Supplementary Material for: Bayesian hidden Markov model analysis of single-molecule force spectroscopy: Characterizing kinetics under measurement uncertainty

John D. Chodera,[1,*] Phillip Elms,[1,2,3] Frank Noé,[4] Bettina Keller,[4] Christian M. Kaiser,[1,5] Aaron Ewall-Wice,[6] Susan Marqusee,[1,7,3] Carlos Bustamante,[1,7,3,5,8,9] and Nina Singhal Hinrichs[10]

[1]*California Institute of Quantitative Biosciences (QB3), University of California, Berkeley, CA 94720, USA*
[2]*Biophysics Graduate Group, University of California, Berkeley, CA 94720, USA*
[3]*Jason L. Choy Laboratory of Single Molecule Biophysics, Institute for Quantitative Biosciences, University of California, Berkeley, CA 94720, USA*
[4]*DFG Research Center Matheon, FU Berlin, Arnimallee 6, 14195 Berlin, Germany*
[5]*Department of Physics, University of California, Berkeley, CA 94720, USA*
[6]*University of Chicago, IL 60637, USA*
[7]*Department of Molecular & Cell Biology, University of California, Berkeley, CA 94720, USA*
[8]*Department of Chemistry, University of California, Berkeley, CA 94720, USA*
[9]*Howard Hughes Medical Institute, University of California, Berkeley, CA 94720, USA*
[10]*Departments of Statistics and Computer Science, University of Chicago, IL 60637, USA*

(Dated: June 11, 2014)

VALIDATION

The confidence interval comparison figure referred to in the main body text appears as Supplementary Figure 1.

PROOF OF STATE HISTORY SAMPLING SCHEME

In the reverse recursion for updating the hidden state sequence in sampling from the BHMM posterior by Gibbs sampling, we *sample* a state sequence by sampling each hidden state from the conditional distribution $s_t \sim P(s_t \mid s_{t+1}, \ldots, s_L)$ starting from $t = L$ and proceeding down to $t = 0$, where the conditional distribution is given by (Eq. 35 in the main text),

$$P(s_t = i \mid s_{t+1}, \ldots, s_L) \propto \begin{cases} \alpha_{ti} / \sum_{j=1}^{M} \alpha_{tj} & t = L \\ \alpha_{ti} T_{i s_{t+1}} / \sum_{j=1}^{M} \alpha_{tj} T_{j s_{t+1}} & t = (L-1), \ldots, 0 \end{cases} \tag{1}$$

It is straightforward to show the result of these sampling steps reconstitutes the probability distribution $P(\mathbf{S}|\mathbf{T}, \mathbf{E}, \mathbf{O})$. We note that the composition of these conditional distributions defines a joint distribution,

$$\begin{aligned} &P(s_0|s_1, \ldots)P(s_1|s_2, \ldots)\cdots P(s_L) = \\ &= \left[\frac{\alpha_{0 s_0} T_{s_0 s_1}}{\sum_{j=1}^{M} \alpha_{0j} T_{j s_1}}\right]\left[\frac{\alpha_{1 s_1} T_{s_1 s_2}}{\sum_{j=1}^{M} \alpha_{1j} T_{j s_2}}\right]\cdots \\ &= \left[\frac{\rho_{s_0}\varphi(o_0|\mathbf{e}_{s_0}) T_{s_0 s_1}}{\sum_{j=1}^{M} \alpha_{0j} T_{j s_1}}\right]\left[\frac{\varphi(o_1|\mathbf{e}_{s_1})\left(\sum_{j=1}^{M} \alpha_{0j} T_{j s_1}\right) T_{s_1 s_2}}{\sum_{j=1}^{M} \alpha_{1j} T_{j s_2}}\right]\cdots \\ &= \frac{\rho_{s_0}\varphi(o_0|\mathbf{e}_{s_0}) T_{s_0 s_1}\cdots\varphi(o_L \mid \mathbf{e}_{s_L})}{\sum_{j=1}^{M} \alpha_{Lj}} \\ &= \frac{\rho_{s_0}\varphi(o_0|\mathbf{e}_{s_0}) T_{s_0 s_1}\cdots\varphi(o_L \mid \mathbf{e}_{s_L})}{P(\mathbf{O}|\mathbf{E})} \end{aligned} \tag{2}$$

where we have used the fact that $P(\mathbf{O} \mid \mathbf{E}) = \sum_{j=1}^{M} \alpha_{tj}$,

$$\begin{aligned} \sum_{j=1}^{M} \alpha_{Lj} &= \sum_{s_L=1}^{M} \varphi(o_L|\mathbf{e}_{s_L}) \sum_{s_{L-1}=1}^{M} \alpha_{(t-1)s_{L-1}} T_{s_{L-1}s_L} \\ &= \sum_{s_{L-1}=1}^{M} \alpha_{(t-1)s_{L-1}} \sum_{s_L=1}^{M} \varphi(o_L|\mathbf{e}_{s_L}) T_{s_{L-1}s_L} \\ &= \sum_{s_0} \rho_{s_0}\varphi(o_0 \mid \mathbf{e}_{s_0}) T_{s_0 s_1}\cdots\varphi(o_L \mid \mathbf{e}_L) \\ &= \sum_{\mathbf{S}} \rho_{s_0}\varphi(o_0 \mid \mathbf{e}_{s_0}) \prod_{t=1}^{L} T_{s_{(t-1)}s_t}\varphi(o_t \mid \mathbf{e}_{s_t}) \\ &= P(\mathbf{O} \mid \mathbf{\Theta}). \end{aligned} \tag{3}$$

where the last step is seen by recursive substitution of the α_{tj}.

COMPARISON WITH THRESHOLD MODEL

To demonstrate that simply dividing the observable into non-overlapping regions with thresholds to define states leads to consistently biased model properties, we tested a simple segmentation scheme on the same three-state synthetic example considered in the main text. Here, a mixture of Gaussians was first fit to the pooled observations (here, forces) according to the scheme described in *Methods* to identify the peaks corresponding to each state. A simple partition (or threshold) was assigned to the crossing point between each Gaussian probability density. This threshold was used to determine changes between states and therefore transitions and lifetimes. A maximum-likelihood transition matrix was then fit to the observed transition counts between states.

The result of this segmentation procedure is shown for trajectories of 10^3 to 10^6 observations in Figure 2. Even by eye, it is clear that there is a discrepancy between changes in state assignment color and changes in kinetic behavior. Small conformational fluctuations that cause the force to transiently cross a threshold, but not commit to the new conformational state, result in the obsevation of spurious extra transitions. These

extra transitions will lead to an underestimate of the state self-transition probabilities and lifetimes, and convergence to incorrect model parameters. This is confirmed by examination of the estimated model properties as a function of trajectory length in Table I. Compare with Table I from the main text, where by 10^5 observations, the estimated mean BHMM model parameters have attained relatively small error compared to the error in the segmentation model even for 10^6 observations.

CHOICE OF OBSERVATION INTERVAL FOR HAIRPIN

To determine that a 1 ms time was an appropriately Markovian time for the observation interval Δt used in HMM analysis, the autocorrelation function of the experimental force trace was computed at the full 50 kHz time resolution for a trap position where only a single conformational state appeared to dominate (Figure 3). An initial nonexponential or multiexponential (with unclear number of components) decay is seen within the first 0.5 ms to approximately one tenth of the initial value, with slow decay following. We therefore chose a 1 ms time resolution as appropriate for describing the slow conformational dynamics between states.

FIGURES

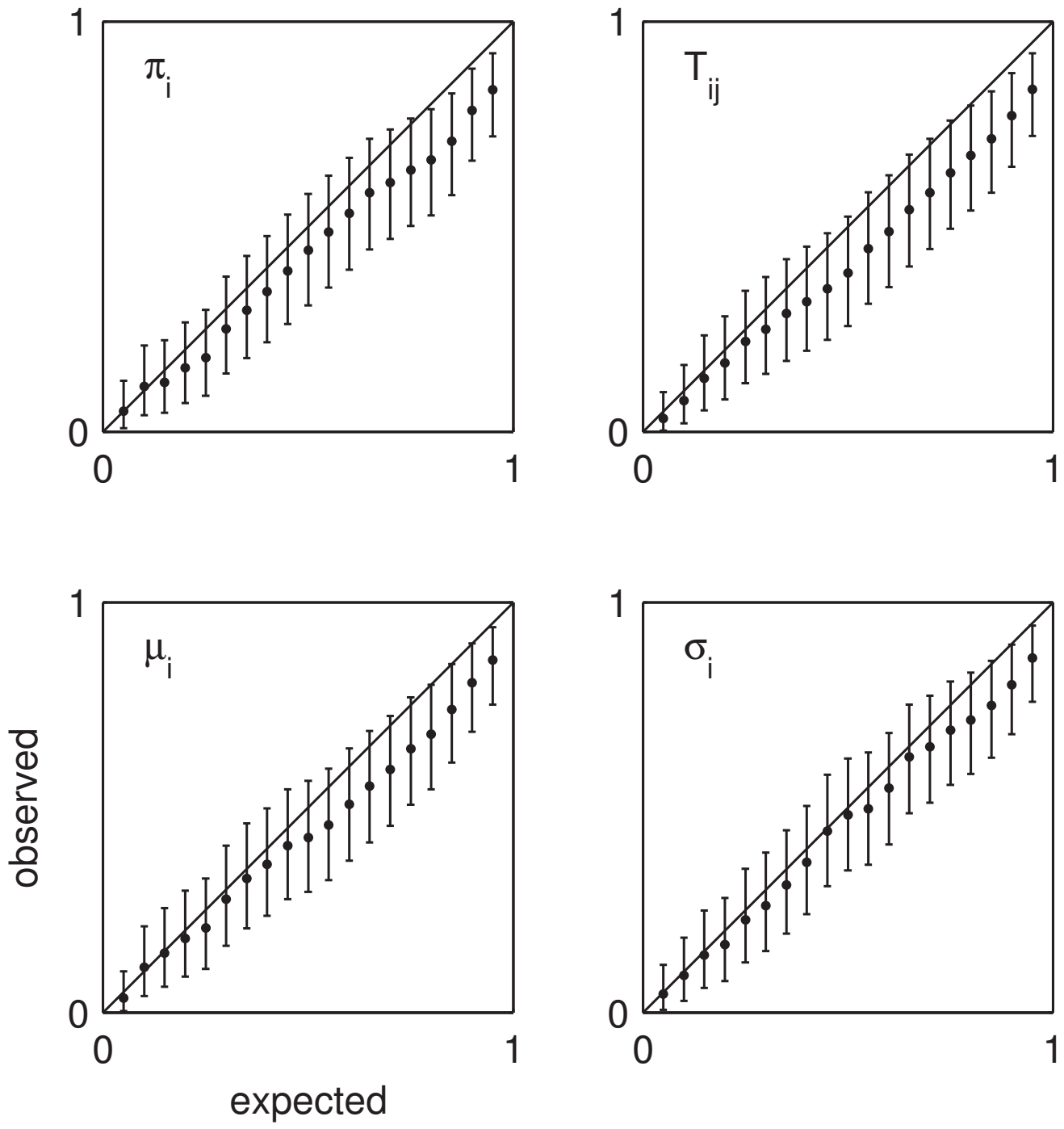

FIG. 1. **Validation of confidence intervals using randomly-generated synthetic test data.** Observed confidence intervals (points) are plotted as a function of the desired confidence intervals for equilibrium probabilities (π_i), transition probabilities (T_{ij}), state means (μ_i), and state standard deviations (σ_i). The black diagonal line indicates perfect agreement between expected and observed confidence intervals, while observed confidence intervals above the diagonal indicate overestimates of the uncertainty, and below the diagonal indicate underestimates. Because only 50 random models were evaluated, error bars denote a 95% confidence interval in the estimated observed confidence intervals.

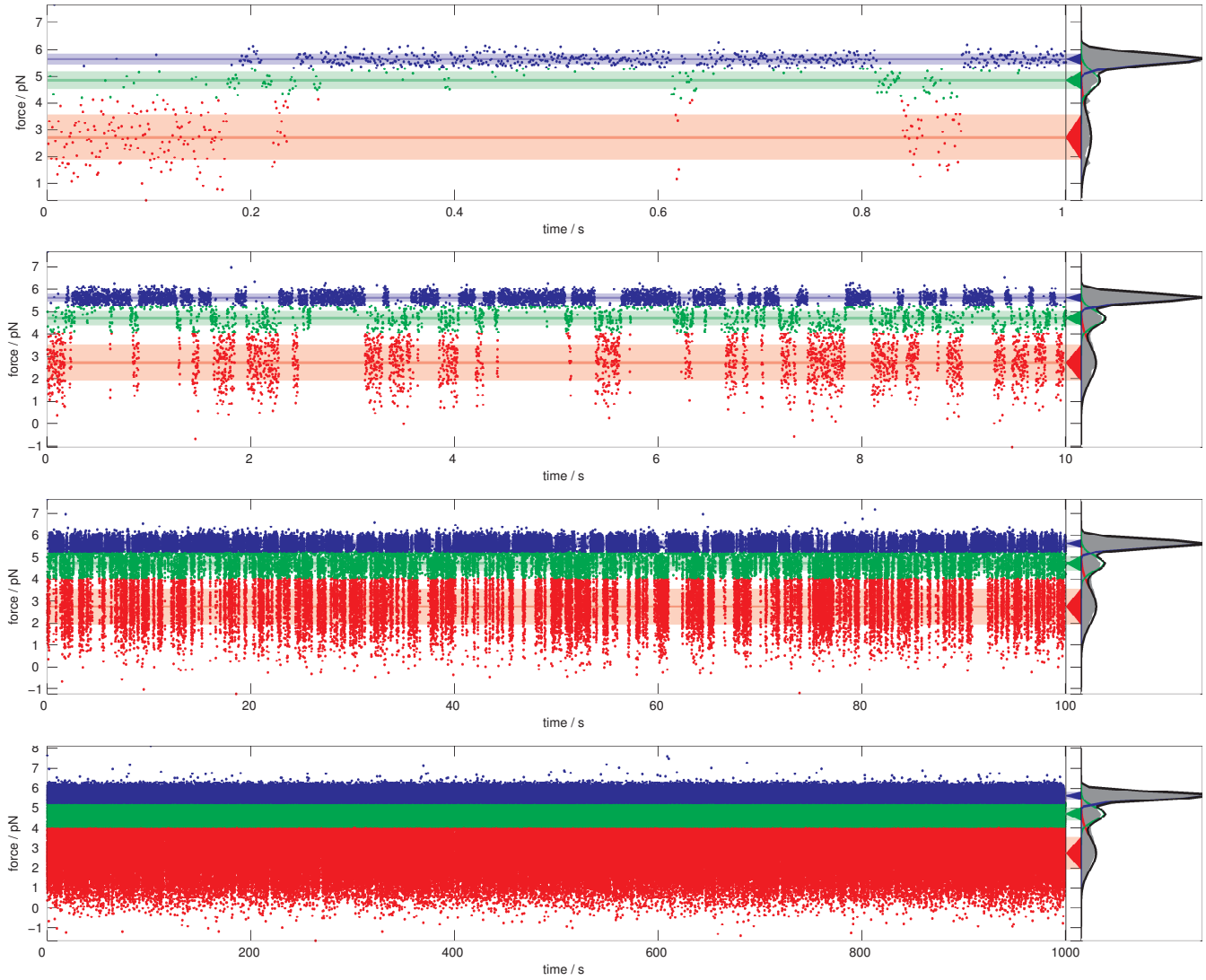

FIG. 2. **Synthetic force trajectory and segmentation state assignments.** Observed samples are colored by their hidden state assignments from a model in which the observable is segmented into regions that correspond to each state, for trajectories of length 10^3 (*top*), 10^4 (*upper middle*), 10^5 (*lower middle*), and 10^6 (*bottom*) observations. The segmentation was derived by first fitting the set of all observed forces to a mixture of Gaussians and then dividing the observations at the crossing points of the Gaussian mixture component probability density functions. Dark horizontal lines terminating in triangles to the right denote state means, while lightly colored bands indicate one standard deviation on either side of the state mean. The gray histogram on the right side shows the total observed probability of samples, while the colored peaks show the contribution from each state, assuming a Gaussian fit to the mean and standard deviation.

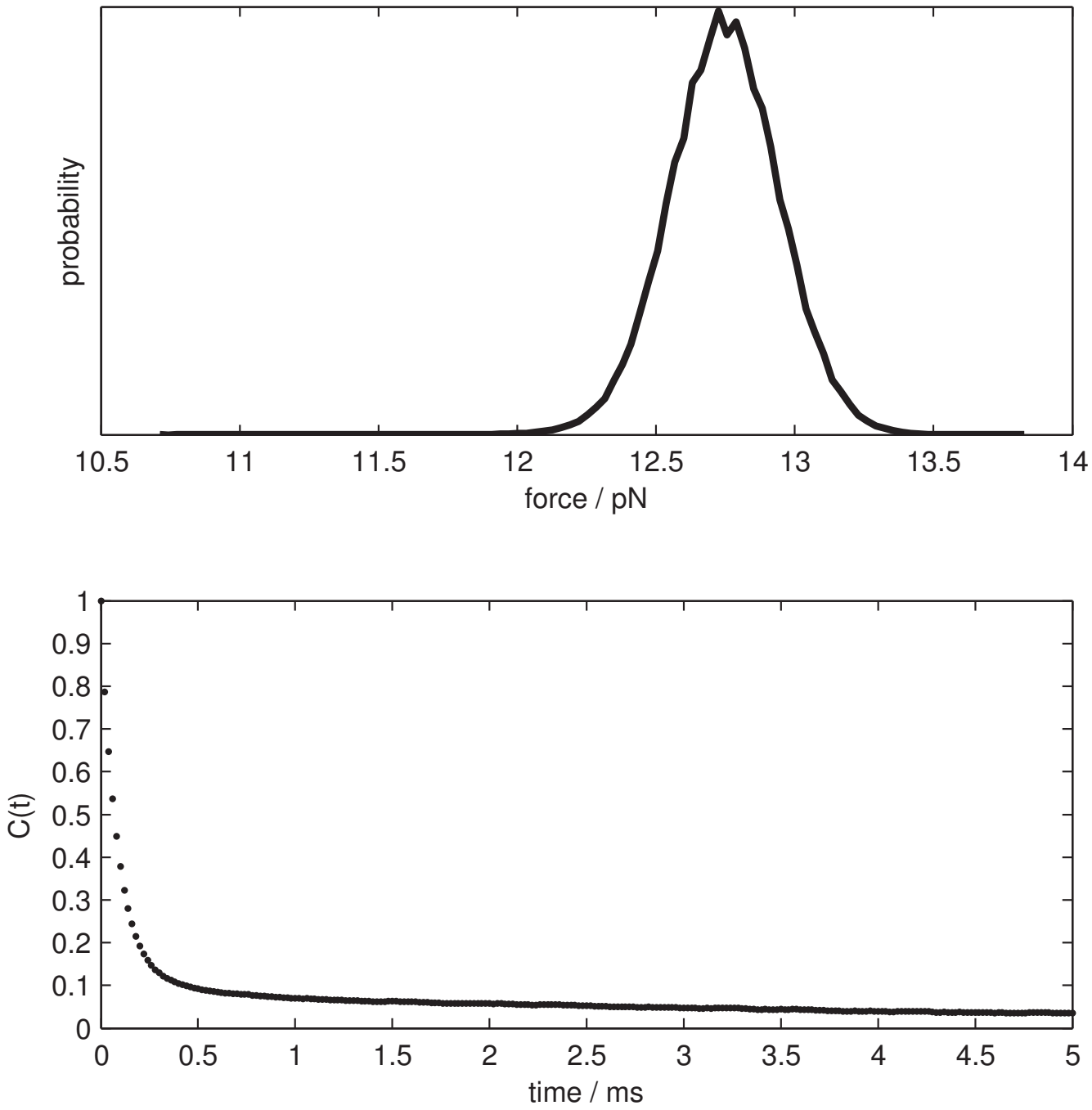

FIG. 3. **Normalized fluctuation autocorrelation function for p5ab hairpin under conditions stabilizing a single conformation.** *Top:* The force histogram for a trap position (different from that in main text Fig. 4) showing that a single conformational state is dominant at this trap configuration. *Bottom:* The normalized integrated autocorrelation function for the force trace, showing rapid decay of non-Markovian contributions within 1 ms.

TABLES

TABLE I. **Estimated segmentation model parameters for synthetic timeseries data**

property	true value	observation length			
		10^3	10^4	10^5	10^6
π_1	0.308	0.214	0.276	0.278	0.267
π_2	0.113	0.155	0.184	0.178	0.174
π_3	0.579	0.632	0.540	0.544	0.559
T_{11}	0.980	0.859	0.822	0.828	0.833
T_{12}	0.019	0.125	0.165	0.158	0.154
T_{13}	0.001	0.016	0.013	0.014	0.014
T_{21}	0.053	0.173	0.247	0.246	0.236
T_{22}	0.900	0.523	0.579	0.564	0.577
T_{23}	0.050	0.304	0.174	0.190	0.187
T_{31}	0.001	0.005	0.007	0.007	0.006
T_{32}	0.009	0.074	0.059	0.062	0.058
T_{33}	0.990	0.920	0.934	0.930	0.935
μ_1	3.000	2.707	2.716	2.746	2.737
μ_2	4.700	4.841	4.707	4.714	4.710
μ_3	5.600	5.636	5.621	5.619	5.618
σ_1	1.000	0.842	0.804	0.806	0.806
σ_2	0.300	0.330	0.332	0.339	0.334
σ_3	0.200	0.201	0.187	0.182	0.183