

# Extracting Kinetic and Stationary Distribution Information from Short MD Trajectories via a Collection of Surrogate Diffusion Models

Christopher P. Calderon<sup>†</sup> \*and Karunesh Arora<sup>‡</sup>

<sup>†</sup> Department of Statistics and Department of Computational and Applied Mathematics, Rice University, Houston, TX 77005-1892, USA.

<sup>‡</sup>Department of Chemistry, Biophysics Program, University of Michigan, 930 North University Avenue, Ann Arbor, MI 48109 USA.

June 8, 2009

## **Abstract:**

Low-dimensional stochastic models can summarize dynamical information and make long time predictions associated with observables of complex atomistic systems. Maximum likelihood based techniques for estimating low-dimensional surrogate diffusion models from relatively short time series are presented. It is found that a heterogeneous population of slowly evolving conformational degrees of freedom modulates the dynamics. This underlying heterogeneity results in a collection of estimated low-dimensional diffusion models. Numerical techniques for exploiting this finding to approximate skewed histograms associated with the simulation are presented. In addition, statistical tests are also used to assess the validity of the models and determine physically relevant sampling information, e.g. the maximum sampling frequency at which one can discretely sample from an atomistic time series and have a surrogate diffusion model pass goodness-of-fit tests. The information extracted from such analyses can possibly be used to assist umbrella sampling computations as well as help in approximating effective diffusion coefficients. The techniques are demonstrated on simulations of Adenylate Kinase.

## **1 Introduction**

A significant understanding of complex biomolecules like proteins and nucleic acids has been obtained through the use of low-dimensional equations approximating the dynamics

---

\*To whom correspondence should be addressed. E-mail:calderon@rice.edu

of these systems.<sup>1,2</sup> Single-molecule experiments and computer simulations are allowing researchers to better understand the physics governing complex biomolecular systems at small length and time scales.<sup>1–16</sup> Advances in nanotechnology are starting to demand higher accuracy from stochastic dynamical approximations of small biomolecular systems. Fortunately, the time series associated with current experiments and simulations contain a rich amount of information related to molecular motion occurring over a broad-range of time scales.<sup>17</sup> However, the presence of a wide range of relevant time scales significantly complicates determining reliable low-dimensional models associated with small scale, but highly complex systems.<sup>18–21</sup> We refer to such approximate models as “surrogate” models, some authors use the term “effective model” in a similar context.<sup>20–22</sup>

In single-molecule experiments, researchers can usually only monitor and/or manipulate a small number of observable quantities that describe the system. This is because simultaneously tracking the position and velocity of many atoms in a system at the same time is very challenging experimentally. Experiments usually do have the luxury of being able to directly measure quantities associated with longer time scales of physical interest, e.g. it is sometimes possible to monitor a protein unfold and refold nearly quasi-statically.<sup>23</sup> On the other hand, atomistic simulations provide detailed descriptions of the dynamics (i.e. monitoring the position and velocity of every particle is possible), but encounter computational limitations. Perhaps the most serious being the small time step size enforced by numerical stability considerations. Current simulations can only reasonably explore  $\approx \mathcal{O}(ns) - \mathcal{O}(\mu s)$  length trajectories in all-atom MD simulations of biomolecules<sup>17,24</sup> due to the time step constraint. Rapid advances in experiments and simulations are likely to further facilitate making comparisons between these two information sources and using both to refine/construct surrogate models.<sup>25–28</sup>

We present surrogate models which can approximate both simulation<sup>29,30</sup> and experimental time series,<sup>31,32</sup> but the focus here is exclusively on simulation data. Throughout, we refer to quantities monitored in our time series as “system observables” (SOs). “Pathwise” statistical methods are used to estimate diffusion models from observed time series. We use the term “pathwise” simply to refer to situations where all statistical inference procedures (estimation, hypothesis tests, etc.) are applied to a single time series. We use all-atom MD simulations to generate multiple trajectories (i.e. a batch of time series). We estimate the parameters of a new surrogate model for each observed time series as opposed to aggregating the time series together to estimate the parameters of a single model. We demonstrate surrogate models which can account for state-dependent noise. This has relevance to several systems because it is known that when a high-dimensional system is summarized with a small number of SOs, that the noise magnitude often varies as a function of the SO.<sup>3,18,29,30,33–36</sup> The information in the surrogate diffusion models can also be related to the effective friction experienced by a particular SO in different portions of phase space.<sup>30,32,37–39</sup> The methods presented make heavy use of recent developments in statistical testing and modeling<sup>29,30,34,40–44</sup> to assess the validity of the estimated models and quantitatively learn about the various multi-scale noise sources. For example, we analyze histograms of a root-mean-square-displacement (rmsd) type SO and determine how

much variability is due to traditional thermal noise and how much is introduced by conformational heterogeneity. The former is usually associated with fast-scale motions whose details are not of interest and the latter with slow-scale collective motions.

Besides quantifying the contributions of various noise sources to the observed variability of the SO, the collection of estimated models can be also be used to help traditional physical chemistry computations. For example, they can be used to reduce the variance of equilibrium statistics and can also be used to quantify how certain factors influence the dynamics and stationary distributions of SOs. This is relevant because it is known that conformational degrees of freedom often can cause skewed distributions of low-dimensional SOs. Correlating information accessible in the laboratory with that in computer simulations has the potential to help in various computational chemistry tasks.<sup>15,31,45-49</sup>

Our techniques are demonstrated on short (525 ps) constrained umbrella sampling trajectories of the enzyme Adenylate Kinase.<sup>50,51</sup> It is shown that surrogate models, calibrated from observational data, can be used to predict and/or refine approximations of stationary distributions associated with a selected SO. The particular SO studied is related to known “active” and “inactive” crystal structures of the enzyme.<sup>51</sup> We demonstrate that the diffusion coefficient<sup>52</sup> can be approximated using the surrogate models and that confidence bands for this quantity can be constructed using a single short time trajectory. We demonstrate that this estimate (and confidence band associated with the estimate) can be obtained with much less data than traditional approaches used in MD.<sup>24,52,53</sup> We also show that an experimentally accessible<sup>49</sup> slowly evolving conformational coordinate correlates with the surrogate model parameters. In this system, taking the state-dependence of the noise magnitude into account,<sup>29,30,33,35-37</sup> as well as the underlying structural inhomogeneity is shown to be important to faithfully approximate the complex high-dimensional atomistic simulation using low-dimensional surrogate diffusion models. The Supp. Info. provides results demonstrating that the same ideas can be applied to approximate the SO associated with longer (10-50 ns length) unconstrained molecular dynamics of the so-called “Engrailed Homeodomain”.<sup>54</sup>

The article is organized as follows: Section 2 provides a theoretical background reviewing our basic motivation some established results from statistical physics. This background helps in physically interpreting the information contained in the collection of surrogate diffusion models we attempt to fit from observations. This section also presents the statistical methods we employ and contains a discussion illustrating how the information extracted from the methods can help computational physical chemistry. Section 3 provides the MD simulation details. Section 4 presents the numerical results and Section 5 concludes.

## 2 Theoretical Background

### 2.1 Data-driven Multiscale Stochastic Modeling

The basic idea of using “short bursts” of simulation times series to estimate effective dynamical models motivated the types of methods we propose.<sup>20,55,56</sup> We obtain the param-

eters of surrogate diffusion models<sup>29,30</sup> from observed time series using maximum likelihood type (ML) techniques. The type of modeling we propose would fall under the label of a “data-driven” modeling procedure. Several other researchers are developing data-driven methods for describing various complex systems ranging from molecular dynamics to weather forecasting.<sup>20,29,30,33,56–59</sup> The basic idea behind a data-driven description is to assimilate information contained in empirical observations, either simulation or experiment, coming from a complex high-dimensional system into a surrogate model.

If accurate surrogate models can be calibrated from short time series then these models can be used for a variety of purposes, e.g. they can be used to simulate sample paths for longer time intervals than those accessible to the MD simulation. Other applications are discussed in the Section 4. These basic ideas are not new to chemical physics and it is well-known that the multiple time-scales associated with the underlying complex process significantly complicates these types of tasks.<sup>18,21,60</sup> Our main contributions to this type of endeavor are associated with showing how modern time series analysis tools can be used to help in quantitatively determining some “coarse-graining” parameters and also determine the goodness-of-fit of surrogate models in a pathwise fashion. We also demonstrate how a *collection* of low-dimensional models can be used to study various all-atom simulations and single-molecule experiments where certain collective degrees of freedom are associated with slow time-scales and influence the estimated surrogate models.<sup>30,31,61,62</sup> The motivation for using a collection of simple low-dimensional models as opposed to more complicated high-dimensional models as surrogates is discussed in detail in Ref. [61] and in the Conclusions.

## 2.2 Generalized Langevin Equations

The generalized Langevin equation (GLE) description has been used to describe the evolution of certain MD trajectories.<sup>21</sup> Although we do not utilize this structure in our models, we introduce it here because several analogies can be drawn to the modeling methods we introduce and can aid in the physical interpretation of our procedure. A generic GLE typically takes the form:

$$\dot{\Phi} = F(\Phi) + \int_0^t K(\Phi(t-s), s)ds + \sqrt{2k_B T}N(t), \quad (1)$$

where  $k_B T$  is Boltzmann’s constant multiplied by the system temperature,  $\Phi$  is the vector of SOs the modeler wishes to dynamically track, a dot above a variable denotes the time derivative,  $F(\cdot)$  represents the *Markovian* contribution to the dynamics,<sup>19,21</sup>  $N(\cdot)$  represents the “orthogonal noise” coming from degrees of freedom not explicitly resolved in the model,<sup>19</sup> and  $K(\cdot, \cdot)$  the so-called “memory kernel”.  $K(\cdot, \cdot)$  is usually interpreted as a *non-Markovian* contribution to the dynamics. The specific functional form of the memory kernel and the orthogonal noise can, in principle, be determined once the full system dynamics are specified using the Mori-Zwanzig projection operator formalism.<sup>19,21</sup> The orthogonal noise is typically constructed to have a zero mean over ensembles and its value

depends on the initial conditions of all degrees of freedom in the system. For a single realization, the orthogonal dynamics can slowly evolve making the “noise” term appear to be a systematic bias when viewed over short time-scales of a single trajectory. Observations of this nature motivated the authors in Ref. [63] to use a long-memory (fractional Gaussian) process to describe the orthogonal noise in experimental data tracking a protein’s slow conformational dynamics.

The AdK system studied is also associated with slow conformational dynamics. However, the surrogate models we propose attempt to approximate  $F(\Phi) + \int_0^t K(\Phi(t-s), s)ds$  using a Markovian term  $\mu(\Phi)$  and use a fairly simple noise process (standard Brownian motion). One of the surrogate models attempts to utilize the so-called “overdamped approximation”.<sup>38</sup> The term “overdamped” is meant to refer to the fact that a particle has a position and velocity, but knowledge of the velocity is not needed to accurately approximate the statistical properties of the particle position. This is a temporal coarse-graining procedure commonly used in statistical physics.<sup>18,38</sup> It has been labeled as somewhat *ad hoc* because it requires quantitative knowledge of the time that one needs to wait between adjacent observations for such an approximation to be valid and this selection is usually based on intuitive physical arguments as opposed to precise mathematical criteria.<sup>21</sup> We demonstrate that down-sampling (or sub-sampling<sup>30,60</sup>) ideas along with statistical hypothesis tests<sup>42,44,64</sup> can be used to help put a quantitative handle on an overdamped approximation. We refer to fast-scale noise which may induce short time memory as “fast-scale memory” throughout the text. Velocity is one possible source, but others like vibrational motion would contribute to this type of fast-scale memory. The modeling of slow conformational degrees of freedom is more subtle. This article focuses on modeling the output of MD simulations, so producing time series where a long-memory process of the type given in Ref. [63] can be estimated is somewhat problematic. This is because in simulations it is difficult to sample for a large temporal amount, so the fitted parameters of a long-memory process would likely contain substantial uncertainty. To approximate the variability induced by slowly evolving conformational degrees of freedom in relatively short time series we use a *collection* of surrogate models. We expand on this point throughout the text.

### 2.3 Time-scale Separation in System Observables

If the  $K(\cdot, \cdot)$  in Eq. 1 is zero everywhere and the noise process is a standard Gaussian white-noise process with “Dirac-delta time correlation”, then this is commonly written as a diffusion type stochastic differential equation (SDE).<sup>65</sup> SDEs have a rich history<sup>66</sup> in the physical sciences, early studies focused primarily on analyzing properties of Fokker-Planck type partial differential equation (PDE) associated with the stochastic process as opposed to the SDE. We utilize the SDE view, sometimes called the pathwise view<sup>21</sup> because we feel it facilitates connecting the physics to the estimated surrogate models.

The data-driven modeling methods we propose assume that the effective dynamics<sup>21</sup> of the underlying high-dimensional atomistic simulations can be accurately captured by a

small set of SOs whose dynamics are governed by the following system of SDEs:

$$\begin{aligned} d\Phi &= \alpha(\Phi, \mathcal{C})dt + \sqrt{2}\sigma(\Phi, \mathcal{C})dW_t^1 \\ d\mathcal{C} &= \beta(\Phi, \mathcal{C})dt + \sqrt{2}\kappa(\Phi, \mathcal{C})dW_t^2, \end{aligned} \quad (2)$$

where the  $W_t^i$  represent standard Brownian motions,<sup>65</sup>  $\Phi$  represents a “fast-scale” coordinate, and  $\mathcal{C}$  a “slow-scale” conformational coordinate. We assume that observations are made on a time-scale shorter than  $\Phi$ ’s characteristic relaxation time and that the dynamics of  $\mathcal{C}$  are associated with much “slower” time-scales than those of  $\Phi$ . The functions  $\alpha(\cdot, \cdot)$  &  $\beta(\cdot, \cdot)$  are referred to as the drift functions and  $\sigma(\cdot, \cdot)$  &  $\kappa(\cdot, \cdot)$  are related to the diffusion matrix.<sup>65</sup>

We do not assume that we have the system of SDEs in Eq. 3 available in closed-form. We only assume that the stochastic dynamics of the higher-dimensional atomistic system can be accurately approximated by this system of SDEs (i.e. evolution rules are more complicated than Eq. 3). Recall that the data-driven approach we use attempts to estimate effective dynamical equations from time series. If a scale separation exists, the dimension of the SDE system can often be reduced.<sup>21,22,55,67</sup> This can significantly facilitate estimation of a surrogate model. Traditional SDE model reduction techniques often ignore the details of the “fast” component and focus on describing the details of the “slow” component’s evolution.<sup>21,55,67,68</sup> In our notation, the modeler would treat  $\Phi$  as a “noise” and focus on stochastic dynamical models that explicitly model  $\mathcal{C}$ . We do the opposite, namely we estimate a scalar SDE of the form:

$$d\Phi = \mu^{\mathcal{C}}(\Phi)dt + \sqrt{2}\sigma^{\mathcal{C}}(\Phi)dW_t^1, \quad (3)$$

to the SO time series. Recall we assume that our time series observations are spaced by a time shorter than the relaxation time of  $\Phi$  (and hence *much* shorter than that of  $\mathcal{C}$ ). If we additionally assume that the local noise,  $\kappa$ , of the slow coordinate  $\mathcal{C}$  is small, then this coordinate will not have time to appreciably change in a short time simulation.  $\mathcal{C}$  can be treated as effective constant that modulates the drift and noise functions observed, e.g.  $\mu^{\mathcal{C}}(\cdot) \equiv \mu(\cdot, \mathcal{C}), \sigma^{\mathcal{C}}(\cdot) \equiv \sigma(\cdot, \mathcal{C})$ .

However, a wide range of  $\mathcal{C}$  are explored, albeit slowly, at “thermodynamic” equilibrium<sup>1</sup>. Ideally, the computational cost associated with a standard long time integration would be moderate so that observing substantial changes in both effective fast and slow components would be possible. In this case, we would attempt to either estimate Eq. 3 directly or use more traditional time-scale separation techniques.<sup>21,22,55,67</sup> With longer times series, we could also entertain using a single more complicated model<sup>63</sup> attempting to capture slow fluctuations and relate this to a memory kernel. Unfortunately, due to

---

<sup>1</sup>By this we mean the stationary distribution associated with all time-scales. Note that the underlying system maybe biased, as in umbrella sampling simulations, but we still refer to the stationary state as “thermodynamic” equilibrium.

computational limitations, generating long time trajectories is problematic in many complex all-atom systems. We propose methods that estimate a collection of SDEs of the form given in Eq. 3. The initial conformation of each MD simulation is drawn randomly from an equilibrium ensemble, so each different time series trajectory is associated with a different  $\mathcal{C}$  value. We demonstrate how to use this collection to assist in computations commonly encountered in computational physical chemistry. Note carefully that the system of SDEs approximating the dynamics in Eq. 3 is usually much smaller in dimension than the full atomistic system. By only modeling one component ( $\Phi$ ), we are applying an additional reduction to the system of SDEs. Reduction of this sort introduces the collection of surrogate model phenomenon we discuss throughout.

Before providing specific details of the assumed surrogate models, we would like to comment on the process we used for assigning “fast” and “slow” variables in AdK. In Ref. [51], there was interest in generating stationary histograms associated with a coordinate quantifying the difference in rmsd of intermediate conformations with respect to the open and closed enzyme states (in what follows we simply call this distance “rmsd type”). Dynamically monitoring this rmsd type quantity in the laboratory is not currently feasible. There were FRET measurements available providing information about a distance between dye labeled residues Lys145 and Ile52.<sup>49</sup> These particular residues were chosen since they lie in two domains of AdK that undergo the largest conformational change between the open to closed state of AdK and help detect functionally relevant motion.<sup>51</sup> In our simulations we do not directly attempt to manipulate this residue distance despite its physical relevance; it has been shown that this residue distance explores a wide range of values, but does so slowly with respect to the time-scale of simulations.<sup>51</sup> As a result we used the rmsd type distance as  $\Phi$  and the distance between the center of mass of residues Lys145 and Ile52 as  $\mathcal{C}$ . In order to enhance sampling of  $\Phi$ , a harmonic biasing potential was introduced (see Section 3). The biasing potential also made a linear effective force approximation seem more plausible in a surrogate model (later we quantitatively tested this assumption). This biasing potential altered the effective underlying energy landscape and also made the dynamics associated with  $\Phi$  “faster” in relation to  $\mathcal{C}$ . Recall we also assumed that the local fluctuation magnitude associated with  $\mathcal{C}$  was small. This assumption was due to the fact that collective conformational degrees of freedom do not typically wildly fluctuate in an enzyme.

The above considerations clearly utilized our physical intuition about the system. In general, measurements of fast  $\Phi$  type coordinates are difficult to accurately monitor and/or control in the laboratory. However, several slowly evolving conformational degrees of freedom can readily be monitored and/or manipulated by novel single-molecule techniques.<sup>3,5-16</sup> We demonstrate that a measurable correlation exists between the selected SOs. The time-scale separation between the correlated SOs is exploited in the methods we report. This is one way in which simulation predictions can be compared to experimental observations.<sup>51</sup>

However, we would like to note that selecting coordinates having a large time scale separation can be difficult if physical intuition alone is used. The problem is even harder if the interest is in analyzing unconstrained simulations. General, data-driven procedures for

identifying “good” variables where a significant time-scale separation exists is challenging, but would be of great help to studying systems where physical intuition is lacking.<sup>69,70</sup> In addition, other more sophisticated types of multiscale approximations can be applied to SDEs like those in Eq. 3 using less restrictive assumptions about the dynamics.<sup>21,56,68,71</sup>

## 2.4 Proposed Functional Form of Surrogate SDE Models

For every observed time series, we proposed two model structures to use along with Eq. 3, namely:

$$\begin{aligned} \text{Model 1 :} & & (4) \\ \mu^c(\Phi) &= \left( A + B(\Phi - \Phi_0) \right), & \sigma^c(\Phi) = C \end{aligned}$$

$$\begin{aligned} \text{Model 2 :} & & (5) \\ \mu^c(\Phi) &= \frac{\sigma^c(\Phi)^2}{k_B T} \left( A + B(\Phi - \Phi_0) \right), & \sigma^c(\Phi) = \left( C + D(\Phi - \Phi_0) \right) \end{aligned}$$

where,  $\Phi_0$  corresponds to a user specified point<sup>2</sup> and  $\theta \equiv (A, B, C, D)$  are parameters ( $D$  is only used in Model 2). Model 1 is known as the Ornstein-Uhlenbeck (OU) model. The drift function can readily be interpreted as coming from a harmonic potential connected to a heat bath whose fluctuations are independent of the state. Model 2 explicitly utilizes the overdamped (OD) Langevin approximation<sup>18,29,30,38</sup> and also takes the noise magnitude’s dependence on the current state into account using a relatively simple model. The estimated parameters can be interpreted as local approximations of the effective force and effective local diffusion coefficient<sup>3</sup>.

## 2.5 Fitting and Testing the Surrogate Models

For discretely sampled time series, the maximum likelihood estimator attempts to find the parameter vector maximizing the logarithm of the joint density associated with the observations,  $\log(p(\Phi_0, \Phi_1, \dots, \Phi_N; \theta))$ , where subscripts denote the time index of the observations. The parameter yielding this maximum is denoted by  $\hat{\theta}$ . For general SDEs, estimating  $\hat{\theta}$  analytically is problematic because the joint density cannot usually be expressed in closed-form. The OU model is appealing because it does admit a closed-form expression for the ML parameter estimate and also yields some other useful diagnostic information. For example, an asymptotic expression for the parameter covariance can be obtained.<sup>43</sup> For the case where we cannot obtain the ML estimate in closed form, we appeal to approximate

<sup>2</sup>For example, this could coincide with the constraint point of an US simulation<sup>51</sup> or the mean value of the observed time series. We use the latter in this article.

<sup>3</sup>Note that if the dynamics can be approximated by the OD model above and the noise magnitude is truly constant, then the parameters of the two different estimated surrogate models can be directly compared and the difference should only be do to different sampling uncertainty magnitudes associated with the parameterization used.

likelihood methods.<sup>40,41</sup> Both ML approximation methods previously cited yielded similar  $\hat{\theta}$  values in the cases explored. A url link to MATLAB scripts illustrating how to obtain  $\hat{\theta}$  for both the OU and OD models given a time series is provided in the Supp. Info.

The ‘‘Q test-statistic’’ developed in Ref. [42] is used to check the validity of the assumed surrogate model. This test is designed to test for temporal dependencies which are atypical for an assumed model. We demonstrate that it can be used to detect if fast-scale memory effects are statistically significant. The test is also appealing because it applicable to both stationary and nonstationary signals. This test also provides us with physically relevant coarse-graining information. We demonstrate how we can use this test to determine the appropriate frequency at which data can be discretely be sampled from a simulation and provide a diffusion model which is not rejected by a hypothesis test. If one is willing to make a stationarity assumption about the time series, more powerful tests can be used<sup>44,64</sup>. We demonstrate the ‘‘ $T_3$ ’’ test statistic of Ref. [64] is useful in assessing the accuracy of a stationary density predicted using a short simulation burst. This test is shown to have better power than the Q-test.

## 2.6 Computing the Stationary Density Associated with Scalar SDEs

Under mild regularity conditions, the stationary density, denoted by  $p^{EQ}(\Phi; \mathcal{C})$ , associated with a scalar diffusion model given in Eq. 3 can be expressed in closed-form using only information contained in the estimated SDE coefficient functions via the relation<sup>72,73</sup> <sup>4</sup>:

$$p^{EQ}(\Phi; \mathcal{C}) = \frac{Z}{(\sigma^{\mathcal{C}}(\Phi))^2} \exp\left(\int_{\Phi_{\text{REF}}}^{\Phi} \mu^{\mathcal{C}}(\Phi')/(\sigma^{\mathcal{C}}(\Phi'))^2 d\Phi'\right) \quad (6)$$

where  $Z$  is a constant used to ensure that the density integrates to unity and  $\Phi_{\text{REF}}$  is a specified constant used as a ‘‘reference point’’. It is assumed that the diffusion process obtained by the ML estimate admits a well-behaved stationary density<sup>5</sup>. Recall that for each time series, we estimate a new set of parameters and hence a new SDE of the form given in Eq. 3. The different trajectories each have unique conformational state initial conditions (i.e. different  $\mathcal{C}$  values) in the underlying detailed atomistic simulations. Each of these estimated models can be used to compute a ‘‘stationary’’ density resulting in a collection of ‘‘stationary’’ densities. Quotes are used in the previous sentence because in each short times series burst the value of  $\mathcal{C}$  is effectively fixed and the  $\Phi$  coordinate fluctuates about a fixed point.  $\mathcal{C}$  determines this fixed point as well as the shape of the ‘‘stationary’’ density of  $\Phi$ . The thermodynamic stationary distribution ( $\equiv \Pi^{EQ}$ ) needs to account for the variabil-

<sup>4</sup>Note the print version contains a typesetting error corrected here.

<sup>5</sup>Evaluating  $p^{EQ}(\Phi; \mathcal{C})$  can encounter technical difficulties if one allows the diffusion coefficient to take a zero value (especially relevant to the OD model). Careful selection of  $\Phi_{\text{REF}}$  along with using a finite support can help in numerically dealing with this issue. However, one must be careful to ensure that all regions of non-negligible probability are accounted for in the finite support. SDE simulation can be used to assist in this type of task. Alternatively, one can modify  $\sigma^{\mathcal{C}}$  to smoothly approach a minimum value  $> 0$  and use an infinite support for  $p^{EQ}(\Phi; \mathcal{C})$ .

ity inherent in  $\mathcal{C}$ . Due to the slow time-scale associated with this coordinate, it is difficult to exhaustively sample phase space in a single simulation trajectory. If we somehow had access to a closed-form expression describing the (thermodynamic) probability density of  $\mathcal{C}$ , denote by  $f(\cdot)$ <sup>6</sup>, we could integrate this quantity out using:

$$\Pi^{EQ}(\Phi) := \int p^{EQ}(\Phi; \mathcal{C}) f(\mathcal{C}) d\mathcal{C} \approx \frac{1}{N} \sum_{i=1}^N p_i^{EQ}(\Phi; \mathcal{C}_i), \quad (7)$$

The right-hand-side of the above represents a Monte Carlo approximation of the continuous integral to the left.  $N$  represents the number of time series batches used to calibrate  $N$  different surrogate models describing  $\Phi$ 's dynamics.  $\mathcal{C}_i$  denotes the temporal average value of  $\mathcal{C}$  observed in time series batch  $i$ . The subscript  $i$  on  $p_i^{EQ}$  denotes using the invariant density obtained for  $\Phi$  associated with  $\mathcal{C}_i$  (using Eq. 6 for each estimated SDE). To more systematically overcome the conformational sampling barrier, one could attempt to generate initial conformations utilizing more sophisticated equilibrium sampling methods,<sup>24,53</sup> however we demonstrate this is not necessary to obtain accurate results in the systems studied here, but may be useful in other applications.

Effectively we are modeling the more complex distribution  $\Pi^{EQ}(\Phi)$  using a mixture of simpler densities. This mixture modeling can also be given a physical interpretation. The thermal noise for a fixed value of  $\mathcal{C}$  induces a certain amount variability in the SO of interest; for each single SDE “ $i$ ” this can be quantified using  $p_i^{EQ}(\Phi; \mathcal{C}_i)$ . The variability induced by conformational heterogeneity can be quantified by looking at how disjoint a collection of  $\{p_i^{EQ}(\Phi; \mathcal{C}_i)\}_{i=1}^N$  are relative to the average quantity<sup>7</sup>.

## 2.7 Relation to Computational Chemistry

One interest in this paper is in approximating the global stationary histogram associated with a  $\Phi$  type coordinate using a small amount of MD time series. Two complications are commonly encountered: (1) Time correlation in MD trajectories can complicate constructing reliable estimates due to statistical dependence,<sup>74</sup> (2) Dynamics induced by “orthogonal coordinates” can induce skewed (non-Gaussian) distribution in the stationary histogram associated with the fast SO of interest. Such skewed distributions are commonly encountered in both experiments and simulations when a low-dimensional SO is modulated by a diverse population of conformational degrees of freedom not explicitly included in the model.<sup>46–48</sup>

Knowledge of the shape of such global stationary distributions is important in a variety of applications, e.g. in umbrella sampling type applications one needs to ensure a high degree of overlap between adjacent sampling windows<sup>51,74</sup> and knowledge of the skewed

<sup>6</sup>We simply assume that this density exists and is statistically independent of the  $\Phi$  variable. Including dependence on  $\Phi$  is in principle possible, but the time-scale separation is large enough to make this coupling fairly weak in the particular systems studied.

<sup>7</sup>To do so one must have an estimate of the inherent uncertainty associated with a finite length discrete time series used to estimate the parameters.

histogram shape can help one in refining the grids used in such computations. The non-Gaussian shape of a work histogram is also important in nonequilibrium free energy computations.<sup>30,47</sup> We demonstrate that that our modeling procedures, utilizing a collection of SDE models can help in predicting such shapes and treat the two issues listed in the above paragraph. We also demonstrate that the time-scale separation between the fast-time scale coordinate  $\Phi$  and the slow conformational coordinate  $\mathcal{C}$  also influences kinetic quantities of interest such as the diffusion coefficient.<sup>33,52</sup>

### 3 Simulation Details

We assign  $\Phi \equiv \Delta D_{rmsd}$  (difference in rmsd of the instantaneous structures from the reference open and closed crystal structure of the enzyme) and  $\mathcal{C} \equiv$  the distance between mass centers of residues Ile-52 and Lys-145 characterizing the dynamics of large-scale conformational transitions in AdK (see Ref. [51] for details). This distance type SO has also been measured in solution using single molecule FRET experiments by Henzler-Wildman *et al.*<sup>49</sup>

As detailed in Ref. [51], the initial path between the open and closed conformations of AdK was generated using the Nudged Elastic Band (NEB) method.<sup>75</sup> Subsequently, 81 configurations obtained from NEB path optimization, separated by the interval of 0.2 Å in  $\Delta D_{rmsd}$  space were subjected to US simulations. During these US simulations, production dynamics of 525ps at 300K was performed from each configuration with a weak restraint of 10 kcal/mol/Å<sup>2</sup> in  $\Delta D_{rmsd}$  (the specified target SO value in each window is denoted by  $\Delta D_{rmsd}^0$ ). No restraints were applied along the conformational SO, ( $\mathcal{C}$ ). Solvent effects were modeled implicitly using GBMV approximation<sup>76</sup> in CHARMM.<sup>77</sup>

For further statistical analysis, 50-100 restrained trajectories of 525ps in length each were performed from the eight starting conformations along the path corresponding to the  $\Delta D_{rmsd}^0$  values of (measured in Å), -5.79, -3.67, -0.01, 1.38, 3.30, 5.34 and 7.02. All trajectories were subjected to similar restraint of 10 kcal/mol/Å<sup>2</sup> along  $\Delta D_{rmsd}$ , but were started with the different initial velocities, assigned randomly. The time series of  $\Delta D_{rmsd}$  and distance  $\mathcal{C}$  were extracted from the trajectories (sampled every 0.15 ps) and used in analysis below.

## 4 Results

### 4.1 Parameter Estimates and Goodness-of-Fit Tests

The ML parameters of the OU model were obtained at each of the 81 different US windows. The measured noise magnitude ( $C$ ) depends significantly on the value of  $\Phi(\equiv \Delta D_{rmsd})$ . This is demonstrated in Fig. 1. Parameter estimates were obtained using three different down-sampling (or sub-sampling) parameters.<sup>60</sup> The down-sampling parameter is an integer represented by “ds”. Knowledge of this parameter is related to temporal coarse-graining; it determines the amount of time used to “average out” certain fast-scale non-

Markovian memory effects.<sup>18,21,34,56,68,71</sup> To get a better physical understanding of this quantity, suppose one is numerically integrating a high-dimensional chaotic deterministic Hamiltonian system using a constant time-step size  $\delta t$ . The output of a discrete error-free integration would be the sequence  $\{p_i\}$  where,  $p_i \equiv -\int_{i\delta t}^{(i+1)\delta t} \nabla_q \mathcal{H}(t) dt$  (using notation from classical mechanics). If we attempted to fit a diffusion approximation directly to the sequence  $\{p_i\}$ , it would likely fail because the fast-scale chaotic motion has not had sufficient time to “mix” and the noise is not a “white noise process” (i.e. temporal correlations exist in the fast-scale noise<sup>67</sup>). Alternatively, if we used the sequence  $\{p_i^{\text{ds}}\}$  where,  $p_i^{\text{ds}} \equiv -\int_{i(\text{ds}\times\delta t)}^{(i+1)(\text{ds}\times\delta t)} \nabla_q \mathcal{H}(t) dt$ , the chaotic motion would have more time to “mix” and would make a diffusion model more plausible. The surrogate models estimated from our MD simulations, although inherently stochastic due to the Langevin thermostat, still exhibit dependence on the down-sampling because systematic forces associated with fast-scale memory still need time to average out.

Next we demonstrate how surrogate models calibrating from short time simulations can be tested for goodness-of-fit. An US point ( $\Delta D_{\text{rmsd}} = 7.0$ ) exhibiting significant state dependence in the noise was analyzed in detail. At this point, 75 MD trajectories were simulated and  $\Phi$  was observed every 0.15 ps (the interval is much larger than the integration step-size of 1.5 fs). For every proposed ds, we estimated 75 surrogate model parameter vectors (both OU and OD). A total of three ds values were tested (ds = 1, 2, 3). The total number of time series observations used to estimate each surrogate parameter vector was fixed to be 350 in each case so the terminal length of the time series depended on ds, however each time series used (regardless of ds) started with the same initial observations to maximize the degree of temporal overlap in the  $\Phi$  time series used parameter estimation.

We utilized both the “Q-test statistic”<sup>42</sup> and the “ $T_3$  test statistic”.<sup>64</sup> It is demonstrated that they both have utility in regards to our applications. Ideal finite sample null distributions associated with a time series size of 350 were obtained using Monte Carlo simulation to generate  $1 \times 10^4$  samples<sup>8</sup>. This sample size was assumed large enough to obtain accurate continuous cumulative distribution function (CDF) approximation of the null. The relatively small batch size of MD simulation samples led us to treat the distribution associated with the test statistics as empirical distribution functions (EDFs). The reference null distribution and various EDFs are plotted in Fig. 2. We shade the plot to highlight the critical region associated with a significance level ( $\alpha$ ) of 10%. The x-intercept of the shaded region is the critical value associated with this level and the percentage of rejected models can be obtained by evaluating the EDF at this value and subtracting this result from unity. Although we shade for  $\alpha = 10\%$ , the plot can be used to assess any  $\alpha$  of interest.

Panel (a) of Fig. 2 plots the Q-test results testing both the OD and OU surrogate models using various ds parameters. The percentage of test statistics rejected for ds=1,2, and 3 was roughly 90%, 15% and 5%, respectively for the OD model and was 95%, 10% and 7.5% for

---

<sup>8</sup>For the  $T_3$  test statistic, we used a bootstrap scheme whereby we used the ensemble average of  $\hat{\theta}$  to generate paths and used this single model to create the null. For large samples sizes, under mild assumptions, this test statistic can be shown to be independent of the underlying data generating mechanism making this test appealing in situations where conformational heterogeneity exists.

the OU model. This suggests that when simulation data of the AdK system is discretely observed, the time between adjacent observations should be  $\geq 0.30$  ps before a “statistically acceptable” diffusion model can be used. Artifacts of fast-scale non-Markovian type effects can readily be detected by the Q-test when one samples more frequently in time than this value using even fairly small time series (here 350 observations per trajectory). The other US windows analyzed (not reported) also indicated that  $ds = 2$  was the appropriate coarse-graining parameter to use; this corresponded to 0.30 ps between observations; all subsequent results used this spacing between time series observations. The main utility of such a pathwise goodness-of-fit analysis is that a very small number of short sample paths can be used to determine the time one needs to wait to let fast-scale non-Markovian effects “average out”.<sup>21,71</sup> One of the appealing features of the Q-test is that the underlying nature of the signal is not important (stationary or non-stationary cases can be treated), but for this generality one pays a price in regards to statistical power. The Q-test performs similarly for the OD and OU test despite there being fairly large state-dependence at this point. Later we demonstrate that ignoring this dependence causes poor predictions related to stationary histogram estimation. It would be useful if we could apply a more powerful pathwise test in order to see how well the two different models perform. The  $T_3$  statistic<sup>64</sup> makes use of a stationarity assumption. Panel B reports the results associated with applying this test. The  $T_3$  tests indicates more clearly demonstrates the OD model fits the observations better (roughly 5% of the OD models were rejected whereas about 20% of the OU were using  $ds = 2$ ). Results reported in the Supp. Info. show that increasing the time series sample size to 700 makes rejection easier in both cases, but the OU is still more strongly rejected at the  $\alpha = 10\%$  significance level.

## 4.2 Diffusion Coefficient Approximation

Approximating the dynamics with a simple process like the OU model is appealing because the ML parameter can be obtained directly (a numerical parameter search is not necessary) and the limiting asymptotic large sample distribution of the parameter estimates is also available in closed-form.<sup>43</sup> In Table 1 we report the the mean and standard deviation of  $C^2$  estimated with the OU model. The OD and OU model predictions for this quantity were nearly identical due to the low state-dependence on the noise, so we focus on the latter<sup>9</sup>. Also, if the OU model accurately captures the data, we can exploit several analytical results for statistical inference purposes. For example, the large sample asymptotic standard deviation of  $C^2$  is computed analytically by exploiting the Gaussian property of this process. Table 1 demonstrates that our surrogate models can approximate such quantities. In the atomistic simulation community, the diffusion coefficient is typically determined by using an empirically measured autocorrelation function to determine the  $\tau$  where correlations are small and then  $C^2$  is computed by using ensemble averaging over temporal blocks.<sup>24,33,35,36</sup> The time one needs to wait between observations can be fairly large when one uses this

---

<sup>9</sup>We studied a case where the state-dependence on the noise is mild to facilitate physical interpretation and to compare to established diffusion coefficient methods used in MD simulation analysis.<sup>24,33,35,36</sup>

approach or variants of it.<sup>52</sup>

Figure 3 plots the autocorrelation of the detailed MD samples analyzed in Table 1. The AC was estimated from 50 batches of 525 ps MD data. The plot also contains the average autocorrelation function, i.e. the function obtained by averaging over the 50 autocorrelations. The 95% confidence bands for zero correlation are also plotted for this time series sample size as dotted lines. The observed autocorrelation was used to find a relaxation time ( $\tau$ ) by fitting the early portion to a single exponential via least squares (resulting in  $\tau \approx 15ps$ ). The diffusion coefficient was then computed using:<sup>33,52</sup>  $\langle (\Phi(t+\tau) - \Phi(t))^2 \rangle / (2\tau) = 1.27 \times 10^{-3}$ , where brackets denote ensemble averaging over non-overlapping temporal blocks. This number was compared to the effective diffusion coefficient predicted by the OU model for various  $ds$  values (results reported Table 1). Our models actually exploit the temporal correlation to get better estimates of such quantities and do not necessarily require long time series observations. Recall that a value of  $ds = 2$  (corresponding to 0.30 ps between observations) performed fairly well in regards to the Q-test on all of the data we observed for AdK and coincidentally this parameter also appears to most closely capture the diffusion coefficient computed via traditional ensemble MD methods.<sup>52</sup> We also report the mean, predicted uncertainty and the measured uncertainty in the estimated  $C^2$  for various  $ds$  values.<sup>10</sup>

Note how in Fig 3 one initially observes a roughly single-exponential decay. This feature allowed us to approximate the effective diffusion coefficient using a single time series of length  $\approx 100ps$ . With this small amount of data, we were even able to compute confidence bands which were fairly accurate. However, closer inspection of the AC signal at longer times reveals it may be more complex than an exponential decay. Complex ACs are common in single-molecule experiments where conformational fluctuations persist for a relatively long-time.<sup>15,16,63</sup> Such artifacts may limit the predictive power of a diffusion coefficient calibrated from a scalar SO observed over fairly short time scales. The diffusion coefficient information reported in Table 1 *does not* attempt to account for complex long time behavior<sup>11</sup>.

The bottom panel of Fig. 4 demonstrates that the estimated noise magnitude ( $C_i$ ) of surrogate model “ $i$ ” correlates with  $C_i$ . The figure also contains some sample trajectories demonstrating the time-scale separation existing between  $C$  and  $\mathcal{C}$ . Panel A of Fig. 4 uses three distinct (color-coded) symbols to identify the three sample trajectories plotted in panels B and C. The fast  $\Phi$  SO oscillates (or rapidly reverts) about the observed temporal mean associated with path  $i$  whereas the slow  $\mathcal{C}$  SO exhibits a slow random walk (i.e. not oscillating about a mean). Given that panel A demonstrates that the intensity of effective thermal noise varies with  $\mathcal{C}$  and this quantity does a random walk through phase space, this influences the distribution of the SO of interest ( $\Phi$ ). This plot provides one fairly simple illustration of how a time-scale separation can affect histograms relevant to thermodynamic

<sup>10</sup>Note that as one waits a longer time, fast-scale non-Markovian effects become less important. However in longer time series, artifacts of the evolution on  $\mathcal{C}$  type coordinates can more readily be measured and in this case it appears to increase the effective diffusion coefficient.

<sup>11</sup>A single OU model predicts an AC with a single exponential rate of decay.

applications.

The time-scale separation can be explored more quantitatively by estimating different stochastic models and analyzing the results. For example, if one estimates the effective force using the OD model for  $\mathcal{C}$  and then compares the result of the same model estimated for  $\Phi$ , the characteristic time-scale associated with each effective force is roughly quantified by looking directly at the  $B$  parameter. In the SDE considered, the parameter  $B$  corresponds to the linear sensitivity of the effective force. For the data observed, the characteristic time-scale associated with  $\mathcal{C}$  ranges from  $\approx 40$ -100 times the length of that associated with  $\Phi$  if the ratio of the linear sensitivities are used to quantify the time-scale gap<sup>12</sup>.

### 4.3 Stationary Histogram Approximation

Finally we present results illustrating how a collection of simple models can be used stationary histogram information usually sought in umbrella sampling type applications. We demonstrate that a collection of surrogate models can account for variability associated with the slow  $\mathcal{C}$  time-scale. Figure 5 reveals that the collection of OU invariant densities seems to accurately capture the general shape of  $\Pi^{EQ}$ . However the cases at the edges of the US simulation points do not approximate the shape of the histograms as well. Using a mixture of OD models (Equation 6) remedies the situation in both cases<sup>13</sup>. Figure 6 plots the resulting density prediction ( $\Pi^{EQ}$ ) along with the measured  $\Phi$  histogram obtained directly from the MD data. The inset plots some sample  $p^{EQ}(\cdot)$  functions measured from these models (these are used to construct  $\Pi^{EQ}$ ). The accuracy obtained by this “mixture of density” plots gives further evidence that a diverse ensemble of  $\mathcal{C}$  values modulate the dynamics of  $\Phi$ . For points near the boundaries the correlation between  $\mathcal{C}$  and  $\Phi$  is stronger than that shown in Fig. 4 (see Supp. Mat. Fig. 2) and accounting for this variability is important if one demands high accuracy in the surrogate model density estimates (referring to both  $p^{EQ}(\cdot)$  and  $\Pi^{EQ}$ ).

Before concluding, we provide a description which hopes to show why a “mixture of densities” can help in approximating the  $\Pi^{EQ}$  of a SO associated with a complex molecule. Accurately approximating  $\Pi^{EQ}$  usually requires one to exhaustively sample phase space, not just a small region explored in single short trajectory. The variability induced by “very fast-degrees” of freedom (e.g. vibrational degrees of freedom and solvent bombardment) is modeled in each single surrogate diffusion using Brownian motion. This along with the drift parameters determines  $p^{EQ}(\cdot)_i$  which provides one with quantitative information about of how fast-scale fluctuations cause variability in  $\Phi$  for a relatively frozen value of  $\mathcal{C}$ . Slow time-scale conformational variability (e.g. that introduced by the distribution of  $\mathcal{C}$ ) is accounted for using a collection of surrogate models. In Ref. [51] it was shown that the free energy profile of  $\mathcal{C}$  was effectively flat. It effectively does a random walk in phase space

---

<sup>12</sup>It should also be noted that when a coupled 2-d model was estimated, the eigenvalues of the effective force (of the coupled system) indicated a similar separation in time scales in the effective forces.

<sup>13</sup>Results for the case near  $\Phi \approx -6$  are given in Supp. Mat. Fig 1.)

when viewed over long time-scales whereas  $\Phi$  is constrained with a harmonic potential. The skewed histogram of  $\Phi$  observed in the single 525 ps run shown in Fig. 6 is an artifact of this modulating effect.

## 5 Conclusions

We have demonstrated that a collection of fairly simple surrogate diffusion models estimated from time series data can accurately capture dynamical features of short constrained AdK simulations. The techniques presented should be thought of as a “post-processing” analysis in which statistical summaries (such as correlation and the invariant distribution of SOs) are obtained by time series techniques. In most cases, the parameters of the diffusion models were modulated by degrees of freedom associated with large-scale conformational changes. The slow SO monitored in AdK is experimentally accessible in solution via single molecule FRET.<sup>49</sup> We also demonstrated that pathwise statistical inference could be used to obtain efficient parameter estimates from temporally correlated MD observations. Application of goodness-of-fit tests helped identify the time needed to wait (a coarse-graining parameter) before fast-scale non-Markovian artifacts “averaged out”.

Information extracted from a collection of surrogate diffusion models can be used to assist free energy computations as well as obtain kinetic information in the form of effective diffusion using a relatively short amount of detailed simulation trajectories in certain situations. Confidence bands and goodness-of-fit tests can be used to check the quality of the approximation without requiring a large number of expensive simulation results. This shows promise in reducing the computational load needed to obtain kinetic and thermodynamic properties of complex biomolecules. These types of methods may also possibly be used to assist established sampling techniques like WHAM, parallel tempering, or metadynamics<sup>24,74,78</sup> where many histograms need to be approximated. If this can be done with shorter time series, computational resources will be free to explore other portions of phase space believed physically relevant. The findings are not isolated to very short constrained simulations; the Supporting Information reports results demonstrating results using longer unconstrained simulations coming from an explicitly solvated protein trajectory obtained from the `dynamomics.org` library.<sup>26</sup> We also see the statistical analysis tools presented here as being useful in data-mining applications.

The surrogate models we appealed to in this article did not explicitly exploit the structure of any underlying governing equations. The proposed models had a phenomenological motivation. The collection of estimated surrogate models did give dynamic and static information about a  $\Phi$  type coordinate over a broad range of phase space not typically explored in a single simulation and did so using SDE models which we could efficiently estimate, quantify the uncertainty in our estimates, and readily interpret in terms of established statistical physics. We could also assess the goodness-of-fit of the estimated models *a posteriori* fashion. If simplified models coming from mathematical model reduction techniques are available, e.g.,<sup>19,71</sup> the parameters of reduced models could possibly be es-

timated from observations. One could also consider attempting to model the dynamics of more SOs and/or utilize the structure of a generalized Langevin equation resulting in more complicated surrogate models. The statistical analysis of such models (estimation and inference) is fairly involved and introduces many new mathematical challenges, but interesting results are being obtained in that direction.<sup>63</sup> Data-driven modeling is particularly attractive because recent advances in single-molecule manipulation methods<sup>1-16</sup> are making a variety of low-dimensional SOs available to dynamically analyze. Synergistically combining data-driven modeling techniques with new and established simulation methods as well as mathematical multiscale analysis shows great promise in providing new insights into complex biological systems.

## 6 Acknowledgments

We thank two anonymous referees for helpful comments which improved the quality of the manuscript. In addition CPC was supported by NSF grants #s DMS 0240058 & ACI-0325081, NIH grant T90 DK070121-04 and obtained partial computational support from the Rice Computational Research Cluster funded by NSF under Grant CNS-0421109, and a partnership between Rice University, AMD and Cray.

**Supporting Information:** A PDF containing supplemental information related to Figs. 4 and 6 of the main text as well as a text file containing a url link to MATLAB scripts demonstrating the parameter estimation procedure. This information is available free of charge via the Internet at <http://pubs.acs.org>.

## References

1. Bustamante, C.; Bryant, Z.; Smith, S. *Nature* **2003**, *421*, 423.
2. Carrion-Vazquez, M.; Oberhauser, A.; Fisher, T.; Marszalek, P.; Li, H.; Fernandez, J. *Prog. Biophys. Mol. Bio.* **2000**, *74*, 63.
3. Stock, G.; Ghosh, K.; Dill, K. *J. Chem. Phys.* **2008**, *128*, 194102.
4. Collin, D.; Ritort, F.; Jarzynski, C.; Smith, S.; Tinoco, Jr., I.; Bustamante, C. *Nature* **2005**, *437*, 231.
5. Min, W.; Gopich, I.; English, B.; Kou, S.; Xie, X.; Szabo, A. *J. Phys. Chem. B* **2006**, *110*, 20093.
6. Smith, S.; Y. Cui.; Bustamante, C. *Science* **1996**, *271*, 795.
7. Rief, M.; Clausen-Schaumann, H.; Gaub, H. *Nat. Struct. Biol.* **1999**, *6*, 346.
8. Clausen-Schaumann, H.; Rief, M.; Tolksdorf, C.; Gaub, H. *Biophys. J.* **2000**, *78*, 1997.
9. Albrecht, C.; Neuert, G.; Lugmaier, R.; Gaub, H. *Biophys. J.* **2008**, *94*, 4766.
10. Lee, G.; Rabbi, M.; Marszalek, R. C. P. *Small* **2007**, *5*, 809.
11. Ke, C.; Humeniuk, M.; S-Gracz, H.; Marszalek, P. *Phys. Rev. Lett.* **2007**, *99*, 018302.
12. Harris, N. C.; Song, Y.; Kiang, C.-H. *Phys. Rev. Lett.* **2007**, *99*, 068101.
13. Dixit, S.; Singh-Zocchi, M.; Hanne, J.; Zocchi, G. *Phy. Rev. Lett.* **2005**, *94*, 118101.
14. Vendruscolo, M.; Dobson, C. *Science* **2006**, *313*, 1586.
15. Liu, S.; Bokinsky, G.; Walter, N.; Zhuang, X. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 12634.
16. Greenleaf, W.; Frieda, K.; Foster, D.; Woodside, M.; Block, S. *Science* **2008**, *319*, 630.
17. Schlick, T.; Skeel, R. D.; Brunger, A. T.; Kale, L. V.; Board, J. A.; Hermans, J.; Schulten, K. *J. Comp. Phys.* **1999**, *151*, 9.
18. Zwanzig, R. Brownian motion and Langevin equations. In *Nonequilibrium Statistical Mechanics*, 1st ed.; Oxford University Press: New York, 2001; pp 3-24, 143.
19. A. J. Chorin, A. K.; Kupferman, R. *Proc. Natl. Acad. Sci. USA* **1998**, *95*, 4094.
20. Kopelevich, D.; Panagiotopoulos, A.; Kevrekidis, I. *J. Chem. Phys.* **2005**, *122*, 044908.
21. Givon, D.; Kupferman, R.; Stuart, A. *Nonlinearity* **2004**, *17*, R55.

22. E, W.; Liu, D.; Vanden-Eijnden, E. *Commun. Pur. Appl. Math.* **2005**, *58*, 1544.
23. Borgia, A.; Williams, P.; Clarke, J. *Annu. Rev. Biochem.* **2008**, *77*, 101.
24. Schlick, T. Molecular dynamics: Basics. In *Molecular Modeling and Simulation: An Interdisciplinary Guide*, 2nd ed.; Marsden, J., Sirovich, L., Wiggins, S., Antman, S., Eds.; Springer-Verlag: New York, 2002; pp 394-406.
25. Sotomayor, M.; Schulten, K. *Science* **2007**, *316*, 1144.
26. Simms, A.; Toofanny, R.; Kehl, C.; Benson, N.; Daggett, V. *Protein. Eng. Des. Sel.* **2008**, *21*, 369.
27. Maragakis, P.; Lindorff-Larsen, K.; Eastwood, M.; Dror, R.; Klepeis, J.; Arkin, I.; Jensen, M.; Xu, H.; Trbovic, N.; Friesner, R.; Palmer, A.; Shaw, D. *J. Phys. Chem. B* **2008**, *112*, 6155.
28. Moffitt, J.; Chemla, Y.; Smith, S.; Bustamante, C. *Annu. Rev. Biochem.* **2008**, *77*, 19.1.
29. Calderon, C. *J. Chem. Phys.* **2007**, *126*, 084106.
30. Calderon, C.; Chelli, R. *J. Chem. Phys.* **2008**, *128*, 145103.
31. Calderon, C.; Chen, W.; Harris, N.; Lin, K.; Kiang, C. *J. Phys.: Condensed Matter* **2008**, *in press*.
32. Calderon, C.; Harris, N.; Kiang, C.-H.; Cox, D. *J. Phys. Chem. B* **2008**, *in press*.
33. Hummer, G. *New J. Phys.* **2005**, *7*, 1.
34. Calderon, C. *Multiscale Mod Sim* **2007**, *6*, 656.
35. Chahine, J.; Oliveira, R.; Leite, V.; Wang, J. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 14646.
36. Snow, C.; Rhee, Y.; Pande, V. S. *Biophys. J.* **2006**, *95*, 078102.
37. Sigg, D.; Qian, H.; Bezanilla, F. *Biophys. J.* **1999**, *76*, 782.
38. Park, S.; Schulten, K. *J. Chem. Phys.* **2004**, *120*, 5946.
39. Khatri, B. S.; Byrne, K.; Kawakami, M.; Brockwell, D.; Smith, D.; Radford, S.; McLeish, T. *Faraday Discuss* **2008**, *139*, 35.
40. Ait-Sahalia, Y. *Econometrica* **2002**, *70*, 223.
41. Jimenez, J.; Ozaki, T. *J Time Ser. Anal.* **2005**, *27*, 77.
42. Hong, Y.; Li, H. *Rev Financ Stud* **2005**, *18*, 37–84.

43. Chen, S.; C.Y., T. submitted *J. Econometrics* , <http://www.stat.iastate.edu/preprint/articles/2006-21.pdf>, (accessed Oct 1, 2008).
44. Chen, S.; Tang, C. *Ann. Stat.* **2008**, *36*, 167.
45. Walther, K.; Brujic, J.; Li, H.; Fernandez, J. *Biophys. J.* **2006**, *90*, 3806.
46. Minh, D.; Bui, J.; Chang, C.; Jain, T.; Swanson, J.; McCammon, J. *Biophys. J. Letters* **2005**, *72*(89), L25.
47. Procacci, P.; S., M.; Barducci, A.; Signorini, G.; Chelli, R. *J. Chem. Phys.* **2006**, *125*, 164101.
48. Paramore, S.; Ayton, G.; Voth, G. *J. Chem. Phys.* **2007**, *14*, 105105.
49. Henzler-Wildman, K.; et al.. *Nature* **2007**, *450*, 06410.
50. Muller, C. W.; Schulz, G. E. *J. Mol. Biol.* **1992**, *224*(1), 159.
51. Arora, K.; Brooks, III, C. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 18496.
52. Socci, N. D.; Onuchic, J. N.; Wolynes, P. G. *J. Chem. Phys.* **1996**, *104*, 5860.
53. Frenkel, D.; Smit, B. Molecular dynamics simulations. In *Understanding Molecular Simulation: From Algorithms to Applications*, 1st ed.; Academic Press: San Diego, CA, 1996; pp 75-88, 377.
54. Beck, D.; Daggett, V. *Biophysical J.* **2007**, *93*, 3382.
55. Kevrekidis, I.; Gear, C.; Hummer, G. *AIChE J.* **2004**, *50*, 474.
56. Givon, D.; Kevrekidis, I.; Kupferman, R. *Commun. Math. Sci.* **2006**, *4*, 707.
57. Horenko, I.; Hartmann, C.; Schütte, C. *Phys. Rev. E* **2007**, *76*, 016706.
58. Evensen, G.; van Leeuwen, P. *Mon. Weather. Rev.* **2000**, *128*, 1852.
59. Vanden-Eijnden, E. *Commun. Pur. Appl. Math* **2003**, *1*, 385.
60. Zhang, L.; Mykland, P.; Ait-Sahalia, Y. *J. Am. Stat. Assoc.* **2005**, *100*, 1394.
61. Calderon, C.; Martinez, J.; Carroll, R.; Sorensen, D. *submitted* **2008**.
62. Calderon, C.; Janosi, L.; Kosztin, I.; Technical Report, TR08-24; CAAM Dept.: Rice University, 2008.
63. Kou, S.; Xie, X. *Phys. Rev. Lett.* **2004**, *93*, 18.

64. Ait-Sahalia, Y.; Fan, J.; Peng, H. Social Science Research Network. <http://ssrn.com/abstract=955820> (accessed Oct 1, 2008).
65. Kloeden, P.; Platen, E. Introduction. In *Numerical Solution of Stochastic Differential Equations*, 1st ed.; Springer-Verlag: Berlin Springer-Verlag, 1999; pp 37.
66. Chandrasekhar, S. *Rev. Mod. Phys.* **1943**, *15*, 1.
67. El-Ansary, M.; Khalil, H. *SIAM J. Control Optim.* **1986**, *24*, 83.
68. Skorokhod, A.V. Asymptotic behavior of systems of stochastic equations containing a small parameter. In *Asymptotic Methods in the Theory of Stochastic Differential Equations*, 1st ed.; Amer Mathematical Society : Providence, RI, 1989; pp 77.
69. Krishnan, J.; Runborg, O.; Kevrekidis, I. *Comp. Chem. Eng.* **2004**, *28*, 557.
70. Erban, R.; Frewen, T.; Wang, X.; Elston, T.; Coifman, R.; Nadler, B.; Kevrekidis, I. *J Chem Phys* **2007**, *126*, 155103.
71. Pavliotis, G. A.; Stuart, A. M. *J. Stat. Phys.* **2007**, *127*, 741.
72. Kutoyants, Y. Diffusion processes and statistical problems. In *Statistical Inference for Ergodic Diffusion Processes*, 1st ed.; Springer: New York, 2004; pp 50.
73. Risken, H. Fokker-Planck equation for one variable. In *The Fokker-Planck Equation*, 2nd ed.; Springer-Verlag: Berlin, 1996; pp 98.
74. Chodera, J.; Swope, W.; Pitera, J.; Seok, C.; Dill, K. *J. Chem. Theory and Comput.* **2007**, *3*, 26.
75. Chu, J.-W.; Trout, B. L.; Brooks, B. R. *J. Chem. Phys.* **2003**, *119*(24), 12708.
76. Michael, S.; Salsbury, Jr., F.; Brooks III, C. *J. Chem. Phys.* **2002**, *116*(24), 10606.
77. Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comp. Chem.* **1983**, *4*, 187.
78. Marsili, S.; Barducci, A.; Chelli, R.; Procacci, P.; Schettino, V. *J. Phys. Chem. B* **2006**, *110*, 14011.

Table 1: Diffusion Coefficient ( $\tilde{D}$ ) Estimation. The effective diffusion coefficient (in asymptotic mean square displacement sense) was computed from the MD data at the point  $\Delta D_{rmsd}^0 = 1.38$ . The value obtained was  $1.27 \times 10^{-3} \text{Å}^2/\text{ps}$  (see text). The diffusion coefficient estimated by the OU models is reported using three different down-sampling rates. Results using a time series of length ( $N$ ) 350 and  $N = 700$  are reported. In each case, results from analyzing 50 batches of time series are summarized by using the mean and empirically measured standard deviation (“Emp Std. Dev.”) of the diffusion coefficient estimated surrogate models (each time series gave one estimate). We also report the large sample uncertainty estimate (“Asymp. Std. Dev.”) of the maximum likelihood estimate. An analytic expression for this quantity is reported in Theorem 3.1.1 of Ref.<sup>43</sup>.

	$\Delta t = 0.15ps$	$\Delta t = 0.30ps$	$\Delta t = 0.45ps$
$\tilde{D} (N = 350)$	$1.0350 \times 10^{-3}$	$1.4387 \times 10^{-3}$	$1.5989 \times 10^{-3}$
Emp Std. Dev. $\tilde{D}$	$1.1912 \times 10^{-4}$	$1.6584 \times 10^{-4}$	$1.7161 \times 10^{-4}$
Asymp. Std. Dev. $\tilde{D}$	$1.1136 \times 10^{-4}$	$1.5481 \times 10^{-4}$	$1.7190 \times 10^{-4}$
$\tilde{D} (N = 700)$	$1.0237 \times 10^{-3}$	$1.4012 \times 10^{-3}$	$1.5564 \times 10^{-3}$
Emp Std. Dev. $\tilde{D}$	$9.7924 \times 10^{-5}$	$1.2283 \times 10^{-4}$	$1.3652 \times 10^{-4}$
Asymp. Std. Dev. $\tilde{D}$	$7.7736 \times 10^{-5}$	$1.0632 \times 10^{-4}$	$1.1810 \times 10^{-4}$

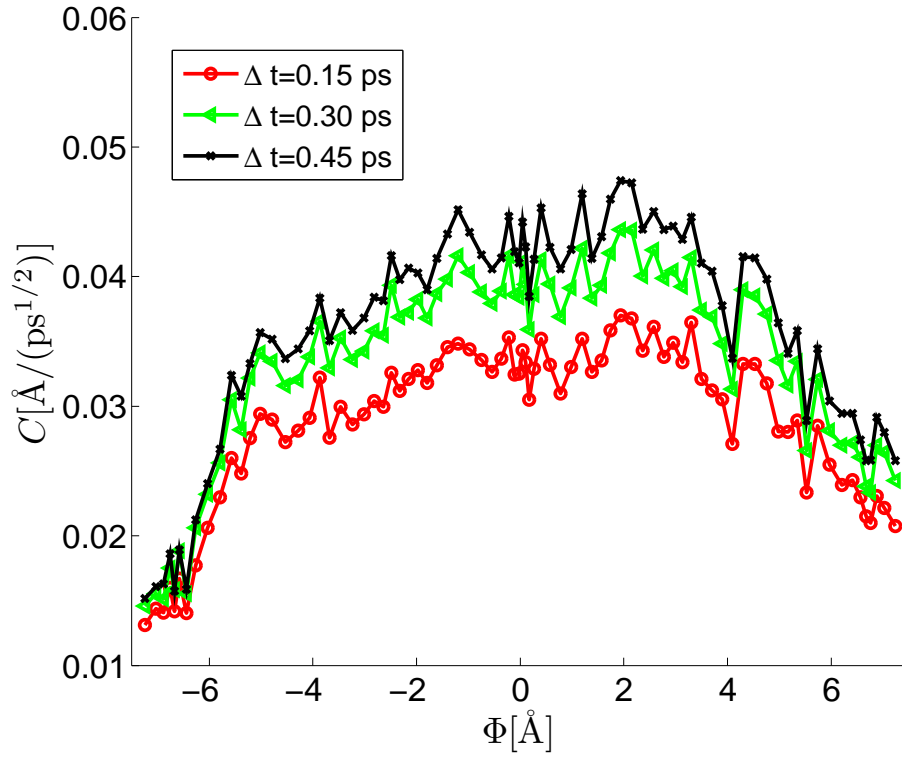


Figure 1: The  $C$  parameter of the Ornstein-Uhlenbeck process was estimated using short time series from 81 different (independent) US windows. The values estimated are denoted by symbols and the purpose of the line connecting the points is only to guide the eye. Each parameter estimate came from a time series containing 350 uniformly spaced entries. Three different  $\Delta t$  values were used. The corresponding time  $\Delta t$  between observations is reported in the legend.

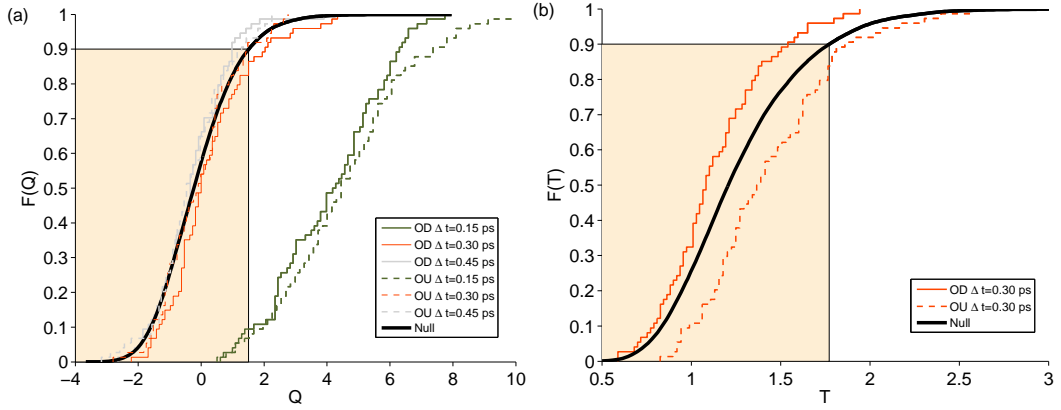


Figure 2: Hypothesis test results. In each panel, the staircase plots correspond to the empirical distribution function (EDF) of the test statistics obtained from batches of 75 time series (each using different  $\Delta t$  values, the corresponding time between observations,  $\Delta t$ , are reported in the legend) and the solid curve corresponds to the distribution of the null computed for a finite sample size of 350 which was the length of each time series analyzed in this plot. The shaded region is used to show the  $\alpha = 0.10$  critical value. The percent of models rejected at this level can be found by noting the point on the x-axis where a color change occurs (denote by  $x_{\text{crit}}$ ) and then evaluating  $1-\text{EDF}(x_{\text{crit}})$ . Panel (a): The Q-test statistic given in Ref. [42] was applied to determine the time needed to wait between observations before an overdamped diffusion model could be applied to simulation data. The surrogate model parameters were estimated for each path and then the Q-test statistic was computed using the data and the estimated model. Panel (b): The  $T_3$  test statistic<sup>64</sup> computed using the same estimated parameters and data.

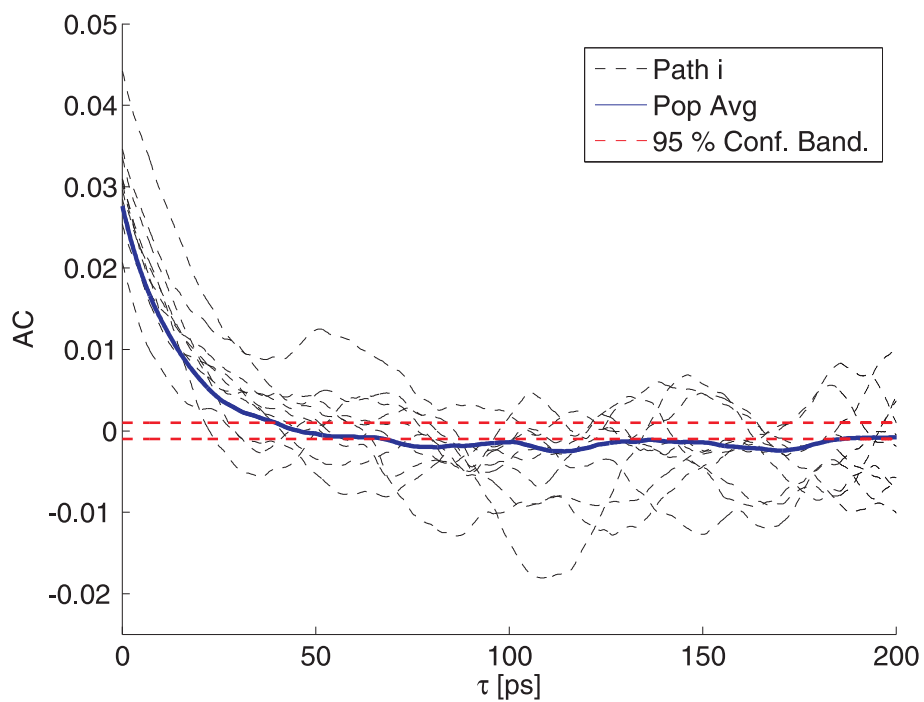


Figure 3: Autocorrelation (AC) measured from MD data taken from US point  $\Delta D_{rmsd}^0 = 7.03$ . The thick line represents the mean AC function obtained using the full 525 time series and estimating the AC for each sample path and then averaging the results. The thin AC labeled as “Path i” are some representative ACs. The thick dotted horizontal lines correspond to the 95 % confidence intervals.



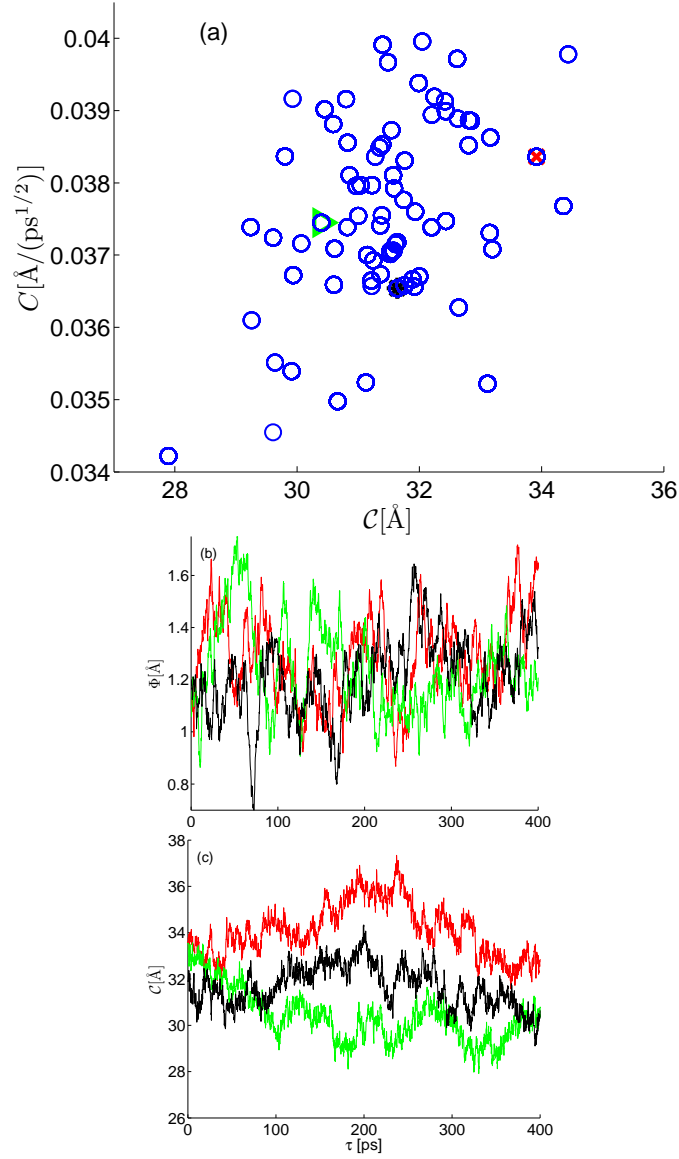


Figure 4: (a) Scatter plot of the estimated  $C$  of the Ornstein-Uhlenbeck (OU) Model vs  $\langle C \rangle$ . The data consists of the estimated OU noise parameter ( $C$ ) obtained using time series consisting of uniformly sampled observations spaced by  $\Delta t = 0.30$ . The noise parameter was estimated for 50 batches of short time series and all US simulations used simulations corresponding to the US constraint point  $\Delta D_{rmsd}^0 = 1.38$  plotted against the (temporal) average value of  $C$  for the corresponding  $\Delta D_{rmsd}$  time series used to estimate the OU parameters. The linear correlation ( $r$ ) between the estimated  $C$  and  $\langle C \rangle$  was found to be 0.34 and the associated p-value was  $1.0 \times 10^{-3}$ . (b) representative time series of the “fast”  $\Delta D_{rmsd}$  coordinate and (c) “slow”  $C$  coordinate. The three color coded trajectories in (b) and (c) correspond to the three color coded symbols in (a).

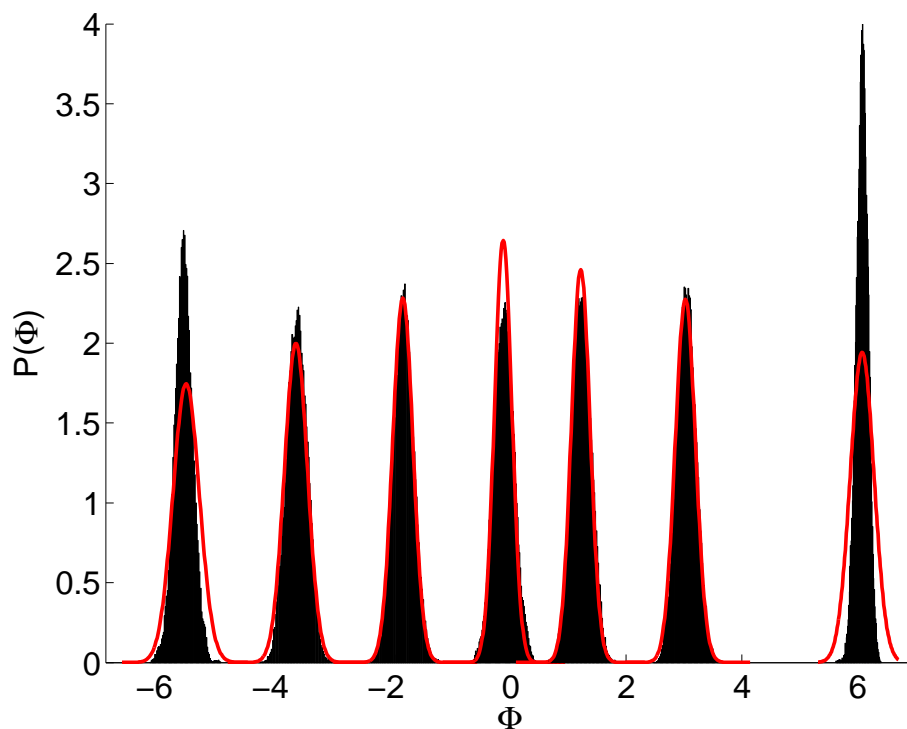


Figure 5: The histogram obtained from running MD simulations using 7 different constraint points are reported. Each data point contains the results from 50 independent simulations run for  $105/ps$  (again uniform sampling with  $\Delta t = 0.30/ps$ ). The prediction of the simple OU model which accounts for the conformational heterogeneity (see text for details) is shown as a solid line. In most cases this crude approximation is accurate, the largest discrepancy here is in the left and rightmost distributions.

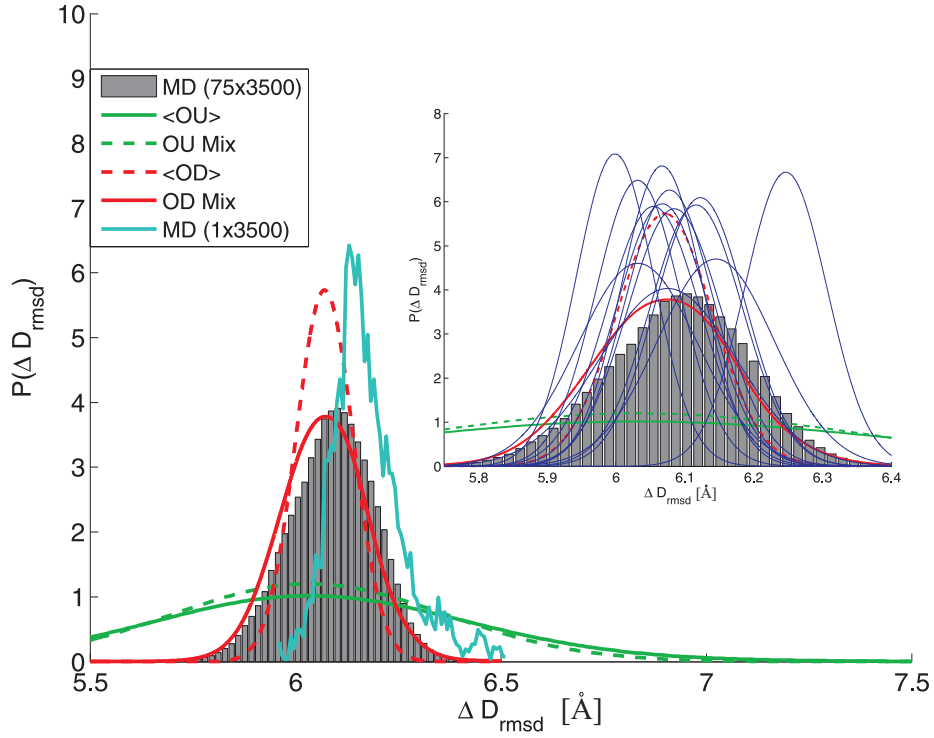


Figure 6: Stationary density estimate focusing on rightmost density shown in Fig 5 (corresponding to  $\Delta D_{rmsd}^0 = 7.03$ ). The result obtained using the Ornstein-Uhlenbeck surrogate (solid red line) was poor. A batch of over-damped models were estimated (from the same times series used to fit the Ornstein-Uhlenbeck models in Fig 5). The solid lines denotes the invariant density obtained by appealing to Equation 7 and the dotted lines represent the invariant density prediction obtained by using  $\langle \theta \rangle$ . The collection of thin blue lines display some representative invariant density predictions (i.e. “ $p_i^{EQ}$ ” in Equation 7). The histogram of  $\Phi$  coming from an ensemble of 75 genuine MD time series of length 525 ps is represented as the jagged line.