Long-timescale predictions from short-trajectory data: A benchmark analysis of the trp-cage miniprotein

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

Abstract

Elucidating physical mechanisms with statistical confidence from molecular dynamics simulations can be challenging owing to the many degrees of freedom that contribute to collective motions. To address this issue, we recently introduced a dynamical Galerkin approximation (DGA) [Thiede *et al. J. Phys. Chem.* **150**, 244111 (2019)], in which chemical kinetic statistics that satisfy equations of dynamical operators are represented by a basis expansion. Here, we reformulate this approach, clarifying (and reducing) the dependence on the choice of lag time. We present a new projection of the reactive current onto collective variables and provide improved estimators for rates and committors. We also present simple procedures for constructing suitable smoothly varying basis functions from arbitrary molecular features. To evaluate estimators and basis sets numerically, we generate and carefully validate a data set of short trajectories for the unfolding and folding of the trp-cage miniprotein, a well-studied system. Our analysis demonstrates a comprehensive strategy for characterizing reaction pathways quantitatively.

Graphical Abstract

---
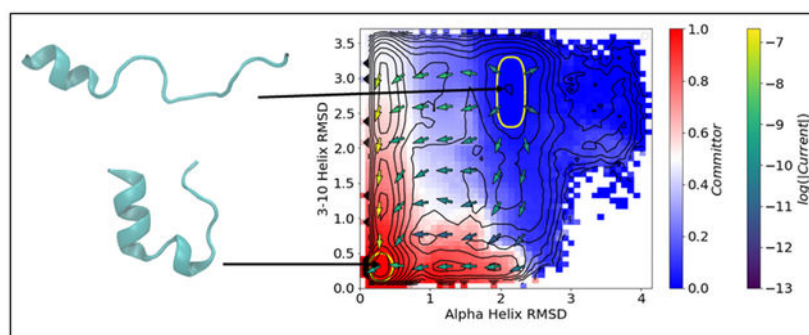
weare@cims.nyu.edu; dinner@uchicago.edu.

Supporting Information Available

Mathematical appendices, asymptotic variances for REUS, and difference maps between DGA and REUS PMFs. The trp-cage data set, together with an implementation of DGA, is available at https://projects.rcc.uchicago.edu/dali/Trp_Cage_Data/.

1 Introduction

Molecular dynamics simulations enable atomic-resolution investigation of complex processes. These investigations are often carried out by direct simulation: the equations of motion are numerically integrated forward in time to generate trajectories (times series of atomic positions and, as needed, momenta) for as long as possible given available computational resources. Since most events of interest occur on timescales longer than those accessible by direct simulation, many enhanced sampling schemes have been developed to allow more extensive interrogation of an event of interest without sacrificing model fidelity. Splitting methods, for example, branch and prune a collection of simultaneously evolving trajectories to promote progress in a small number of order parameters (or collective variables, CVs).[1–6] Regardless of whether trajectory data are generated by direct simulation or enhanced sampling, an essential question remains: How can these data be analyzed to yield new understanding about the process under study?

We recently introduced dynamical Galerkin approximation (DGA)[7] to analyze trajectory data generated by direct simulation, as well as many enhanced sampling schemes. In this approach, conditional expectations such as committor functions are cast as solutions to equations involving the operator determining the statistics of the underlying process, its transition operator. The solution to the equation is then approximated as a linear combination of basis functions. This approach builds on an extensive literature from the last decade that shows that the eigenvalues and eigenvectors of the transition operator can be approximated from trajectory data, subject to a Markov assumption.[8–21] These spectral estimation methods aim to characterize the slowest dynamical features of the system (e.g., transitions between metastable states) as eigenvectors corresponding to the largest eigenvalues of the transition operator. When the goal is to study a particular event of interest, the indirect relationship between the eigenvectors of the transition operator and the specific event of interest is a weakness of the spectral estimation approach. Indeed, for many complex systems the true slowest dynamical features of the system are too slow to be of any physical interest.

In contrast, the aim of DGA is *not* to extract spectral information. Instead DGA aims to compute statistics that directly characterize a particular event under study. For example, when transitions between particular metastable states are of interest, the statistics that DGA yields can be combined within the framework of transition path theory (TPT)[22–24] to obtain reactive fluxes and in turn reaction mechanisms. Because DGA analyzes short trajectory fragments, it can be used to process the data generated by many splitting schemes. Alternatively, the trajectory data can be generated by seeding initial conditions for short direct simulations throughout state space.

Though most often employed as a spectral estimation tool, Markov State Models (MSMs)[25–27] and related methods[28] have also been used to approximate TPT quantities. DGA can be viewed as an extension of these MSM variants to a more general class of target quantities and to more general representations (basis set expansions) of those quantities. However, even with parameters chosen as in MSMs, the DGA estimators introduced in this

article improve upon their MSM counterparts in several ways including reduced dependence on the crucial lag time parameter and lower variance estimates of certain TPT quantities.

In our previous study,[7] we compared diffusion map[29] and indicator basis sets for predicting mean first-passage times and committors for the Müller-Brown model[30] and the folding of the protein Fip35 using six long equilibrium trajectories from D. E. Shaw Research.[31] Because there were only a few folding events within those trajectories, it was difficult to assess the performance of the method. One goal of the present study is to generate a protein folding data set that enables robust application of the approach and to compare different basis sets and estimators systematically.

To this end, we study the trp-cage miniprotein, a 20-residue fast-folding artificial sequence (asn-leu-tyr-ile-glu-trp-leu-lys-asp-gly-gly-pro-ser-ser-gly-arg-pro-pro-pro-ser) that has been studied extensively both experimentally and computationally.[32–38] In solution at 298 K, the protein folds on a 4 μs timescale and unfolds on a 12 μs timescale,[32] which makes these processes difficult but not impossible to simulate directly. In particular, D. E. Shaw Research produced a 208 μs equilibrium simulation of the K8A mutant of trp-cage using the Anton supercomputer with the CHARMM 22* force field.[36] Although, like the Fip35 data, this trajectory contains relatively few folding events, it has been the subject of previous MSM[37] and variational approach for Markov processes (VAMP) studies.[38] These earlier studies serve as valuable points of comparison and enable us to identify CVs that provide good control over sampling. Though DGA does not depend directly on any choice of CV, its performance is strongly affected by the quality of the data set of sampled trajectories. We use our chosen set of CVs together with enhanced sampling methods to generate a new data set comprised of many short trajectories that are distributed evenly throughout the CV space.

In this article we reformulate DGA in terms of the transition operator of the underlying Markov process. This has two primary advantages relative to our previous formulation in terms of the generator of the process.[7] First, it clarifies the role of lag time in DGA estimates, showing that correctly constructed estimators should have no dependence on lag time in the infinite-basis, infinite-sampling limit. Second, the formulation in terms of the transition operator leads directly to estimators that correctly account for boundary conditions by stopping underlying trajectories appropriately. Using our improved DGA estimators we introduce new estimators for TPT reaction rates and reactive currents. To make computation of the reactive current tractable and the result readily interpretable, we introduce a projection formula for the reactive current onto a CV space which allows us to assign relative weights to transition paths in arbitrary CV spaces. We also introduce a new procedure for constructing a basis set from arbitrary molecular features (here, primarily pairwise distances between C_a atoms, though we also explore CVs with delay embedding) and compare it with two basis sets that are used widely in the MSM literature: indicator functions on molecular features and indicator functions on time-lagged independent component analysis (TICA) coordinates.[39–41] We show that our DGA estimators with selected basis sets can robustly yield remarkably good agreement with published results for committors and pathways, even though the total simulation time of our trp-cage data set is only 30 μs, with a maximum trajectory length of 30 ns. The projection of the reactive currents on CVs facilitates both visualization and quantification of information about pathways, enabling immediate

identification of the defining properties of transition states. This makes our approach an efficient one for exploring mechanisms.

2 Long time phenomena from short trajectory data

In this section, we introduce key dynamical statistics and explain how they can be defined in terms of an evolution operator (Section 2.1). An emphasis on forms that lead directly to practical and accurate numerical estimators causes several departures from the standard presentation of this material. We present our approach for solving the operator equations numerically by Galerkin (basis expansion) approximation[7] (Section 2.2), and distinguish forward-in-time statistics (Section 2.2.1) from backward-in-time statistics involving the adjoint of the evolution operator (Section 2.2.2); this is followed by a discussion of basis sets (Section 2.2.3) and an approach for constructing an approximately Markovian process when the molecular representation does not adequately capture the dynamics (delay embedding, Section 2.2.4). Finally, guided by TPT, we combine the dynamical statistics estimated by DGA to yield approximations of reaction rates and currents. (Section 2.3).

2.1 The transition operator and Feynman-Kac representation

The dynamics of a Markov process $X(t)$ can be encoded in its associated transition operator, $\mathcal{T}^t$, which specifies the evolution of the expectation of a function f over some interval of time $t \geq 0$:

$$\mathcal{T}^t f(x) = \mathbb{E}[f(X(t)) \mid X(0) = x].$$

(1)

The time index t can be continuous or discrete. The transition operator (also known as the Koopman operator), and in particular its eigenvectors and eigenvalues, are the key quantities in well-established methods for discovering slowly decorrelating features of a Markov process.[42] The transition operator is also central to the DGA approach.[7] However, in DGA, instead of estimating the spectrum of the transition operator, the goal is to solve linear equations representing certain conditional expectations.

In ref. 7, we presented DGA in terms of the generator which, for a continuous time process is defined by the limit:

$$\mathcal{L} f(x) = \lim_{t \to 0} \frac{\mathcal{T}^t f(x) - f(x)}{t}.$$

(2)

For a discrete time process the limit is removed and t in (2) is replaced by the unit of a single time step. A presentation in terms of the generator has the advantage that it results in very concise equations for quantities of interest. For example, consider the (forward) committor, $q_+(x)$, which is the probability of entering a product state B before a reactant state A starting from $x \notin A \cup B$:

$$q_+(x) = \mathbb{P}[X(T_{A \cup B}) \in B \mid X(0) = x],$$

(3)

where $T_{A\cup B} = \min\{t \geq 0 : X(t) \in A\cup B\}$ is the time of first entrance into $A\cup B$. For $x \in A$, $q_+(x) = 0$, and, for $x \in B$, $q_+(x) = 1$. The committor satisfies the Feynman-Kac relation

$$\mathscr{L}q_+(x) = 0 \ \text{ for } \ x \notin A \cup B, \quad q_+(x) = \mathbb{1}_B(x) = \begin{cases} 1, \ x \in B \\ 0, \ x \notin B \end{cases} \ \text{ for } \ x \in A \cup B \qquad (4)$$

(see Eqs. (18) and (19) of ref. 7).

In this article we choose to work directly with the transition operator instead of the generator because it facilitates the implementation of numerical formulas. It also greatly simplifies our description of TPT and clarifies the relationship between DGA and the well-established VAC approach to approximating spectral properties of the transition operator (see ref. 42). In the case of the committor, we integrate (4) until a chosen time τ to obtain the equivalent form of the Feynman-Kac relation,

$$\mathscr{T}_{A \cup B}^{\tau}q_+(x) - q_+(x) = 0 \ \text{ for } \ x \notin A \cup B, \ \ q_+(x) = \mathbb{1}_B(x) \ \text{ for } \ x \in A \cup B. \qquad (5)$$

In this expression we have introduced the notation $\mathscr{T}_{A \cup B}^{t}$ for the transition operator of the stopped process $X(t \wedge T_{A\cup B})$, i.e.,

$$\mathscr{T}_{A \cup B}^{t}f(x) = \mathbb{E}\left[f(X(t \wedge T_{A \cup B})) \mid X(0) = x\right]. \qquad (6)$$

Here and below $t \wedge T_{A\cup B} = \min\{t, T_{A\cup B}\}$, indicating that the evolution process does not proceed beyond escape.

For a more general domain D and $T_{D^c} = \min\{t \geq 0 : X(t) \notin D\}$, the conditional expectation

$$u(x) = \mathbb{E}\left[b(X(T_{D^c})) - \int_0^{T_{D^c}} a(X(t))dt \mid X(0) = x\right] \ \text{ for } \ x \in D \qquad (7)$$

solves the equation

$$\mathscr{T}_{D^c}^{\tau}u(x) - u(x) = \int_0^{\tau} \mathscr{T}_{D^c}^{t}a(x)dt \ \text{ for } \ x \in D, \quad u(x) = b(x) \ \text{ for } \ x \notin D. \qquad (8)$$

To obtain (5) for the committor, choose $D = (A\cup B)^c$, $b = \mathbb{1}_B$, and $a = 0$. In (7) and (8) we assume for simplicity that $a(x) = 0$ for $x \notin D$. For a discrete-time process the time integral in these expressions should be interpreted as a sum.

Crucially, (8) holds for any choice of $\tau \geq 0$ including relatively small values. For very large values of τ, (8) converges to (7). However, in most cases of interest, the escape time T_{D^c} is very large, making estimation of u in (7) by direct simulation of sample trajectories of $X(t)$ prohibitively expensive. In the context of DGA, the significance of (8) is that it expresses u in terms of an expectation over short trajectories. The catch is that (8) must be "inverted" to solve for u.

2.2 Dynamical Galerkin Approximation (DGA)

We now describe a Galerkin approach to approximating conditional expectations from short trajectory data. We first introduce a "guess" function ψ that satisfies the boundary conditions (i.e., $\psi(x) = b(x)$ for $x \notin D$). Our approximation has the form

$$u(x) \approx \psi(x) + \sum_{j=1}^{n} \phi_j(x) v_j, \tag{9}$$

where $\{\phi_j(x)\}$ is a set of n basis functions satisfying $\phi_j(x) = 0$ for $x \notin D$, and v is a vector of n coefficients.

2.2.1 Forward-in-time predictions—We begin by approximating predictions of quantities forward-in-time as in (7) by expanding the solution u of (8) at a particular user chosen value of τ called the lag time. While the solution u itself is independent of τ in (8), the quality of our approximation of u with a finite basis may depend on the choice of lag time (even in the absence of sampling error). A similar phenomenon has recently been explained in detail in the context of the VAC algorithm.[42] Substituting (9) into (8), multiplying by ϕ_i and integrating over the distribution of sampled points μ to form the inner product $\langle f, g \rangle = \int f(x) g(x) \mu(dx)$, we obtain the linear system of equations:

$$\left(C^\tau - C^0 \right) v = r^\tau, \tag{10}$$

with matrices $C^s \in \mathbb{R}^{n \times n}$ for $s = 0, \tau$,

$$C_{ij}^s = \left\langle \phi_i, \mathscr{T}_D^s c \phi_j \right\rangle_\mu \tag{11}$$

and vector $r^\tau \in \mathbb{R}^n$,

$$r_i^\tau = \left\langle \phi_i, \psi(x) - \mathscr{T}_D^\tau c \psi(x) + \int_0^\tau \mathscr{T}_D^t c a(x) dt \right\rangle_\mu. \tag{12}$$

Given (11) and (12), (10) can be readily solved for v by standard methods of linear algebra.

In models that represent molecules with high fidelity, (11) and (12) cannot be evaluated directly because a closed form of $\mathscr{T}_D^\tau c$ is not known. DGA overcomes this issue by approximating the action of the transition operator using short molecular dynamics trajectories: if $X(0)$ is a sample drawn from μ and $\{X(t)\}_{t=0}^\tau$ is a trajectory segment of length τ starting from $X(0)$, then we can estimate C_{ij}^s (for $s = 0, \tau$) and r_i^τ as

$$C_{ij}^s \approx \frac{1}{M} \sum_{m=1}^{M} \phi_i \left(X^{(m)}(0) \right) \phi_j \left(X^{(m)}(s \wedge T_D c) \right) \tag{13}$$

$$r_i^\tau \approx \frac{1}{M} \sum_{m=1}^{M} \phi_i\left(X^{(m)}(0)\right)\left(\psi\left(X^{(m)}(0)\right) - \psi\left(X^{(m)}(\tau \wedge T_{D^c})\right) + \Delta \sum_{p=0}^{N-1} h\left(X^{(m)}(p\Delta)\right)\right) \quad (14)$$

where m indexes trajectory segments, is the sampling interval, and N satisfies $N\Delta = \tau \wedge T_{D^c}$. To avoid overhead, it is advantageous to generate trajectories much longer than τ (but still much shorter than typical values of T_{D^c}) and use a rolling window to generate short trajectories of length τ. We further note that in practice configurations are not saved at every molecular dynamics step. This limits the resolution of both the lag time and the stopping time, which we take to be the time of the first saved configuration outside the domain D.

2.2.2. Adjoints, the steady state, and backward-in-time predictions—To compute many important quantities we need not only to solve equations involving the transition operator but also equations involving its adjoint $\left(\mathcal{T}^t\right)_\mu^\dagger$ in the μ-weighted inner product, which by definition satisfies

$$\int f(x)\mathcal{T}^t g(x)\mu(dx) = \int g(x)\left(\mathcal{T}^t\right)_\mu^\dagger f(x)\mu(dx). \quad (15)$$

One such equation is for the change of measure $w = d\pi/d\mu$, which can be used to reweight from the sampling distribution μ to the stationary distribution π:

$$\int f(x)\pi(dx) = \int f(x)w(x)\mu(dx), \quad (16)$$

assuming μ and w are normalized such that $\int w(x)\mu(dx) = 1$. Owing to the time translational invariance of averages over the stationary distribution π, (15), and (16), the change of measure satisfies the equation

$$\left(\mathcal{T}^\tau\right)_\mu^\dagger w(x) - w(x) = 0. \quad (17)$$

(17) can be solved analogously to (8), but, in this case, there are no boundary conditions. The introduction of a basis leads to a linear system of equations of the form

$$\left(\bar{C}^\tau - \bar{C}^0\right)^\top v = 0, \quad (18)$$

with $\bar{C}^s$ (for $s = 0, \tau$) differing from C^s only in the choice of basis (which is no longer restricted to D) and the use of $\mathcal{T}^s$ in place of $\mathcal{T}^s_{D^c}$; $\top$ denotes the transpose. We note that by including $\phi_1(x) = 1$ in the basis we can guarantee that the equation for v has a solution. Given an approximate w, (16) can be computed as

$$\int f(x)\pi(dx) \approx \sum_{j=1}^{M} f\big(X^{(j)}(0)\big)w\big(X^{(j)}(0)\big),\tag{19}$$

with the weights normalized such that

$$\sum_{j=1}^{M} w\big(X^{(j)}(0)\big) = 1.\tag{20}$$

That the change of measure can be estimated from short nonequilibrium trajectory data was previously observed in ref. 16.

Another important quantity expressible in terms of an equation involving an adjoint of the transition operator is the backwards committor

$$q_-(x) = \mathbb{P}[X(-T^-_{A\,\cup\,B}) \in A \mid X(0) = x], \quad T^-_{A\,\cup\,B} = \min\{t \ge 0 : X(-t) \in A \cup B\},\tag{21}$$

for $x \notin A \cup B$, where $X(-t)$, $t \ge 0$ is the steady-state backward-in-time process governed by the transition operator

$$\mathcal{T}^{-t}f(x) = \big(\mathcal{T}^t\big)^\dagger_\pi f(x) = \frac{1}{w(x)}\big(\mathcal{T}^t\big)^\dagger_\mu[fw](x)\tag{22}$$

(the last equality can be verified using (15)). The backward committor is the probability that a trajectory currently at position x last came from the reactant state A rather than the product state B. It satisfies the Feynman-Kac relation

$$\mathcal{T}^{-\tau}_{A\,\cup\,B}q_-(x) - q_-(x) = 0 \ \text{ for } \ x \notin A \cup B, \quad q_-(x) = \mathbb{1}_A(x) \ \text{ for } \ x \in A \cup B.\tag{23}$$

Consistent with our definition of $\mathcal{T}^t_{A\,\cup\,B}$ above, $\mathcal{T}^{-t}_{A\,\cup\,B}$ is the transition operator for the steady-state backward-in-time process stopped upon first entrance in $A \cup B$.

To expand and approximate q_- according to the DGA recipe described above, we need to estimate μ-weighted inner products involving $\mathcal{T}^{-t}_{A\,\cup\,B}$. To that end we note that, as long as $g = 0$ on $A \cup B$,

$$\langle g, \mathcal{T}^{-t}_{A\,\cup\,B}f\rangle_\mu = \int \mathbb{E}\left[f(X(S_{A\,\cup\,B}(t)))\frac{g(X(t))}{w(X(t))} \mid X(0) = x\right]w(x)\mu(dx)\tag{24}$$

where

$$S_{A\,\cup\,B}(t) = \max\{s \le t : X(s) \in A \cup B\}\tag{25}$$

(with $S_{A\cup B}(t) = 0$ if $X(s) \notin A \cup B$ for all $0 \le s \le t$). We provide a derivation of (24) in Appendix A. Just as for the forward committor, we expect that use of a sampling measure μ with high resolution in transition regions will lead to higher approximation accuracy (i.e.,

better ability of a finite basis to capture the dynamics). However, in our experience the factor of $w^{-1}(X(t))$ in (24) leads to significant sampling errors for larger values of t. For our backward committor calculation we therefore weight inner products by π, using the formula

$$\left\langle g, \mathcal{T}_{A \cup B}^{-t} f\right\rangle_\pi = \int \mathbb{E}[f(X(S_{A \cup B}(t)))g(X(t)) \mid X(0) = x]w(x)\mu(dx). \tag{26}$$

(26) allows inner products involving $\mathcal{T}_{A \cup B}^{-t}$ to be computed using forward trajectories of X initiated according to μ, i.e., exactly the same ingredients required to make forward-in-time predictions by DGA.

Following our procedure for forward quantities outlined in Section 2.2, given a guess function ψ satisfying $\psi = 1$ on A and $\psi = 0$ on B and basis functions ϕ_j that are zero on $A \cup B$, we can build an approximation

$$q_-(x) \approx \psi(x) + \sum_{j=1}^{n} \phi_j(x)v_j \tag{27}$$

by solving

$$\left(C^{-\tau} - C^0\right)v = r^{-\tau} \tag{28}$$

with

$$C_{ij}^{-\tau} = \left\langle \phi_i, \mathcal{T}_{A \cup B}^{-\tau} \phi_j \right\rangle_\pi = \int \mathbb{E}\left[\phi_j(X(S_{A \cup B}(\tau)))\phi_i(X(\tau)) \mid X(0) = x\right]w(x)\mu(dx) \tag{29}$$

and

$$\begin{aligned} r_i^{-\tau} &= \left\langle \phi_i, \psi \right\rangle_\pi - \left\langle \phi_i, \mathcal{T}_{A \cup B}^{-\tau} \psi \right\rangle_\pi \\ &= \int \mathbb{E}\left[(\psi(X(\tau)) - \psi(X(S_{A \cup B}(\tau))))\phi_i(X(\tau)) \mid X(0) = x\right]w(x)\mu(dx) \end{aligned} \tag{30}$$

where the second equality in each display follows from (26).

Along with the forward committor q_+ and the stationary change of measure w, the backward committor is a key ingredient of TPT. In Section 2.3 we describe how DGA estimates of these quantities can be combined with TPT to reveal key properties of steady-state transition paths from the reactant state A to the product state B. However, before that, we complete our presentation of DGA with a discussion of molecular representations and basis sets, with emphasis on those that we employ in the present study to analyze trp-cage miniprotein unfolding and folding.

2.2.3 Basis functions—A key determinant of the performance of DGA is the choice of basis set. Constructing a basis set that respects the boundary conditions of the problem and captures the dynamics with relatively few functions requires care. Here we discuss how

we generated the basis sets that we compare later in our numerical experiments, and explain why we chose them over alternatives.

In addition to the choice of functions, there is also a choice of molecular representation (i.e., the features that serve as inputs to the functions). Although molecular dynamics trajectories are generally recorded as sequences of Cartesian coordinates, the inputs to the basis functions are generally internal coordinates. This removes the effects of trivial translations and rotations, and it can improve the statistics. The internal coordinates that we use are pairwise distances between all C_α atoms, except those pairs which are less than three sequence positions apart; for trp-cage, there are 153 such distances. In other words, the process $X(t)$ to which we apply DGA (and TPT) is the length 153 vector of pairwise distance values. In our tests we found that including additional features, such as backbone dihedral angles, did not improve performance. We assume that the reactant state A and product state B of interest can be characterized in terms of these variables. We construct basis functions of these variables that satisfy the homogeneous boundary condition on the domain $D = (A \cup B)^c$.

In this work, we compare three choices of basis set: indicator functions on the pairwise distances, indicator functions constructed on the top 10 TICA coordinates[39–41] computed from the pairwise distances at a lag time of 0.5 ns, and smooth functions of pairwise distances that satisfy the boundary conditions. We refer to these henceforth as the distance indicator, TICA indicator, and modified distance basis sets. We constructed the distance indicator and TICA indicator basis sets and their guess functions as follows:

1. For the distance indicator basis set, we constructed 200 indicator functions by mini-batch k-means clustering as implemented in PYEMMA on the values of the 153 pairwise distances. For the TICA indicator basis set, the clustering was performed on the top 10 TICA coordinates constructed on the pairwise distances.

2. We retained all resulting indicator functions with non-zero regions fully contained in $(A \cup B)^c$ as the basis set. We split any indicator functions with non-zero regions overlapping with A or B, and we redefined them to be non-zero only in the portions in $(A \cup B)^c$. For the change of measure calculations, boundary conditions are not present, so we used all indicator functions unmodified.

3. For the forward committor calculation we took the the guess function to be $\psi(x) = \mathbb{1}_B(x)$. For the backward committor calculation we took the guess function to be $\psi(x) = \mathbb{1}_A(x)$.

With an indicator basis, the DGA and MSM estimator (with appropriate state definitions) of the forward committor q_+ and change of measure w become similar.[7] We note however that the DGA (as formulated here) and MSM approaches diverge both in DGA's use of stopped trajectories and in the way q_+ and w (and q_-) are used to estimate TPT quantities as described in Section 2.3.

We constructed the distance basis set and its guess function as follows:

1. We computed d_A and d_B as the distance in feature space (i.e. in 153-dimensional Euclidean space) to the sampled points in states A and B, respectively.

2. We set $h(x) = d_A d_B/(d_A + d_B)^2$, which obeys the homogeneous boundary conditions by construction.

3. We computed basis functions obeying the boundary conditions by multiplying each coordinate of the pairwise distance vector x by $h(x)$: $\phi_i(x) = x_i h(x)$. For the change of measure calculation, we use $\phi_i = x_i$ and add the constant function into the set of chosen features.

4. To remove any linear dependencies introduced by enforcing the boundary conditions, and to ensure numerical stability, we orthogonalized the basis set ϕ_i with respect to the sampling measure (up to sampling error) using a singular value decomposition.

5. For the forward committor calculation we took the guess function to be $\psi(x) = d_B^2/(d_A + d_B)^2$. For the backward committor calculation we took the guess function to be $\psi(x) = d_A^2/(d_A + d_B)^2$.

Although here we use the backbone pairwise distances, we note that this construction procedure could be used to generate basis sets obeying the homogeneous boundary conditions for a choice of variables other than the pairwise distances such as dihedral angles, radial basis functions, or soft indicator functions.

The indicator and TICA basis sets are the most widely used in the MSM literature. Various alternatives have been proposed specifically in the context of spectral estimation.[43–46] In our previous work,[7] we considered a basis set based on diffusion maps.[29] Due to the size of our trp-cage data set (~10^6 datapoints), the $O(N^3)$ scaling of the matrix diagonalization associated with the diffusion map proved prohibitively computationally costly without subsampling and out of sample extension.

2.2.4   Delay Embedding—Application of DGA as described so far assumes that the underlying process $X(t)$ is Markovian; the conditional expectations that DGA seeks to approximate are not fully defined if $X(t)$ is not Markovian. Yet, in the previous section we described an approach to building a basis set for DGA consisting of functions of only a subset of the full collection of variables (selected pairwise distances). Though the dynamics of this subset are not strictly Markovian, in Section 4 we show that, at least in the specific context of the trp-cage system, the remaining degrees of freedom relax sufficiently fast that DGA yields accurate results.

However, in some circumstances, one may only have access to a small number of variables that are insufficient to specify the dynamics. This situation is typical when the data are from an experiment. In this case, we can construct a more expressive representation of the system from time-lagged images, i.e., if $X(t)$ is not itself Markovian we can instead apply DGA to the augmented process $\overline{X}(t) = (X(t - M\delta), X(t - (M - 1)\delta), ..., X(t))$.[7] For large enough M one can expect $\overline{X}(t)$ to be nearly Markovian. State space augmentation was also used in the history-augmented MSM (haMSM) approach of ref. 47 to obtain accurate

MFPT estimates at all lag times. Our approach differs in that we construction our basis on the delay embedded space, whereas in the haMSM approach the transition probabilities are conditioned on visiting multiple clusters in sequence. In principle, one can explicitly include memory as defined by the Mori-Zwanzig formalism,[48] though our delay-embedding approach is computationally more straightforward because it does not require choosing a form for the memory kernel and then estimating it from data, both of which are quite challenging.[49,50]

In Section 4.5, we show that delay embedding can significantly improve DGA estimates when a small number of CVs is used to characterize molecular configurations. Writing the values of the CVs at time t as the vector $X(t)$, we construct the delay embedded process $\overline{X}(t)$. We then construct a basis set following the recipe in Section 2.2.3 for the modified distance basis, but replacing X with $\overline{X}$. We then extend other functions f of the CV space to the delay-embedded space by $f(\overline{X}(t)) = f(X(t - \lfloor M/2 \rfloor \delta))$. This allows us to extend the states A and B (which can both be defined in terms of the CVs) as well as the functions a and b in (7). We then apply DGA as outlined above directly on the delay-embedded space.

2.3 Reaction rates and currents

Estimates of rates from simulations are frequently of interest because they can be compared directly with experimental measurements, and they can provide indirect information about mechanisms. TPT in principle provides not just rate estimates but reactive currents or fluxes, which provide direct information about mechanisms. However, previous calculations of reactive current have been limited to toy models and depictions of the reactive flux between metastable states can been difficult to interpret. Working within the TPT framework and building upon DGA approximations of w, q_+, and q_-, in this section we introduce robust estimates of the reaction rate and of an easily interpretable projection of the reactive current onto CVs (as opposed to over the network of metastable states).

There are various expressions for the rate in TPT. One approach is based on the rate at which trajectories transition from A to B, R_{AB}. If U is any set for which $A \subset U$ and $B \subset U^c$ then, for a continuous time process,

$$
\begin{aligned}
R_{AB} &= \lim_{t \to 0} \frac{1}{t} \int \left(\mathbb{1}_U(x) \mathscr{T}^t[\mathbb{1}_{U^c} q_+](x) - \mathbb{1}_{U^c}(x) \mathscr{T}^t[\mathbb{1}_U q_+](x) \right) q_-(x) \pi(dx) \\
&= \lim_{t \to 0} \frac{1}{t} \int \left(\mathbb{1}_U(x) \mathscr{T}^t q_+(x) - \mathscr{T}^t[\mathbb{1}_U q_+](x) \right) q_-(x) \pi(dx),
\end{aligned}
\tag{31}
$$

where the second line is obtained by noting that $\mathbb{1}_{U^c}(x) = 1 - \mathbb{1}_U(x)$. Here and below, for a discrete time process the limit is removed and t is replaced by the unit of a single time step.

Expression (31) simply counts trajectories with forward crossings of the surface dividing U and U^c, weighted by their probabilities that they start in A and end in B. Consequently, when using this formula to estimate rates from data, only those trajectories that cross the surface dividing U and U^c contribute. Because these trajectories are generally a small fraction of the data, this results in relatively large variances in estimates. We can obtain considerably better estimates by considering the isocommittor surfaces: $U(z) = \{ x : q_+(x)$

z, x ∈ D} for $z \in (0, 1)$, and noting that R_{AB} is independent of z. Integrating (31) with respect to z,[22] then exchanging integrals over z with applications of $\mathscr{T}^t$ and noting that $\int_0^1 \mathbb{1}_{[0, z]}(q_+(x))dz = 1 - q_+(x)$, we find that

$$
\begin{aligned}
R_{AB} &= \lim_{t \to 0} \frac{1}{t} \iint_0^1 \left\{ \mathbb{1}_{[0, z]}(q_+(x))\mathscr{T}^t q_+(x) - \mathscr{T}^t[q_+ \mathbb{1}_{[0, z]}(q_+)](x) \right\} dz\, q_-(x)\pi(dx) \\
&= \lim_{t \to 0} \frac{1}{t} \int \left\{ [1 - q_+(x)]\mathscr{T}^t q_+(x) - \mathscr{T}^t[q_+(1 - q_+)](x) \right\} q_-(x)\pi(dx) \\
&= \lim_{t \to 0} \frac{1}{t} \int \left(\mathscr{T}^t q_+^2(x) - q_+(x)\mathscr{T}^t q_+(x) \right) q_-(x)\pi(dx),
\end{aligned}
\tag{32}
$$

where we have made use of the fact that the integral of the Heaviside function (which enters through the indicator functions) is the ramp function. This expression for R_{AB} immediately suggests the estimator:

$$
R_{AB} \approx \frac{1}{t} \sum_i q_+\left(X^{(i)}(t \wedge T_{A \cup B})\right)\left[q_+\left(X^{(i)}(t \wedge T_{A \cup B})\right) - q_+\left(X^{(i)}(0)\right)\right]q_-\left(X^{(i)}(0)\right)w\left(X^{(i)}(0)\right)
\tag{33}
$$

for some small choice of t. Note the use of stopped trajectories in (33). For very small values of t the inclusion of the stopping time $T_{A \cup B}$ has no impact. However, in our numerical experiments we find that use of stopped trajectories improves the accuracy of (33) and, in particular, (37) below, for most choices of t. Given an estimate of R_{AB}, the rate constant is

$$
k_{AB} = \frac{R_{AB}}{\sum_i q_-\left(X^{(i)}(0)\right)w\left(X^{(i)}(0)\right)}.
\tag{34}
$$

The denominator in 34 is the mean of the backward committor, which is the fraction of time the system spends having last visited state A.

As noted above, we can also use simulations to understand how reactive trajectories flow through a CV space. One way to do this is to partition the space into discrete states and then estimate the reactive fluxes between pairs of states.[24] However, the resulting directed graph can be complicated and difficult to interpret. When the sample paths are continuous the reactive flux between neighboring values in CV space is can be summarized as a single vector field in CV space. If θ is a vector-valued CV and ds is a bin in CV space of volume $|ds|$, the reactive current at point s is

$$
\begin{aligned}
J_{AB}^\theta(s) &= \lim_{t, |ds| \to 0} \frac{1}{2t|ds|} \int \left(\mathscr{T}^t[\theta q_+](x) - \theta(x)\mathscr{T}^t q_+(x) \right)\mathbb{1}_{\{\theta \in ds\}}(x)q_-(x)\pi(dx) \\
&\quad + \left(\mathscr{T}^t\left[\mathbb{1}_{\{\theta \in ds\}}\theta q_+\right](x) - \theta(x)\mathscr{T}^t\left[\mathbb{1}_{\{\theta \in ds\}}q_+\right](x) \right)q_-(x)\pi(dx)
\end{aligned}
\tag{35}
$$

In Appendices B and C, we show that $J_{AB}^\theta(s) = \int J_{AB} \cdot \nabla\theta(x)\delta(\theta(x) - s)\pi(dx)$, and we establish that the projected reactive current satisfies

$$\int_{\partial C^\theta} J^\theta_{AB}(s) \cdot n_C\theta d\sigma_C\theta = \int_{\partial C} J_{AB} \cdot n_C d\sigma_C, \tag{36}$$

where C^θ is any region of CV space such that its inverse image (under the CV mapping) in the full configuration space, C, contains A and does not intersect B. To estimate J^θ_{AB} from trajectory data we have the following estimator:

$$
\begin{aligned}
J^\theta_{AB}(s) &\approx \frac{1}{2t|ds|} \sum_{i=1}^{M} q_+\big(X^{(i)}(t \wedge T_{A \cup B})\big)\big(\theta\big(X^{(i)}(t \wedge T_{A \cup B})\big) - \theta\big(X^{(i)}(0)\big)\big) \\
&\times \mathbb{1}_{\theta \in ds}\big(X^{(i)}(0)\big)q_-\big(X^{(i)}(0)\big)w\big(X^{(i)}(0)\big) \\
&+ \frac{1}{2t|ds|} \sum_{i=1}^{M} q_+\big(X^{(i)}(t)\big)\big(\theta\big(X^{(i)}(t)\big) - \theta\big(X^{(i)}(S_{A \cup B}(t))\big)\big) \\
&\times \mathbb{1}_{\theta \in ds}\big(X^{(i)}(t)\big)q_-\big(X^{(i)}(S_{A \cup B}(t))\big)w\big(X^{(i)}(0)\big)
\end{aligned}
\tag{37}
$$

Note that the lag time t in (33) and (37) need not be the same as the lag time τ used to estimate the committors q_+ and q_-. Even with perfect sampling and a perfect basis, estimates of TPT quantities will depend on t, in contrast to τ. Several considerations are involved in the choice of t. For larger values of t (33) and (37) incur significant bias due to poor approximation of the $t \to 0$ limit in (32) and (35). On the other hand, for small values of t, we found that (33) and (37) suffer large statistical errors. Alternative estimators for the rate and reactive current based on expressions that are exact at any lag time up to error from the discrete sampling interval are given in Appendix D. A full analysis of error sources is beyond the scope of this work, and in practice we choose a lag time that gives reasonable results for the change of measure and reasonable smoothness in the vector field.

3 Simulation methods and choices

In this section, we specify the computational procedure to generate and analyze the data set for the unfolding and folding of trp-cage. We describe preparing the system and its underlying dynamics (Section 3.1), choosing collective variables based on their ability to distinguish metastable states (Section 3.2), generating and validating the data set of short trajectories (3.3), and defining the unfolded and folded states (Section 3.4).

3.1 System setup

Unless otherwise noted, all molecular dynamics simulations were performed with GROMACS 5.1.4[51] and PLUMED 2.3[52–54] using the CHARMM36m force field[55–57] in the NVT ensemble at 300 K using the Langevin thermostat with a temperature coupling constant of 10 ps^{-1} applied to all atoms, and a time step of 2 fs. Bonds to hydrogen atoms were constrained using the LINCS algorithm.[58] Electrostatic interactions were computed using particle-mesh Ewald summation with a cutoff of 1.2 nm. Lennard-Jones interactions were switched off from 1.0 to 1.2 nm using the default GROMACS switching function.

The system was prepared from an NMR structure of trp-cage (PDB code 1L2Y[59]). The protein was solvated in a 50 Å cubic box with the TIP3P water model[60] using CHARMM-

GUI 3.0.[61,62] 10 K^+ and 11 Cl^- ions were added, bringing the system to charge neutrality and 150 mM KCl. The energy of the system was minimized until the maximum force was below 1000 kJ/mol nm. The system was then equilibrated for 1 ns in the NVT ensemble with position restraints (using a 1 fs timestep), 10 ns in the NPT ensemble with harmonic restraints on non-hydrogen atom positions (force constant 400 kj/mol nm^2 for backbone atoms and 40 kj/mol nm^2 for side chain atoms.) and a Parrinello-Rahman barostat with a pressure coupling constant of 5 ps^{-1}, 5 ns in the NPT ensemble without position restraints, and then 10 ns in the NVT ensemble without position restraints. The cubic box length was determined from the restraint-free NPT equilibration run to be 4.48 nm and fixed at that value after that run.

3.2 Choice of CVs

The performance of DGA rests on having a data set with good sampling of all states that contribute to the reaction mechanism. As mentioned in the Introduction, the available physically weighted molecular dynamics data for trp-cage[36] contain few unfolding and folding transitions. We thus sought to use enhanced sampling methods to generate a data set with improved representation outside the stable states. To this end, we evaluated CVs for their ability to control sampling and resolve the unfolded and folded states.

Based on previous studies,[34,38] we considered five CVs:

1. The radius of gyration of the C_α atoms (R_g);

2. The root mean squared deviation (RMSD) of all C_α atoms from their positions in an equilibrated structure ($RMSD_{full}$);

3. The RMSD of the C_α atoms of residues 2 to 9, which make up the α helix in the native state ($RMSD_{hx}$);

4. The RMSD of the C_α atoms of residues 11 to 15, which make up the 3–10 helix in the native state ($RMSD_{3-10}$);

5. The end-to-end distance (d).

R_g, $RMSD_{full}$, and $RMSD_{hx}$ were used in ref. 34, and $RMSD_{3-10}$ was used in ref. 38 (there defined only to residue 14), where they found that it was able to resolve several metastable states identified by spectral clustering.

To explore how these collective variables change as trp-cage unfolds, we ran a series of Adiabatic Bias Molecular Dynamics (ABMD)[63] simulations to drive unfolding from the equilibrated native structure. ABMD uses a ratchet-and-pawl-like bias to trap spontaneous fluctuations that move the system forward in selected CVs. By applying ABMD with different combinations of the CVs above, we found that $RMSD_{full}$ and $RMSD_{3-10}$ yielded reasonable control of the system and enabled exploration of all metastable states characterized in previous studies.

3.3 Generation of the DGA data set

To initialize a data set of short trajectories for DGA, we defined a grid of 64 points in the space of $RMSD_{full}$ and $RMSD_{3-10}$ (Figure 1). We then used 64 independent ABMD

simulations to steer the system to each of these points from the final structure from the equilibration simulations described in Section 3.1. We ran each ABMD simulation for 1 ns, saving the structure every 5 ps; the force constants were 1.25 kJ/(mol Å^2) and 1.0 kJ/(mol Å^2) for RMSD$_{full}$ and RMSD$_{3-10}$, respectively. From the set of all recorded structures, we chose the 64 structures closest to the targets and equilibrated each for 1 ns with a harmonic restraint with the same force constants as in the AMBD simulations. From each of the resulting structures, we then launched 14 free simulations (with different random number generator seeds) of length 30 ns each, saving structures every 5 ps.

From this data set, we computed all possible two-dimensional potentials of mean force (PMFs) involving the CVs listed in Section 3.2. We compared these PMFs with corresponding ones from replica exchange umbrella sampling (REUS). Based on the DGA PMFs, we used the RMSD of the a helix (RMSD$_{hx}$), and the RMSD of the 3–10 helix (RMSD$_{3-10}$), and the end-to-end distance (d) to control the sampling. REUS window centers were placed on a uniform $8 \times 8 \times 8$ grid of these three CVs, with RMSD$_{hx}$ ranging from 0.3 to 2.8 Å, RMSD$_{3-10}$ ranging 0.3 to 3.3 Å, and d ranging from 6 to 38 Å. This grid fully covered the relevant areas of CV space identified by previous simulations. The force constants for the harmonic potentials for each window were 29.2 kJ/(mol · Å^2) for RMSD$_{hx}$, 20.3 kJ/(mol · Å^2) for RMSD$_{3-10}$, and 0.178 kJ/(mol · Å^2) for d, following ref. 64. To initialize each window, structures were taken from the DGA database that were closest to each window center. The built-in replica exchange functionality of GROMACS was used to create a three-dimensional replica exchange procedure, where structures from nearby windows were periodically exchanged.[65] Every window was first simulated for 100 ps, with swaps attempted between adjacent windows in d space (i.e., window centers with the same RMSD$_{hx}$ and RMSD$_{3-10}$ values, but neighboring d values) every 10 ps. This was repeated for a total of three 100 ps iterations, with the second and third iterations proposing swaps between neighboring windows in RMSD$_{hx}$ and RMSD$_{3-10}$, respectively. This 300-ps procedure was repeated until a total simulation time of 10 ns was reached for each window, with structures saved every 10 ps. Following this protocol, structures were exchanged across all of the three-dimensional grid, with exchange probabilities in the range 10–60%. The PMF was constructed by using the Eigenvector Method for Umbrella Sampling (EMUS)[66] extended to REUS.[67] The REUS simulations were run until the asymptotic variance of the PMF dropped below 0.1 ($k_B T$)^2 (Figure S1).

The REUS PMFs suggested that the initial DGA data set did not adequately sample configurations with RMSD$_{hx} > 1.5$ Å (Figure S3, note the lack of sampling toward the upper right areas of the plots compared with those in Figure 2). In this case several of the basins are missing, and the RMSD over all bins is $> 1.3 \, k_B T$. Therefore, we selected 64 more points from a grid with RMSD$_{hx} > 1.5$ Å and a range of end-to-end distances from our short trajectory data set. From each of these points, we released two new free molecular dynamics simulations of length 30 ns (Figure 1B). With these additional trajectories, we obtained good agreement between DGA and REUS PMFs. Adding the extra sampling improved the PMFs involving RMSD$_{hx}$ the most, but other PMFs were also noticeably improved. The data set used for all further DGA calculations thus contains a total of 1024 trajectories, each of length 30 ns, with structures saved every 5 ps.

3.4 State definitions

We found that PMFs projected onto only global measures of unfolding ($RMSD_{full}$, R_g, and d) did not have clearly identifiable unfolded basins (Figures 1 and 2). By contrast, the PMF on the CVs tracking secondary structure ($RMSD_{hx}$ and $RMSD_{3-10}$) had clearly identifiable unfolded and folded basins, as well as several intermediates. Based on this analysis, we took the unfolded state to be

$$\frac{\left| RMSD_{hx} - 2.15 \text{ Å} \right|^3}{0.008 \text{ Å}^3} + \frac{\left| RMSD_{310} - 2.8 \text{ Å} \right|^3}{0.125 \text{ Å}^3} < 1. \tag{38}$$

The folded state is

$$\frac{\left(RMSD_{hx} - 0.3 \text{ Å} \right)^2}{0.0289 \text{ Å}^2} + \frac{\left(RMSD_{310} - 0.3 \text{ Å} \right)^2}{0.04 \text{ Å}^2} < 1 \text{ and } d < 17 \text{ Å}. \tag{39}$$

We included the end-to-end distance constraint on the folded state to exclude structures which are extended but have the secondary structure intact.

Heterogeneous structures contribute to the unfolded state, making it challenging to define, and there is no guarantee that the choices above are optimal in any sense. Because we expect unfolding and folding to be among the slowest motions of the system, an alternative would be to define the states in terms of the slowest mode of the system identified by a dimensionality-reduction algorithm. However, data-driven state definitions are often difficult to interpret physically, despite their theoretical justifications. Furthermore, data-driven state definitions can be difficult to incorporate into sampling algorithms. We thus use physical CVs for path sampling, stratification, and state definitions, and we then check for consistency with a data-driven state choice.

Figure 3 shows that the slowest mode of the system identified by TICA applied to the DGA data set correlates with the PMF and switches between low and high values in going between the unfolded and folded states. Here and going forward, all functions we project onto CVs are conditional averages of the form $\int f(x)\delta(\theta(x) - s)\pi(dx)/\int \delta(\theta(x) - s)\pi(dx)$. We estimate these by binning our CV space into bins, and for each bin ds, plotting:

$$\frac{\int f(x)\delta(\theta(x) - s)\pi(dx)}{\int \delta(\theta(x) - s)\pi(dx)} \approx \frac{\sum_i f\left(X^{(i)}(0)w(X^{(i)}(0))\mathbb{1}_{\theta \in ds}\left(\left(X^{(i)}(0)\right)\right)}{\sum_i w\left(X^{(i)}(0)\right)\mathbb{1}_{\theta \in ds}\left(\left(X^{(i)}(0)\right)\right)}. \tag{40}$$

We furthermore show in Section 4.2 that this mode correlates with the committor. We thus feel that $RMSD_{hx}$ and $RMSD_{3-10}$ enable the clearest two-dimensional projection of the reaction and present most of our results in terms of these CVs. In addition to the unfolded and folded states, we define four intermediate states U1, U2, L1, and L2 shown on Figure 3. In the next section, we apply our DGA and TPT formalism to show that trp-cage can fold along an upper path through intermediates U1 and U2, or a lower path through L1 and L2.

4 Trp-cage analysis

In this section, we evaluate how the three basis sets described in Section 2.2.3 (indicator functions of pairwise distances, indicator functions of TICA coordinates, and pairwise distances modified to satisfy the boundary conditions) impact the performance of DGA for estimating PMFs, rates, committors, and reactive currents for the unfolding and folding of the trp-cage miniprotein. Where possible, we compare our results with references obtained by independent means.

4.1 Comparison of PMFs

Figure 2 shows PMFs computed on each pair of the physically motivated CVs with DGA with the modified distance basis set. The corresponding PMFs from REUS are shown in Figure S2; difference maps comparing the results obtained with the two methods and three basis sets are shown in Figures S4, S5, and S6. All of the main basins identified by REUS are present in the DGA PMFs, and there is good quantitative agreement between REUS and DGA, with RMSDs of $< 1\ k_B T$ for all three basis sets (that said, of these, the distance indicator basis set results in the largest deviations). Consistent with their agreement with the REUS PMF, the three DGA PMFs are in agreement with each other. We did observe that REUS tends to give slightly flatter PMFs than DGA with all three basis sets. In principle, there are two sources of error in the DGA PMFs: (i) approximation error from representing the true change of measure with a basis expansion and (ii) estimation (sampling) error. Analysis of error in DGA will be the subject of future work. Error in US is discussed in refs. 66, 67, and 68.

We found that the projection onto the $\mathrm{RMSD}_{\mathrm{hx}}$ and RMSD_{3-10} coordinates was best able to separate the pathways and states of interest, so we now focus on this projection. Figure 3 indicates the folded (lower left) and unfolded (upper right) basins, as well four intermediates. The intermediates define two pathways, which we label upper (with intermediates U1 and U2) and lower (with intermediates L1 and L2). Table 1 gives the five CV values for each of the six states.

To understand the characteristics of the intermediate states, we turn to Figure 4, which shows the solvent-accessible surface area (SASA) of trp-6 on the left, and pro-12 on the right. We find that the U2 intermediate state is characterized by partial solvation of the hydrophobic core, measured by the SASA of trp-6, and nearly full detachment of pro-12. Furthermore, the U2 state is significantly more extended than the lower pathway intermediates as measured both by R_g and end-to-end distance. In addition to being more compact, with near-native R_g values, L1 and L2 have near-native trp-6 and pro-12 SASA values, suggesting the hydrophobic core is fully formed. These intermediate states can be mapped to those previously reported in the literature. Bolhuis and Jurazek[33] identified three folding intermediates. Our U1 and U2 intermediates roughly map onto their P_d and I intermediates, and our L1 and L2 intermediates roughly map onto their L intermediate. U1 and U2 also correspond to states S_7 and S_0 identified by Sidky *et al.*[38]

4.2 Comparison of committors

We next calculated both forward and backward committors using DGA with the three basis sets and lag times ranging from 0.5 ns to 12 ns (Figure 5 and Figure S7). As they should, the backward committors mirror the forward committors, so we focus our discussion on the latter. The timescale of trp-cage folding is on the order of 5 μs from both experiment[32] and simulation,[34] thus both our trajectory lengths (30 ns) and lag times are several orders of magnitude shorter than the motions of interest, providing an appropriate setting in which we expect DGA to show benefits.

In contrast to the PMFs, we found the committors to be sensitive to the choice of basis set (and associated guess function). The modified distance basis set, in addition to being substantially faster to construct as it avoids slow and unstable high-dimensional clustering, is less prone to discontinuities at the boundary than the distance indicator function basis set. The TICA indicator function basis set performs similarly to the modified distance basis set and has the advantage over the distance indicator basis set that clustering on the lower-dimensional subspace is significantly faster and more stable.

For a given basis set, we found relatively little variation in the committors across lag times. This is in contrast to variational approach for conformational dynamics (VAC) algorithm, where the results can strongly depend on the lag time[42] (although this can be mitigated by using multiple lag times[69]). We postpone a full investigation of DGA's error properties, and in particular its dependence on the choice of lag time, to future work.

Because we expect unfolding and folding to be among the slowest motions of the system, we can validate the DGA committors by comparing them with the slowest mode of the system identified by TICA. Comparing Figures 3 and 5 shows that the largest TICA eigenvector (estimated with a lag time of 0.5 ns) correlates almost perfectly with the estimated committors obtained with the modified distance basis set, when projected onto $RMSD_{hx}$ and $RMSD_{3-10}$. The agreement between these two independent calculations furthermore suggests that the physically motivated CVs capture the behavior detected by the data-driven method. In this projection, we see that the transition states fall where the SASA of trp-6 (Figure 4) changes rapidly.

As an additional validation, we used DGA with the modified distance basis set and a lag time of 0.5 ns (Figure 6) to compute committors on the CVs used by Juraszek and Bolhuis.[33] When projected onto RMSD and $RMSD_{hx}$, the positions of the transition states in Figure 4 of ref. 33 fall in areas estimated to have $q_+ = 0.5$ (white in Figure 6). The traditional shooting approach employed in ref. 33 is quite computationally costly and provides information about only a limited number of structures. Our ability to capture the transition states thus makes clear the benefit of DGA. We discuss DGA's ability to provide mechanistic information further in the next section.

4.3 Reactive currents

We computed reactive currents for the three basis sets using the estimator in (37) and the committors from the the previous section (Figure 7). For this calculation, we use the shortest lag time of 0.5 ns for both the committor and reactive current, though in principle they

could be chosen separately. As previously, we primarily present our results projected onto $RMSD_{hx}$ and $RMSD_{3-10}$. Overall the results for the three basis sets are similar, though the distance indicator basis set exhibits greater noise around $(RMSD_{hx}, RMSD_{3-10}) = (1.3$ Å, 1.3 Å$)$, consistent with the plateau in the committor in this region.

The currents, which provide information directly about dynamics, confirm the presence of two paths for the folding process: an upper path with formation of the α helix prior to formation of the 3–10 helix, and a lower path with the order of these events transposed. The upper path proceeds through intermediates U1 and U2, with folding beginning with formation of the α helix and partial desolvation of trp-6, followed by full formation of the 3–10 helix. The lower path proceeds through L1 and L2, with folding beginning with collapse into the L2 intermediate with no α helix, but the hydrophobic core fully formed, followed by formation of the α helix. Both of these paths correspond to troughs in the PMFs on these CVs.

Previous studies have found multiple pathways resembling the ones we find here. Kim et al.[70] used diffusion maps to identify two pathways: one with tertiary contacts forming first, followed by α helix formation, and another with the order transposed. Jurazek and Bolhius came to similar conclusions using transition path sampling.[33]

An advantage of the reactive current is that we can use it to assign weights to the two paths. By computing the relative flux crossing $RMSD_{3-10} = 1.8$ Å with either $RMSD_{hx} < 1.4$ Å (upper pathway) or $RMSD_{hx} > 1.4$ Å (lower pathway), we conclude that 88% of the reactive paths proceed by first forming the α helix, and then the 3–10 helix and hydrophobic core (i.e., the upper pathway). Although we are not aware of a previous estimate of the reactive current for this system, we can compare these numbers to the frequencies with which transition path sampling sampled the pathways in ref. 33. There, Juraszek and Bolhuis observed the pathway in which tertiary contacts form first (i.e., the lower pathway) 80% of the time. The difference may be due to different CV and state definitions (Jurazek and Bolhuis[33] used 5 CVs in their state definitions, whereas we consider only $RMSD_{3-10}$ and $RMSD_{hx}$) or force field and setup differences.

4.4   Rates

Finally, we computed rates using the estimator in (31). We present our results as inverse rates (unfolding and folding times) to make comparisons to lag times and trajectory lengths clear. As mentioned previously, these times are expected to be on the order of microseconds. In particular, Juraszek and Bolhuis used transition interface sampling to estimate inverse unfolding and folding rates of 1.2 μs and 0.4 μs,[34] though as noted previously those results are for a different model.

All three basis sets gave rate estimates that were within an order of magnitude of those numbers (Figure 8). However, the results for the distance indicator basis were markedly faster. Furthermore, in all three cases, the inverse rate exhibited significant dependence on lag time. We do not show lag times >12 ns since they suffer from pronounced statistical error due to the limitations of our short-trajectory data set. Our analysis of the trajectory of the K8A mutant suggests the need for a lag time of at least 100 ns (consistent with ref.

38), though as discussed in the Introduction, those data do not contain a sufficient number of unfolding and folding events to obtain accurate rate estimates. Juxtaposed with the lack of sensitivity to lag time for the committor and reactive current, these observations suggest that DGA's strength is in its ability to give statistical insight into mechanisms with relatively little data, but that rates may be more efficiently computed by methods that directly sample relevant statistics such as stratification schemes.[5]

4.5 Demonstration of delay embedding

As described in Section 2.2.4, delay embedding can be used to construct an approximately Markovian process when the feature space does not fully capture the dynamics. To illustrate this idea using our trp-cage data set, we restrict the feature space to the five physical CVs and apply DGA with the modified distance basis set on either the feature space itself or the delay-embedded feature space. Figure 9 shows the reactive currents and committors resulting from DGA on these two spaces. We find that the committor and current constructed from the delay embedded representation largely agree with the DGA result constructed on the 153 pairwise distances. Without delay embedding, we find several qualitative disagreements, in particular the U2 state has a committor value close to zero, and the reactive current does not resolve the two pathways since many of the arrows point directly towards the folded state.

5 Conclusions

In this paper, we have cast the dynamical Galerkin approximation (DGA)[7] for computing chemical kinetic statistics from short trajectories in terms of the stopped transition operator. This formulation can be immediately translated into expressions that can be applied to simulation data. It also clarifies the role of the lag time, showing that estimates of conditional expectations computed by DGA are exact in the infinite basis and data limit, independent of the choice of lag time.

To evaluate DGA's performance, we generated and carefully validated a data set of short trajectories for the unfolding and folding of the trp-cage miniprotein, a well-characterized system. We used umbrella sampling to validate our short trajectory data set by comparing the resulting PMFs. Quantitative agreement between the PMFs was observed, suggesting that our short trajectory data set had sufficient sampling to compute dynamical statistics. The PMF calculations furthermore enabled us to rapidly assess different combinations of CVs for their abilities to separate metastable states. The α helix RMSD and 3–10 helix RMSD in particular allowed us to resolve intermediates to a greater degree than found in previous studies.

We next applied DGA to compute forward and backward committors between the unfolded and folded states. We evaluated a number of competing estimators for the backward committor and found that one based on forward trajectories weighted by the stationary distribution gave the best results. The committors by themselves are not able to identify reaction pathways or transition states, but they can be combined according to transition path theory to extract this information. Specifically, we introduce a new estimator for the TPT rate, and a projection formula and corresponding estimator for the reactive current in a CV

space. Our projected reactive current allows us to easily resolve and visualize the pathways that the system takes in arbitrary CV spaces, and even lets us assign relative weights to these pathways. Acquiring this kind of mechanistic information has previously been possible only through transition path sampling and related methods; such methods do not as readily allow exploration of CVs and state definitions because the sampling is linked directly to them.

We introduced a simple procedure that takes an arbitrary set of molecular features and adapts them to produce a basis set that satisfies the homogeneous boundary conditions. Using pairwise distances as the molecular features, we compared the performance of such a basis set with indicator functions on the molecular features and indicator functions on TICA coordinates. Other basis constructions such as diffusion maps and radial basis functions are possible, and we expect that the best choice will be system dependent. We applied our DGA and TPT formalism to our data set, and identified intermediate states and pathways which have been previously reported in the literature, providing further validation of our methods. We found that the estimates of the TPT rate, while on the same order of magnitude as previous estimates, nevertheless show significant dependence on lag time. Finally, we showed that delay embedding can be an effective strategy for constructing a molecular representation with approximately Markovian dynamics from a low-dimensional feature space.

Our results suggest several interesting directions for future investigation. We have seen that in our trp-cage application the choice of lag time has only a modest effect on DGA estimates of conditional expectations, while TPT quantities, in particular the rate, depend sensitively on lag time. Recently, we showed that integrating over lag times for VAC improves the robustness of that method.[69] It will be interesting to see if an analogous strategy can improve rate estimates from DGA. An in depth mathematical study of DGA's error and its dependence on lag time along the lines of our previous analysis of VAC[42] is also in order. By showing how DGA's results depend on the sampling measure, such an analysis could lead to a practical scheme for targeting sampling to selected regions, just as our analysis of US[71,72] did.[67] This will be particularly important for systems that are not amenable to the strategy that we took in the present study of using REUS for identifying regions of CV space that require more sampling.

Though DGA has performed well in our tests so far, looking ahead to larger and more complex systems, it may become necessary to move away from a Galerkin approach and toward more flexible representations of the kinetic functions we seek to approximate. This would be consistent with a trend toward using neural networks to represent eigenfunctions in spectral estimation.[17,38,69] Indeed, some of the first estimates of committors from data used neural networks.[73,74] Introducing this higher level of representational flexibility while maintaining the reliability we observe in our trp-cage application of DGA will be a challenge.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgement

References

(1). Huber GA; Kim S Weighted-ensemble Brownian dynamics simulations for protein association reactions. Biophysical Journal 1996, 70, 97–110. [PubMed: 8770190]

(2). Dickson A; Dinner AR Enhanced sampling of nonequilibrium steady states. Annual review of physical chemistry 2010, 61, 441–459.

(3). Guttenberg N; Dinner AR; Weare J Steered transition path sampling. The Journal of Chemical Physics 2012, 136, 234103, Publisher: American Institute of Physics. [PubMed: 22779577]

(4). Aristoff D; Bello-Rivas JM; Elber R A Mathematical Framework for Exact Milestoning. Multiscale Modeling & Simulation 2016, 14, 301–322, Publisher: Society for Industrial and Applied Mathematics. [PubMed: 27239166]

(5). Dinner AR; Mattingly JC; Tempkin JOB; Koten BV; Weare J Trajectory Stratification of Stochastic Dynamics. SIAM Review 2018, 60, 909–938, Publisher: Society for Industrial and Applied Mathematics. [PubMed: 34650314]

(6). Allen RJ; Frenkel D; ten Wolde PR Simulating rare events in equilibrium or nonequilibrium stochastic systems. The Journal of chemical physics 2006, 124, 024102. [PubMed: 16422566]

(7). Thiede EH; Giannakis D; Dinner AR; Weare J Galerkin approximation of dynamical quantities using trajectory data. The Journal of Chemical Physics 2019, 150, 244111. [PubMed: 31255053]

(8). Schmid PJ Dynamic mode decomposition of numerical and experimental data. Journal of Fluid Mechanics 2010, 656, 5–28, Publisher: Cambridge University Press.

(9). Prinz J-H; Wu H; Sarich M; Keller B; Senne M; Held M; Chodera JD; Schütte C; Noé F Markov models of molecular kinetics: Generation and validation. The Journal of chemical physics 2011, 134, 174105. [PubMed: 21548671]

(10). Nüske F; Keller BG; Pérez-Hernández G; Mey ASJS; Noé F Variational Approach to Molecular Kinetics. Journal of Chemical Theory and Computation 2014, 10, 1739–1752. [PubMed: 26580382]

(11). Williams MO; Kevrekidis IG; Rowley CW A Data–Driven Approximation of the Koopman Operator: Extending Dynamic Mode Decomposition. Journal of Nonlinear Science 2015, 25, 1307–1346.

(12). Husic BE; Pande VS Markov State Models: From an Art to a Science. Journal of the American Chemical Society 2018, 140, 2386–2396, Publisher: American Chemical Society. [PubMed: 29323881]

(13). Klus S; Nüske F; Koltai P; Wu H; Kevrekidis I; Schütte C; Noé F Data-driven model reduction and transfer operator approximation. Journal of Nonlinear Science 2018, 28, 985–1010.

(14). Pande VS; Beauchamp K; Bowman GR Everything you wanted to know about Markov State Models but were afraid to ask. Methods (San Diego, Calif.) 2010, 52, 99–105.

(15). Eisner T; Farkas B; Haase M; Nagel R Operator Theoretic Aspects of Ergodic Theory; Springer, 2015; Vol. 272.

(16). Wu H; Nüske F; Paul F; Klus S; Koltai P; Noé F Variational Koopman models: slow collective variables and molecular kinetics from short off-equilibrium simulations. The Journal of Chemical Physics 2017, 146, 154104, arXiv: 1610.06773. [PubMed: 28433026]

(17). Mardt A; Pasquali L; Wu H; Noé F VAMPnets for deep learning of molecular kinetics. Nature Communications 2018, 9, 1–11, Number: 1 Publisher: Nature Publishing Group.

(18). Nüske F; Schneider R; Vitalini F; Noé F Variational tensor approach for approximating the rare-event kinetics of macromolecular systems. The Journal of Chemical Physics 2016, 144, 054105. [PubMed: 26851906]

(19). Prinz J-H; Chodera JD; Noé F Spectral rate theory for two-state kinetics. Physical Review X 2014, 4, 011020. [PubMed: 25356374]

(20). Del Moral P Feynman-Kac Formulae; Springer, 2004.

(21). Karatzas I; Shreve S Brownian Motion and Stochastic Calculus; Springer Science & Business Media, 2012; Vol. 113.

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

(23). E. W; Vanden-Eijnden E Towards a Theory of Transition Paths. Journal of Statistical Physics 2006, 123, 503.

(24). Metzner P; Schütte C; Vanden-Eijnden E Transition Path Theory for Markov Jump Processes. Multiscale Modeling & Simulation 2009, 7, 1192–1219, Publisher: Society for Industrial and Applied Mathematics.

(25). Noé F; Schütte C; Vanden-Eijnden E; Reich L; Weikl TR Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. Proceedings of the National Academy of Sciences 2009, 106, 19011–19016.

(26). Liu Y; Hickey DP; Minteer SD; Dickson A; Calabrese Barton S Markov-State Transition Path Analysis of Electrostatic Channeling. The Journal of Physical Chemistry C 2019, 123, 15284–15292, Publisher: American Chemical Society.

(27). Meng Y; Shukla D; Pande VS; Roux B Transition path theory analysis of c-Src kinase activation. Proceedings of the National Academy of Sciences 2016, 113, 9193–9198.

(28). Elber R; Bello-Rivas JM; Ma P; Cardenas AE; Fathizadeh A Calculating isocommittor surfaces as optimal reaction coordinates with milestoning. Entropy 2017, 19, 219. [PubMed: 28757794]

(29). Coifman RR; Lafon S Diffusion maps. Applied and Computational Harmonic Analysis 2006, 21, 5–30.

(30). Müller K; Brown LD Location of saddle points and minimum energy paths by a constrained simplex optimization procedure. Theoretica chimica acta 1979, 53, 75–93.

(31). Shaw DE; Maragakis P; Lindorff-Larsen K; Piana S; Dror RO; East-wood MP; Bank JA; Jumper JM; Salmon JK; Shan Y; Wriggers W Atomic-Level Characterization of the Structural Dynamics of Proteins. Science 2010, 330, 341–346, Publisher: American Association for the Advancement of Science Section: Research Article. [PubMed: 20947758]

(32). Qiu L; Pabit SA; Roitberg AE; Hagen SJ Smaller and Faster: The 20-Residue Trp-Cage Protein Folds in 4 s. Journal of the American Chemical Society 2002, 124, 12952–12953. [PubMed: 12405814]

(33). Juraszek J; Bolhuis PG Sampling the multiple folding mechanisms of Trp-cage in explicit solvent. Proceedings of the National Academy of Sciences 2006, 103, 15859–15864, Publisher: National Academy of Sciences Section: Biological Sciences.

(34). Juraszek J; Bolhuis PG Rate Constant and Reaction Coordinate of Trp-Cage Folding in Explicit Water. Biophysical Journal 2008, 95, 4246–4257, Publisher: Elsevier. [PubMed: 18676648]

(35). Marinelli F; Pietrucci F; Laio A; Piana S A kinetic model of trp-cage folding from multiple biased molecular dynamics simulations. PLoS Comput Biol 2009, 5, e1000452. [PubMed: 19662155]

(36). Lindorff-Larsen K; Piana S; Dror RO; Shaw DE How Fast-Folding Proteins Fold | Science. Science 2011, 334, 517–520. [PubMed: 22034434]

(37). Deng N.-j.; Dai W; Levy RM How Kinetics within the Unfolded State Affects Protein Folding: An Analysis Based on Markov State Models and an Ultra-Long MD Trajectory. The Journal of Physical Chemistry B 2013, 117, 12787–12799, Publisher: American Chemical Society. [PubMed: 23705683]

(38). Sidky H; Chen W; Ferguson AL High-Resolution Markov State Models for the Dynamics of Trp-Cage Miniprotein Constructed Over Slow Folding Modes Identified by State-Free Reversible VAMPnets. The Journal of Physical Chemistry B 2019, 123, 7999–8009. [PubMed: 31453697]

(39). Molgedey L; Schuster HG Separation of a mixture of independent signals using time delayed correlations. Physical review letters 1994, 72, 3634. [PubMed: 10056251]

(40). Pérez-Hernández G; Paul F; Giorgino T; De Fabritiis G; Noé F Identification of slow molecular order parameters for Markov model construction. The Journal of chemical physics 2013, 139, 07B604_1.

(41). Schwantes CR; Pande VS Improvements in Markov state model construction reveal many non-native interactions in the folding of NTL9. Journal of chemical theory and computation 2013, 9, 2000–2009. [PubMed: 23750122]

(42). Webber RJ; Thiede EH; Dow D; Dinner AR; Weare J Error bounds for dynamical spectral estimation. arXiv:2005.02248 [physics] 2020, arXiv: 2005.02248.

(43). Schütte C; Noé F; Lu J; Sarich M; Vanden-Eijnden E Markov state models based on milestoning. The Journal of chemical physics 2011, 134, 05B609.

(44). Vitalini F; Noé F; Keller B A basis set for peptides for the variational approach to conformational kinetics. Journal of chemical theory and computation 2015, 11, 3992–4004. [PubMed: 26575895]

(45). Boninsegna L; Gobbo G; Noé F; Clementi C Investigating molecular kinetics by variationally optimized diffusion maps. Journal of chemical theory and computation 2015, 11, 5947–5960. [PubMed: 26580713]

(46). Weber M; Fackeldey K; Schütte C Set-free Markov state model building. The Journal of Chemical Physics 2017, 146, 124133. [PubMed: 28388134]

(47). Suárez E; Adelman JL; Zuckerman DM Accurate Estimation of Protein Folding and Unfolding Times: Beyond Markov State Models. Journal of Chemical Theory and Computation 2016, 12, 3473–3481, Publisher: American Chemical Society. [PubMed: 27340835]

(48). Zwanzig R Nonequilibrium Statistical Mechanics; Oxford University Press, 2001.

(49). Chorin AJ; Hald OH; Kupferman R Optimal prediction with memory. Physica D: Nonlinear Phenomena 2002, 166, 239–257.

(50). Guttenberg N; Dama JF; Saunders MG; Voth GA; Weare J; Dinner AR Minimizing memory as an objective for coarse-graining. The Journal of Chemical Physics 2013, 138, 094111. [PubMed: 23485281]

(51). Abraham MJ; Murtola T; Schulz R; Páll S; Smith JC; Hess B; Lindah E Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX 2015, 1–2, 19–25.

(52). Bonomi M; Branduardi D; Bussi G; Camilloni C; Provasi D; Raiteri P; Donadio D; Marinelli F; Pietrucci F; Broglia RA; Parrinello M PLUMED: A portable plugin for free-energy calculations with molecular dynamics. Computer Physics Communications 2009, 180, 1961–1972.

(53). Tribello GA; Bonomi M; Branduardi D; Camilloni C; Bussi G PLUMED 2: New feathers for an old bird. Computer Physics Communications 2014, 185, 604–613.

(54). PLUMED, Promoting transparency and reproducibility in enhanced molecular simulations. Nature Methods 2019, 16, 670–673. [PubMed: 31363226]

(55). MacKerell AD; Bashford D; Bellott M; Dunbrack RL; Evanseck JD; Field MJ; Fischer S; Gao J; Guo H; Ha S; Joseph-McCarthy D; Kuchnir L; Kuczera K; Lau FTK; Mattos C; Michnick S; Ngo T; Nguyen DT; Prodhom B; Reiher WE; Roux B; Schlenkrich M; Smith JC; Stote R; Straub J; Watanabe M; Wiórkiewicz-Kuczera J; Yin D; Karplus M All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins †. The Journal of Physical Chemistry B 1998, 102, 3586–3616. [PubMed: 24889800]

(56). Best RB; Zhu X; Shim J; Lopes PEM; Mittal J; Feig M; MacKerell AD Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone ϕ, ψ and Side-Chain χ1 and χ2 Dihedral Angles. Journal of Chemical Theory and Computation 2012, 8, 3257–3273. [PubMed: 23341755]

(57). Huang J; Rauscher S; Nawrocki G; Ran T; Feig M; De Groot BL; Grubmüller H; MacKerell AD CHARMM36m: An improved force field for folded and intrinsically disordered proteins. Nature Methods 2016, 14, 71–73. [PubMed: 27819658]

(58). Hess B; Bekker H; Berendsen HJC; Fraaije JGEM LINCS: A linear constraint solver for molecular simulations. Journal of Computational Chemistry 1997, 18, 1463–1472.

(59). Neidigh JW; Fesinmeyer RM; Andersen NH Designing a 20-residue protein. Nature Structural Biology 2002, 9, 425–430. [PubMed: 11979279]

(60). Jorgensen WL; Chandrasekhar J; Madura JD; Impey RW; Klein ML Comparison of simple potential functions for simulating liquid water. The Journal of Chemical Physics 1983, 79, 926–935.

(61). Jo S; Kim T; Iyer VG; Im W CHARMM-GUI: A web-based graphical user interface for CHARMM. Journal of Computational Chemistry 2008, 29, 1859–1865. [PubMed: 18351591]

(62). Lee J; Cheng X; Swails JM; Yeom MS; Eastman PK; Lemkul JA; Wei S; Buckner J; Jeong JC; Qi Y; Jo S; Pande VS; Case DA; Brooks CL; MacKerell AD; Klauda JB; Im W CHARMM-GUI Input Generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM Simulations Using the CHARMM36 Additive Force Field. Journal of Chemical Theory and Computation 2016, 12, 405–413. [PubMed: 26631602]

(63). Marchi M; Ballone P Adiabatic bias molecular dynamics: A method to navigate the conformational space of complex molecular systems. Journal of Chemical Physics 1999, 110, 3697–3702.

(64). Park S; Kim T; Im W Transmembrane Helix Assembly by Window Exchange Umbrella Sampling. Physical Review Letters 2012, 108, 108102. [PubMed: 22463457]

(65). Sugita Y; Kitao A; Okamoto Y Multidimensional replica-exchange method for free-energy calculations. Journal of Chemical Physics 2000, 113, 6042–6051.

(66). Thiede EH; Van Koten B; Weare J; Dinner AR Eigenvector method for umbrella sampling enables error analysis. Journal of Chemical Physics 2016, 145.

(67). Antoszewski A; Feng C-J; Vani BP; Thiede EH; Hong L; Weare J; Tok-makoff A; Dinner AR Insulin Dissociates by Diverse Mechanisms of Coupled Unfolding and Unbinding. The Journal of Physical Chemistry B 2020, 124, 5571–5587. [PubMed: 32515958]

(68). Dinner AR; Thiede EH; Koten BV; Weare J Stratification as a General Variance Reduction Method for Markov Chain Monte Carlo. SIAM/ASA Journal on Uncertainty Quantification 2020, 1139–1188, Publisher: Society for Industrial and Applied Mathematics. [PubMed: 34611500]

(69). Lorpaiboon C; Thiede EH; Webber RJ; Weare J; Dinner AR Integrated VAC: A robust strategy for identifying eigenfunctions of dynamical operators. arXiv:2007.08027 [physics] 2020, arXiv: 2007.08027.

(70). Kim SB; Dsilva CJ; Kevrekidis IG; Debenedetti PG Systematic characterization of protein folding pathways using diffusion maps: Application to Trp-cage miniprotein. The Journal of Chemical Physics 2015, 142, 085101, Publisher: American Institute of Physics. [PubMed: 25725756]

(71). Thiede EH; Van Koten B; Weare J; Dinner AR Eigenvector method for umbrella sampling enables error analysis. The Journal of Chemical Physics 2016, 145, 084115. [PubMed: 27586912]

(72). Thiede E; Van Koten B; Weare J Sharp Entrywise Perturbation Bounds for Markov Chains. SIAM Journal on Matrix Analysis and Applications 2015, 36, 917–941. [PubMed: 26491218]

(73). Ma A; Dinner AR Automatic Method for Identifying Reaction Coordinates in Complex Systems. The Journal of Physical Chemistry B 2005, 109, 6769–6779. [PubMed: 16851762]

(74). Hu J; Ma A; Dinner AR A two-step nucleotide-flipping mechanism enables kinetic discrimination of DNA lesions by AGT. Proceedings of the National Academy of Sciences 2008, 105, 4615–4620.
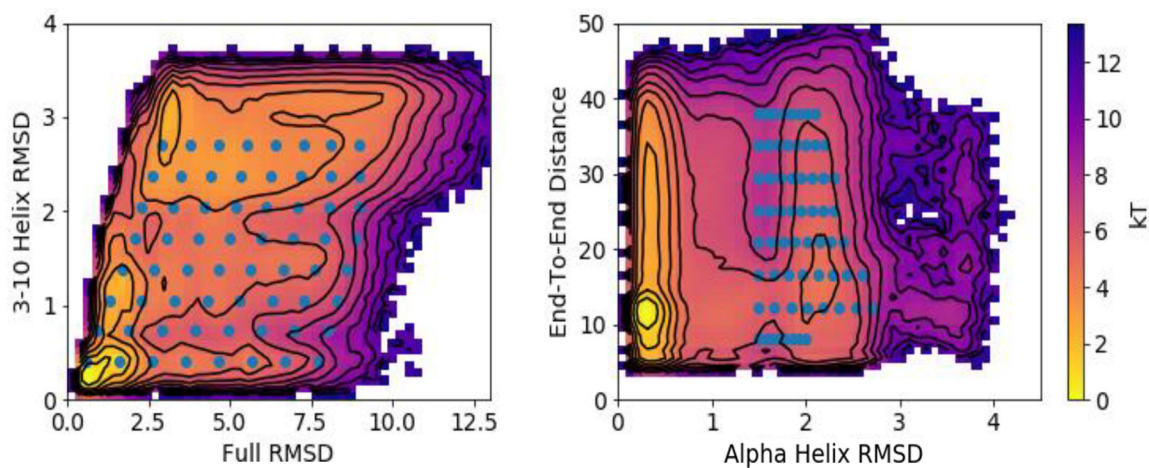
Figure 1:
Initialization points for the data set of short trajectories. ABMD targets (symbols) are overlaid on DGA PMFs (color scale and contours, spaced every 1 $k_B T$) for the CVs used for steering. (left) The initial 64 ABMD targets were based on RMSD$_{full}$ and RMSD$_{3-10}$; 14 free simulations of length 30 ns were launched from e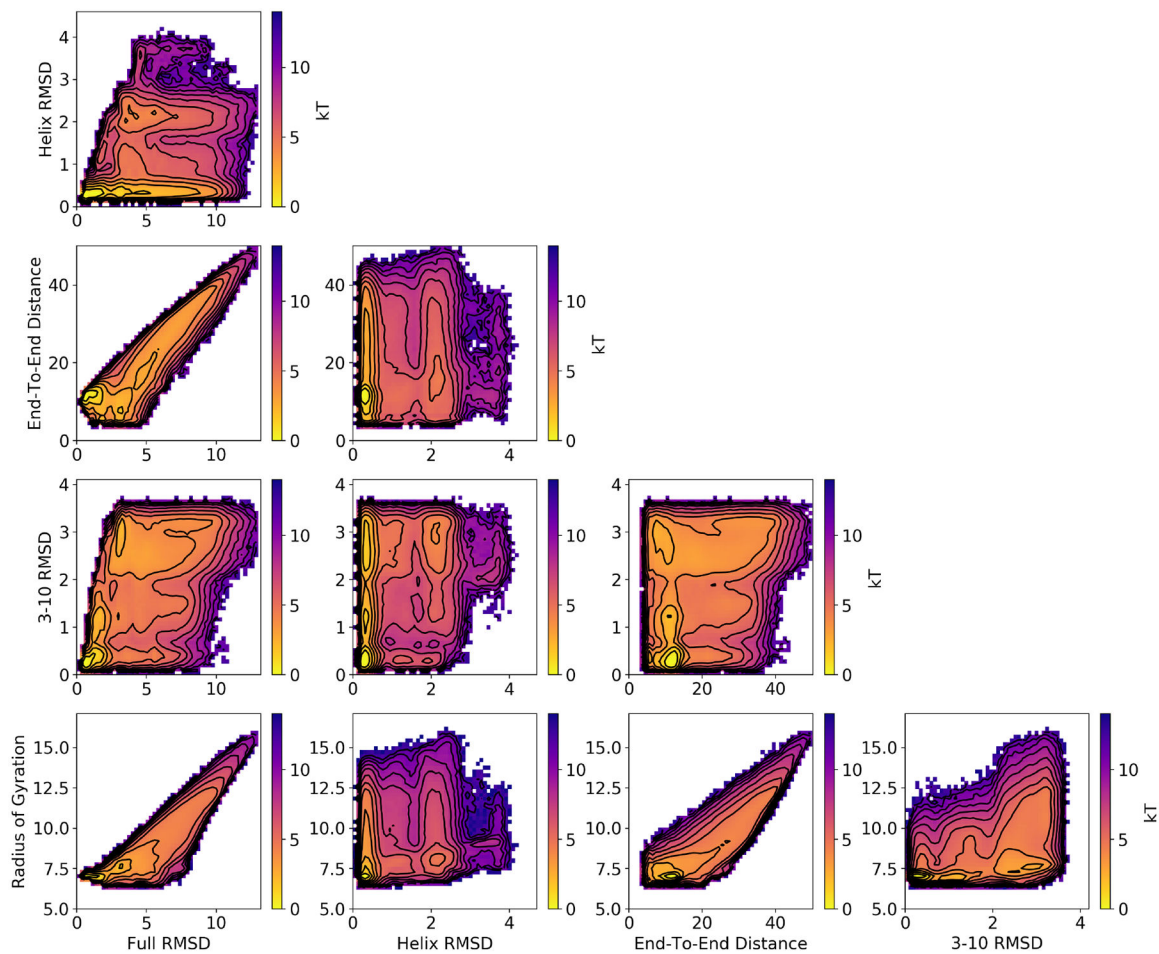ach of the structures resulting from these ABMD simulations. (right) 64 ABMD targets in RMSD$_{hx}$ and end-to-end distance added to ensure adequate sampling of the unfolded state; 2 free simulations of length 30 ns were launched from each of the structures resulting from these ABMD simulations.

Figure 2:
PMFs for the indicated CVs. Results shown are computed by DGA with the modified distance basis set and a lag time of 0.5 ns. We use a 50×50 grid to compute each PMF. Similar results are obtained with other basis sets and REUS; see Figures S4, S5, and S6.
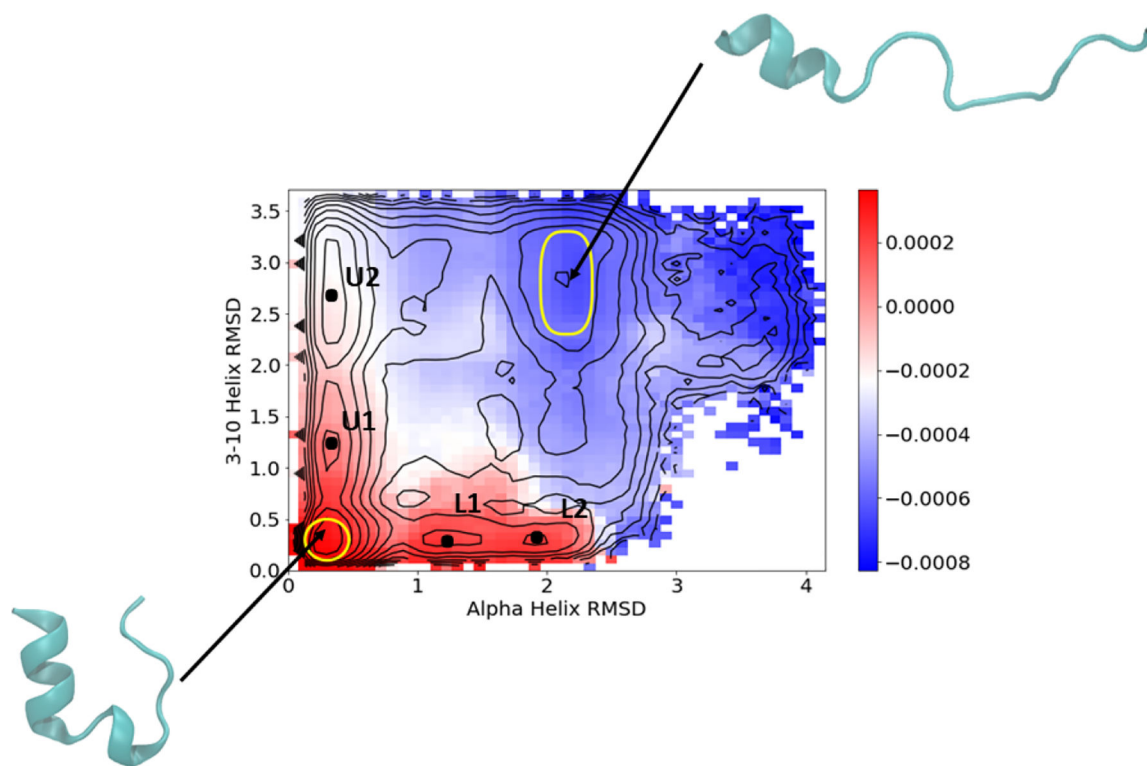
Figure 3:
Top nontrivial TICA eigenvector averaged on the $RMSD_{hx}$ and $RMSD_{3-10}$ CVs with physical weighting. The unfolded and folded states are indicated in yellow with representative structures. Intermediate states in Table 1 are marked and labeled.
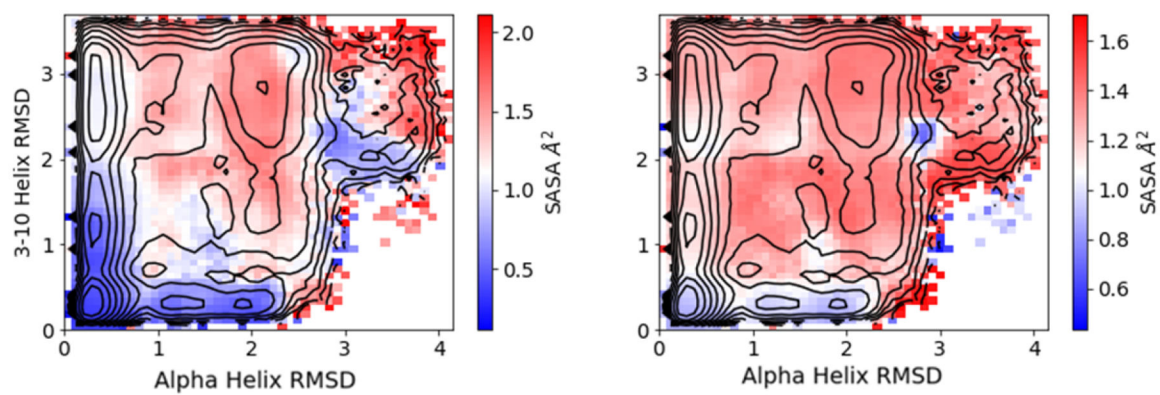
Figure 4:
Equilibrium average solvent accessible surface area (SASA) projected onto the $RMSD_{hx}$ and $RMSD_{3-10}$ CVs for (left) trp-6 and (right) proline-12.
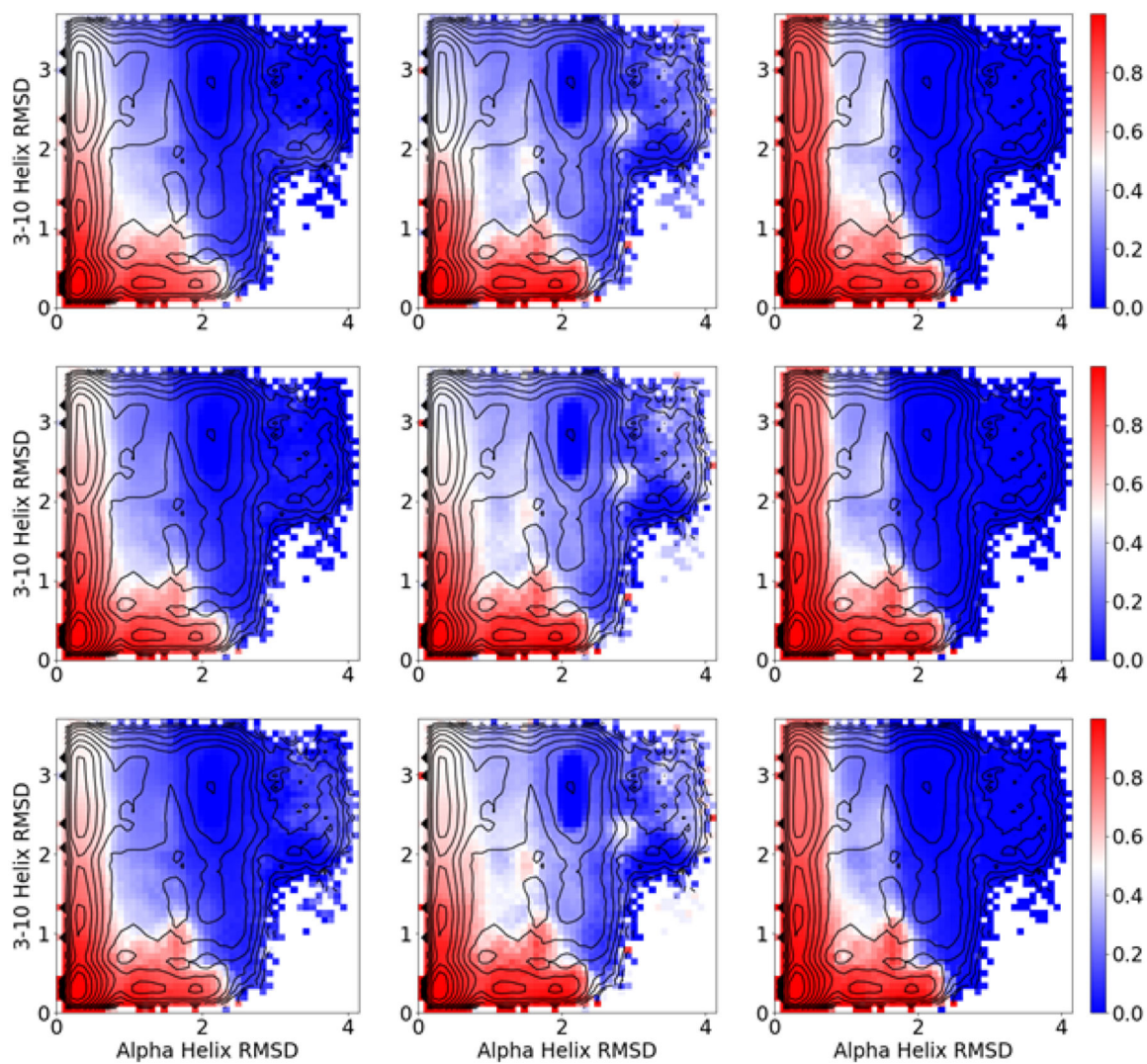
Figure 5:
DGA forward committors. Left, middle, and right columns are computed with the modified distance, distance indicator, and TICA indicator basis sets, respectively. Top, middle, and bottom rows are computed with lag times of 0.5, 2.5, and 7.5 ns, respectively.
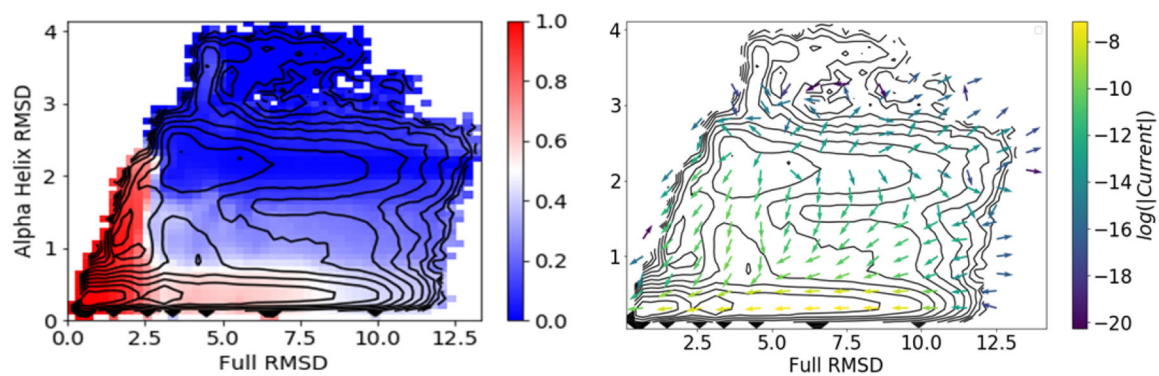
Figure 6:
Forward committor (left) and reactive current (right) projected onto the $RMSD_{hx}$ and full RMSD CVs used in ref. 33. Results shown are computed with the modified distance basis set and a lag time of 0.5 ns.
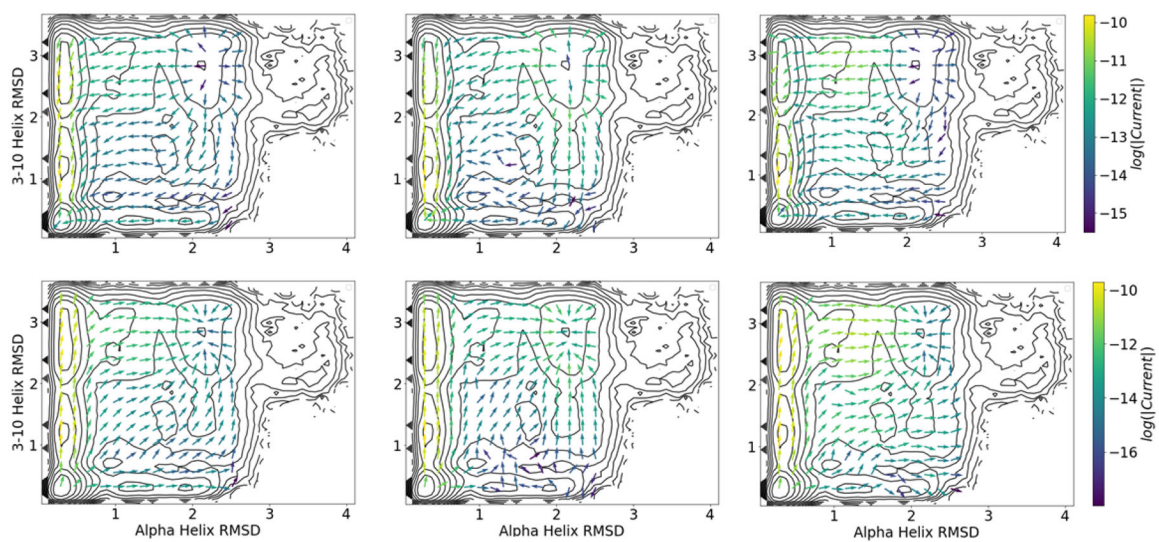
Figure 7:
Folding (top) and unfolding (bottom) reactive current projected onto the RMSD_{hx} and RMSD_{3-10} CVs using (37) with the three choices of basis set. Left, middle, and right columns are computed with the modified distance, distance indicator, and TICA indicator basis sets, respectively. All computations use a lag time of 0.5 ns.
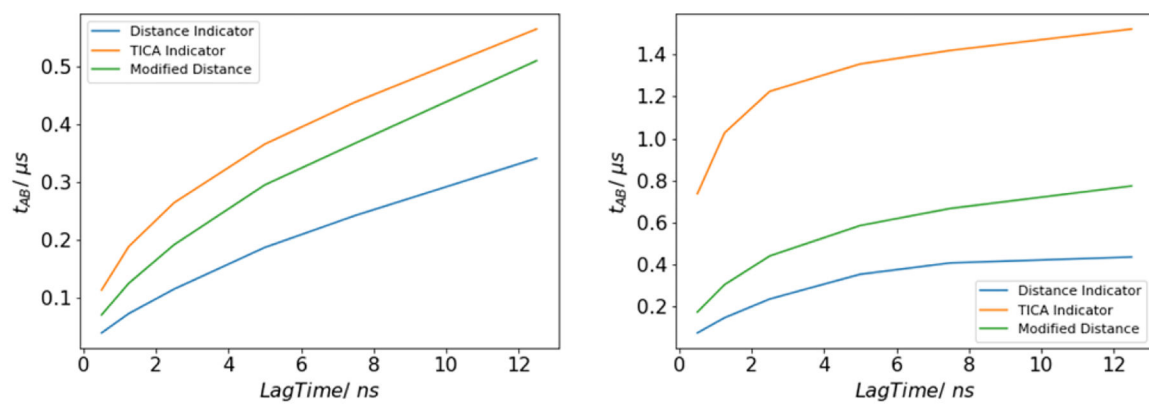
Figure 8:
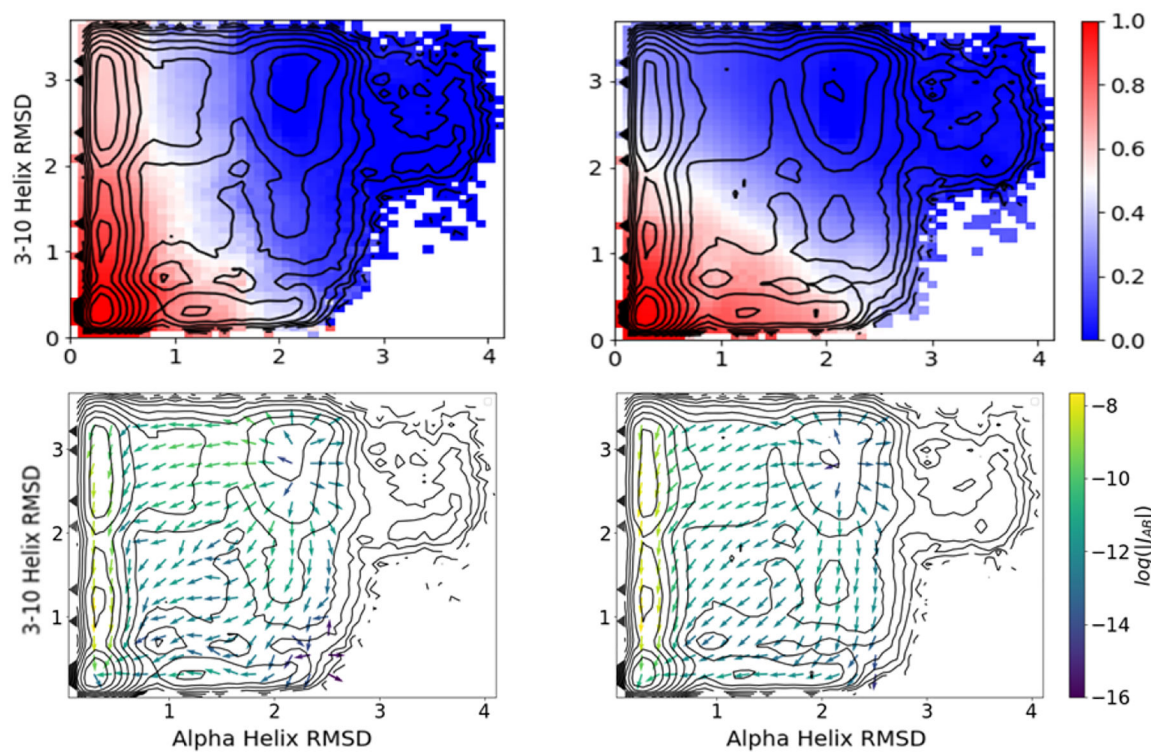Inverse rates estimated for folding (left) and unfolding (right).

Figure 9:
Comparison of DGA estimates for the forward committor (top) and reactive current for folding (bottom) with the modified distance basis set on a feature space restricted to the five physical CVs (right) and a delay-embedded feature space (left). The delay-embedded results are obtained with a delay of $\delta = 0.125$ ns, $N = 40$ images, and a DGA lag time of 0.5 ns.

Table 1:

CV values for metastable states.

State	RMSD$_{full}$/Å	RMSD$_{hx}$/Å	d/Å	RMSD$_{3-10}$/Å	R_g/Å
Folded	1.1	0.30	11.1	0.30	7.0
Unfolded	5.8	2.1	20.2	2.8	9.2
U1	2.4	0.34	13.1	1.2	7.3
U2	5.2	0.34	19.3	2.8	8.8
L1	2.2	1.2	9.5	0.30	7.2
L2	2.6	1.9	14.5	0.30	7.3