Biophysical Journal, Volume 97

**Supporting Material**

# Learning Rates and States from Biophysical Time Series: A Bayesian Approach to Model Selection and Single-Molecule FRET Data

Jonathan E. Bronson, Jingyi Fei, Jake M Hofman, Ruben L. Gonzalez, Jr., and Chris H. Wiggins

# S Supporting material

## S.1 $\chi^2$ and maximum likelihood

Minimization of squared loss is most commonly derived in the natural sciences by asserting that 'error', the difference between parameterized model prediction and experimental data, is additive, normally distributed, and independent for each example (here indexed by $i$):

$$y_i = f_\vartheta(x_i) + \xi_i; \;\; \xi_i \sim \mathcal{N}(\xi|0,\sigma). \tag{10}$$

This notation emphasizes that the model $f$ depends on parameters $\vartheta$, and the $\sim$ indicates the distribution from which the error $\xi_i$ on the $i^{\text{th}}$ observation is drawn (i.e., the Gaussian or normal distribution and variance $\sigma$). Assuming independent and identically distributed observations, the probability of all the $N$ data $\mathbf{y} = \{y_i\}_{i=1}^{i=N}$ is then the *likelihood*

$$L = p(\mathbf{y}|\vartheta) = \Pi_{i=1}^{i=N}\mathcal{N}(y_i - f_\vartheta(x_i)|0,\sigma) = \frac{\mathrm{e}^{-\chi^2}}{(\sqrt{2\pi\sigma})^N} \tag{11}$$

with the usual $\chi^2 = \sum_{i=1}^{i=N}(y_i - f_\vartheta(x_i))^2/\sigma^2$ arising as a linear term in the logarithm of the likelihood $\ell$:

$$\ell \equiv \ln L = -\chi^2 + \frac{N}{2}\ln 2\pi\sigma. \tag{12}$$

Minimization of $\chi^2$, is thus derived from the more general principle of ML: the parameters $\vartheta_*$ chosen are those which are the most likely.

## S.2 "BIC": an intuition-building heuristic

Often, explicit calculation of $p(\mathbf{y}|K)$ is computationally difficult, and one resorts to approximation. For example, if the likelihood $p(\mathbf{y}|\vec{\vartheta}, K)$ is sharply and uniquely peaked as a function of $\vec{\vartheta}^K$, meaning that there is one unique maximum, Schwartz (34) suggested a pair of approximations: (i) Taylor expansion of $\ell(\vec{\vartheta})$ (from Eq. 12) and Laplace approximation of the integral; and (ii) replacing the second derivative of $\ell(\vec{\vartheta})$ by its asymptotic behavior in the limit $\{K, N\} \to \infty$. The first approximation reads

$$p(\mathbf{y}|K) = \int d^K\vec{\vartheta} \mathrm{e}^{\ell(\vec{\vartheta})} p(\vec{\vartheta}|K) \approx \mathrm{e}^{\ell_*} p(\vec{\vartheta}_*|K) \sqrt{\left|\frac{2\pi}{H}\right|} \tag{13}$$

where $\ell_* = \ell(\vartheta_*)$ is the ML over all parameters $\vartheta$, and the $K \times K$ matrix $H$, also termed the Hessian, is the matrix of derivatives (evaluated at $\vec{\vartheta}_*$)

$$H_{\alpha\beta} \equiv \frac{\partial^2 \ell(\vartheta)}{\partial\vartheta_\alpha \partial\vartheta_\beta}. \tag{14}$$

In the case of $N$ independent data points the derivative of $\ell$ is a sum of $N$ independent terms, and the determinant of the Hessian scales as $N^K$ in the limit of infinite data $N$ and infinitely many $K$ equally-important parameters $\vartheta_\alpha$. Under this pair of asymptotic approximations, then,

$$p(\mathbf{y}|K) \approx \mathrm{e}^{\ell_*} p(\vec{\vartheta}_*|K) \sqrt{\left|\frac{2\pi}{H}\right|} \approx C(K, N) \mathrm{e}^{(\ell_* - (K/2)\ln N)}. \tag{15}$$

The exponent is sometimes referred to as the *Bayesian Information Criterion* or BIC; for clarity it is worth noting, though, that it does not depend on the prior (the most common meaning of the adjective 'Bayesian' in statistics) and that it is derived without any appeal to or use of information theory. The usage of such an algebraic expression alone, ignoring the possible dependence of terms lumped into $C(K, N)$ (i.e., treating $C(K, N)$ as a constant) is a simple[1], intuitive, and appealing approach to model selection. The increase in $\ell_*$ as $K$ increases is penalized by the term $-(K/2)\ln N$, selecting the optimal model indexed by $K_*$, the maximizer of the BIC.

In the case of FRET data the likelihood is complicated by the presence of a hidden state $z_i$ (the discrete conformational state of the molecule which gives rise to the observed FRET ratio), meaning that the evidence $p(\mathbf{y}|K)$ has the richer formulation (suppressing the cluttering superscripts $K$ on the hidden and manifest variables $z$ and $\vartheta$, respectively)

$$p(\mathbf{y}|K) = \sum_{\mathbf{z}} \int d^K\vec{\vartheta} p(\mathbf{y}, \mathbf{z}|\vec{\vartheta}, K) p(\mathbf{z}|\vec{\vartheta}, K) p(\vec{\vartheta}|K). \tag{16}$$

This rich structure renders completely inappropriate the assumptions of the BIC derivation above:

[1]Note that, although use of the BIC obviates determining many facets of one's model and its relation to the data, we still need to know the error bars $\sigma$, which appear in $\ell$.

among other problems, the hidden variables will be modeled by a Markovian dynamic, coupling each of the example data (and thus violating the assumption of $N$ independent data); and the permutation symmetry of the labels on these violates the assumption that the likelihood is sharply and singly peaked – rather there are $K!$ such peaks from the possible re-labelings of the states.

## S.3 Proof of variational relation

We provide a proof of the variational relation in Eq. 6. We start with the desired quantity, the evidence $p(\mathbf{y}|K)$, and multiply by one,

$$\ln p(\mathbf{y}|K) \quad = \quad \left[\sum_{\boldsymbol{z}} \int d\vartheta q(\mathbf{z}, \vec{\vartheta})\right] \ln p(\mathbf{y}|K), \tag{17}$$

valid for any normalized probability distribution $q(\mathbf{z}, \vec{\vartheta})$. We then use the definition of conditional probability to write

$$p(\mathbf{y}, \vec{\vartheta}, \mathbf{z}|K) = p(\mathbf{z}, \vec{\vartheta}|\mathbf{y}, K)p(\mathbf{y}|K). \tag{18}$$

We use this to rewrite the argument of the logarithm and multiply by one yet again:

$$\begin{aligned}
\ln p(\mathbf{y}|K) &= \sum_{\boldsymbol{z}} \int d\vartheta q(\mathbf{z}, \vec{\vartheta}) \ln \left[\frac{p(\mathbf{y}, \vec{\vartheta}, \mathbf{z}|K)}{p(\mathbf{z}, \vec{\vartheta}|\mathbf{y}, K)}\right] && (19)\\
&= \sum_{\boldsymbol{z}} \int d\vartheta q(\mathbf{z}, \vec{\vartheta}) \ln \left[\frac{p(\mathbf{y}, \vec{\vartheta}, \mathbf{z}|K)q(\mathbf{z}, \vec{\vartheta})}{q(\mathbf{z}, \vec{\vartheta})p(\mathbf{z}, \vec{\vartheta}|\mathbf{y}, K)}\right] && (20)\\
&= \sum_{\boldsymbol{z}} \int d\vartheta q(\mathbf{z}, \vec{\vartheta}) \ln \left[\frac{p(\mathbf{y}, \vec{\vartheta}, \mathbf{z}|K)}{q(\mathbf{z}, \vec{\vartheta})}\right] + \sum_{\boldsymbol{z}} \int d\vartheta q(\mathbf{z}, \vec{\vartheta}) \ln \left[\frac{q(\mathbf{z}, \vec{\vartheta})}{p(\mathbf{z}, \vec{\vartheta}|\mathbf{y}, K)}\right] && (21)
\end{aligned}$$

where in the last line we have separated logarithm to decompose the integral into two parts. We recognize the rightmost term as the Kullback-Leibler divergence between $q(\mathbf{z}, \vec{\vartheta})$ and $p(\mathbf{z}, \vec{\vartheta}|\mathbf{y}, K)$,

$$D_{KL}(q(\mathbf{z}, \vec{\vartheta})||p(\mathbf{z}, \vec{\vartheta}|\mathbf{y}, K)) \quad = \quad \sum_{\boldsymbol{z}} \int d\vartheta q(\mathbf{z}, \vec{\vartheta}) \ln \left[\frac{q(\mathbf{z}, \vec{\vartheta})}{p(\mathbf{z}, \vec{\vartheta}|\mathbf{y}, K)}\right] \tag{22}$$

and define the remaining term as the *free energy*,

$$F[q(\mathbf{z}, \vec{\vartheta})] = -\sum_{\boldsymbol{z}} \int d\vartheta q(\mathbf{z}, \vec{\vartheta}) \ln \left[\frac{p(\mathbf{y}, \vec{\vartheta}, \mathbf{z}|K)}{q(\mathbf{z}, \vec{\vartheta})}\right], \tag{23}$$

which results in the variational relation presented in Eq. 6,

$$\ln p(\mathbf{y}|K) \quad = \quad -F[q(\mathbf{z}, \vec{\vartheta})] + D_{KL}(q(\mathbf{z}, \vec{\vartheta})||p(\mathbf{z}, \vec{\vartheta}|\mathbf{y}, K)). \tag{24}$$

This completes the proof of the variational relation and offers several insights.

The first is that the free energy is strictly bounded by the log-evidence, as the Kullback-Leibler (KL) divergence is a non-negative quantity, proven through an application of Jensen's inequality (an extension of the definition of convexity). Thus we have reduced the problem of approximating

the evidence to that of finding the distribution $q(\mathbf{z}, \vec{\vartheta})$ which is "closest" to the true (and intractable) posterior $p(\mathbf{z}, \vec{\vartheta}|\mathbf{y}, K)$ in the KL sense. As per Eq. 24, we see that this is equivalent to minimizing the free energy $F[q(\mathbf{z}, \vec{\vartheta})]$ as a functional of $q(\mathbf{z}, \vec{\vartheta})$. This observation motivates the VBEM algorithm, in which a specific factorization for $q(\mathbf{z}, \vec{\vartheta})$ is chosen as to make calculation of $F[q(\mathbf{z}, \vec{\vartheta})]$ tractable (here, that $q(\mathbf{z}, \vec{\vartheta}) = q(\mathbf{z})q(\vec{\vartheta})$), and iterative coordinate ascent is performed to find a local minimum.

In addition, we provide motivation for the term "free energy", rewriting Eq. 23 by decomposing the logarithm:

$$F[q(\mathbf{z}, \vec{\vartheta})] = -\sum_{\boldsymbol{z}} \int d\vartheta q(\mathbf{z}, \vec{\vartheta}) \ln \left[ \frac{p(\mathbf{y}, \vec{\vartheta}, \mathbf{z}|K)}{q(\mathbf{z}, \vec{\vartheta})} \right] \tag{25}$$

$$= -\sum_{\boldsymbol{z}} \int d\vartheta q(\mathbf{z}, \vec{\vartheta}) \ln p(\mathbf{y}, \vec{\vartheta}, \mathbf{z}|K) + \sum_{\boldsymbol{z}} \int d\vartheta q(\mathbf{z}, \vec{\vartheta}) \ln q(\mathbf{z}, \vec{\vartheta}). \tag{26}$$

Recognizing the negative log-probability in the first term as an energy (as in the Boltzmann distribution) and the second term as the information entropy (35) of $q(\mathbf{z}, \vec{\vartheta})$, i.e.

$$E \equiv -\ln p(\mathbf{y}, \vec{\vartheta}, \mathbf{z}|K) \tag{27}$$

$$S \equiv -\sum_{\boldsymbol{z}} \int d\vartheta q(\mathbf{z}, \vec{\vartheta}) \ln q(\mathbf{z}, \vec{\vartheta}). \tag{28}$$

Thus we can rewrite 23 as

$$F = \langle E \rangle - TS, \tag{29}$$

(with unit "temperature" T) where the angled brackets denote expectation under the variational distribution $q(\mathbf{z}, \vec{\vartheta})$. This familiar form from statistical physics offers the following interpretation: in approximating the evidence (and posterior), we seek to minimize the free energy by finding a distribution $q(\mathbf{z}, \vec{\vartheta})$ that balances minimizing the energy and maximizing entropy.

We encourage the reader to enjoy the texts (20) and (19) for more pedagogical discussions of variational methods.

## S.4 2D inference

Instead of analyzing the 1D FRET ratio, it is also possible to model the donor and acceptor molecule intensities directly. Such analysis can be accomplished by treating the donor and acceptor signals as a 2D vector, which is then fit by a 2D Gaussian. (The VBEM solution to the HMM with multidimensional Gaussian observables is solved in (23), and requires only a minor change to the code used in the rest of this manuscript.) Because information is necessarily lost by transforming the 2D donor / acceptor signal into a 1D ratio, the 2D data may yield more information about the FRETing complex and, therefore, better inference. However, it is also possible that the 2D data provide information only about the photophysics rather than the biophysics, i.e., that the only biophysically meaningful quantity is the donor / acceptor transfer efficiency reflected in the FRET ratio. While it is outside the scope of this paper to assess the relative merits of 1D and 2D FRET analysis, it is worth considering whether evidence could be used to evaluate the relative accuracies of inferences performed in 1 or 2 dimensions.

Intuitively, one should not expect evidence to be an appropriate evaluative quantity in this situation: evidence allows one to select between competing models for a fixed data set, i.e. a selection between $p(D|M_1)$ and $p(D|M_2)$ (which is how evidence is used in the rest of this manuscript). Here, we are asking the for $p(D_1|M_1)$ versus $p(D_2|M_2)$, where $D_1$ and $D_2$ are the 1D FRET ratio and donor / acceptor data. Because the space of 2D data sets is so much larger than the space of 1D data sets, the evidence for the 2D inference may be lower, regardless of the quality of the inference.

To test this hypothesis, the following data set was devised. Ten $K = 3$ traces were generated as described in Sec. S.5.4, each of length $T = 200$, with 1D FRET state means of $\sim 0.11 \pm 0.014FRET$. The traces were then replicated 5 times, and both the donor and acceptor signals were mollified by multiplication with the a linearly decreasing envelope function $A(t) = 1 - (t/T)S$, where $S = \{0, 0.15, 0.30, 0.45, 0.60\}$. When $S = 0$, the 2D traces should have more information than do the 1D FRET transformations. By the time $S = 0.60$, the 2D traces should be poorly described by 2D Gaussian HMMs (since the means of the donor / acceptor signals at the end of the traces are 40% of their original values), but the 1D traces will still look the same as when $S = 0$ (since the multiplying factor cancels out of the FRET ratio). A sample trace is shown in Fig. S.1.

The results are shown in Fig. S.2. The mean and standard deviation of $\ln(p(D|M))$ for each set of traces are shown in Table S.1. Consistent with expectations, as S increases from 0 to 0.60, the 1D inference is unchanged, but the accuracy of the inferred trajectory and the sensitivity to transitions decrease for the 2D inference. Inspection of individual traces shows that when $S = 0$, 2D inference is better or equal to 1D inference, by all four accuracy metrics in Fig. S.2, for 9 out of 10 traces (data not shown). When $S = 0.60$, accuracy of the inferred trajectory and the sensitivity to transitions is worse for 2D inference than for 1D inference for all traces (specificity of transitions is slightly better for 2D inference, but that is a result of missing transitions in the data).

Regardless of the quality of inference, $\ln(p(D|M))$ for the 2D inference is lower than for the 1D inference, consistent with our intuition about the far greater number of possible 2D datasets, with $\log p(D_1|M_1) \approx 40$ and $\log p(D_2|M_2) \approx -500$. Similar results were observed for all synthetic data sets we have tested (other data sets not shown) and reflect that evidence *cannot* be used to

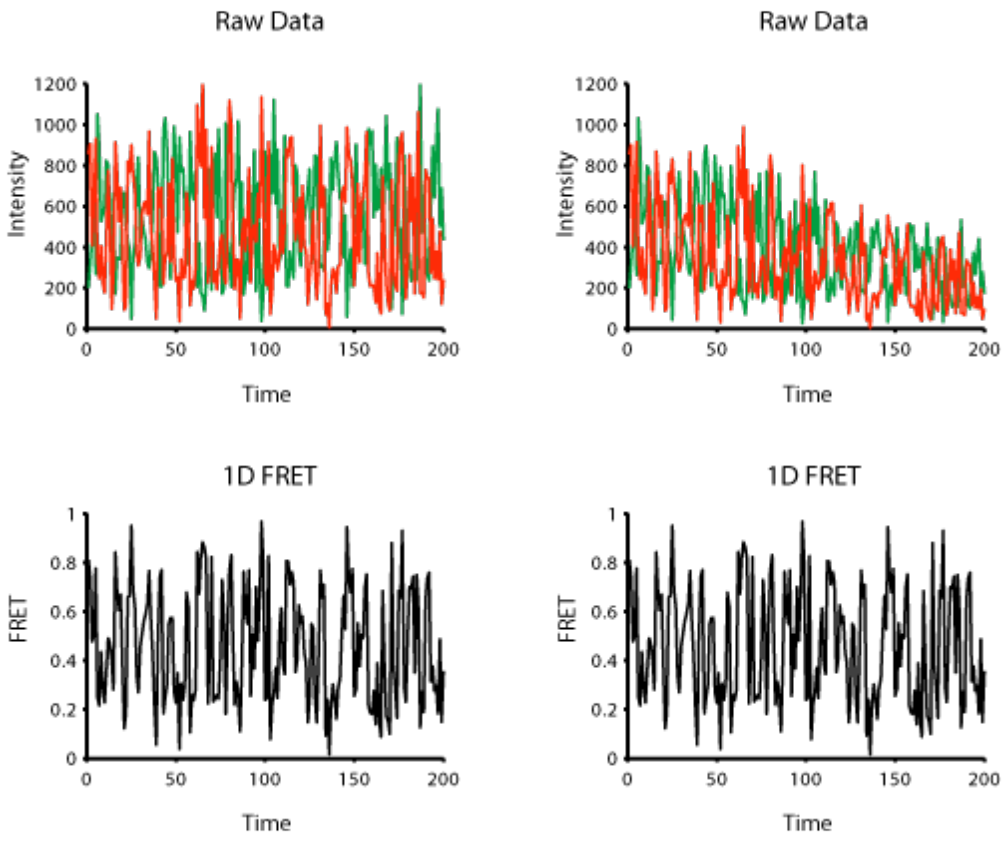


Figure S.1: (Top left) one of the $S = 0$ traces fit using a 2D Gaussian HMM and (bottom left) the 1D FRET transform. (Right) the same trace, but multiplied by the $A(t) = 1 - 0.60(t/T)$ vector. By the end of the 2D trace the means of the donor and acceptor signal intensities are 40% of their original values while the 1D transformation (bottom right) is unchanged.

assess the quality of 1D versus 2D data inference. It should be noted, however, that evidence based model selection does work for choosing between different 2D models of varying complexity, as evidenced in Fig. S.2, since the comparison is again between $p(D|M_1)$ and $p(D|M_2)$.

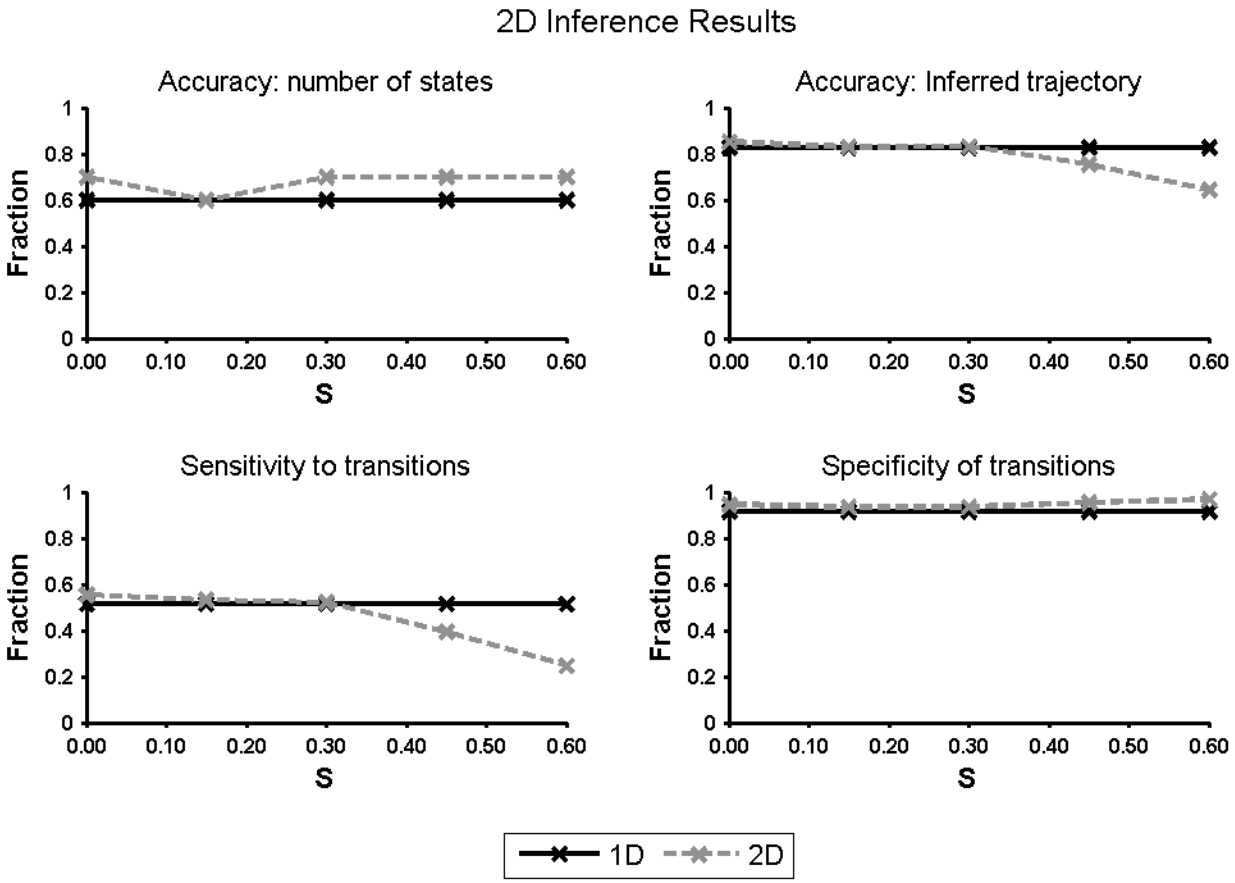


Figure S.2: Accuracy of 1D inference versus 2D inference. Accuracy metrics are the same as those used in Fig. 2.

Table S.1: $\ln(p(D|M))$ for 1D and 2D inference.

| S: | 0 | 0.15 | 0.30 | 0.45 | 0.60 |
|---|---|---|---|---|---|
| 1D | $45.13 \pm 18.27$ | $45.13 \pm 18.27$ | $45.13 \pm 18.27$ | $45.13 \pm 18.27$ | $45.13 \pm 18.27$ |
| 2D | $-501.20 \pm 19.80$ | $-502.47 \pm 18.69$ | $-509.10 \pm 16.45$ | $-515.51 \pm 16.22$ | $-504.58 \pm 18.26$ |

## S.5 Methods

### S.5.1 ML inference settings

Following the HaMMy user manual, ML analyses use $K_{max} + 2$ states, where $K_{max}$ is either $K_0$ (the true number of states) in the case of synthetic data or simply 3 in the case of experimental data, as 1D FRET histograms suggest two biophysical states and one photophysical state: the photobleached state. No additional complexity control was applied to the resulting parameters inferred from individual traces. The default guess for the initial distribution of the means $\mu_z$ was used, i.e., uniform spacing between $0$ and $1$ FRET.

Also consistent with default settings, we use the parameters inferred using only one set of initial parameter-guesses. Note that this differs from the usual implementation of expectation-maximization as a technique for performing ML (*cf.,* (19)). Expectation-maximization (the maximization technique used in both HaMMy and VBFRET) provably converges to a local optimum, and therefore the maximization typically is performed using many random restarts for parameter values. One possible reason to avoid this procedure is the inescapable pathology of ML for real-valued emissions (*e.g.,* in FRET data) and for which the width of each state is an inferred parameter: the optimization is ill-posed since the case in which one observation is assigned to a state of 0 uncertainty is infinitely likely (*cf.,* (19) Ch. 9: "These singularities provide another example of the severe overfitting that can occur in a maximum likelihood approach. We shall see that this difficulty does not occur if we adopt a Bayesian approach.").

### S.5.2 ME inference settings

In analyzing synthetic and experimental data with ME, we attempt each choice of $K = 1, 2, \ldots, K_{\max}+2$ with $K_{\max}$ as above. For synthetic data, 25 random initial guesses were used for each of the traces; for experimental data, 100 initializations were used (though, in our experience, little or no change in the optimization was found after 25 initializations). As with all local optimization techniques, including expectation maximization in ML or in ME, we use the parameters which give the optimum over all restarts (here, the set of parameters specifying the approximating distribution $q$ which gives the maximum evidence $p(\mathbf{y}|K)$).

### S.5.3 Rate constant calculations

Rates for the smFRET$_{\mathrm{L1-L9}}$ experimental data, both for ME and ML analyses, were extracted as previously described (6, 33). First, the set of all idealized traces over all times is histogrammed into 50 bins, evenly spaced between $-0.2$ and $1.2$ FRET. The counts in the resulting histogram are given to Origin 7.0, which learns a Gaussian mixture model via expectation-maximization, using user-supplied initial guesses for the three means (we used $\mu = (0, 0.35, 0.55)$ FRET). Origin returns true means and variances for each of the 3 states. From these variances the width at half-max for each mixture is determined, defining three acceptable ranges of fret values. (For this experiment, these ranges had widths of approximately $.05$ FRET. We next re-scan the idealized traces and, for each transition from one acceptable range to another, record the dwell time (the

total time spent within the range; any number of inferred transition within one accepted range are ignored, effectively smoothing of overfit idealized traces). The cumulative distribution of dwell times from a given state is now given to Origin 7.0 to infer the most likely parameters, asserting exponential decay. The inverse of the inferred time constant is the rate constant reported for that state.

### S.5.4 Generating synthetic data

Synthetic data were generated in MATLAB. Rather than testing the inference on data generated precisely by the emissions model (one in which the scalar FRET signal is taken to be normally-distributed in each state), we challenge the inference by using a slightly more realistic distribution: one that is normally-distributed in each of the two florophore colors. That is, each synthetic trace was created from a hidden Markov model with 2D Gaussian output (representing the two florophore colors). The 2D data $\boldsymbol{x}_1, \boldsymbol{x}_2$ were then FRET transformed using $\boldsymbol{f} = \boldsymbol{x}_2/(\boldsymbol{x}_1 + \boldsymbol{x}_2)$; points such that $f \notin (0, 1)$ were discarded.

The 2D Gaussians are chosen so that, in any state $z$, the sum of the means is 1000 ($\mu_z^1 + \mu_z^2 = 1000\ \forall z$), roughly corresponding to our experimental data. Variances were drawn from a uniform distribution centered at each dimension's mean over a range given by 10% of the mean. The two components were allowed a nonzero covariance, also drawn from a uniform distribution centered at 0, with a range given by $1/2$ the smaller of the two means. We emphasize that these choices are intended both to be consistent with the smFRET$_{\mathrm{L1-L9}}$ and smFRET$_{\mathrm{L1-tRNA}}$ data and *not* to match the algebraic expressions in the priors used below, which would be a less challenging inference task (model specification identically matching the generative process).

Increasingly noisy traces were generated by multiplying the covariance matrix of each hidden state by a constant. Ten constants, chosen log-linearly between 1 and 100, were used. The mean standard deviation of the FRET state noise in the resulting 1D traces varied from, approximately, $0.02 < \sigma < 1.4$.

## S.6 Priors

### S.6.1 Mathematical expressions for priors

To calculate the model evidence, we treat the components of $\vec{\vartheta}$ as random variables. The vector $\vec{\pi}$ and each row of A are modeled as Dirichlet distributions:

$$p(\vec{\pi}) = \frac{\Gamma(\sum_{k=1}^{K} u_\pi^k)}{\prod_{k=1}^{K} \Gamma(u_\pi^k)} \prod_{k=1}^{K} \pi_k^{u_\pi^k - 1} \tag{30}$$

$$p(a_{j1}, ..., a_{jK}) = \frac{\Gamma(\sum_{k=1}^{K} u_a^{jk})}{\prod_{k=1}^{K} \Gamma(u_a^{jk})} \prod_{k=1}^{K} a_{jk}^{u_a^{jk} - 1} \tag{31}$$

The probabilities for each pair of $\mu_k$ and $\lambda_k$ are modeled jointly as a Gaussian-Gamma distribution:

$$p(\mu_k, \lambda_k) = \sqrt{\frac{u_\beta^k \lambda_k}{2\pi}} e^{-\frac{1}{2} u_\beta^k \lambda_k (\mu_k - u_\mu^k)^2} \frac{1}{\Gamma(u_v^k/2)} (2u_W^k)^{-u_v^k/2} \lambda_k^{(u_v^k/2)-1} e^{-\frac{\lambda_k}{2u_W^k}}. \tag{32}$$

The terms $\vec{u}_\pi$, $\vec{u}_a$, $\vec{u}_\beta$, $\vec{u}_\mu$, $\vec{u}_v$, and $\vec{u}_W$ are called the *hyperparameters* for the probability distributions over $\vec{\vartheta}$.

### S.6.2 Hyperparameter settings

Hyperparameters for VBFRET were set so as to give distributions consistent with experimental data and to influence the inference as weakly as possible: $u_\pi^k = 1$, $u_a^{jk} = 1$, $u_\beta^k = 0.25$, $u_m^k = 0.5$, $u_v^k = 5$ and $u_W^k = 50$, for all values of k. Qualitatively, these hyperparameters priors correspond to probability distributions over the hidden states such that it is most probable that the hidden states are equally likely to be occupied and equally likely to be transitioned to. Quantitatively, they yield $\langle \mu_k \rangle = 0.5$ and typical $\sigma \approx 0.08$, consistent with experimental observation. ($1/\sqrt{\mathrm{mode}(\lambda_k)} = 1/\sqrt{150} \approx 0.08 \; \forall k$).

### S.6.3 Sensitivity to hyperparameter settings

One standard approach (36, 37) to sensitivity analysis is to halve and double hyperparameters and recompute the evidence for different models. The sensitivity of ME inference on hyperparameter settings was investigated on both experimental and synthetic data. First, the two and three state traces from Fig. 2 and Fig. S.7 were reanalyzed with all the hyperparameters set to one half their default values and twice their default values (Figs. S.3, S.4, S.5, S.6). One hyperparameter, the prior on the mean of each Gaussian, was not changed during this analysis, since its value is set to 0.5 based on a symmetry argument.

The results show a relative insensitivity to the hyperparameter values over the settings considered. The largest difference in inference accuracy between the different settings was for the noisy, slow-transitioning traces shown in Fig. S.6, when the hyperparameters were doubled. Interest-

ingly, these traces are harder to resolve than the two state traces but not as difficult to resolve as the noisy, fast-transitioning three state traces. A possible explanation for this behavior is that the two state trace results are insensitive to hyperparameter settings because the data are easy enough to resolve and the noisy, fast-transitioning three state traces are insensitive to hyper parameter settings because they are too hard to resolve. The noisy, slow-transition states are on the border of being resolvable, so using a prior that more closely matches the true parameters of the model yields more accurate results. Additionally, the three state, slow-transition data has the highest probability of having a sparsely populated state (i.e. one that is only present for a few time steps in a trace). When $\sigma$ is large, these sparsely populated states become harder to identify as distinct states, which may explain why $p(|\hat{\mathbf{z}}| = |\mathbf{z}_0|)$ decreases more than $p(\hat{\mathbf{z}} = \mathbf{z}_0)$, sensitivity or specificity .

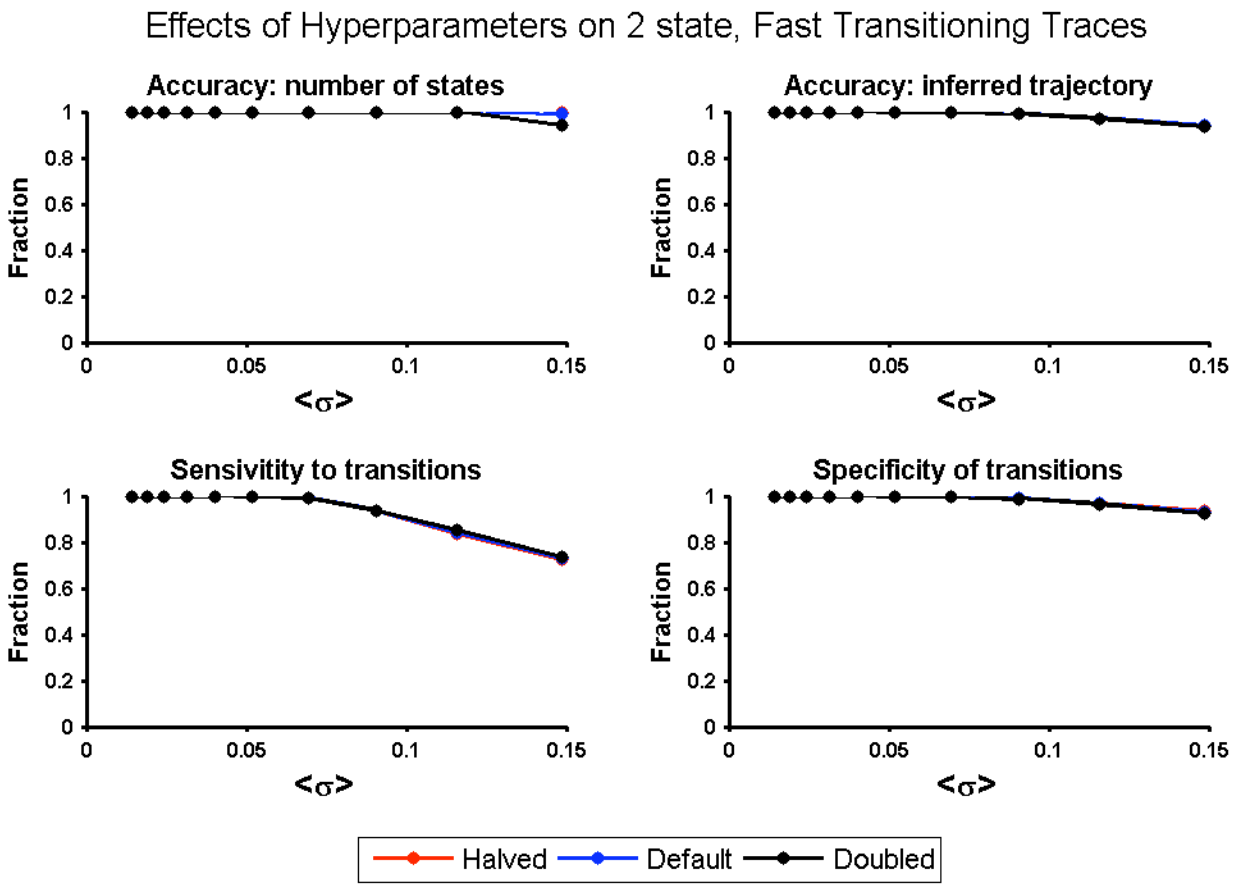


Figure S.3: Effects of hyperparameter settings on fast-transitioning, two state traces.

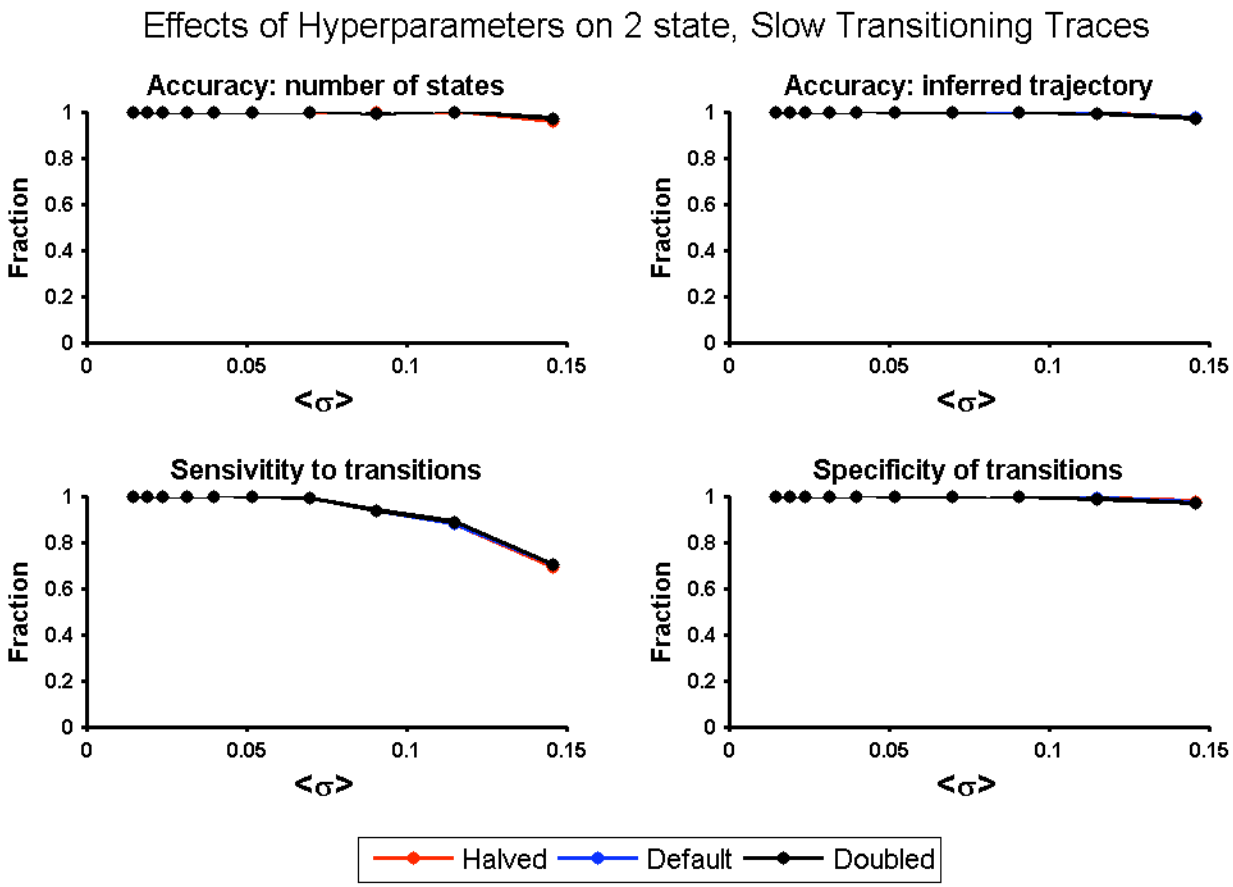


Figure S.4: Effects of hyperparameter settings on slow-transitioning, two state traces.

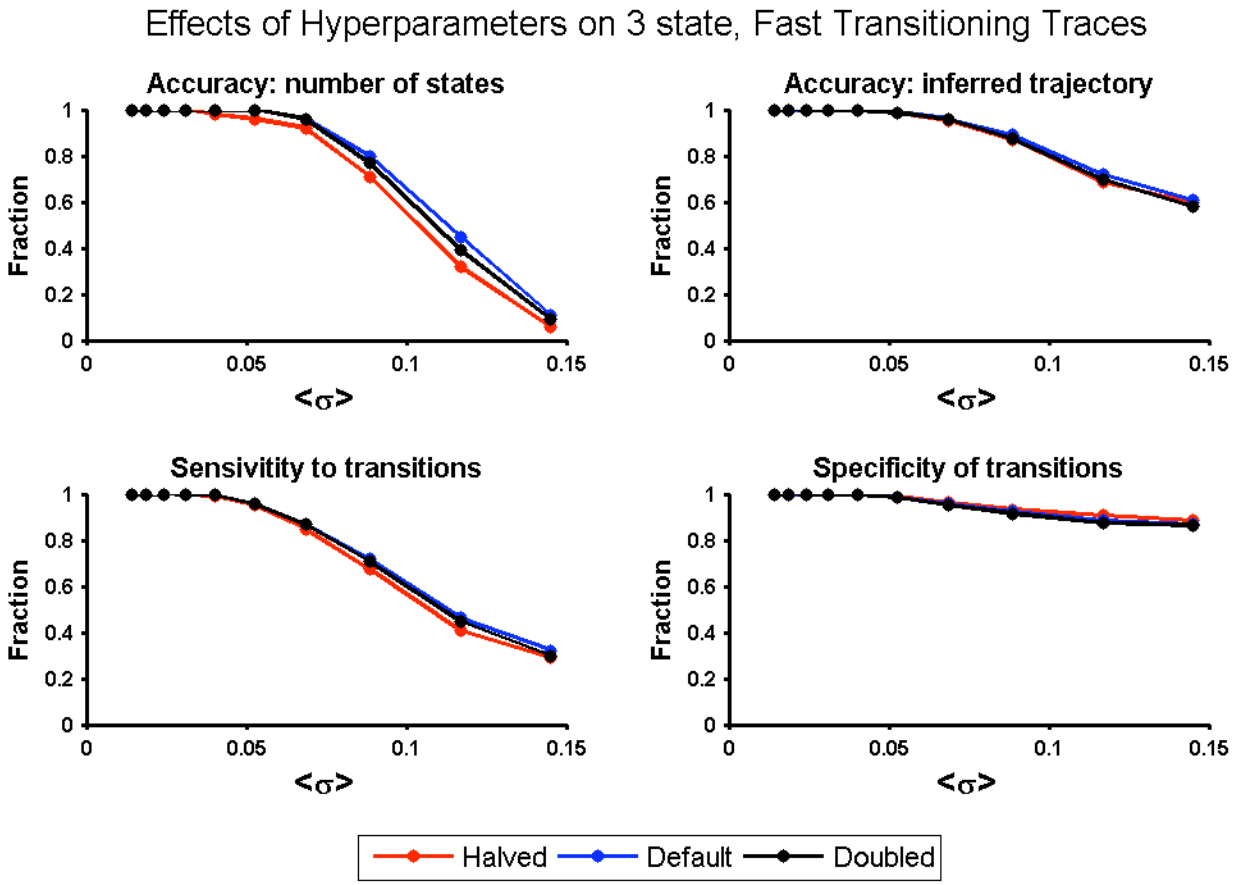


Figure S.5: Effects of hyperparameter settings on fast-transitioning, three state traces.

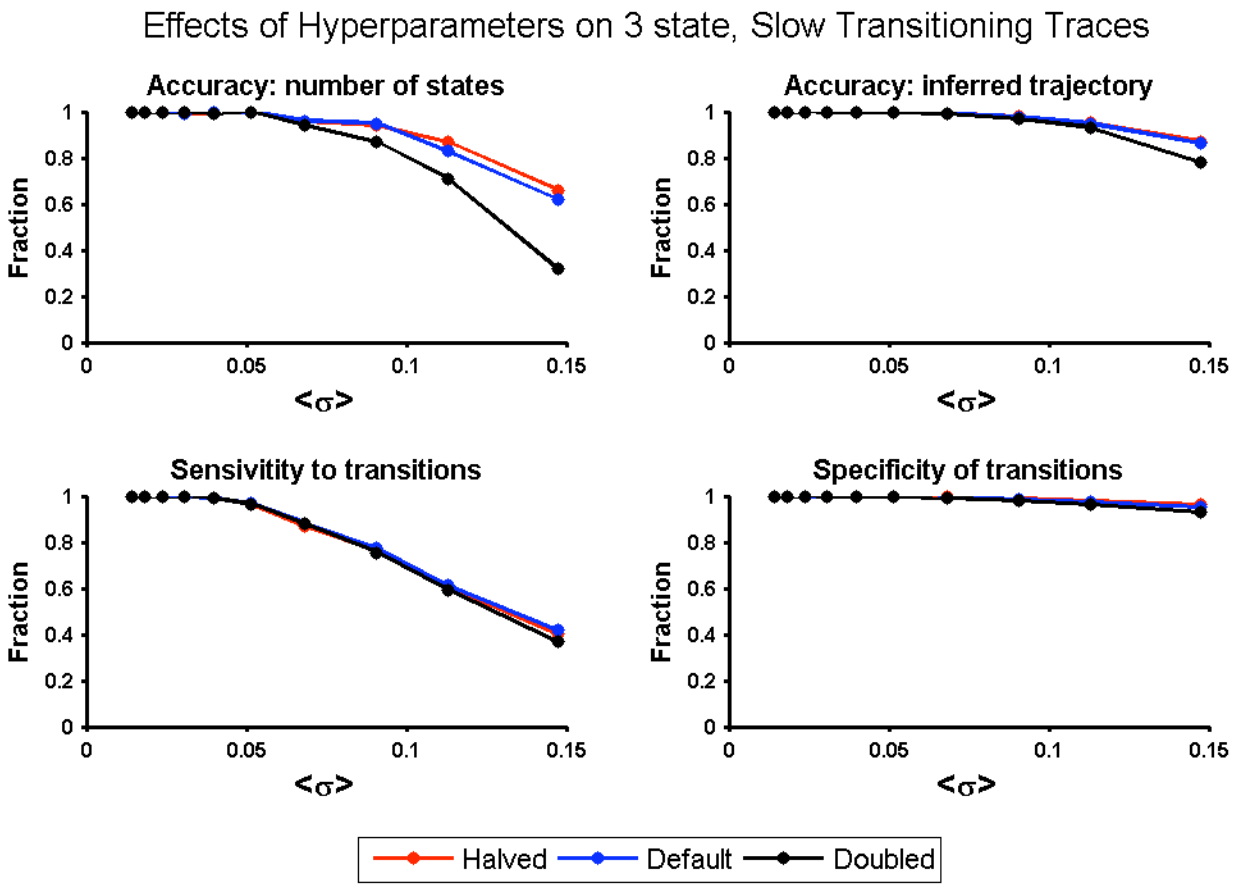


Figure S.6: Effects of hyperparameter settings on slow-transitioning, three state traces.

To investigate further the effects of the hyperparameter settings on ME inference, the experimental data from Table 1 were reanalyzed using a more strongly diagonal transition matrix prior (Table S.2). In this second prior, the diagonal terms of the transition matrix were set to 1 and the off-diagonal terms were set to 0.05, loosely corresponding to a prior belief that the ribosome was 10x more likely to remain in its current state than transition to a new one. For all of the data, the transition rates calculated with both hyperparameter settings are within error of each other for all transition rates.

Table S.2: Effect of hyperparameters on transition rate inference

| Data set* | Settings | $k_{\mathrm{close}}$ | $k_{\mathrm{open}}$ |
|---|---|---|---|
| $\mathrm{PMN_{Phe}}^{\dagger}$ | Default | $0.66 \pm 0.05$ | $1.0 \pm 0.2$ |
| | Diagonal | $0.66 \pm 0.04$ | $1.0 \pm 0.2$ |
| $\mathrm{PMN_{fMet}}^{\ddagger}$ | Default | $0.53 \pm 0.08$ | $1.7 \pm 0.3$ |
| | Diagonal | $0.52 \pm 0.09$ | $1.7 \pm 0.1$ |
| $\mathrm{PMN_{fMet+EFG}}$ | Default | $3.1 \pm 0.6$ | $1.3 \pm 0.2$ |
| $(1\ \mu M)^{\S}$ | Diagonal | $2.8 \pm 0.5$ | $1.3 \pm 0.1$ |
| $\mathrm{PMN_{fMet+EFG}}$ | Default | $2.6 \pm 0.6$ | $1.5 \pm 0.1$ |
| $(0.5\ \mu M)^{\S}$ | Diagonal | $2.6 \pm 0.5$ | $1.4 \pm 0.1$ |

* Rates reported here are the average and standard deviation from three or four independent data sets. Rates were not corrected for photobleaching of the fluorophores.

† $\mathrm{PMN_{Phe}}$ was prepared by adding the antibiotic puromycin to a post-translocation complex carrying deacylated-$\mathrm{tRNA^{fMet}}$ at the E site and fMet-Phe-$\mathrm{tRNA^{Phe}}$ at the P site, and thus contains a deacylated-$\mathrm{tRNA^{Phe}}$ at the P site.

‡ $\mathrm{PMN_{fMet}}$ was prepared by adding the antibiotic puromycin to an initiation complex carrying fMet-$\mathrm{tRNA^{fMet}}$ a the P site, and thus contains a deacylated-$\mathrm{tRNA^{fMet}}$ at the P site.

§ 1.0 $\mu M$ and 0.5 $\mu M$ EF-G in the presence of 1 mM GDPNP (a non-hydrolyzable GTP analog) were added to $\mathrm{PMN_{fMet}}$, respectively.

## S.7 Synthetic validation – 2 and 4 state traces

Synthetic data for 2 FRET state traces (fast- and slow-transitioning, smFRET state means at 0.3 and 0.7 FRET) and 4 FRET state traces (fast-transitioning only, smFRET state means at 0.21, 0.41, 0.61 and 0.81 FRET) were generated and analyzed exactly as the traces in Fig. 2. The results are qualitatively similar to those in Fig. 2. Inference accuracy begins to decrease at a lower noise level as more FRET states are added to the traces. This should not be surprising, though, since the states are more closely spaced as the number of states increases, and therefore should be harder to resolve. Results for $K > 4$ state traces follow the same trend as those for $K = 2, 3, 4$ (data not shown).

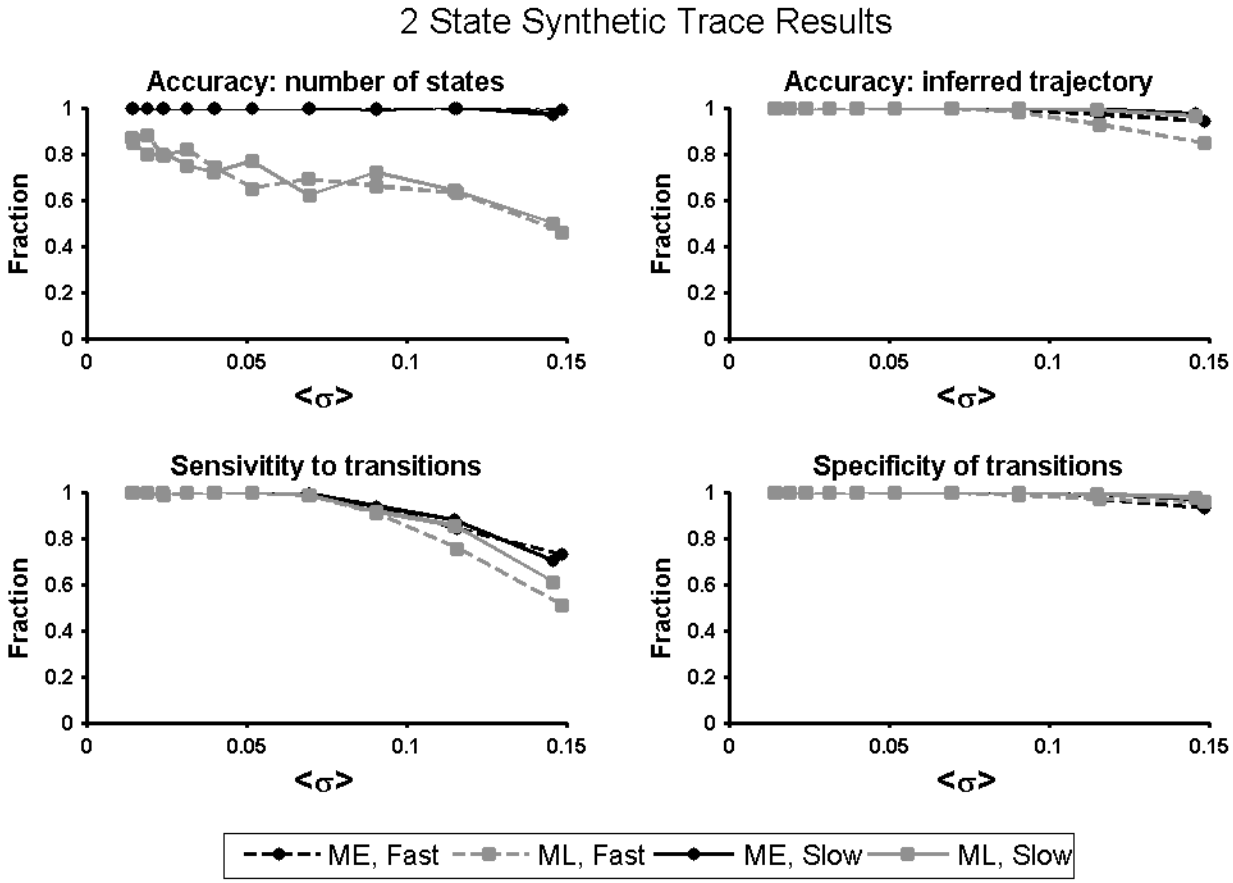


Figure S.7: Synthetic results for two state traces.

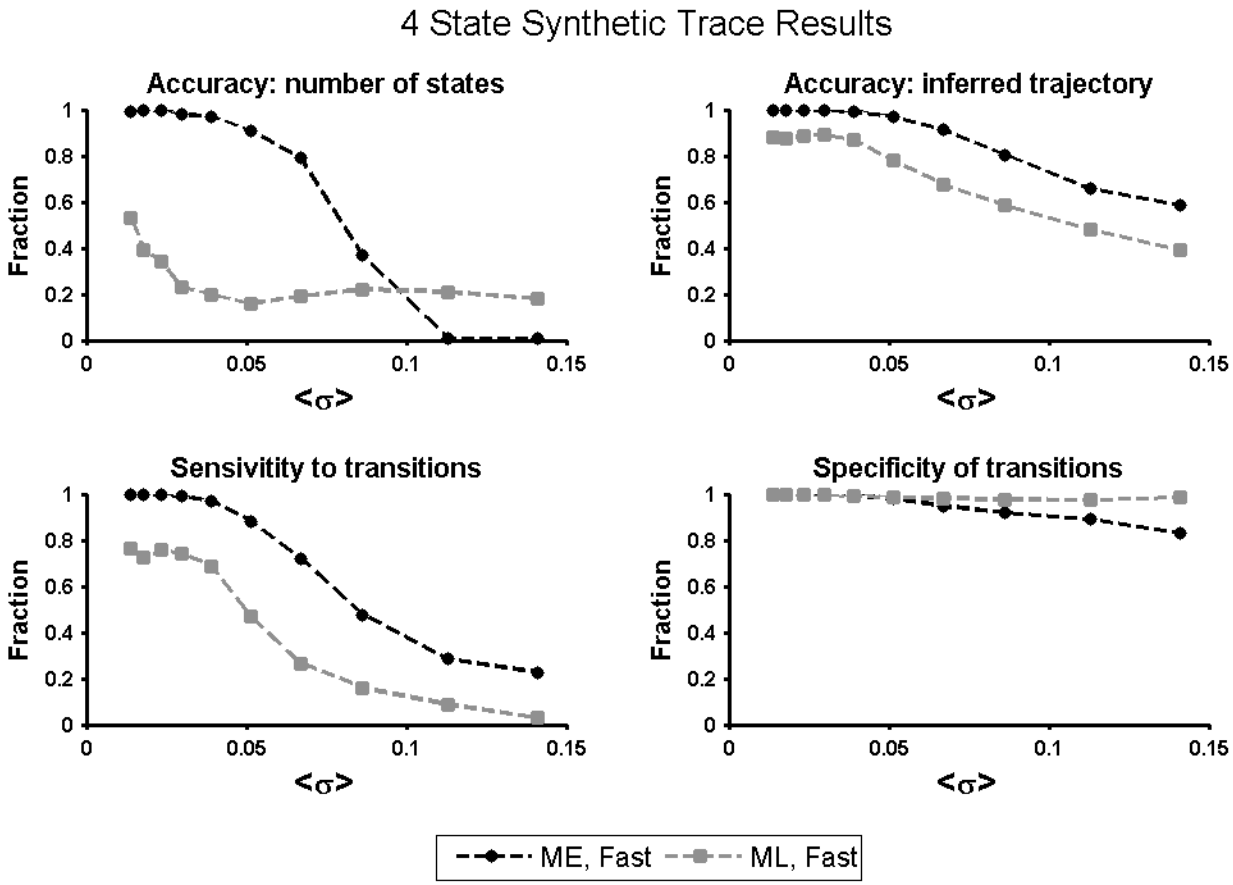


Figure S.8: Synthetic results for four state traces.

## S.8 Proposed method to correct camera blurring

Single data point artifacts caused by stochastic photophysical fluctuations of fluorophore intensity are a well known and common problem in smFRET data (12). These artifacts can be corrected for by applying smoothing algorithms or rolling averages over the data (27, 31) or ignoring FRET states with a dwell time of one time point (6). The artifacts we encounter here are different in nature, since they result from time binning the data rather than a photophysical fluctuation in donor/acceptor signal intensity and, therefore, should be corrected for using a different approach. The algorithm we propose performs a second round of ME inference on the data, using the idealized traces from the first round of ME inference to make the following modification to the raw data: data which could have resulted from time-averaging artifacts (i.e. events lasting exactly one data point and occurring between two distinct idealized values) were moved to the idealized value closest to the value of the suspected time-averaging artifact (the assumption here is that that a single $f_{\mathrm{mid}}$ data point should be considered part of the "real" FRET state that the molecular complex spent the most time in during that transitioning time point). We performed this algorithm on the smFRET$_{\mathrm{L1-tRNA}}$. The TDP for this "cleaned" data shows the blur state at $f_{\mathrm{mid}}$ is virtually eliminated, yielding a result that is wholly consistent with that generated by ML (Fig. S.11). In general, however, it should be cautioned that a *bona fide* intermediate FRET state may well exist and be buried under a strongly-populated blur state. Unless this intermediate FRET state is positively identified and somehow separated from the blur state (i.e. by obtaining data at an increased integration time), eliminating or ignoring FRET states with dwell times exactly equal to one time point may risk overlooking a *bona fide* intermediate FRET state. We note that the vbFRET software package which we have made available allows the user the opportunity to run this second round of ME analysis with possible blur states detected and cleaned as described above.

## S.9 Supporting figures

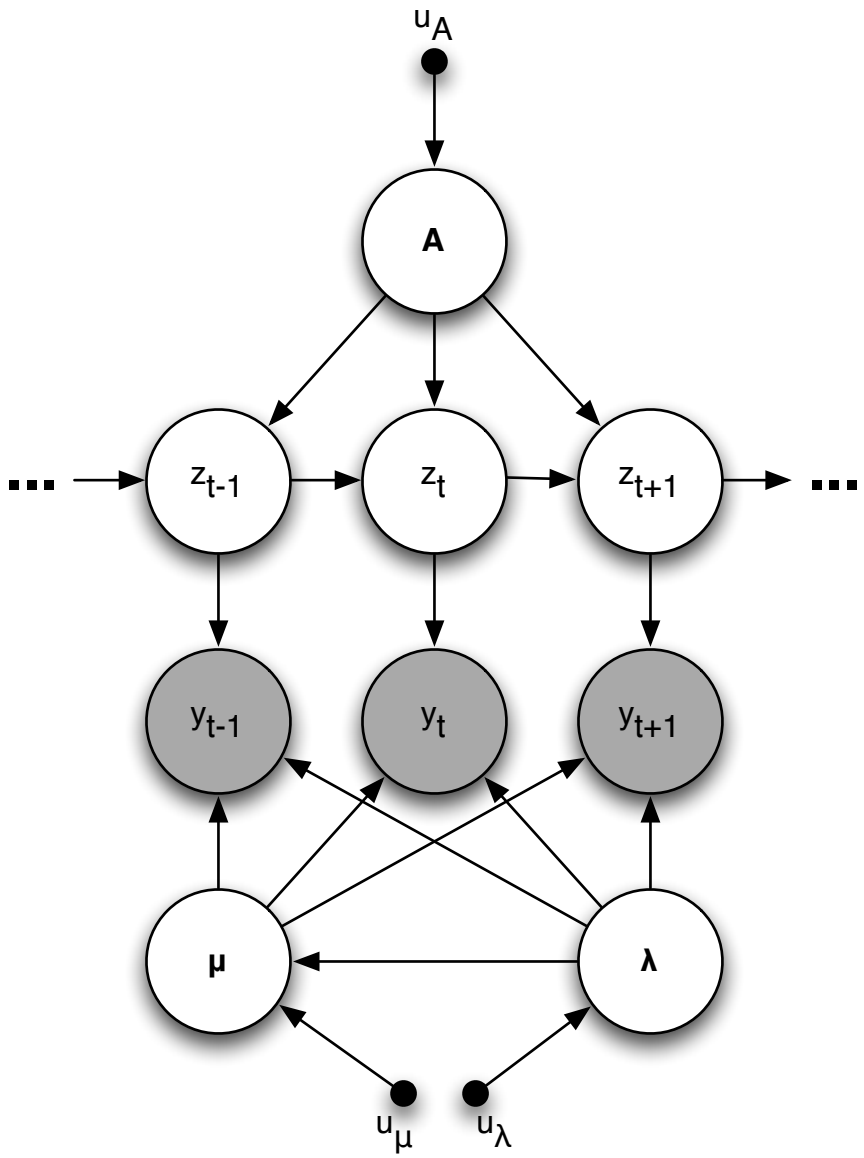


Figure S.9: Graphical model representation of the HMM, corresponding to the factorization of the probability distribution in Eq. 7. Each vertical slice represents a time slice $t = 1, \ldots, T$, for which there is an observed FRET ratio $y_t$, given a hidden conformational state $z_t \in 1, \ldots, K$. Transitions between conformational states are represented by the dependencies between $z_t$ and $z_{t-1}$. Parameters are also shown as random variables, with arrows indicating the dependence of the observed and hidden variables. Parameters for the probability distributions over parameters (Sec. S.6.2) are shown as solid black circles.

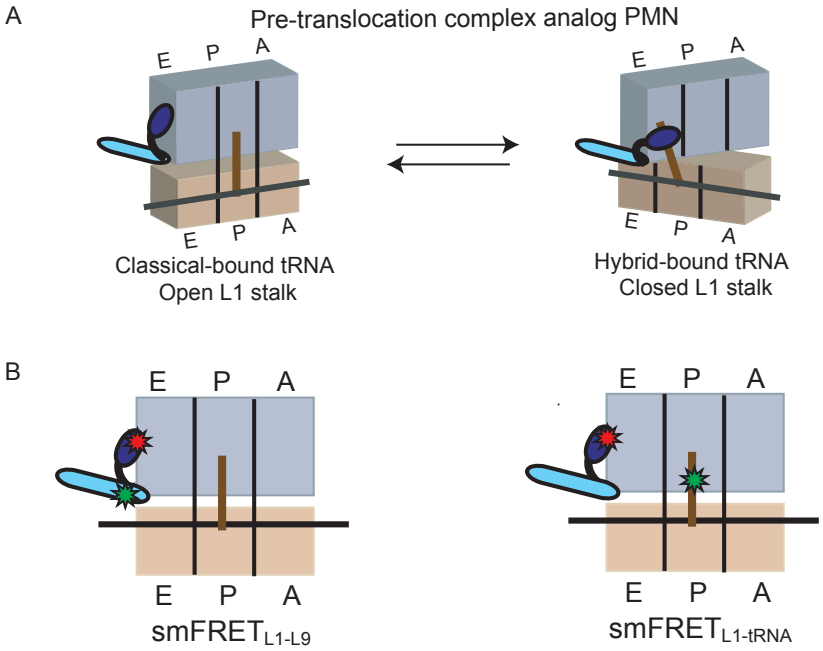


Figure S.10: Conformational rearrangements within a pre-translocation (PMN) complex and smFRET labeling schemes. (A) Cartoon representation of a PMN complex analog. The small and large ribosomal subunits are shown in tan and lavender, respectively, with the L1 stalk depicted in dark blue, and ribosomal protein L9 in cyan. The aminoacyl-, peptidyl- and deacylated-tRNA binding sites are labeled as A, P and E, respectively, and the P-site tRNA is depicted as a brown line. PMN complex analogs are generated by adding the antibiotic puromycin to a post-translocation complex carrying a deacylated-tRNA at the E site and a peptidyl-tRNA at the P site. The resulting PMN complex analog exists in a thermally-driven dynamic equilibrium between two major conformational states in which the P-site tRNA fluctuations between classical and hybrid configurations correlate with the L1 stalk fluctuations between open and closed conformations. (B) Two labeling schemes have been developed in order to investigate PMN complex dynamics using smFRET. PMN complexes are cartooned as in (A) with Cy3 and Cy5 depicted as green and red stars, respectively. $\text{smFRET}_{\text{L1}-\text{L9}}$ (left), which involves a Cy5 label on ribosomal protein L1 within the L1 stalk and a Cy3 label on ribosomal protein L9 at the base of the L1 stalk, reports on the intrinsic conformational dynamics of the L1 stalk. $\text{smFRET}_{\text{L1}-\text{tRNA}}$ (right), which involves a Cy5 label on ribosomal protein within the L1 stalk as in $\text{smFRET}_{\text{L1}-\text{L9}}$ and a Cy3 label on the P-site tRNA, reports on the formation and disruption of a direct interaction between the closed L1 stalk and the hybrid bound P-site tRNA.

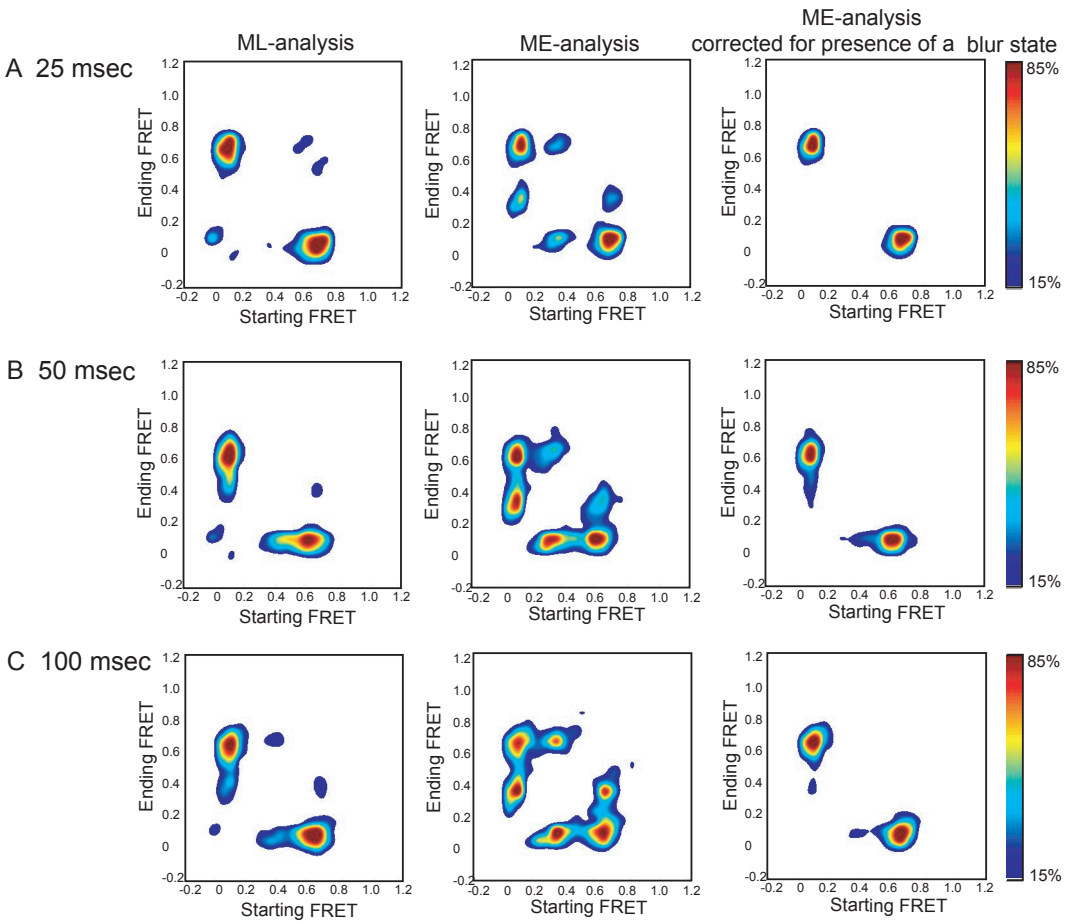


Figure S.11: Transition density plots (TDP) of smFRET$_{\mathrm{L1-tRNA}}$ PMN$_{\mathrm{fMet+EF-G}}$ derived from ME and ML analysis with different CCD integration times. TDPs are contour plots showing the kernel density estimation of the transitions in idealized traces (with starting and ending FRET values of the transitions as the X and Y axes, respectively). Note that transitions to short-lived or nearby states count with equal weight as those to long-lived states in a TDP. This should not be confused with a time-density plot, which illustrates the probability of observing a pair of experimental values at two different times $p(y(t), y(t + \delta t))$, which can be made from the FRET data themselves without appealing to statistical inference. The plots show ML (left), ME (middle) and ME analysis corrected for the presence of a blur state (right) Contours are plotted from tan (lowest population) to red (highest population). Different CCD integration times were used for recoding these data sets: (A) 25 msec, (B) 50 msec, and (C) 100 msec. For interpretation of the significance of these TDPs, *cf.* Sec. 5.