, *dotted grey lines:* mean of the bootstrapping sample, *dashed colored lines:* predicted from the hidden Markov model. (**b**) Lifetime distribution of the individual states calculated from the maximum-likelihood paths.

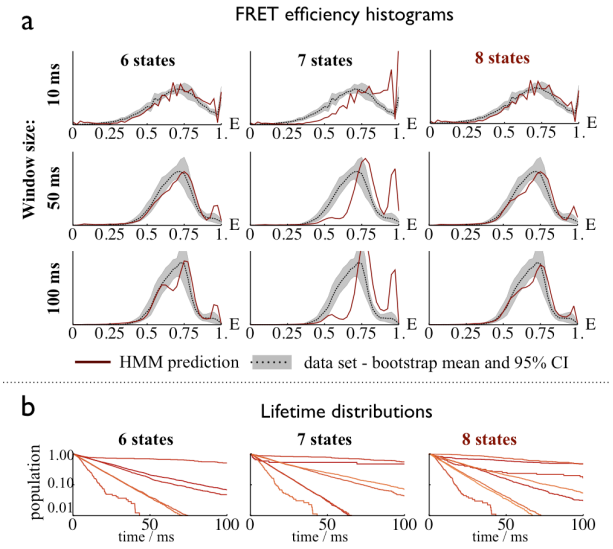

Figure S6: Validation of the hidden Markov models for c(Mg^{2+}) = 0.0 mM. (**a**) Time-window-dependent FRET efficiency histograms at window lengths of 10 ms, 50 ms, and 100 ms. *grey area:* 95% confidence interval of the data set as calculated from a bootstrapping sample, *dotted grey lines:* mean of the bootstrapping sample, *dashed colored lines:* predicted from the hidden Markov model. (**b**) Lifetime distribution of the individual states calculated from the maximum-likelihood paths.

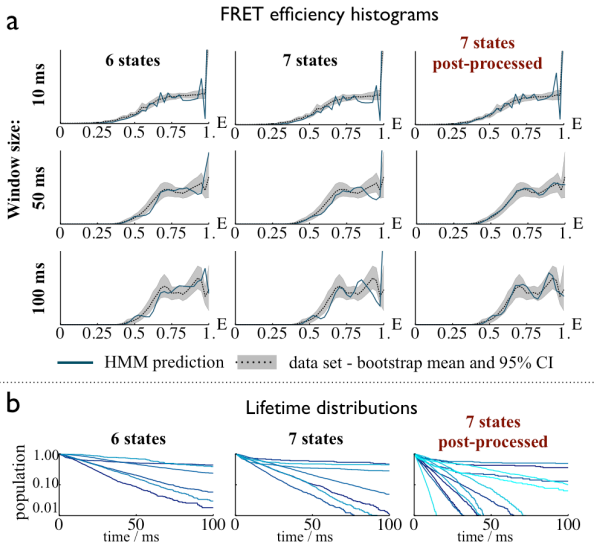

Figure S8: Validation of the hidden Markov models for c(Mg^{2+}) = 40.0 mM. (**a**) Time-window-dependent FRET efficiency histograms at window lengths of 10 ms, 50 ms, and 100 ms. *grey area:* 95% confidence interval of the data set as calculated from a bootstrapping sample, *dotted grey lines:* mean of the bootstrapping sample, *dashed colored lines:* predicted from the hidden Markov model. (**b**) Lifetime distribution of the individual states calculated from the maximum-likelihood paths.

Table S3: Number of states at the stages of the HMM generation protocol. ✓ = validation passed × = validation failed.

construct	I	I	I	II
c(Mg^{2+})	0.0 mM	5.0 mM	40.0 mM	5 mM
HMM optimization	8	8	7	7
HMM validation	✓	✓	×	✓
HMM refinement: splitting of states	7	7	13	7
HMM validation	✓	✓	✓	✓
HMM refinement: merging of states	7	7	9	7
HMM interpretation	7	7	9	7

The 13-state model contained some rapidly-converting states with similar FRET efficiencies. In order to simplify the interpretation, these states were subsequently lumped as described in Sec. "Lumping together of rapidly interconverting states" obtaining the final 9-state model reported in the main manuscript. Fig. S13b shows which states were joined with respect to the 13-state model Table S3 summarizes the numbers of HMM states obtained in the different data processing steps.

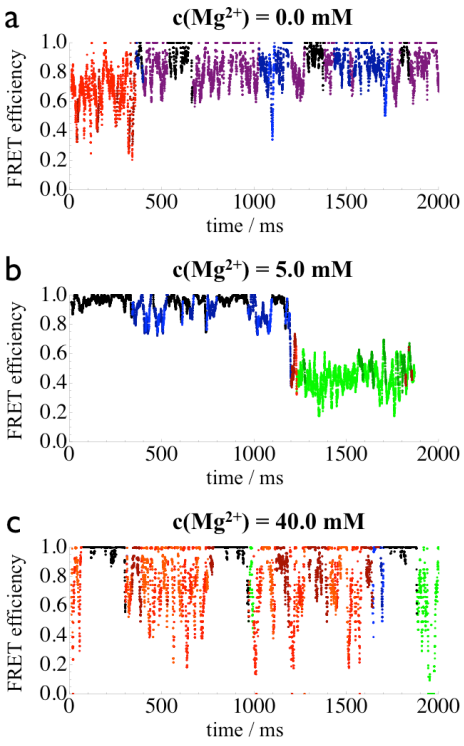

Figure S9: Example of FRET efficiency traces (averaging window 10 ms) from the three data sets colored according to the Viterbi path of corresponding HMM (**a**): c(Mg^{2+}) = 0.0 mM, 7-state model (**b**): c(Mg^{2+}) = 5.0 mM, 7-state model (**c**): c(Mg^{2+}) = 40.0 mM, 13-state model. The coloring of the states is the same as in Fig. 2 in the main text.

4.6 Detailed kinetic profile of the hidden Markov models

Figs. S10-S13 show additional kinetic analyses of the HMMs of Diels-Alderase.

The kinetic network graphs are constructed directly from the transition matrix of the corresponding HMM. The discs represent the hidden states. The colors of the circles corresponds to colors in the state decomposition plots shown in the panel below or in the main text. The edges of the graphs represent the correlation between two states i and j: $c_{ij}(\Delta t) = \pi_i t_{ij}(\Delta t)$, where $\Delta t = 10\,\mu s$ is the lag time of the hidden Markov model, π_i is the equilibrium probability of the ith state, and t_{ij} is the matrix element of the transition matrix representing the transition probability from i to j within time Δt. The thickness of the edge is proportional to $\log_{10}(c_{ij}) - \log_{10}(d)$, where d is the threshold in the respective graph.

The eigenvectors of the transition matrix of a Markov model can be interpreted as kinetic processes. We present the eigenvectors as color-coded projections onto the states of the HMM. Each kinetic process shuttles probability density between regions of the state space to which the corresponding eigenvector assigns positive values (red) to region to which it assign neg-

ative values (blue) (or vice versa depending on the initial probability distribution). States which are assigned a value of zero (white) by a particular eigenvector are not affected by the corresponding kinetic process.

The state decomposition plots shown in Figs. S13 and Fig. S12 are constructed as described in the main text. Discs represent the hidden states and are positioned according to the FRET efficiency and the state lifetime of the corresponding hidden state. The disc area represents the equilibrium probability of the states. Discs of similar color are kinetically close to each other as revealed by a kinetic cluster analysis of the eigenvectors of the transition matrix.[12] States which kinetically merge into larger subensembles at different timescales are shown as colored areas.

The validation test for the 7-state model for 40.0 mM Mg^2+ shown in Fig. S12 is similar to the validation tests shown in Fig. 4 of the main text but reveals that the FRET histograms predicted from the 7-state HMM do not agree with those estimated directly from the data.

4.7 Proposed secondary structure motifs

To obtain a structural interpretation of the HMM states, ten minimum energy secondary structures of the Diels-Alderase were calculated using the barriers server of the Vienna RNA WebServer.[26–28] The temperature was set to 25 °C. Only minima with a barrier higher than 0.1 kcal/mol were distinguished, and the algorithm was forced to avoid isolated base pairs. RNA parameters (Turner model[29]) were used to calculate the free energies. The secondary structures with lowest energy are shown in Fig. 8 of the main text, These settings correspond to the following RNAsubopt/barriers command line:

```
RNAsubopt -d2 --noLP -T 25 -s -e 40 < sequence.fa
> RNAsubopt.out

barriers --rates -G RNA-noLP --max 10 --minh 0.1
<
RNAsubopt.out > barriers.out
```

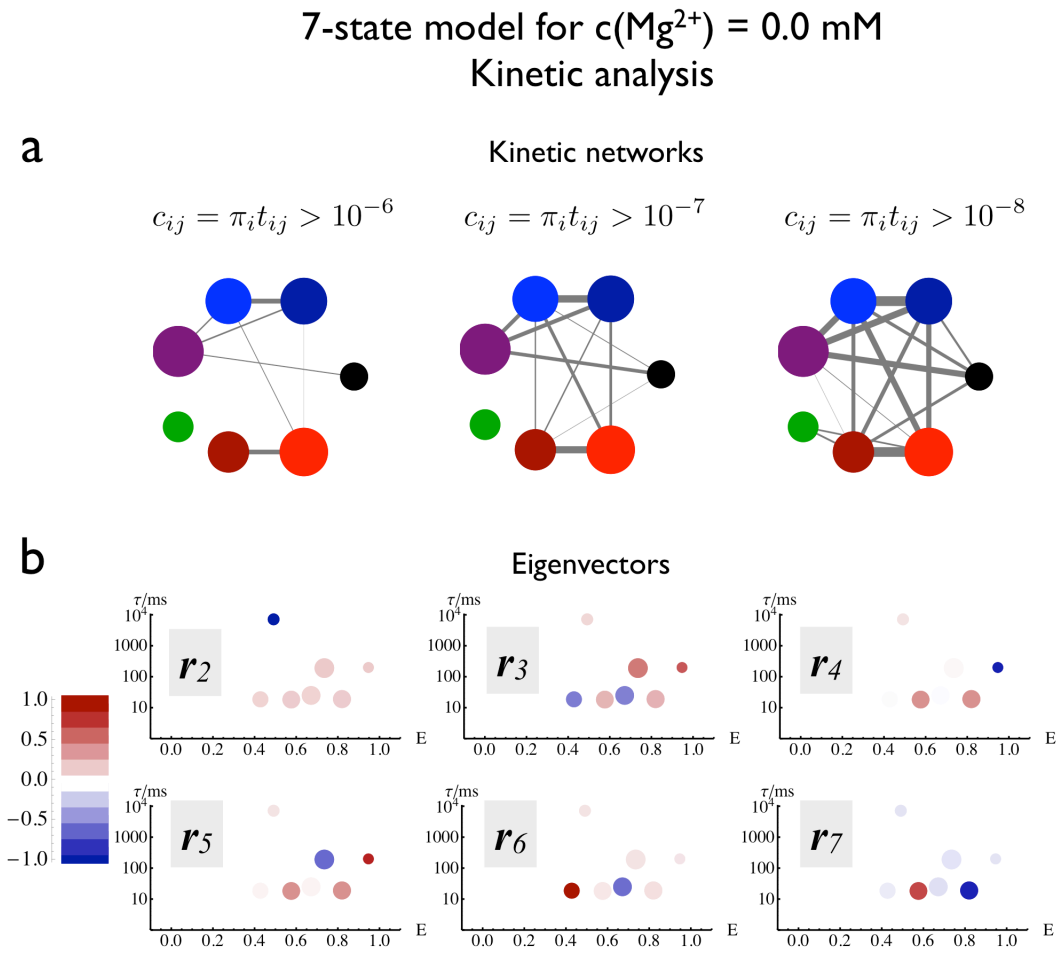

Figure S10: Kinetic analysis of the 7-state hidden Markov model for c(Mg^{2+}) = 0.0 mM obtained by removing the blinking state from the original 8-state model. (**a**) Kinetic networks at different thresholds for the cross-correlations (π-weighted transition probabilities). (**b**) Eigenvectors of the transition matrix of the hidden Markov model projected into the state definition plot.

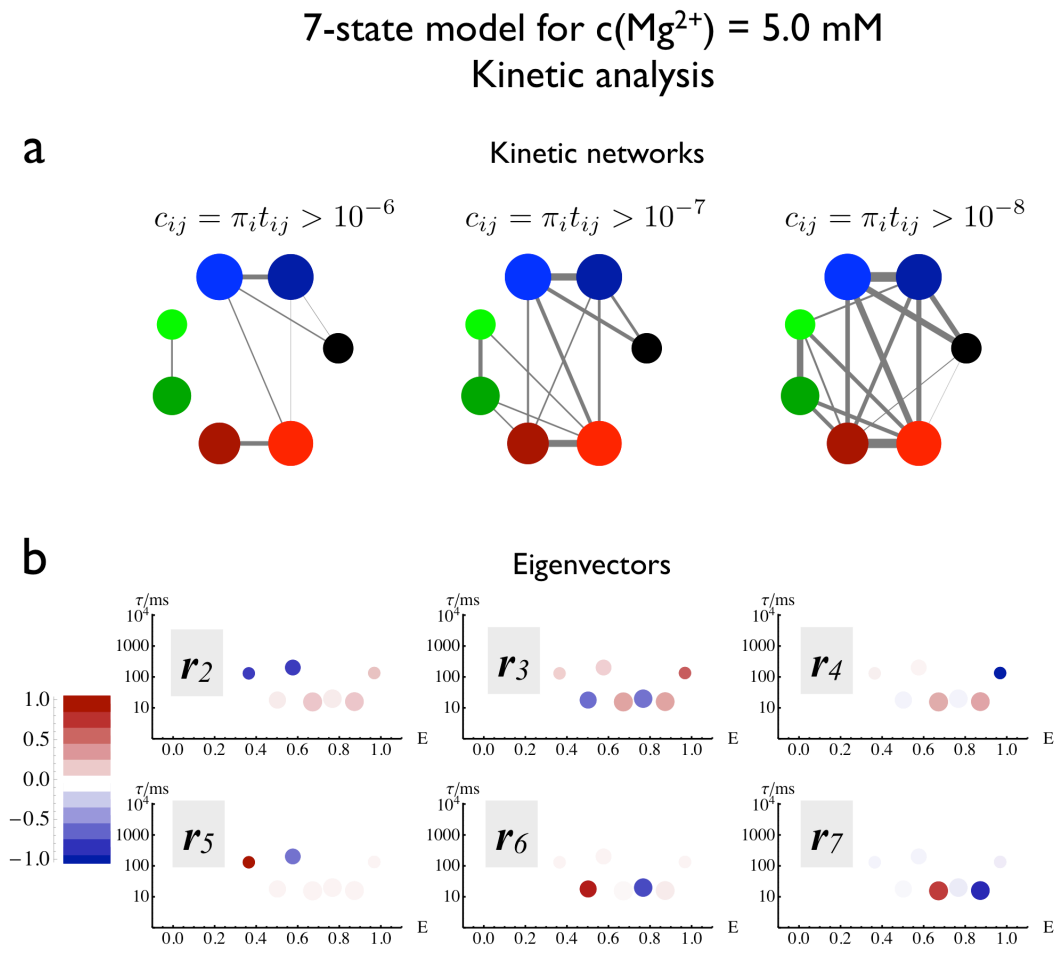

Figure S11: Kinetic analysis of the 7-state hidden Markov model for c(Mg^{2+}) = 5.0 mM obtained by removing the blinking state from the original 8-state model. (**a**) Kinetic networks at different thresholds for the cross-correlations (π-weighted transition probabilities). (**b**) Eigenvectors of the transition matrix of the hidden Markov model projected into the state definition plot.

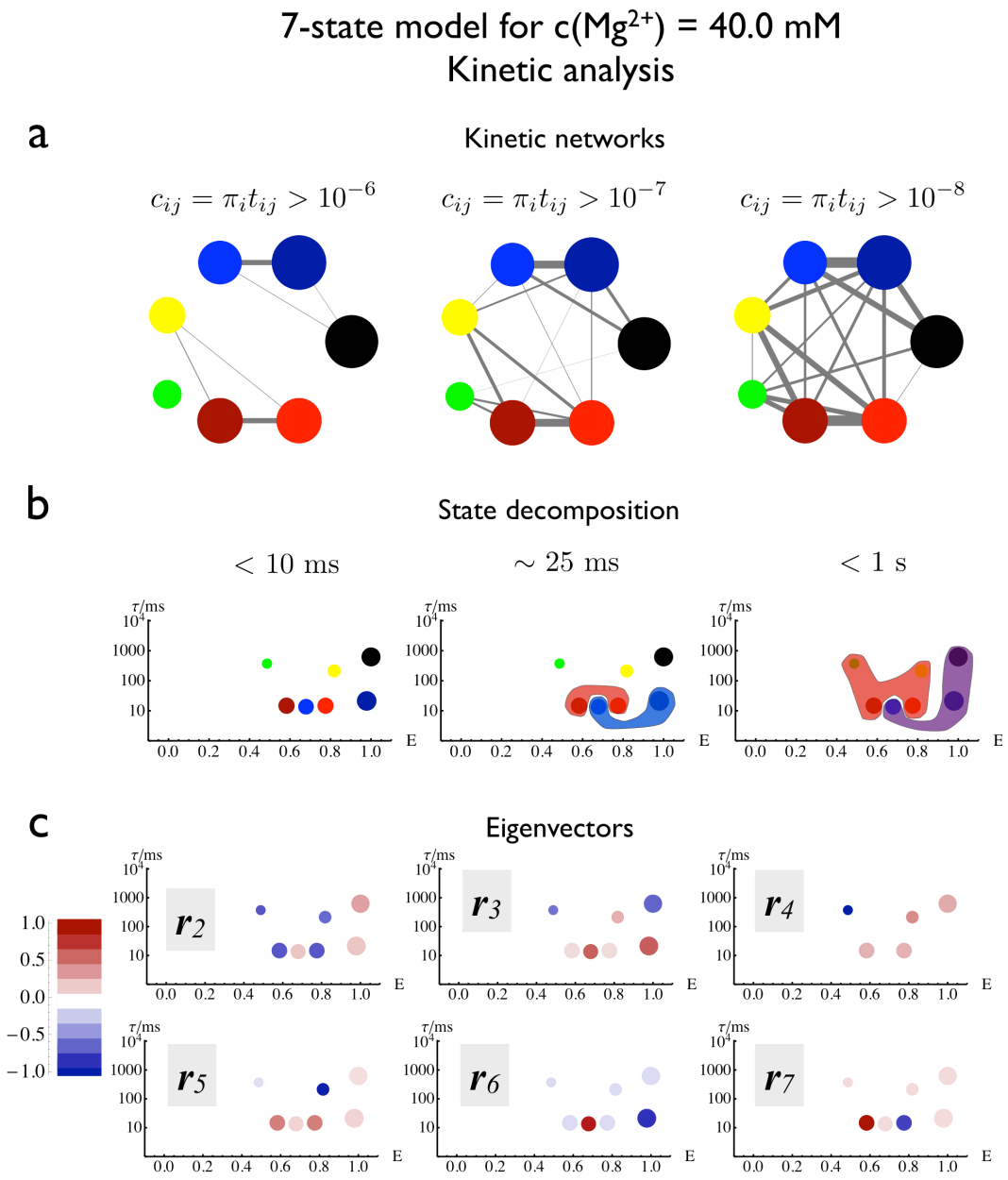

Figure S12: Kinetic analysis of the 7-state hidden Markov model for c(Mg^{2+}) = 40.0 mM. (**a**) Kinetic networks at different thresholds for the cross-correlations (π-weighted transition probabilities). (**b**) State parameters of the hidden Markov model and state decomposition at different timescales. (**c**) Eigenvectors of the transition matrix of the hidden Markov model projected into the state definition plot.

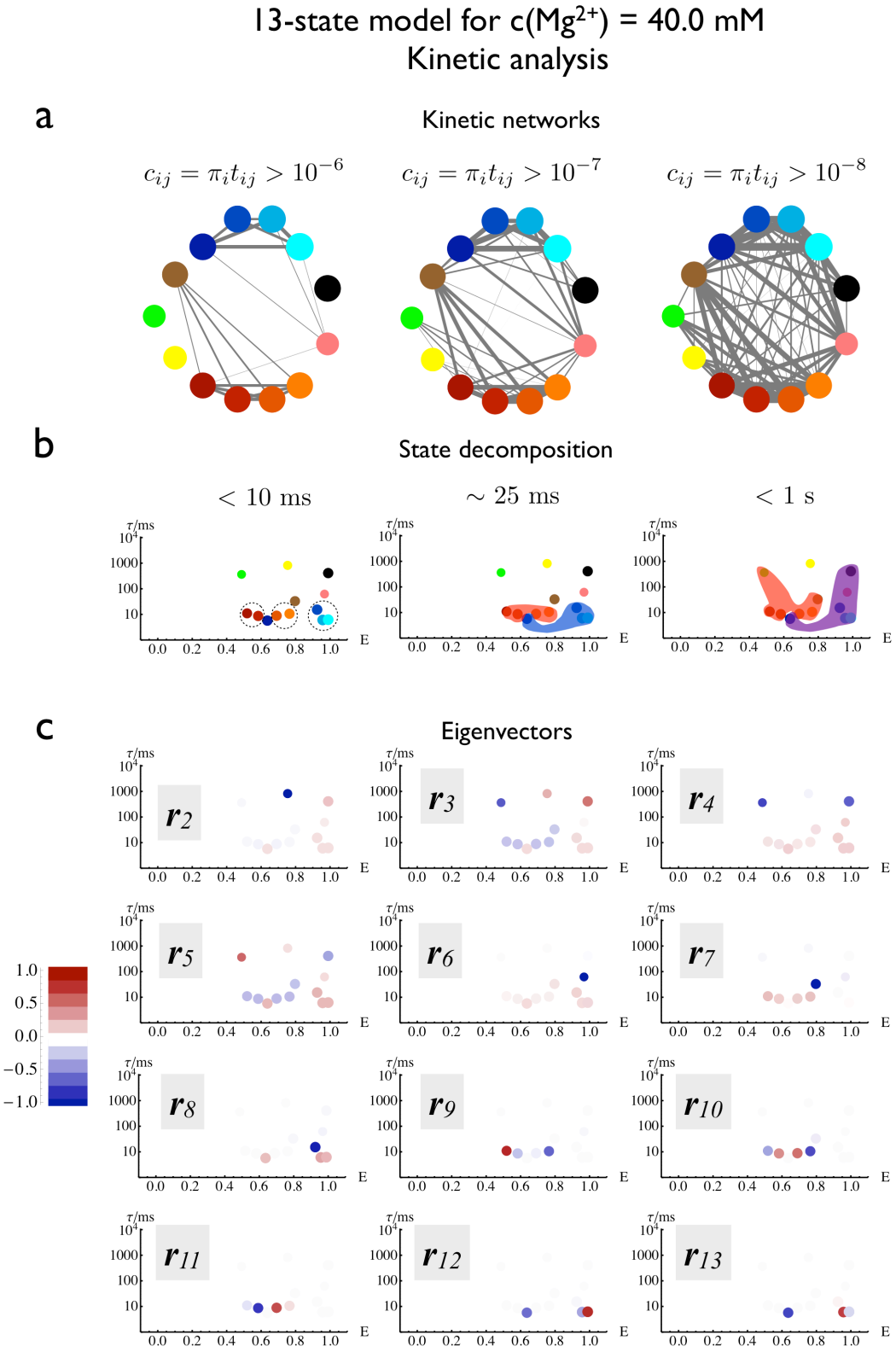

Figure S13: Kinetic analysis of the 13-state hidden Markov model for c(Mg^{2+}) = 4.0 mM. (**a**) Kinetic networks at different thresholds for the cross-correlations (π-weighted transition probabilities). (**b**) State parameters of the hidden Markov model and state decomposition at different timescales. Dashed circles in the first plot show which states are merged to obtain the 9-state model which is discussed in the main text. (**c**) Eigenvectors of the transition matrix of the hidden Markov model projected into the state definition plot.

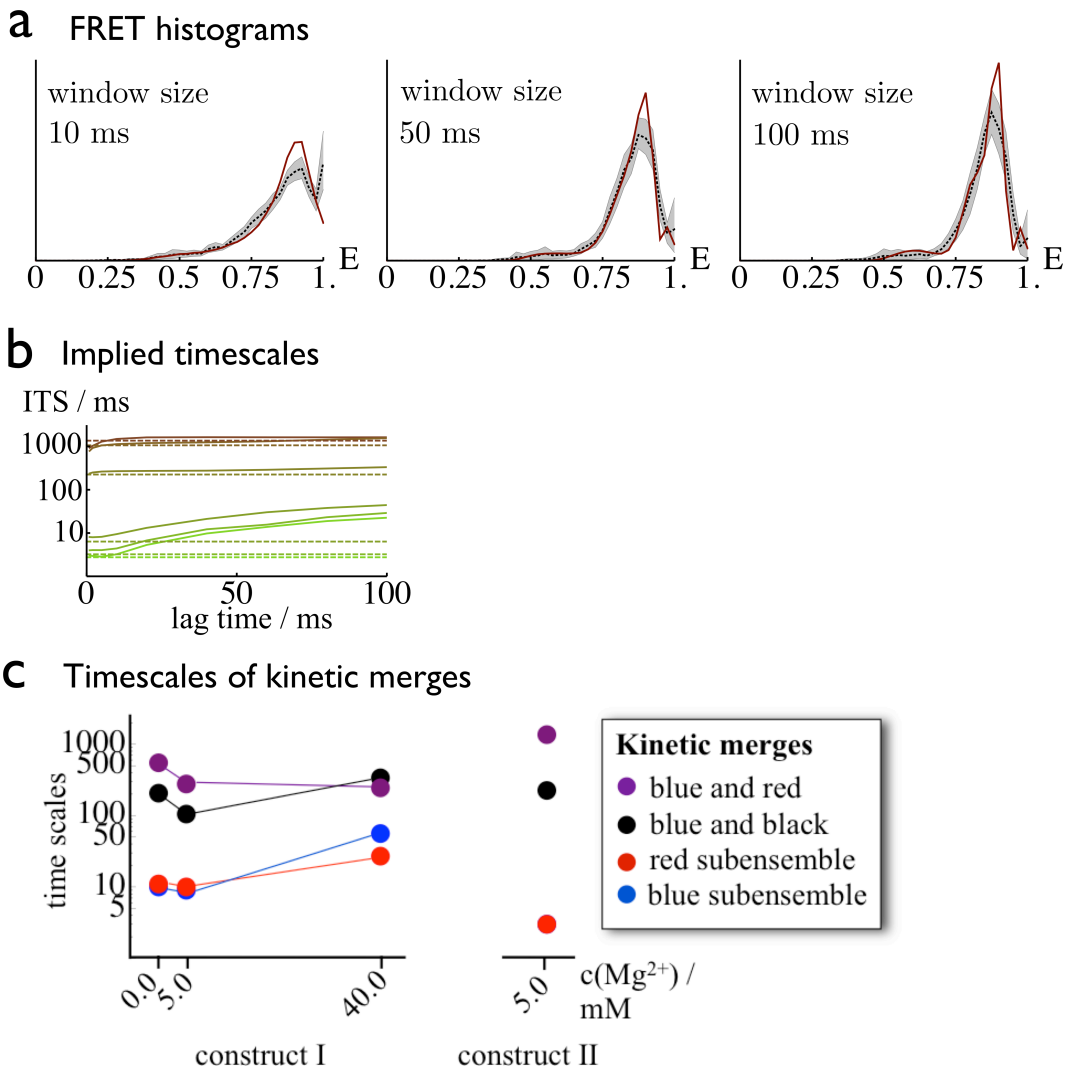

Figure S14: Validation of the HMM for construct II and comparison to the HMMs of construct I. (**a**) Time-window-dependent FRET efficiency histograms for window lengths of 10 ms, 50 ms, and 100 ms. *red line:* Predicition from the hidden Markov model, *grey area / dotted black lines:* Estimation from smFRET data set (bootstrapping mean / 95% confidence interval). (**b**) Implied timescales, indicating that the long-time kinetics of the hidden paths is Markovian and converges to timescales similar to those found in the HMM. The divergence of the shortest timescales at large lag times is expected and due to numerics. **(c)** Comparison of timescales of transition processes found in different Mg^{2+} concentrations and labeling constructs.

References

[1] M. Jäger, A, Kiel, D.P. Herten, and F.A. Hamprecht, Chemphyschem, 10 (2009), pp. 2486–2495

[2] I.V. Gopich, and A. Szabo, Proc Natl Acad Sci USA, 109 (2012), pp. 7747–7752.

[3] I.V. Gopich, and A. Szabo, J Phys Chem B, 113 (2009), pp. 10965–10973.

[4] P. Metzner, I. Horenko, and C. Schütte, Phys Rev E Stat Nonlin Soft Matter Phys, 76 (2006), p. 066702.

[5] D. Crommelin and E. Vanden-Eijnden, Multiscale Modeling Simul., 7 (2009), pp. 1751-1778.

[6] L. R. Rabiner, Proc. Of The IEEE, 77 (1989), pp. 257–286.

[7] L. E. Baum, T. Petrie, G. Soukes, N.A. Weiss, Ann. Math. Stat., 41 (1970), pp. 164.

[8] J.-H. Prinz, H. Wu, M. Sarich, B. Keller, M. Fischbach, M. Held, J.D. Chodera, C. Schütte, and F. Noé, J. Chem. Phys., 134 (2011), pp. 174105.

[9] M.Senne, B. Trendelkamp-Schroer, A.S.J. Mey, C. Schütte, and F. Noé, J. Chem. Theory Comput., 8 (2012), pp. 2223-223.

[10] W.C. Swope, J.W. Pitera, and F. Suits, J. Phys. Chem. B, 108 (2004), pp. 6571–6581.

[11] S. Kube, and M. Weber, J. Chem. Phys., 126 (2007), p. 024103.

[12] P. Deuflhard, and M. Weber, Lin. Alg. Appl., 398 (2005), pp. 161–184.

[13] W. E, and E. Vanden-Eijnden, J. Stat. Phys., 123 (2006), pp. 503-523.

[14] P. Metzner, C. Schütte, E. Vanden-Eijnden, Multiscale Modeling Simul., 7 (2009), pp. 1192-1219.

[15] F. Noé, C. Schütte, E. Vanden-Eijnden, L. Reich, T.R. Weikl, Proc. Natl. Acad. Sci. USA, 106 (2009), pp. 19011-19016.

[16] A.Y. Kobitski, A. Nierth, M. Helm, A. Jäschke, and G.U. Nienhaus, Nucleic Acids Res., 35 (2007), pp. 2047–2059.

[17] B. Seelig, A. Jäschke, Chem Biol, 6 (1999), pp. 167–176.

[18] B. Seelig, S. Keiper, F. Stuhlmann, and A. Jäschke, Angew Chem Int Ed Engl, 39 (2000), pp. 4576.

[19] A. Serganov, S. Keiper, L. Malinina, V. Tereshko, E. Skripkin, C. Hobartner, A. Polonskaia, A.T. Phan, R. Wombacher, R. Micura, Z. Dauter, A. Jäschke, and D.J. Patel, Nat Struct Mol Biol, 12 (2005), pp 218–224.

[20] V. Manoharan, B. Fuertig, A. Jäschke, and H. Schwalbe, J Am Chem Soc, 131 (2009), pp. 6261–6270.

[21] S. Keiper, D. Bebenroth, B. Seelig, E. Westhof, A. Jäschke, Chem Biol, 11 (2004), pp. 1217–1227.

[22] A. Nierth, M. Singer, and A. Jäschke, Chem Commun, 46 (2010), pp. 7975–7977.

[23] R. Wombacher, and A. Jäschke, J Am Chem Soc, 130 (2008) p. 8594.

[24] T. Berezniak, M. Zahran, P. Imhof, A. Jäschke, and J.C. Smith, J Am Chem Soc, 132 (2010) pp. 12587–12596.

[25] S. Kraut, D. Bebenroth, A. Nierth, A. Kobitski, G.U. Nienhaus, and A. Jäschke, Nucleic Acids Res 40 (2012) pp. 1318-1330.

[26] C. Flamm, I.L. Hofacker, P. Stadler, and M. Wolfinger, Z. Phys. Chem., 216 (2002), pp. 1-19.

[27] M.T. Wolfinger, W.A. Svrcek-Seiler, C. Flamm, I.L. Hofacker, and P.F. Stadler, J. Phys. A, 37 (2004), 4731

[28] A.R. Gruber, R. Lorenz, S.H. Bernhart, R. Neuboeck, and I.L. Hofacker, Nucleic Acids Res, 36 (2008), W70-W74.

[29] D.H. Mathews, J. Sabina, M. Zuker, and D.H. Turner, J Mol Biol, 288 (1999), pp. 911-940.