Supplementary material for "Splitting probabilities as a test of reaction coordinate choice in single-molecule experiments"

John D. Chodera[1, *] and Vijay S. Pande[2, †]

[1] *California Institute of Quantitative Biosciences (QB3), University of California, Berkeley, CA 94720*
[2] *Department of Chemistry, Stanford University, Stanford, CA 94305*

(Dated: July 13, 2011)

EMPIRICAL SPLITTING PROBABILITY COMPUTATION

Computation of the empirical splitting probability $\hat{p}_A(x)$ (defined in Eq. 3 of the main text) along with associated statistical error $\delta\hat{p}_A(x)$ proceeded using the following scheme based on timeseries autocorrelation analysis [1, 2]. First, the observed range of x seen in the discrete-time trajectory $x_t \equiv x(n\Delta t)$ was computed as $[x_{\min}, x_{\max}]$, and this interval was split into 100 histogram bins of equal width. A histogram function $\phi_i(x)$ was defined for each bin, assuming the value of unity if x is in bin i, and zero otherwise. For each bin i, two new timeseries are defined:

$$\begin{aligned} a_t &\equiv \phi_i(x_t)\, c_A(t) \\ b_t &\equiv \phi_i(x_t) \end{aligned} \tag{1}$$

where $c_A(t)$ is the discrete-time analog of the hitting function defined in Eq. 5 of the main text.

The means of the new timeseries were computed:

$$\hat{A} \equiv \frac{1}{T}\sum_{t=1}^{T} a_t \;\; ; \;\; \hat{B} \equiv \frac{1}{T}\sum_{t=1}^{T} b_t \tag{2}$$

and the sample covariances

$$\begin{aligned} \hat{\sigma}_A^2 &\equiv \frac{1}{T}\sum_{t=1}^{T}(a_t - \hat{A})^2 \\ \hat{\sigma}_B^2 &\equiv \frac{1}{T}\sum_{t=1}^{T}(b_t - \hat{B})^2 \\ \hat{\sigma}_{AB}^2 &\equiv \frac{1}{T}\sum_{t=1}^{T}(a_t - \hat{A})(b_t - \hat{B}) \end{aligned} \tag{3}$$

from which the squared statistical errors were computed

$$\begin{aligned} \delta^2\hat{A} &\equiv \frac{\hat{\sigma}_A^2}{T/g_A} \\ \delta^2\hat{B} &\equiv \frac{\hat{\sigma}_B^2}{T/g_B} \\ \delta\hat{A}\delta\hat{B} &\equiv \frac{\hat{\sigma}_{AB}^2}{T/g_{A;B}} \end{aligned} \tag{4}$$

where the statistical inefficiencies g_A, g_B, and $g_{A;B}$ were computed as described in Ref. [3].

Finally, the empirical splitting probability $\hat{p}_A(x)$ and the statistical uncertainty $\delta\hat{p}_A(x)$ for bin i are computed using standard propagation of error (as in Ref. [3]), obtaining

$$\begin{aligned} \hat{p}_A(x) &= \frac{\hat{A}}{\hat{B}} \\ \delta\hat{p}_A(x) &= \left[\frac{\hat{A}}{\hat{B}}\right]^2 \left[\frac{\delta^2\hat{A}}{\hat{A}^2} + \frac{\delta^2\hat{B}}{\hat{B}^2} - 2\frac{\delta\hat{A}\delta\hat{B}}{\hat{A}\hat{B}}\right] \end{aligned} \tag{5}$$

Complete trajectories, histograms, estimated PMFs, and empirical and PMF-derived splitting probabilities for the model system data are shown in Supplementary Fig. 1.

COORDINATE-DEPENDENT DIFFUSION

To compute position-dependent diffusion constants, the scheme of Best and Hummer was used [4]. The region in between absorbing boundary conditions was discretized into 50 bins of equal size in the corresponding coordinate, and an initial guess at the rate matrix made. The number of statistically independent transition counts N_{ij} between each pair of bins i and j for an observation interval of one sampling time was computed by dividing the total number of observed transitions in the trajectory x_t, $t = 0, \ldots, T$, by the statistical inefficiency g_i of the timeseries of the corresponding indicator function for histogram bin i,

$$b_t^{(i)} \equiv \phi_i(x_t) \;\; , \;\; t = 0, \ldots, T \tag{6}$$

The Bayesian posterior was then sampled from by choosing to perturb either $K_{i,i+1}$ or $K_{i,i-1}$ with equal probability. Perturbations were of the form

$$\begin{aligned} K'_{ij} &= K_{ij} e^{(2\Delta - 1)} \\ K_{ii} &= K_{ii} + K_{ij} - K'_{ij} \end{aligned} \tag{7}$$

where Δ is a random number drawn uniformly on the interval $[0, 1]$. An initial 20 000 sampling iterations were discarded to burn-in, followed by another 20 000 production sampling iterations, with samples stored every 1 000 iterations.

For each transition matrix sample $\mathbf{K}$, the potential of mean force and diffusion constant were computed as describes in Ref. [4],

$$\begin{aligned} F_i &= -k_B T \ln \frac{\pi_i}{\delta x} \\ D_i &= \frac{(\delta x)^2}{2}\left[K_{(i-1)i}\left(\frac{\pi_{i-1}}{\pi_i}\right)^{1/2} + K_{i(i+1)}\left(\frac{\pi_i}{\pi_{i+1}}\right)^{1/2}\right] \end{aligned} \tag{8}$$

where δx is the histogram bin width and p_i is the stationary probability distribution computed from $\mathbf{T} \equiv e^{\mathbf{K}\tau}$ such that $\pi^{\mathrm{T}}\mathbf{T} = \pi^{\mathrm{T}}$. We note that the diffusion constant D_i is an interpolation of the diffusion constants $D_{i-1/2}$ and $D_{i+1/2}$ defined in Ref. [4], in order to provide an estimate of the diffusion constant in bin i. For each sampled set of $\{F_i, D_i\}$, the splitting probability was computed using Eq. 2 of the main text.

This analysis was performed for both the model system described in the main text (results shown in Supplementary Fig. 2) and the DNA hairpin data described in the main text (Supplementary Fig. 3, with some panels reproduced in the main text as Fig. 4).

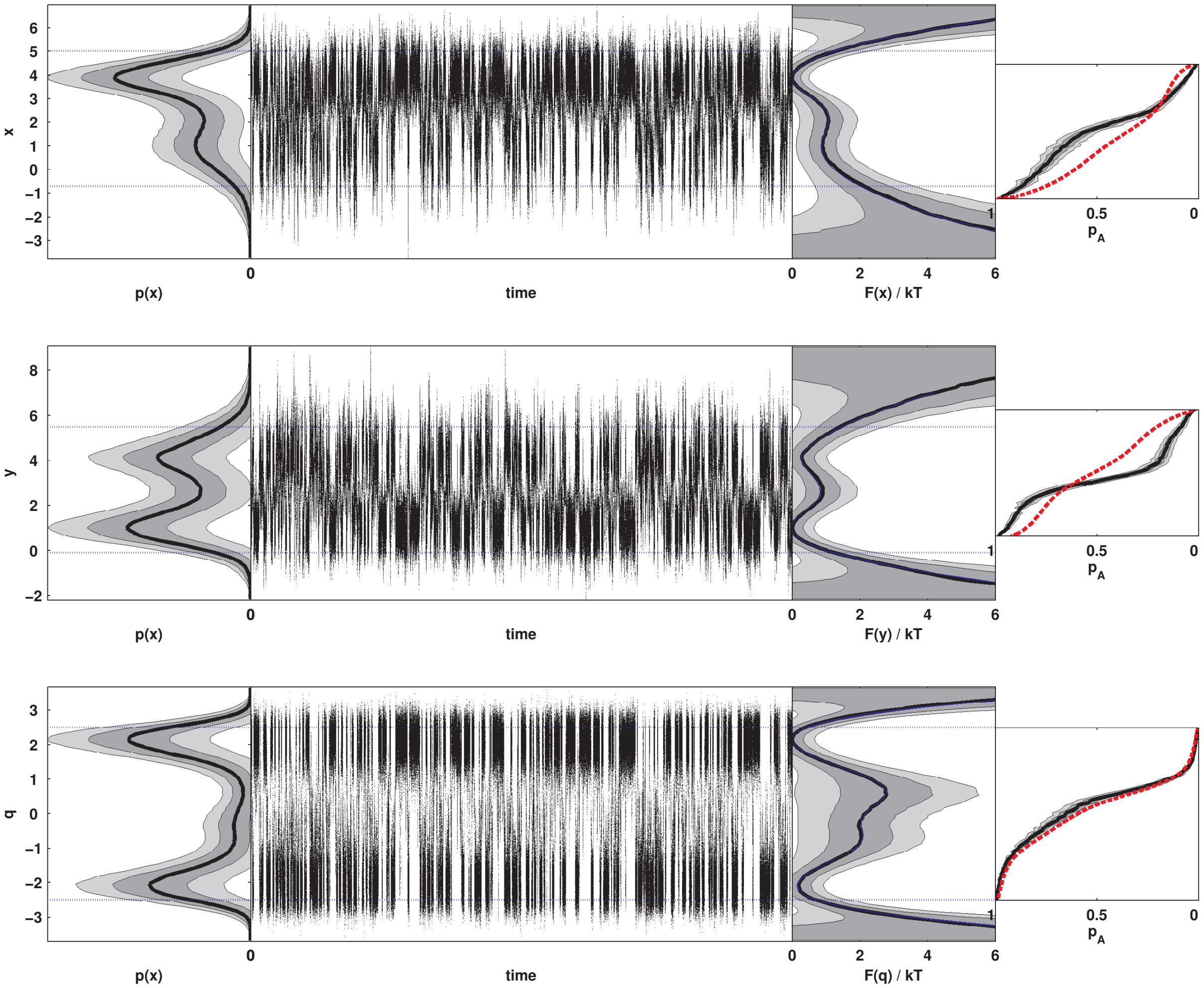

FIG. 1. **Splitting probability tests for poor and good choices of reaction coordinate in two-dimensional model system.** Splitting probability analysis for poor (top and middle) and good (bottom) choices of reaction coordinate for the two-dimensional model system depicted in Fig. 1 of the main text. For each figure, the observed trajectory is shown in the large middle panel, with the observed histogram $p(x)$ and derived potential of mean force $F(x)$ flanking it to the left and right, respectively. The rightmost panel shows the coordinate-dependent average splitting probability $p_A(x)$ computed from the PMF and diffusion profile (red dashed line) and empirically estimated from the observed trajectory (black solid line). For all computed quantities, dark black lines represent expectations, dark shading represents a 68% confidence interval, and light shading a 95% confidence interval.

[*] jchodera@berkeley.edu
[†] Corresponding author; pande@stanford.edu

[1] W. C. Swope, H. C. Andersen, P. H. Berens, and K. R. Wilson, J. Chem. Phys., **76**, 637 (1982).

[2] W. Janke, in *Quantum Simulations of Complex Many-Body Systems: From Theory to Algorithms*, Vol. 10, edited by J. Grotendorst, D. Marx, and A. Murmatsu (John von Neumann Institute for Computing, Jülich, Germany, 2002) pp. 423–445.

[3] J. D. Chodera, W. C. Swope, J. W. Pitera, C. Seok, and K. A. Dill, J. Chem. Theor. Comput., **3**, 26 (2007).

[4] R. B. Best and G. Hummer, Proc. Natl. Acad. Sci. USA, **107**, 1088 (2010).

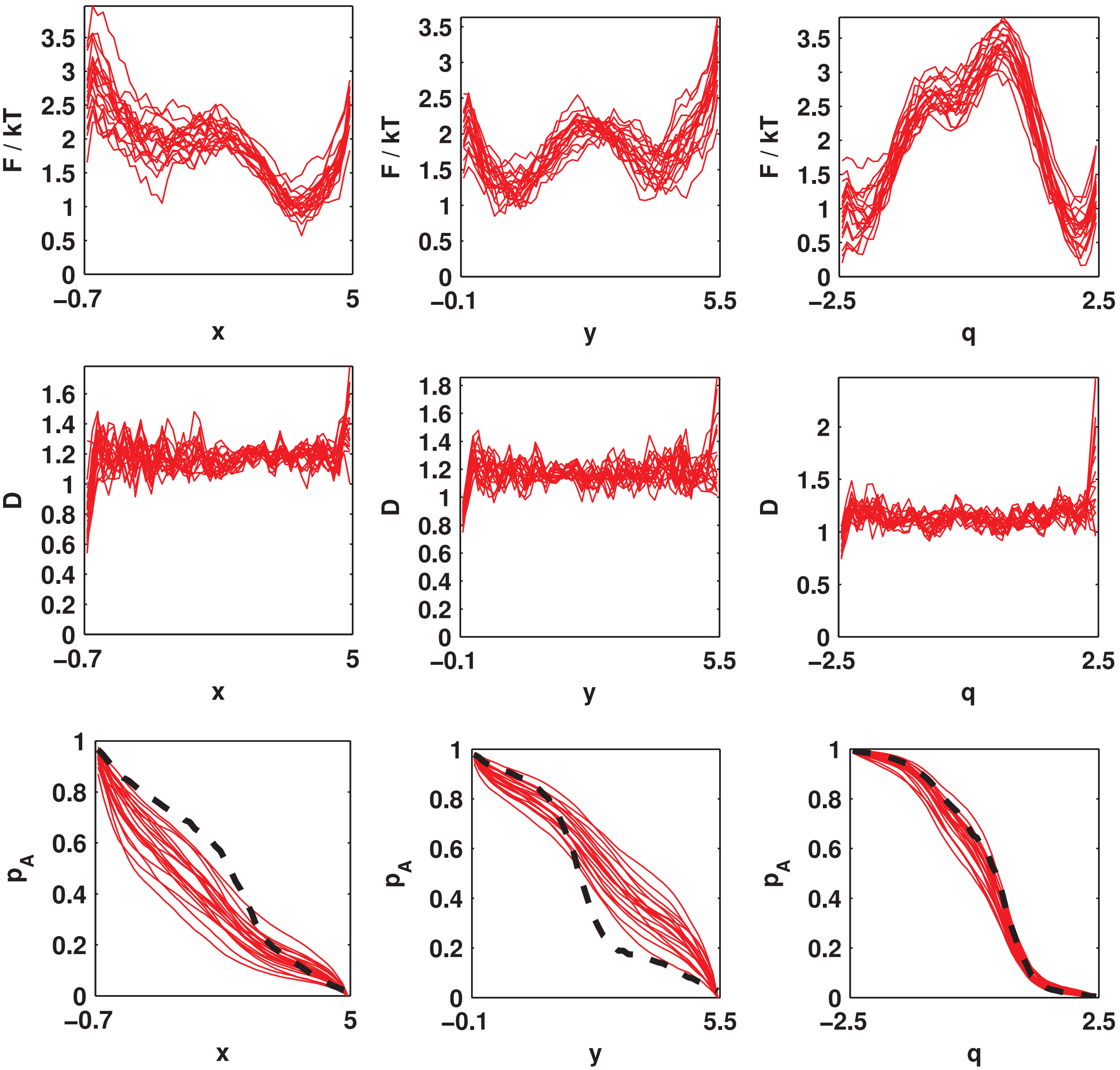

FIG. 2. **Splitting probability tests for poor and good choices of reaction coordinate in two-dimensional model system, incorporating position-dependent diffusion.** Splitting probability analysis for poor (x, left; y, middle) and good (q, right) choices of reaction coordinate for the two-dimensional model system depicted in Fig. 1 of the main text. Red solid lines depict the 20 splitting probability profiles $p_A(s)$ computed from the PMFs and diffusion constants sampled from the Bayesian posterior, while the thick dashed black line depicts the empirical splitting probability profile $\hat{p}_A$.

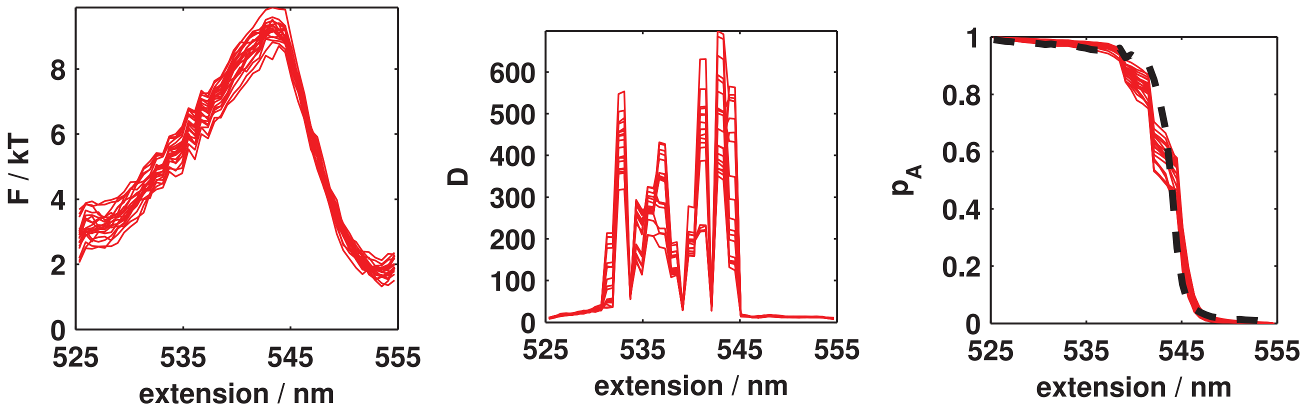

FIG. 3. **PMF, position-dependent diffusion constant, and splitting probability test incorporating position-dependent diffusion for DNA hairpin.** All three plots were generated using the Best-Hummer Bayesian sampling scheme to sample the diffusion constant and PMF from the Bayesian posterior. *Left:* Potential of mean force (PMF) in extension coordinate, in nm. *Middle:* Position-dependent diffusion constant, in nm/s^2. *Right:* Splitting probability test in incorporating position-dependent diffusion constant, with red solid lines denoting the splitting probability p_A computed from the PMF and diffusion constant, and the dashed black line denoting the empirical splitting probability $\hat{p}_A$.