# Supporting Information for Long time-scale predictions from short-trajectory data: A benchmark analysis of the trp-cage mini-protein

John Strahan,[†] Adam Antoszewski,[†] Chatipat Lorpaiboon,[†] Bodhi P. Vani,[†] Jonathan Weare,[*,‡] and Aaron R. Dinner[*,†]

[†]*Department of Chemistry, University of Chicago, Chicago, IL 60637, USA*

[‡]*Courant Institute of Mathematical Sciences, New York University, New York, New York 10012, USA*

E-mail: weare@cims.nyu.edu; dinner@uchicago.edu

## A Backward-in-time inner products

In this appendix we provide an elementary derivation of (24) which is key to our estimates of inner products involving the stopped backward-in-time transition operator $\mathcal{T}^t_{A\cup B}$. For the purposes of this derivation we assume that $X(t)$ is a discrete time process (so that $t$ is a non-negative integer) with probability density $p(x(1)|x(0))$ for transition from $x(0)$ to $x(1)$ and stationary density $\pi$. The steady state backward-in-time process $X(-t)$ then has transition density

$$q(x(0)|x(1)) = \frac{p(x(1)|x(0))\pi(x(0))}{\pi(x(1))}. \tag{S1}$$

From this expression we immediately find that

$$\begin{aligned}&\pi(x(t))q(x(t-1)|x(t))q(x(t-2)|x(t-1))\cdots q(x(0)|x(1))\\&\qquad=\pi(x(0))p(x(t)|x(t-1))p(x(t-1)|x(t-2))\cdots p(x(1)|x(0))\end{aligned}\tag{S2}$$

relating the steady state backward-in-time path density to the steady state forward-in-time path density. Therefore, for any path function $F(x(0), x(1), \ldots, x(t))$ and any density $\mu$ (equivalent to $\pi$) we find (recalling that here $w = \pi/\mu$) that

$$\begin{aligned}&\int \mathbb{E}\left[F(X(0), X(-1), \ldots, X(-t)) \,|\, X(0) = x\right] \mu(x)dx\\&\qquad= \int \frac{F(x(0), \ldots, x(-t))}{w(x(0))}\pi(x(0))q(x(-1)|x(0))\cdots q(x(-t)|x(-t+1))dx(0)\cdots dx(-t)\\&\qquad= \int \frac{F(x(t), \ldots, x(0))}{w(x(t))}\pi(x(t))q(x(t-1)|x(t))\cdots q(x(0)|x(1))dx(t)\cdots dx(0)\\&\qquad= \int \mathbb{E}\left[\frac{F(X(t), X(t-1), \ldots, X(0))}{w(X(t))}\,\middle|\, X(0) = x\right] w(x)\mu(x)dx.\end{aligned}\tag{S3}$$

We will use (S3) to find an expression for

$$\langle g, \mathcal{T}_{A\cup B}^{-t} f\rangle = \int \mathbb{E}\left[g(x)f(X(-(T_{A\cup B}^{-} \wedge t))) \,|\, X(0) = x\right] \mu(x)dx\tag{S4}$$

in terms of the forward-in-time process. If we choose

$$F(X(0), X(-1), \ldots, X(-t)) = g(X(0))f(X(-(T_{A\cup B}^{-} \wedge t))),\tag{S5}$$

then

$$\langle g, \mathcal{T}_{A\cup B}^{-t} f\rangle = \int \mathbb{E}\left[F(X(0), X(-1), \ldots, X(-t)) \,|\, X(0) = x\right] \mu(x)dx.\tag{S6}$$

In terms of the forward process

$$F(X(t), X(t-1), \ldots, X(0)) = f(X(S_{A\cup B}(t)))g(X(t)),\tag{S7}$$

where we remind the reader of (25):

$$S_{A\cup B}(t) = \max\{s \le t : X(s) \in A \cup B\}$$

with $S_{A\cup B}(t) = 0$ if $X(s) \notin A \cup B$ for all $0 \le s \le t$.

Applying (S3) with this choice of $F$ yields (24):

$$\left\langle g, \mathcal{T}_{A\cup B}^{-t} f \right\rangle = \int \mathbb{E}\left[ f(X(S_{A\cup B}(t))) \frac{g(X(t))}{w(X(t))} \,\middle|\, X(0) = x \right] w(x)\mu(dx).$$

# B A formula for the reactive current

It has been shown[1] that for a diffusion with generator

$$\mathcal{L}f(x) = \sum_j b_j(x)\frac{\partial f}{\partial x_j}(x) + \frac{1}{2}\sum_{ij} a_{ij}(x)\frac{\partial^2 f}{\partial x_i \partial x_j}(x) \tag{S8}$$

the reactive current is the vector field given by

$$\begin{aligned}(J_{AB})_i &= q_+(x)q_-(x)J_i + \\ &\quad \pi(x)q_-(x)\sum_j a_{ij}(x)\frac{\partial q_+}{\partial x_j}(x) - \pi(x)q_+(x)\sum_j a_{ij}(x)\frac{\partial q_-}{\partial x_j}(x),\end{aligned} \tag{S9}$$

where $J$ is the equilibrium current:

$$J_i = \pi(x)b_i(x) - \sum_j \frac{\partial[\pi a_{ij}]}{\partial x_j}(x). \tag{S10}$$

To project the current onto a CV space of interest, we take the dot product with $\nabla\theta$ for any smooth CV $\theta$ and, using the identity

$$J_i \cdot \nabla f(x) = \frac{\pi(x)}{2}\left(\mathcal{L}f(x) - \mathcal{L}_\pi^\dagger f(x)\right), \tag{S11}$$

which follows from direct manipulations, we can write

$$J_{AB} \cdot \nabla\theta(x) = \frac{\pi(x)}{2}\left(q_-(x)\mathcal{L}[q_+\theta](x) - q_+(x)\mathcal{L}^\dagger_\pi[q_-\theta](x)\right) \tag{S12}$$

for $x \in (A \cup B)^{\mathrm{c}}$. This formula is not useful computationally since it still contains a backward-in-time generator. To compute statistics from data, we need to formulate their estimators as expectations against the stationary distribution since this (1) permits the use of the adjoint relation to clear away backward transition operators and (2) is consistent with our reweighting scheme. To this end, we define the projected reactive current as

$$J^\theta_{AB}(s) = \int J_{AB}(x) \cdot \nabla\theta(x)\delta(\theta(x) - s)dx = \lim_{|ds|\to 0} \frac{1}{|ds|} \int_{\{\theta(x)\in ds\}} J_{AB}(x) \cdot \nabla\theta(x)dx, \tag{S13}$$

where $ds \in (A \cup B)^{\mathrm{c}}$ is an infinitesimal region of CV space with $s \in ds$, and $\{x : \theta(x) \in ds\}$ does not intersect $A \cup B$. Using (S12) and the fact that $\mathcal{L}q_+ = 0$ and $\mathcal{L}^\dagger_\pi q_- = 0$ on $(A \cup B)^{\mathrm{c}}$, we have

$$\begin{aligned} J^\theta_{AB}(s) &= \lim_{|ds|\to 0} \frac{1}{|ds|} \int \mathbb{1}_{\{\theta(x)\in ds\}} \frac{\pi(x)}{2}\left(q_-(x)\mathcal{L}[q_+\theta](x) - q_+(x)\mathcal{L}^\dagger_\pi[q_-\theta](x)\right) dx \\ &= \lim_{|ds|\to 0} \frac{1}{|ds|} \int \mathbb{1}_{\{\theta(x)\in ds\}} \frac{\pi(x)}{2}\Big(q_-(x)\mathcal{L}[q_+\theta](x) - q_-(x)\mathcal{L}q_+(x)\theta(x) \\ &\qquad\qquad - q_+(x)\mathcal{L}^\dagger_\pi[q_-\theta](x) + q_+(x)\mathcal{L}^\dagger_\pi q_-(x)\theta(x)\Big)dx. \end{aligned} \tag{S14}$$

Writing this expression in terms of the transition operator and canceling terms, we find that

$$\begin{aligned} J^\theta_{AB}(s) &= \lim_{t,|ds|\to 0} \frac{1}{2t\,|ds|} \int \mathbb{1}_{\{\theta(x)\in ds\}}\pi(x)\Big(q_-(x)\mathcal{T}^t[q_+\theta](x) - q_-(x)\mathcal{T}^t q_+(x)\theta(x) \\ &\qquad\qquad - q_+(x)(\mathcal{T}^t)^\dagger_\pi[q_-\theta](x) + q_+(x)(\mathcal{T}^t)^\dagger_\pi q_-(x)\theta(x)\Big)dx \\ &= \lim_{t,|ds|\to 0} \frac{1}{2t\,|ds|} \int \pi(x)q_-(x)\Big(\mathbb{1}_{\{\theta(x)\in ds\}}\left(\mathcal{T}^t[q_+\theta](x) - \mathcal{T}^t q_+(x)\theta(x)\right) \\ &\qquad\qquad + \left(\mathcal{T}^t[q_+\theta\,\mathbb{1}_{\{\theta\in ds\}}](x) - \mathcal{T}^t[q_+\mathbb{1}_{\{\theta\in ds\}}](x)\theta(x)\right)\Big)dx, \end{aligned} \tag{S15}$$

where the second equality follows from the definition of the adjoint $(\mathcal{T}^t)^\dagger_\pi$.

Expression (S15) for $J^\theta_{AB}(s)$ can be directly translated into an estimator for computing from short-trajectory data:

$$\begin{aligned} J^\theta_{AB}(s) \approx \frac{1}{2t|ds|}\sum_{i=1}^{M} & q_+(X^{(i)}(t))\left(\theta(X^{(i)}(t)) - \theta(X^{(i)}(0))\right) \\ & \times q_-(X^{(i)}(0))\mathbb{1}_{\theta\in ds}(X^{(i)}(0))w(X^{(i)}(0)) \\ + \frac{1}{2t|ds|}\sum_{i=1}^{M} & q_+(X^{(i)}(t))\left(\theta(X^{(i)}(t)) - \theta(X^{(i)}(0))\right) \\ & \times q_-(X^{(i)}(0))\mathbb{1}_{\theta\in ds}(X^{(i)}(t))w(X^{(i)}(0)). \end{aligned} \tag{S16}$$

Finally, without affecting the $t \to 0$ limit, we can stop our trajectories when they exit or enter $A \cup B$, yielding the estimator

$$\begin{aligned} J^\theta_{AB}(s) \approx \frac{1}{2t|ds|}\sum_{i=1}^{M} & q_+(X^{(i)}(t\wedge T_{A\cup B}))\left(\theta(X^{(i)}(t\wedge T_{A\cup B})) - \theta(X^{(i)}(0))\right) \\ & \times q_-(X^{(i)}(0))\mathbb{1}_{\theta\in ds}(X^{(i)}(0))w(X^{(i)}(0)) \\ + \frac{1}{2t|ds|}\sum_{i=1}^{M} & q_+(X^{(i)}(t))\left(\theta(X^{(i)}(t)) - \theta(X^{(i)}(S_{A\cup B}(t)))\right) \\ & \times q_-(X^{(i)}(S_{A\cup B}(t)))\mathbb{1}_{\theta\in ds}(X^{(i)}(t))w(X^{(i)}(0)) \end{aligned} \tag{S17}$$

which, in our experience, outperformed (S16) for larger values of $t$. Note that we could have canceled additional terms in (S15) to yield a more concise estimator. However, we found that the estimator (S17) gave less noisy results.

# C Reactive current on a CV space

We now establish that our projected reactive current gives the flux over surfaces in CV space. We assume that our CVs are smooth and that, for some subset $C^\theta$ of CV space with

smooth boundary, the set $C = \{x : \theta(x) \in C^\theta\}$ contains $A$ and does not intersect $B$. We will establish that for such a subset,

$$\int_{\partial C^\theta} J^\theta_{AB}(s) \cdot n_{C^\theta} d\sigma_{C^\theta} = \int_{\partial C} J_{AB} \cdot n_C d\sigma_C. \tag{S18}$$

Here $n_{C^\theta}$ is the outward pointing normal vector to the boundary $\partial C^\theta$ of $C^\theta$, $n_C$ is the normal vector to the boundary $\partial C$ of $C$, $\sigma_{C^\theta}$ is the surface measure on $\partial C^\theta$ and, $\sigma_C$ is the surface measure on $\partial C$. The significance of (S18) is that it shows that our definition of $J^\theta_{AB}$ preserves reactive flux across surfaces in the CV space so that statistics of reactive paths could, in principle, be computed directly from $J^\theta_{AB}$.

Let $f_\delta$ be a smooth function on CV space that is equal to 1 on $C^\theta$ and equal to 0 for $x$ a distance of more than $\delta$ from $C^\theta$. Applying the divergence theorem and integrating by parts we find that

$$\begin{aligned}
\int_{\partial C^\theta} J^\theta_{AB}(s) \cdot n_{C^\theta} d\sigma_{C^\theta} &= \int_{C^\theta} \mathrm{div} J^\theta_{AB}(s) ds \\
&= \lim_{\delta \to 0} \int f_\delta(s) \, \mathrm{div} J^\theta_{AB}(s) ds \\
&= -\lim_{\delta \to 0} \int J^\theta_{AB}(s) \cdot \nabla f_\delta(s) ds.
\end{aligned} \tag{S19}$$

Inserting our definition of $J^\theta_{AB}$ we find that

$$\begin{aligned}
\int_{\partial C^\theta} J^\theta_{AB}(s) \cdot n_{C^\theta} d\sigma_{C^\theta} &= -\lim_{\delta \to 0} \sum_j \int \int J_{AB}(x) \cdot \nabla \theta_j(x) \delta(\theta(x) - s) \frac{\partial f_\delta(s)}{\partial s_j} dx ds \\
&= -\lim_{\delta \to 0} \sum_j \int J_{AB}(x) \cdot \nabla \theta_j(x) \frac{\partial f_\delta}{\partial s_j}(\theta(x)) dx.
\end{aligned} \tag{S20}$$

Using the chain rule the last expression can be rewritten as

$$\int_{\partial C^\theta} J^\theta_{AB}(s) \cdot n_{C^\theta} d\sigma_{C^\theta} = -\lim_{\delta \to 0} \int J_{AB}(x) \cdot \nabla f_\delta(\theta(x)) dx. \tag{S21}$$

Integrating by parts, taking the $\delta \to 0$ limit, and applying the divergence theorem again yields (S18).

# D Rate and Current Estimators for Long Lag Times

The estimators for the reaction rate and the reactive current presented in the main text incur significant bias at longer lag times. Here, we derive estimators which at all lag times converge to the TPT quantities with perfect sampling, committors, and change of measure.

We introduce the notation

$$S^+_{A\cup B}(t) = \min\{s \geq t : X(s) \in A \cup B\} \tag{S22}$$

$$S^-_{A\cup B}(t) = \max\{s \leq t : X(s) \in A \cup B\} \tag{S23}$$

for forward and backward stopping times starting at time $t$.

For the reaction rate estimator, we start with equation (32). Because $X(0)$ is integrated over the stationary distribution, we can shift the term inside the expectation by an arbitrary time $s$. Averaging over $0 \leq s < \tau$ then yields

$$\begin{aligned}
R_{AB} &= \lim_{t\to 0} \frac{1}{t} \int (\mathcal{T}^t q_+^2(x) - q_+(x)\mathcal{T}^t q_+(x)) q_-(x)\pi(dx) \\
&= \lim_{t\to 0} \frac{1}{t} \int \mathbb{E}[q_+(X(t))(q_+(X(t)) - q_+(X(0)))q_-(X(0)) \mid X(0) = x]\pi(dx) \\
&= \lim_{t\to 0} \frac{1}{t} \int \mathbb{E}[q_+(X(s+t))(q_+(X(s+t)) - q_+(X(s)))q_-(X(s)) \mid X(0) = x]\pi(dx) \\
&= \lim_{t\to 0} \frac{1}{t} \int \frac{1}{\tau} \int_0^\tau \mathbb{E}[q_+(X(s+t))(q_+(X(s+t)) - q_+(X(s))) \\
&\qquad\qquad q_-(X(s)) \mid X(0) = x] ds\, \pi(dx) \\
&= \lim_{t\to 0} \frac{1}{t} \int \frac{1}{\tau} \int_0^\tau \mathbb{E}[q_+(X(\tau \wedge S^+_{A\cup B}(s+t)))(q_+(X(s+t)) - q_+(X(s))) \\
&\qquad\qquad q_-(X(0 \vee S^-_{A\cup B}(s))) \mid X(0) = x] ds\, \pi(dx)
\end{aligned} \tag{S24}$$

where (S24) follows from (S3) and the equations solved by the forward and backward committors. This suggests the estimator

$$R_{AB} \approx \frac{1}{\tau} \sum_i \sum_{p=0}^{\frac{\tau}{\Delta}-1} q_+(X^{(i)}(\tau \wedge S^+_{A\cup B}((p+1)\Delta)))(q_+(X^{(i)}((p+1)\Delta)) - q_+(X^{(i)}(p\Delta))) \\ q_-(X^{(i)}(0 \vee S^-_{A\cup B}(p\Delta)))w(X^{(i)}(0)). \tag{S25}$$

In contrast to the estimator in equation (33), we have taken the limit over the sampling interval $\Delta$ rather than the lag time, which is now $\tau$ in this expression. This is the same as the approximation made in equation (14).

To further relate this estimator to equation (33), we manipulate it to obtain a similar form. To do that, we require two identities. First, we express forward stopping times in terms of backward stopping times. We can derive this by case analysis (Figure S8):

1. If $X(0 \vee S^-_{A\cup B}(s)) \notin A \cup B$, then $X(r) \notin A \cup B$ for $0 \leq r \leq s$. This condition is equivalent to that of $X(s \wedge S^+_{A\cup B}(0)) \notin A \cup B$.

2. If $X(0 \vee S^-_{A\cup B}(s)) \in A \cup B$ and $X(s) \in A \cup B$, the process stopped immediately.

3. If $X(0 \vee S^-_{A\cup B}(s)) \in A \cup B$ and $X(s) \notin A \cup B$, the process hit $A \cup B$ after propagating backward in time. The time $0 \vee S^-_{A\cup B}(s)$ can be found by finding a time $r \in [0, s)$ such that $X(t) \notin A \cup B$ for all times $t \in (r, s]$.

Together, for all $s \geq 0$, this yields

$$\begin{aligned} f(X(0 \vee S^-_{A\cup B}(s))) &= \mathbb{1}_{(A\cup B)^c}(X(s \wedge S^+_{A\cup B}(0)))f(X(0)) + \mathbb{1}_{A\cup B}(X(s))f(X(s)) \\ &\quad + \lim_{t\to 0} \frac{1}{t} \int_0^s \mathbb{1}_{(A\cup B)^c}(X(s \wedge S^+_{A\cup B}(r+t)))\mathbb{1}_{A\cup B}(X(r))f(X(r))\, dr \\ &\approx \mathbb{1}_{(A\cup B)^c}(X(s \wedge S^+_{A\cup B}(0)))f(X(0)) + \mathbb{1}_{A\cup B}(X(s))f(X(s)) \\ &\quad + \sum_{p=0}^{\frac{s}{\Delta}-1} \mathbb{1}_{(A\cup B)^c}(X(s \wedge S^+_{A\cup B}((p+1)\Delta)))\mathbb{1}_{A\cup B}(X(p\Delta))f(X(p\Delta)). \end{aligned} \tag{S26}$$

Second, we have the identity

$$S^+_{A\cup B}(r) = \lim_{t\to 0} S^+_{A\cup B}(s+t) \approx S^+_{A\cup B}(s+\Delta) \text{ if } X(s \wedge S^+_{A\cup B}(r)) \in (A\cup B)^c \text{ and } s \geq r. \quad \text{(S27)}$$

This follows from the definition of the stopping time: $S^+_{A\cup B}(r) = \min\{t' \geq r : X(t') \in A\cup B\}$, and so if $X(t') \notin A \cup B$ for $r \leq t' \leq s$, then $\min\{t' \geq r : X(t') \in A \cup B\} = \min\{t' > s : X(t') \in A \cup B\}$.

We now apply equations (S26) and (S27) to equation (S25), yielding

$$
\begin{aligned}
R_{AB} \approx \frac{1}{\tau}\sum_i \sum_{p=0}^{\frac{\tau}{\Delta}-1} & q_+(X^{(i)}(\tau \wedge S^+_{A\cup B}(0)))(q_+(X^{(i)}((p+1)\Delta)) - q_+(X^{(i)}(p\Delta))) \\
& \mathbb{1}_{(A\cup B)^c}(X^{(i)}(p\Delta \wedge S^+_{A\cup B}(0)))q_-(X^{(i)}(0))w(X^{(i)}(0)) \\
+ \frac{1}{\tau}\sum_i \sum_{p=0}^{\frac{\tau}{\Delta}-1} & q_+(X^{(i)}(\tau \wedge S^+_{A\cup B}((p+1)\Delta)))(q_+(X^{(i)}((p+1)\Delta)) - q_+(X^{(i)}(p\Delta))) \\
& \mathbb{1}_{A\cup B}(X^{(i)}(p\Delta))q_-(X^{(i)}(p\Delta))w(X^{(i)}(0)) \\
+ \frac{1}{\tau}\sum_i \sum_{p=0}^{\frac{\tau}{\Delta}-1}\sum_{r=0}^{p-1} & q_+(X^{(i)}(\tau \wedge S^+_{A\cup B}((r+1)\Delta)))(q_+(X^{(i)}((p+1)\Delta)) - q_+(X^{(i)}(p\Delta))) \\
& \mathbb{1}_{(A\cup B)^c}(X^{(i)}(p\Delta \wedge S^+_{A\cup B}((r+1)\Delta)))\mathbb{1}_{A\cup B}(X^{(i)}(r\Delta)) \\
& q_-(X^{(i)}(r\Delta))w(X^{(i)}(0)) \\
= \frac{1}{\tau}\sum_i \sum_{p=0}^{\frac{\tau\wedge S^+_{A\cup B}(0))}{\Delta}-1} & q_+(X^{(i)}(\tau \wedge S^+_{A\cup B}(0)))(q_+(X^{(i)}((p+1)\Delta)) - q_+(X^{(i)}(p\Delta))) \\
& q_-(X^{(i)}(0))w(X^{(i)}(0)) \\
+ \frac{1}{\tau}\sum_i \sum_{p=0}^{\frac{\tau}{\Delta}-1} & q_+(X^{(i)}(\tau \wedge S^+_{A\cup B}((p+1)\Delta)))(q_+(X^{(i)}((p+1)\Delta)) - q_+(X^{(i)}(p\Delta))) \\
& \mathbb{1}_{A\cup B}(X^{(i)}(p\Delta))q_-(X^{(i)}(p\Delta))w(X^{(i)}(0)) \\
+ \frac{1}{\tau}\sum_i \sum_{r=0}^{\frac{\tau}{\Delta}-2} \sum_{p=r+1}^{\frac{\tau\wedge S^+_{A\cup B}((r+1)\Delta)}{\Delta}-1} & q_+(X^{(i)}(\tau \wedge S^+_{A\cup B}((r+1)\Delta))) \\
& (q_+(X^{(i)}((p+1)\Delta)) - q_+(X^{(i)}(p\Delta))) \\
& \mathbb{1}_{A\cup B}(X^{(i)}(r\Delta))q_-(X^{(i)}(r\Delta))w(X^{(i)}(0)) \\
= \frac{1}{\tau}\sum_i & q_+(X^{(i)}(\tau \wedge S^+_{A\cup B}(0)))(q_+(X^{(i)}(\tau \wedge S^+_{A\cup B}(0))) - q_+(X^{(i)}(0)))q_-(X^{(i)}(0))w(X^{(i)}(0)) \\
+ \frac{1}{\tau}\sum_i \sum_{r=0}^{\frac{\tau}{\Delta}-1} & q_+(X^{(i)}(\tau \wedge S^+_{A\cup B}((r+1)\Delta)))(q_+(X^{(i)}(\tau \wedge S^+_{A\cup B}((r+1)\Delta))) - q_+(X^{(i)}(r\Delta)) \\
& \mathbb{1}_{A\cup B}(X^{(i)}(r\Delta))q_-(X(r\Delta))w(X^{(i)}(0)) \qquad \text{(S28)}
\end{aligned}
$$

This is equation (33) with an additional term

$$\frac{1}{\tau}\sum_i \sum_{r=0}^{\frac{\tau}{\Delta}-1} q_+(X^{(i)}(\tau \wedge S^+_{A\cup B}((r+1)\Delta)))(q_+(X^{(i)}(\tau \wedge S^+_{A\cup B}((r+1)\Delta))) - q_+(X^{(i)}(r\Delta)))$$
$$\mathbb{1}_{A\cup B}(X^{(i)}(r\Delta))q_-(X^{(i)}(r\Delta))w(X^{(i)}(0)) \tag{S29}$$

which adds the contributions from trajectories leaving $A \cup B$ within the lag time $\tau$, which becomes significant at longer lag times since a larger portion of the trajectories hit (or start in) and then exit $A \cup B$.

We proceed analogously for the reactive current estimator. We start with expression (35) and integrate to obtain

$$\begin{aligned}
J^\theta_{AB}(s) &= \lim_{t,|ds|\to 0} \frac{1}{2t|ds|} \int (\mathcal{T}^t[\theta q_+](x) - \theta(x)\mathcal{T}^t q_+(x))\mathbb{1}_{\{\theta\in ds\}}(x)q_-(x)\pi(dx) \\
&\qquad + (\mathcal{T}^t[\mathbb{1}_{\{\theta\in ds\}}\theta q_+](x) - \theta(x)\mathcal{T}^t[\mathbb{1}_{\{\theta\in ds\}}q_+](x))q_-(x)\pi(dx) \\
&= \lim_{t,|ds|\to 0} \frac{1}{2t|ds|} \int \mathbb{E}[q_+(X(r+t))(\theta(X(r+t)) - \theta(X(r)))q_-(X(r)) \\
&\qquad (\mathbb{1}_{\{\theta\in ds\}}(X(r+t)) + \mathbb{1}_{\{\theta\in ds\}}(X(r))) \mid X(0) = x]\pi(dx) \\
&= \lim_{t,|ds|\to 0} \frac{1}{2t|ds|} \int \frac{1}{\tau}\int_0^\tau \mathbb{E}[q_+(X(r+t))(\theta(X(r+t)) - \theta(X(r)))q_-(X(r)) \\
&\qquad (\mathbb{1}_{\{\theta\in ds\}}(X(r+t)) + \mathbb{1}_{\{\theta\in ds\}}(X(r))) \mid X(0) = x]dr\,\pi(dx) \\
&= \lim_{t,|ds|\to 0} \frac{1}{2t|ds|} \int \frac{1}{\tau}\int_0^\tau \mathbb{E}[q_+(X(\tau \wedge S^+_{A\cup B}(r+t)))(\theta(X(r+t)) - \theta(X(r))) \\
&\qquad q_-(X(0 \vee S^-_{A\cup B}(r)))(\mathbb{1}_{\{\theta\in ds\}}(X(r+t)) - \mathbb{1}_{\{\theta\in ds\}}(X(r))) \mid X(0) = x]dr\,\pi(dx),
\end{aligned} \tag{S30}$$

which suggests the estimator

$$J^{\theta}_{AB}(s) \approx \frac{1}{2\tau|ds|}\sum_i \sum_{p=0}^{\frac{\tau}{\Delta}-1} q_+(X^{(i)}(\tau \wedge S^+_{A\cup B}((p+1)\Delta)))(\theta(X^{(i)}((p+1)\Delta)) - \theta(X^{(i)}(p\Delta)))$$
$$q_-(X^{(i)}(0 \vee S^-_{A\cup B}(p\Delta)))(\mathbb{1}_{\{\theta\in ds\}}(X^{(i)}((p+1)\Delta)) + \mathbb{1}_{\{\theta\in ds\}}(X^{(i)}(p\Delta)))w(X^{(i)}(0)). \tag{S31}$$

We note that if we integrate over $s$ and substitute $\theta(x) = q_+(x)$, we obtain the reaction rate estimator (S25). As with equation (S25), this estimator converges to the reactive current with perfect sampling, committors, and change of measure. In comparison to equation (37), we find that this estimator has larger variance with limited data. This results from the contribution of $\theta(X^{(i)}((p+1)\Delta)) - \theta(X^{(i)}(p\Delta))$ being assigned locally to $X((p+1)\Delta)$ and $X(p\Delta)$, whereas equation (37) assigns $\theta(X^{(i)}(\tau \wedge S^+_{A\cup B}(0))) - \theta(X(0))$ to $X(0)$ and $\theta(X(\tau)) - \theta(X(0 \vee S^-_{A\cup B}(\tau)))$ to $X(\tau)$ which leads to mixing over longer length scales. Thus, when using (S31), we recommend smoothing the reactive current vector field using a kernel density estimate. Specifically, to estimate the reactive current at a point $s$, we evaluate

$$\frac{\int e^{-|s-s'|^2/2\sigma^2} J^{\theta}_{AB}(s')ds'}{\int e^{-|s-s'|^2/2\sigma^2} ds'}, \tag{S32}$$

where $\sigma$ is chosen to be the smallest value where the vector field is smooth.

Using equations (S26) and (S27), we can express this estimator using a similar procedure

as above for the rate as

$$
\begin{aligned}
J^{\theta}_{AB}(s) \approx \frac{1}{2\tau|ds|}\sum_i q_+(X^{(i)}(\tau\wedge S^+_{A\cup B}(0)))q_-(X^{(i)}(0))w(X^{(i)}(0)) \\
&\sum_{p=0}^{\frac{\tau\wedge S^+_{A\cup B}(0)}{\Delta}-1}(\theta(X^{(i)}((p+1)\Delta))-\theta(X^{(i)}(p\Delta))) \\
&\qquad (\mathbb{1}_{\{\theta\in ds\}}(X^{(i)}((p+1)\Delta))+\mathbb{1}_{\{\theta\in ds\}}(X^{(i)}(p\Delta))) \\
+\frac{1}{2\tau|ds|}\sum_i\sum_{r=0}^{\frac{\tau}{\Delta}-1} q_+(X^{(i)}(\tau\wedge S^+_{A\cup B}((r+1)\Delta)))q_-(X^{(i)}(r))\mathbb{1}_{A\cup B}(X^{(i)}(r\Delta))w(X^{(i)}(0)) \\
&\sum_{p=r}^{\frac{\tau\wedge S^+_{A\cup B}((r+1)\Delta)}{\Delta}-1}(\theta(X^{(i)}((p+1)\Delta))-\theta(X^{(i)}(p\Delta))) \\
&\qquad (\mathbb{1}_{\{\theta\in ds\}}(X^{(i)}((p+1)\Delta))+\mathbb{1}_{\{\theta\in ds\}}(X^{(i)}(p\Delta))). \qquad \text{(S33)}
\end{aligned}
$$

Likewise, by expressing the forward stopping time in terms of the backward stopping time as

$$
\begin{aligned}
f(X(\tau\wedge S^+_{A\cup B}(s))) &= \mathbb{1}_{(A\cup B)^c}(X(s\vee S^-_{A\cup B}(\tau)))f(X(\tau))+\mathbb{1}_{A\cup B}(X(s))f(X(s)) \\
&\quad +\lim_{t\to 0}\frac{1}{t}\int_s^{\tau}\mathbb{1}_{(A\cup B)^c}(X(s\vee S^-_{A\cup B}(r-t)))\mathbb{1}_{A\cup B}(X(r))f(X(r))\,dr \\
&\approx \mathbb{1}_{(A\cup B)^c}(X(s\vee S^-_{A\cup B}(\tau)))f(X(\tau))+\mathbb{1}_{A\cup B}(X(s))f(X(s)) \\
&\quad +\sum_{p=\frac{s}{\Delta}+1}^{\frac{\tau}{\Delta}}\mathbb{1}_{(A\cup B)^c}(X(s\vee S^-_{A\cup B}((p-1)\Delta)))\mathbb{1}_{A\cup B}(X(p\Delta))f(X(p\Delta))
\end{aligned} \qquad \text{(S34)}
$$

for $s\le\tau$ and using the identity

$$
S^-_{A\cup B}(r)=\lim_{t\to 0}S^-_{A\cup B}(s-t)\approx S^-_{A\cup B}(s-\Delta) \text{ if } X(s\vee S^-_{A\cup B}(r))\notin A\cup B \text{ and } s\le r \qquad \text{(S35)}
$$

we obtain

$$J^{\theta}_{AB}(s) \approx \frac{1}{2\tau|ds|}\sum_i q_+(X^{(i)}(\tau))q_-(X^{(i)}(0\vee S^-_{A\cup B}(\tau)))w(X^{(i)}(0))$$
$$\sum_{p=\frac{0\vee S^-_{A\cup B}(\tau)}{\Delta}+1}^{\tau}(\theta(X^{(i)}(p\Delta))-\theta(X^{(i)}((p-1)\Delta)))$$
$$(\mathbb{1}_{\{\theta\in ds\}}(X^{(i)}(p\Delta))+\mathbb{1}_{\{\theta\in ds\}}(X^{(i)}((p-1)\Delta)))$$
$$+\frac{1}{2\tau|ds|}\sum_i\sum_{r=1}^{\frac{\tau}{\Delta}} q_+(X^{(i)}(r\Delta))\mathbb{1}_{A\cup B}(X^{(i)}(r\Delta))q_-(X^{(i)}(0\vee S^-_{A\cup B}((r-1)\Delta)))w(X^{(i)}(0))$$
$$\sum_{p=\frac{0\vee S^-_{A\cup B}((r-1)\Delta)}{\Delta}+1}^{r}(\theta(X^{(i)}(p\Delta))-\theta(X^{(i)}((p-1)\Delta)))$$
$$(\mathbb{1}_{\{\theta\in ds\}}(X^{(i)}(p\Delta))+\mathbb{1}_{\{\theta\in ds\}}(X^{(i)}((p-1)\Delta))). \tag{S36}$$

The reactive current estimator (37) is obtained by averaging the first term of equation (S33) together with the first term of equation (S36), after performing the approximations $\mathbb{1}_{\{\theta\in ds\}}(X^{(i)}(s)) \approx \mathbb{1}_{\{\theta\in ds\}}(X^{(i)}(0))$ and $\mathbb{1}_{\{\theta\in ds\}}(X^{(i)}(s)) \approx \mathbb{1}_{\{\theta\in ds\}}(X^{(i)}(\tau))$ for the terms from equations (S33) and (S36), respectively. The terms that we gain here relative to (37) in the main text are similar in origin to those introduced into the reaction rate estimator in this section. For equation (S33), they account for the portions of the trajectories which leave $A\cup B$ within $\tau$; the same holds for equation (S36), but for the time-reversed trajectory. In effect, equation (37) ignores half the contribution of the trajectories which hit $A\cup B$ and half the contribution of the time-reversed trajectories which hit $A\cup B$.

# References

(1) Vanden-Eijnden, E. *Computer Simulations in Condensed Matter Systems: From Materials to Chemical Biology Volume 1*; Springer, 2006; pp 453–493.

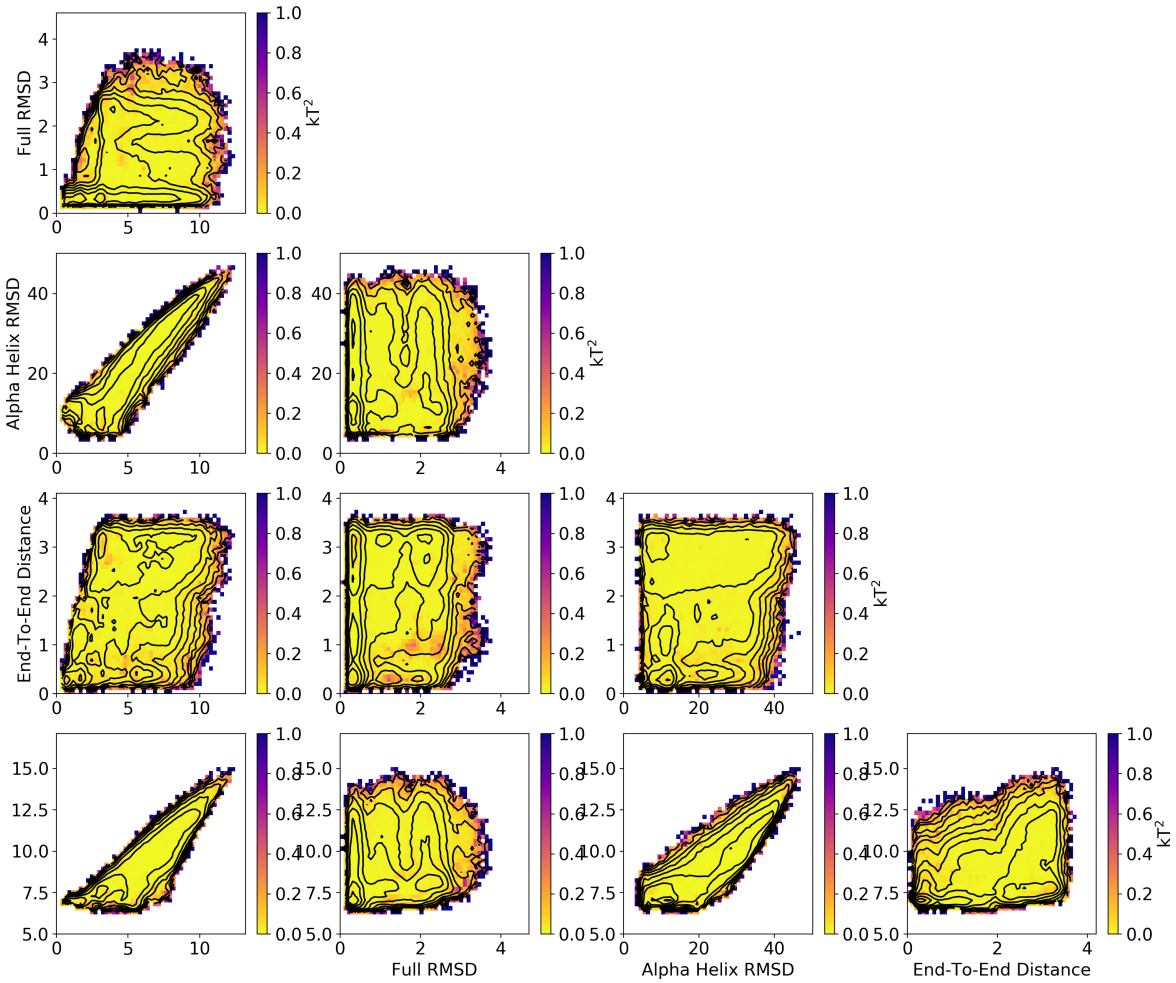


Figure S1: EMUS asymptotic variance for REUS PMFs.

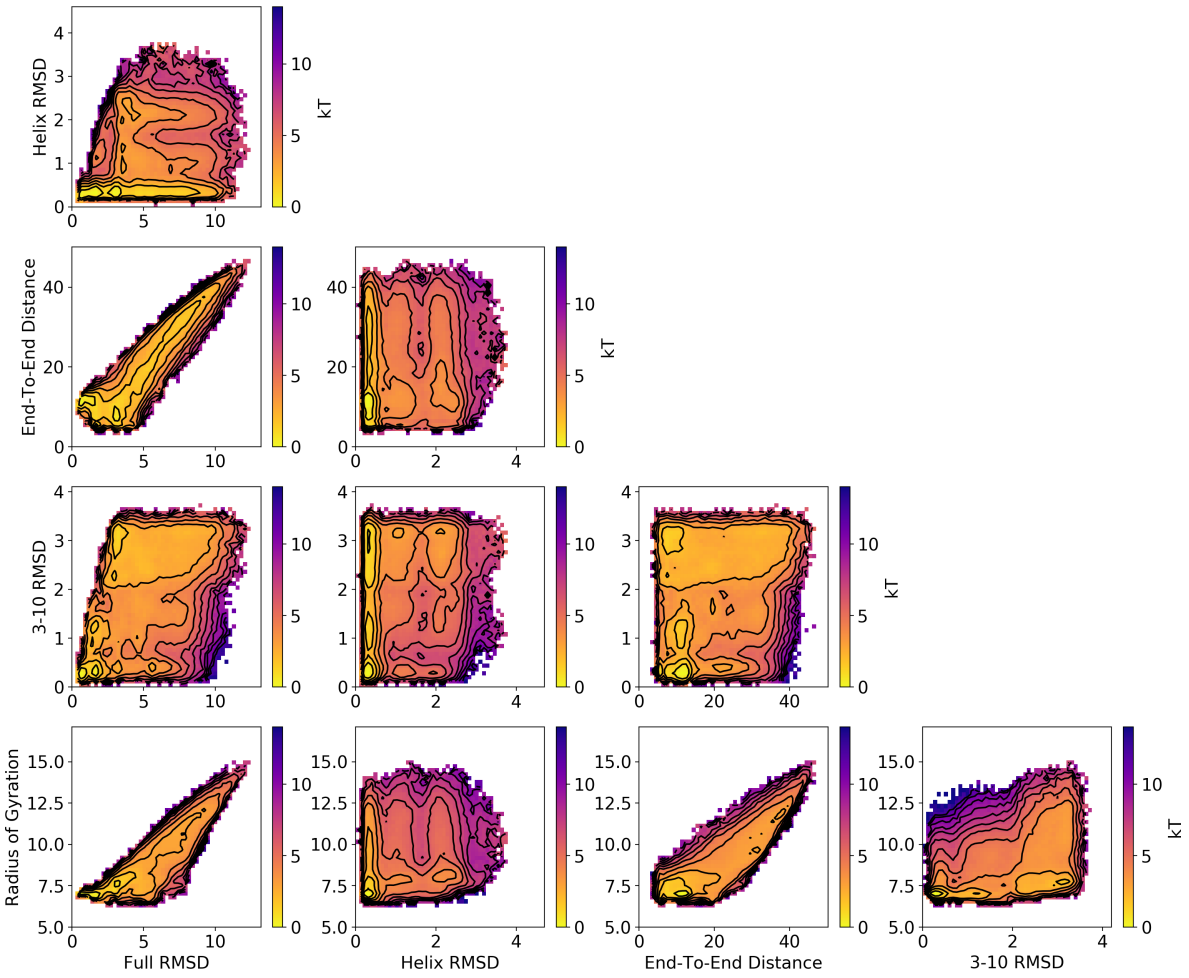


Figure S2: REUS PMFs.

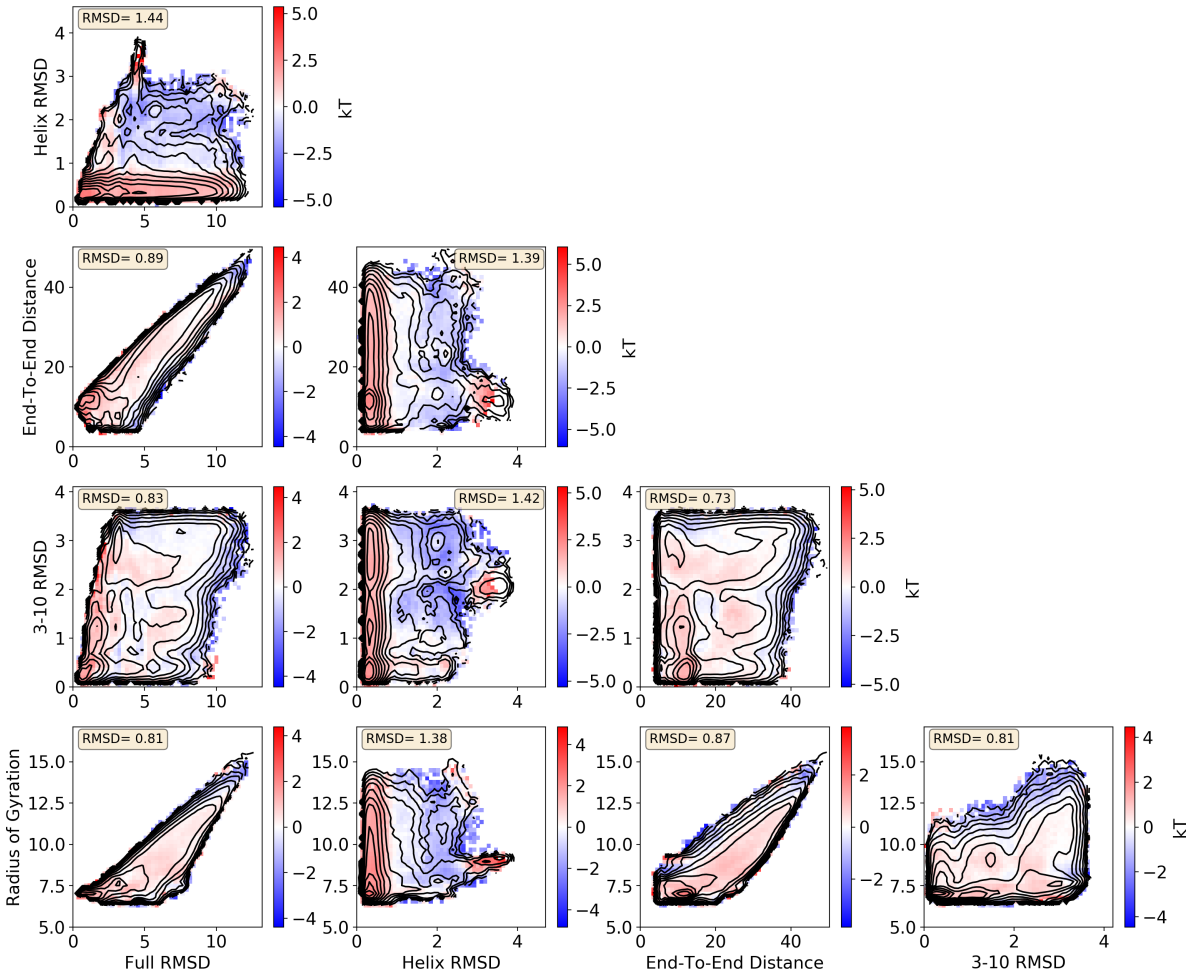


Figure S3: Difference between DGA with the modified distance basis set without the $\alpha$ helix resampling and REUS.

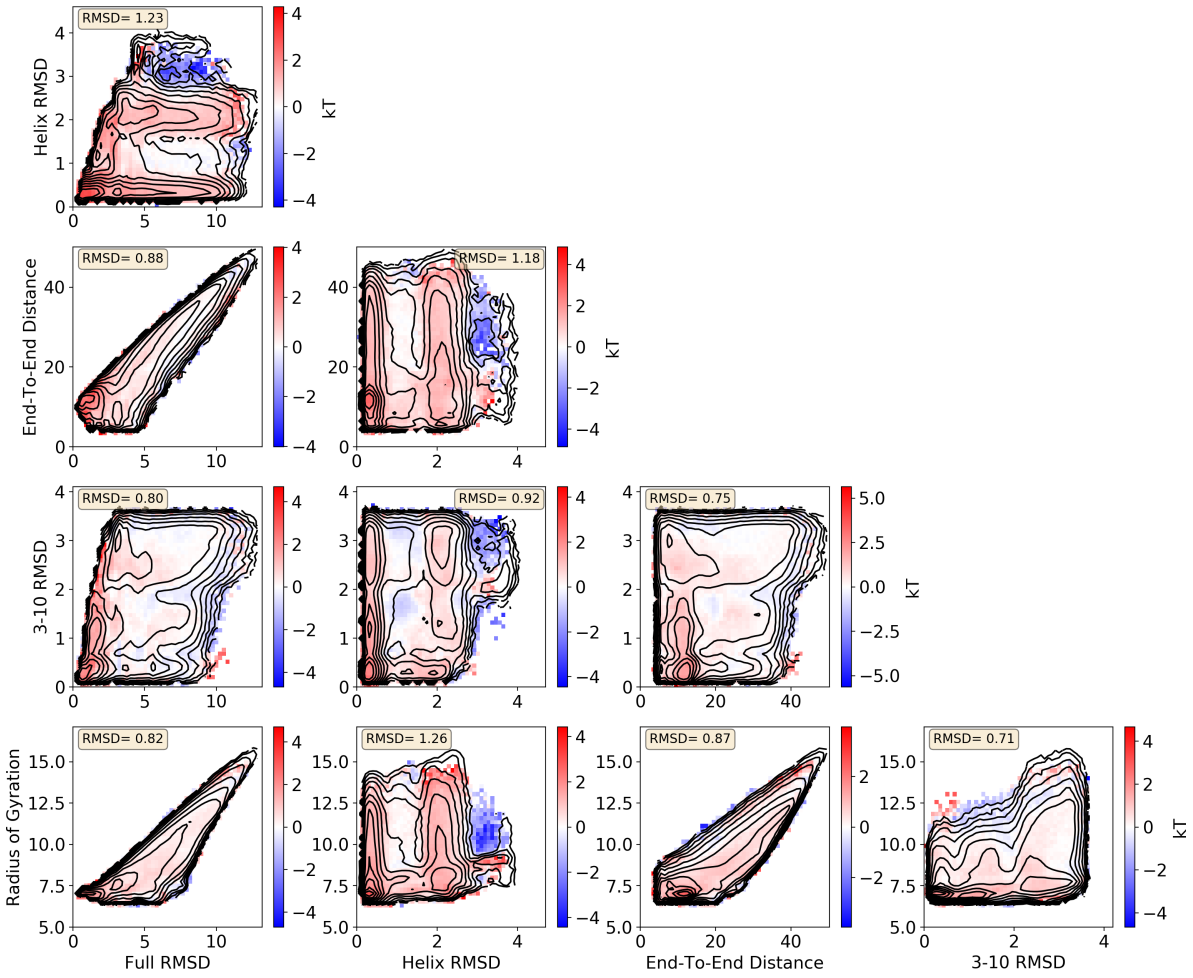


Figure S4: Difference between the PMF from DGA with the distance indicator basis set and the PMF from REUS.

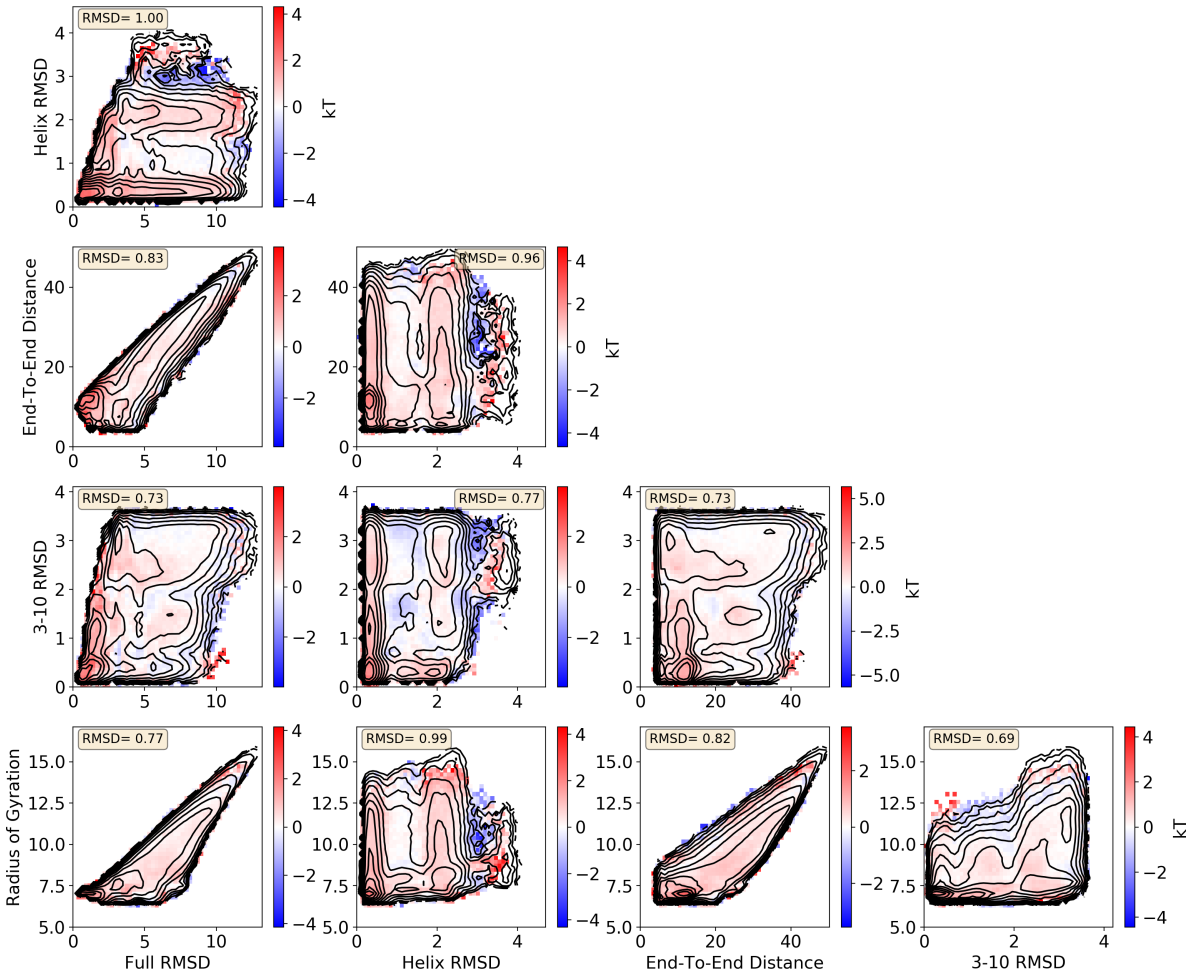


Figure S5: Difference between the PMF from DGA with the TICA indicator basis set and the PMF from REUS.

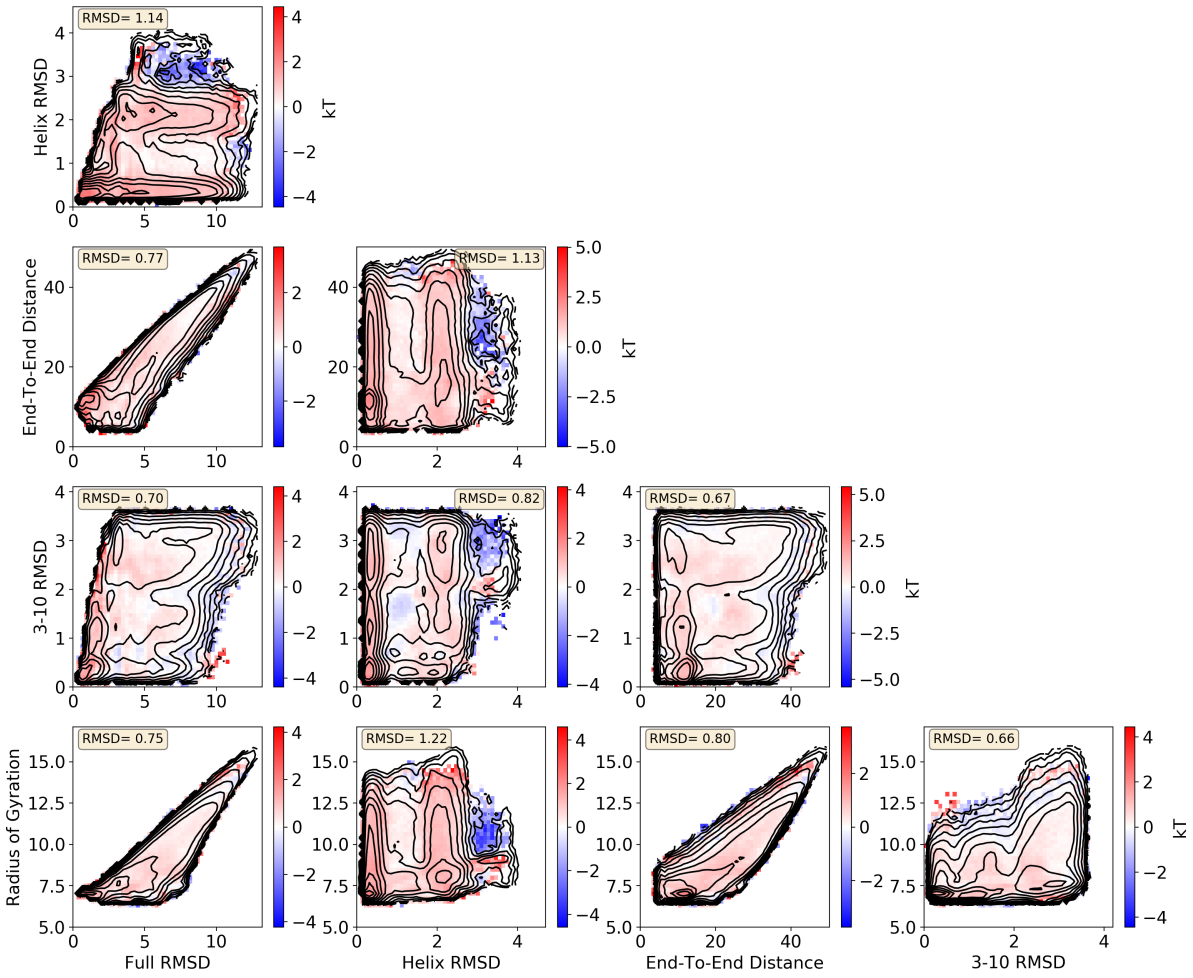


Figure S6: Difference between the PMF from DGA with the modified distance basis set and the PMF from REUS.

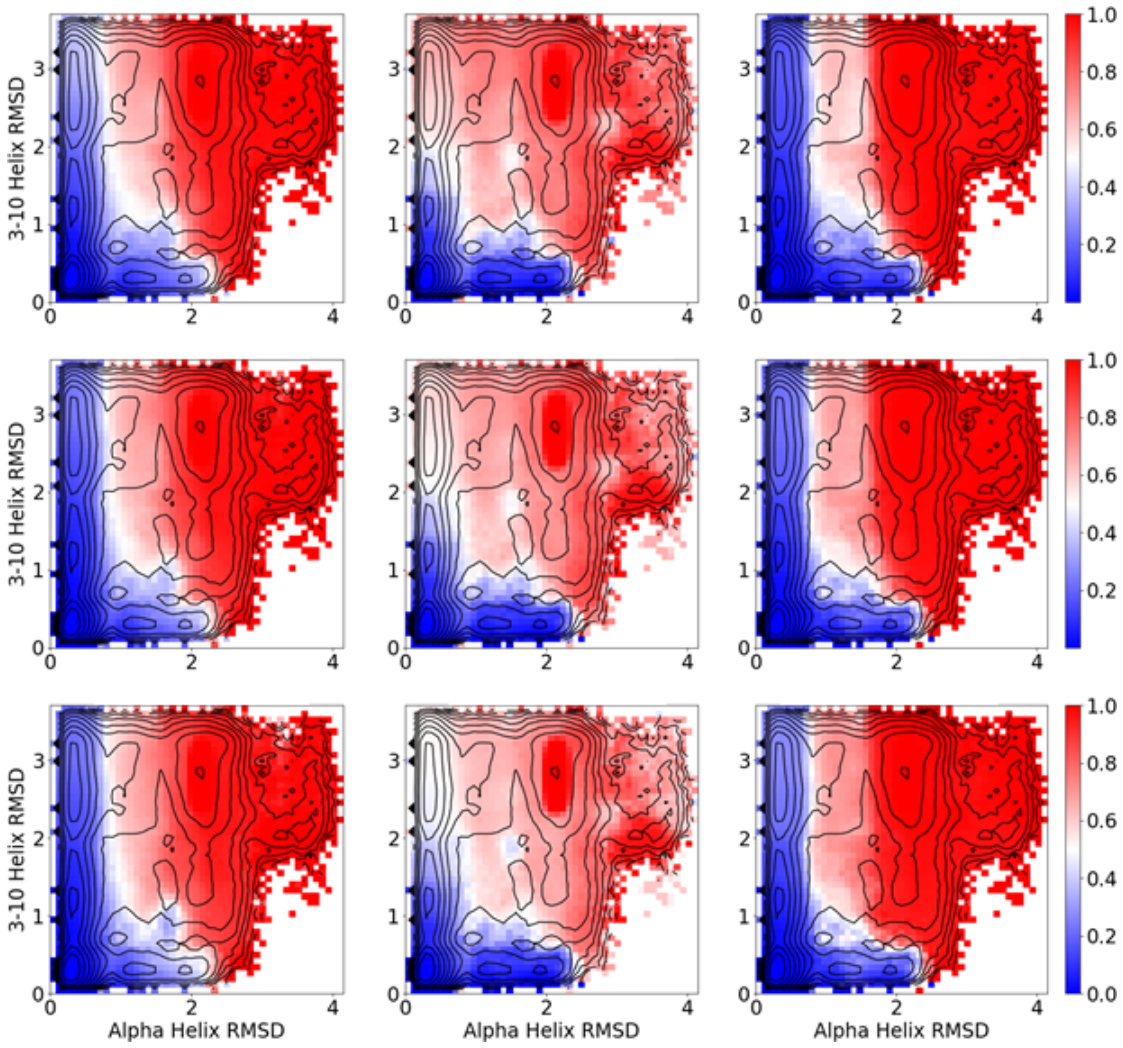


Figure S7: DGA backward committors. Left, middle, and right columns are computed with the modified distance, distance indicator, and TICA indicator basis sets, respectively. Top, middle, and bottom rows are computed with lag times of 0.5, 2.5, and 7.5 ns, respectively.

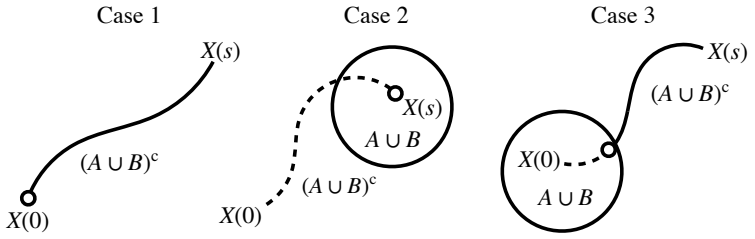


Figure S8: Case analysis for expressing the backward stopping time in terms of forward stopping times.